Acoustics Only Models - Neg Binomial

Yvonne Barkley

10/4/2020

Load libraries

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

Research question:

What environmental variables characterize sperm whale habitat?

Hypothesis: Sperm whales are found in deep, productive offshore waters.

Include more details about what to expect in this document

Load universal variables

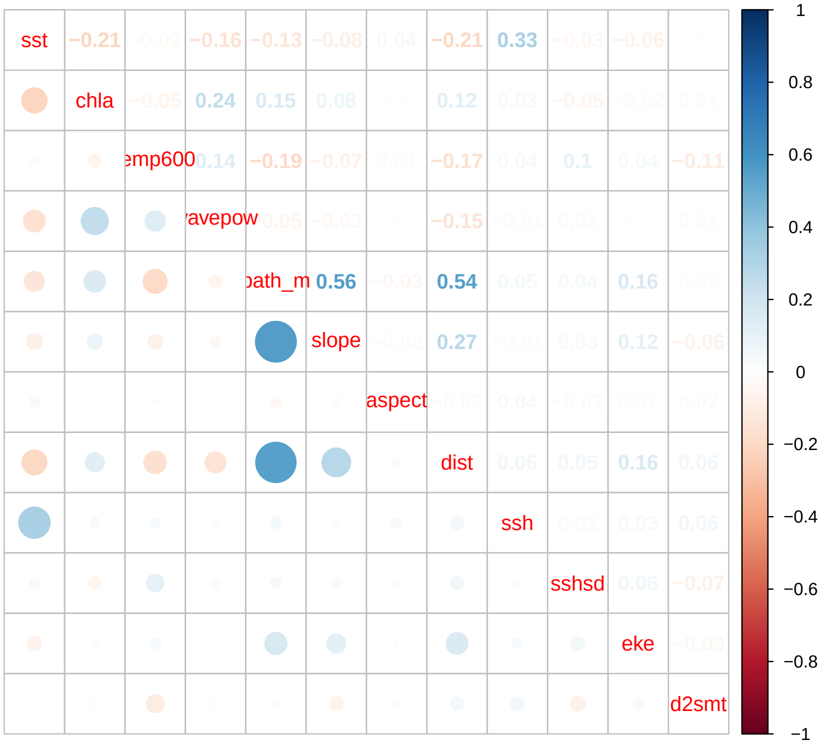
```
# Values used for file and directory names
survey = "AllSurveys"
gridsize = 25
loctype = "AcOnly"
loctype2 = "Ac"
```

Load data from 'models/data' folder

```
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)
PmScaled <- subset(PmScaled, chla <= 10)  #some outliers in a handful of absences
```

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
```

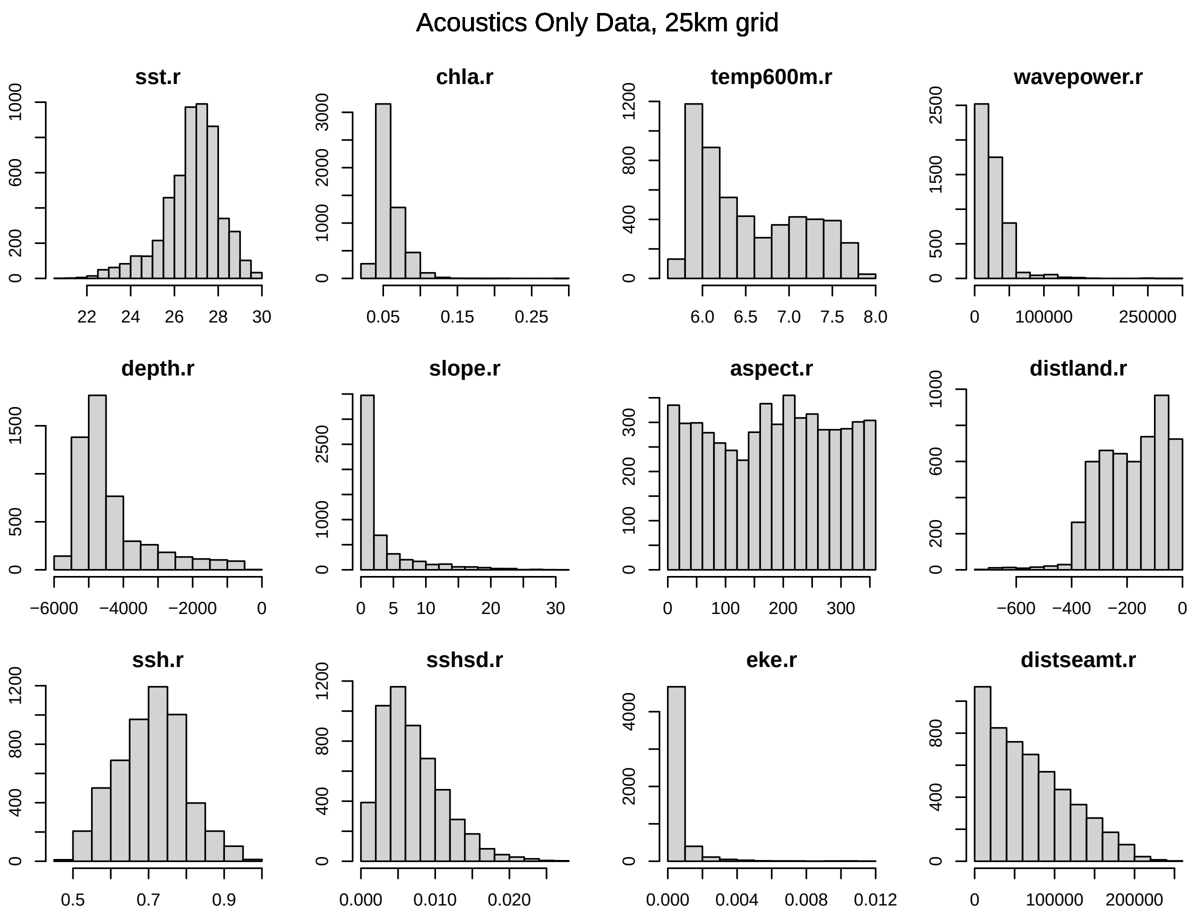
KS tests

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 absent.

Data Visualization

Histograms showing the general distribution of each environmental predictor for the entire dataset.

```r
par(mfrow = c(3, 4), mar = c(3, 3, 2, 1), oma = c(0, 0, 3, 1))

dataSet = PmScaled  #raw values

loopVec <- 30:41  #columns from PmScaled to plot

for (j in loopVec) {

    datPlot <- dataSet[, c(1, j)]

    hist(datPlot[, 2], main = colnames(datPlot)[2], ylab = "frequency",
        xlab = "")
    # plot(datPlot[,2], datPlot[,1], ylab = 'Whales', xlab =
    # colnames(datPlot)[2])
    mtext(paste0("Acoustics Only Data, ", gridsize, "km grid"),
        side = 3, line = 1, outer = TRUE, cex = 1, font = 1)

}
```



```r
# dev.off()
```

```r
# take the log of some variables that are more skewed
PmScaled$chla.log <- log(PmScaled$chla.r)
PmScaled$eke.log <- log(PmScaled$eke.r)
PmScaled$wavepow.log <- log(PmScaled$wavepower.r)


# plot them
dataSet = PmScaled  #raw values

loopVec <- 57:59  #columns from PmScaled to plot

par(mfrow = c(1, 3), mar = c(3, 3, 2, 1), oma = c(0, 0, 3, 1))

for (j in loopVec) {

    datPlot <- dataSet[, c(1, j)]

    hist(datPlot[, 2], main = colnames(datPlot)[2], ylab = "frequency",
        xlab = "")
    # plot(datPlot[,2], datPlot[,1], ylab = 'Whales', xlab =
    # colnames(datPlot)[2])
    mtext(paste0("Acoustics Only Data, ", surveynum, ", ", gridsize,
        "km grid"), side = 3, line = 1, outer = TRUE, cex = 1,
        font = 1)

}
```

Data Splitting

Split the data into train and test sets

```r
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
seed = 1

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)  # & loc == 1) #for S999 versions
    listPres <- splitdf(pres1, seed)  #output is list for train and test

    # subset for absences and split 70/30
    abs0 <- filter(trSub, pa == 0)
    listAbs <- splitdf(abs0, seed)  #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, "_S999b.rda")))
saveRDS(testAcOnly, here::here(paste0("output/models/", loctype,
    "/data/Test_", gridsize, "km_", loctype2, "_S999b.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 negative binomial distribution to model the response variable for comparison with the Tweedie distribution. 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.

25 km spatial scale

- NEGATIVE BINOMIAL DISTRIBUTION
- 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

```
# seed 1
trainS999 <- readRDS(here::here(paste0("output/models/", loctype,
    "/data/Train_", gridsize, "km_", loctype2, "_S999b.rda")))
testS999 <- readRDS(here::here(paste0("output/models/", loctype,
    "/data/Test_", gridsize, "km_", loctype2, "_S999b.rda")))
```

Model Selection

SET 1

Full Models

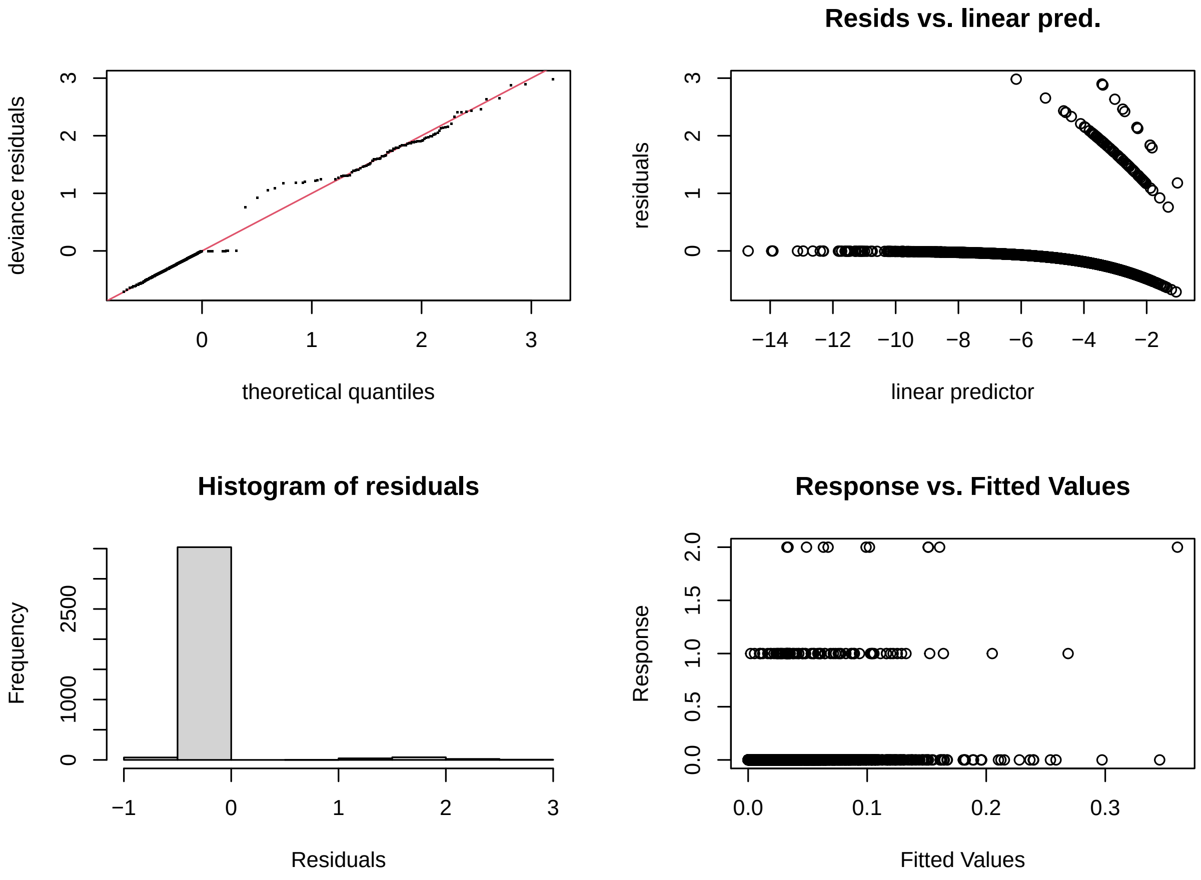
Estimate the smoothing parameters for each predictor variable using restricted maximum likelihood (method = 'REML') + does not include spatial smoother + does not include slope or aspect due to the variation between left and right

```
# * Does NOT include sighted acoustic encounters OR spatial
# smoother
require(mgcv)
nbS999 <- gam(pa ~ s(bath_m, k = 3) + s(dist, 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 = trainS999, family = nb,
    link = "log", select = TRUE, method = "REML")
summary(nbS999)
```

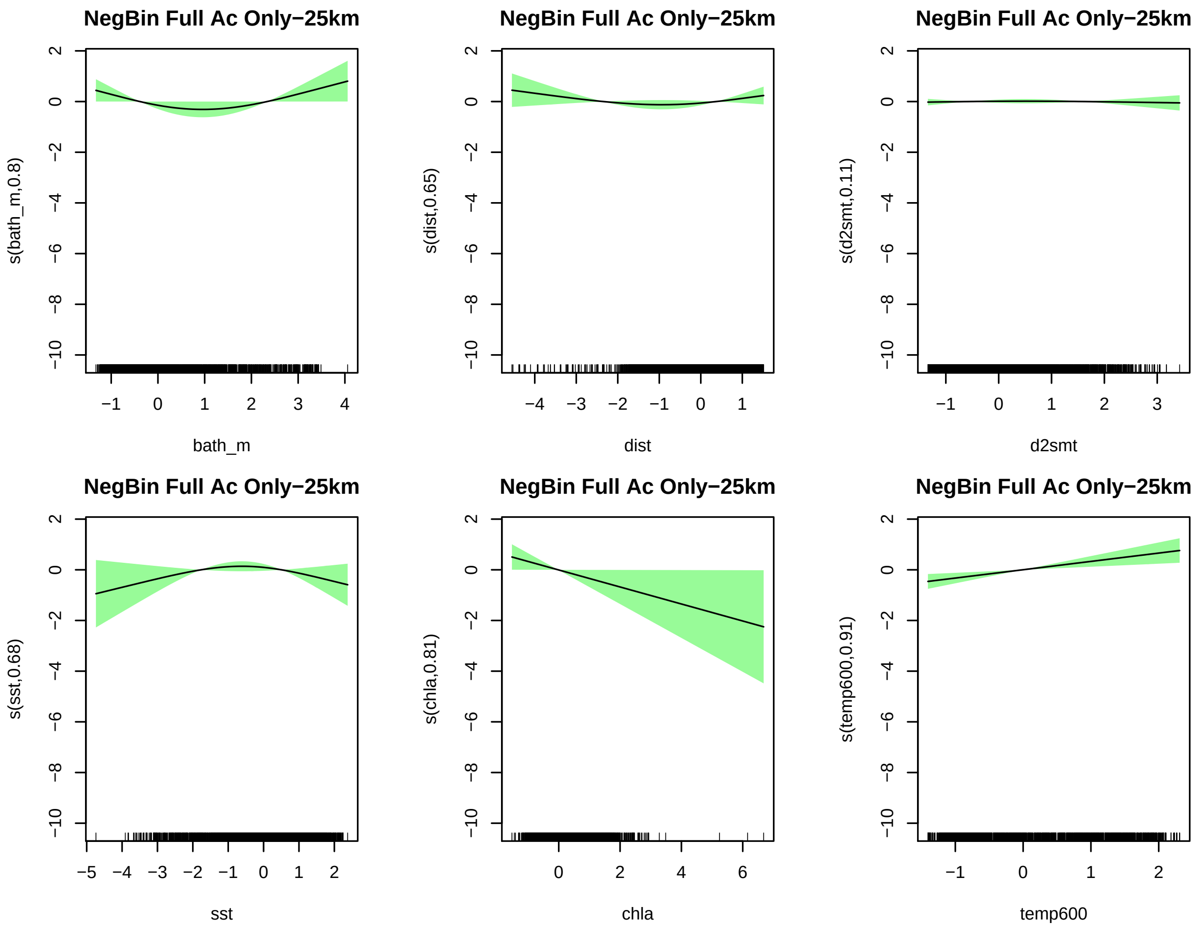
```
Family: Negative Binomial(0.429)
Link function: log

Formula:
pa ~ s(bath_m, k = 3) + s(dist, 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 z value Pr(>|z|)
(Intercept) -22.9302     0.1195  -191.9   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                 edf Ref.df Chi.sq  p-value
s(bath_m)  8.004e-01      2  4.022 0.023926 *
s(dist)    6.496e-01      2  1.846 0.089364 .
s(d2smt)   1.108e-01      2  0.121 0.294713
s(sst)     6.827e-01      2  2.014 0.082015 .
s(chla)    8.125e-01      2  4.062 0.023203 *
s(temp600) 9.086e-01      2  9.910 0.000738 ***
s(ssh)     9.258e-01      2 12.247 0.000267 ***
s(sshsd)   7.250e-01      2  2.670 0.054397 .
s(eke)     8.597e-01      2  5.345 0.012116 *
s(wavepow) 7.488e-06      2  0.000 0.944846
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.0414   Deviance explained = 10.5%
-REML = 424.19  Scale est. = 1         n = 3659
```

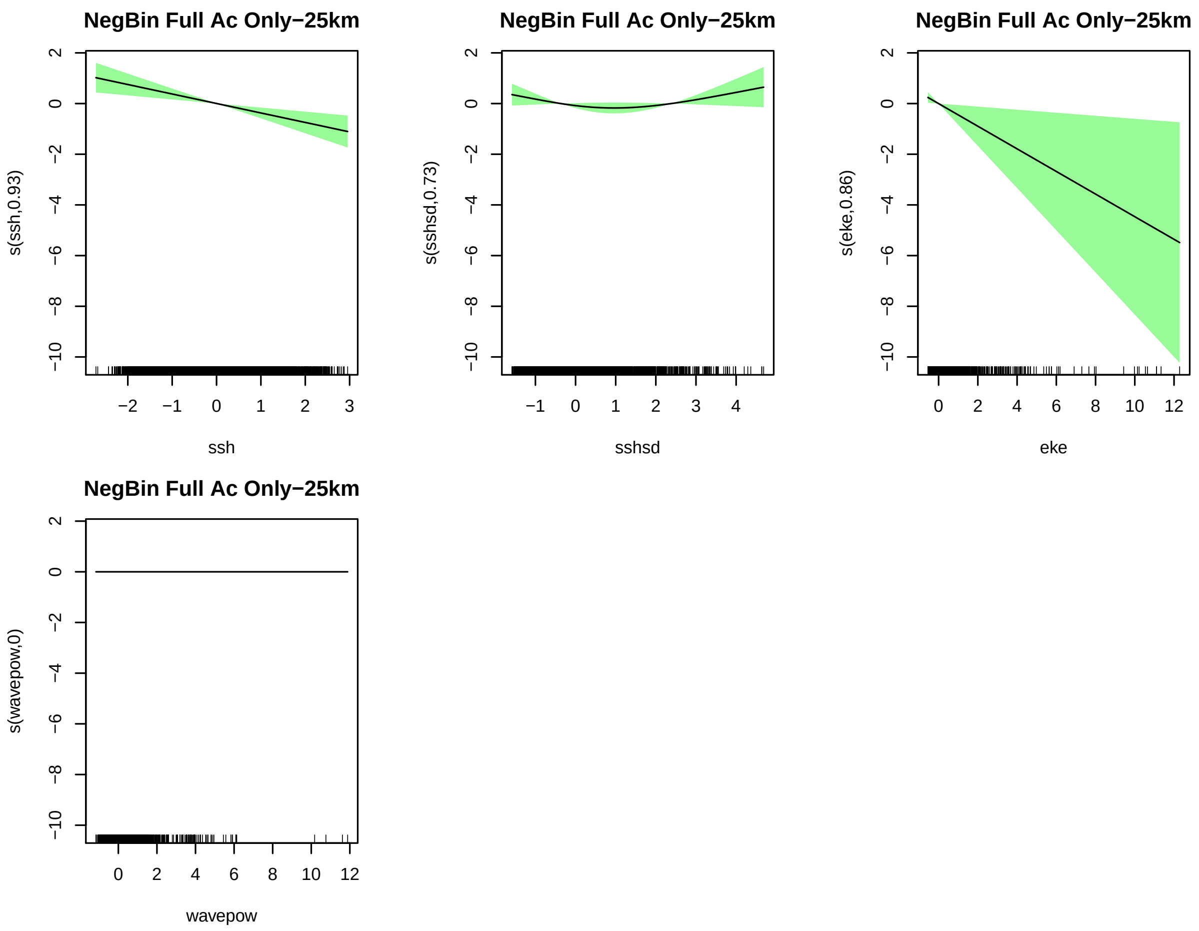
```
AIC(nbS999)
```

```
[1] 844.8141
```



```
Method: REML   Optimizer: outer newton
full convergence after 14 iterations.
Gradient range [-6.672544e-05,1.57489e-05]
(score 424.1868 & scale 1).
Hessian positive definite, eigenvalue range [1.516209e-06,4.191952].
Model rank =  21 / 21

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 8.00e-01    0.82   0.015 *
s(dist)    2.00e+00 6.50e-01    0.85   0.655
s(d2smt)   2.00e+00 1.11e-01    0.85   0.640
s(sst)     2.00e+00 6.83e-01    0.81   0.015 *
s(chla)    2.00e+00 8.12e-01    0.84   0.360
s(temp600) 2.00e+00 9.09e-01    0.73  <2e-16 ***
s(ssh)     2.00e+00 9.26e-01    0.84   0.180
s(sshsd)   2.00e+00 7.25e-01    0.82   0.040 *
s(eke)     2.00e+00 8.60e-01    0.84   0.220
s(wavepow) 2.00e+00 7.49e-06    0.78  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```


NegBin Full Ac Only−25km
s(bath_m,0.8)
2
0
−2
−4
−6
−8
−10
−1
0
1
2
3
4
bath_m
NegBin Full Ac Only−25km
s(dist,0.65)
−4
−3
−2
−1
0
1
dist
NegBin Full Ac Only−25km
s(d2smt,0.11)
−1
0
1
2
3
d2smt
NegBin Full Ac Only−25km
s(sst,0.68)
−5
−4
−3
−2
−1
0
1
2
sst
NegBin Full Ac Only−25km
s(chla,0.81)
0
2
4
6
chla
NegBin Full Ac Only−25km
s(temp600,0.91)
−1
0
1
2
temp600



Checking Tweedie for comparison

```r
# * Does NOT include sighted acoustic encounters OR spatial
# smoother
require(mgcv)
twS999 <- gam(pa ~ s(bath_m, k = 3) + s(dist, 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 = trainS999, family = tw,
    link = "log", select = TRUE, method = "REML")
summary(twS999)
```

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

Formula:
pa ~ s(bath_m, k = 3) + s(dist, 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) -22.9376     0.1157  -198.3   <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)  8.239e-01      2 2.342 0.016406 *
s(dist)    6.391e-01      2 0.907 0.089417 .
s(d2smt)   2.366e-01      2 0.152 0.254782
s(sst)     7.598e-01      2 1.492 0.045248 *
s(chla)    8.409e-01      2 2.518 0.013254 *
s(temp600) 9.193e-01      2 5.684 0.000324 ***
s(ssh)     9.328e-01      2 6.870 0.000122 ***
s(sshsd)   7.977e-01      2 2.014 0.024321 *
s(eke)     8.695e-01      2 2.945 0.008894 **
s(wavepow) 9.707e-05      2 0.000 0.634825
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.0434   Deviance explained = 9.45%
-REML = 303.69  Scale est. = 1.0402    n = 3659
```

```r
AIC(twS999)
```

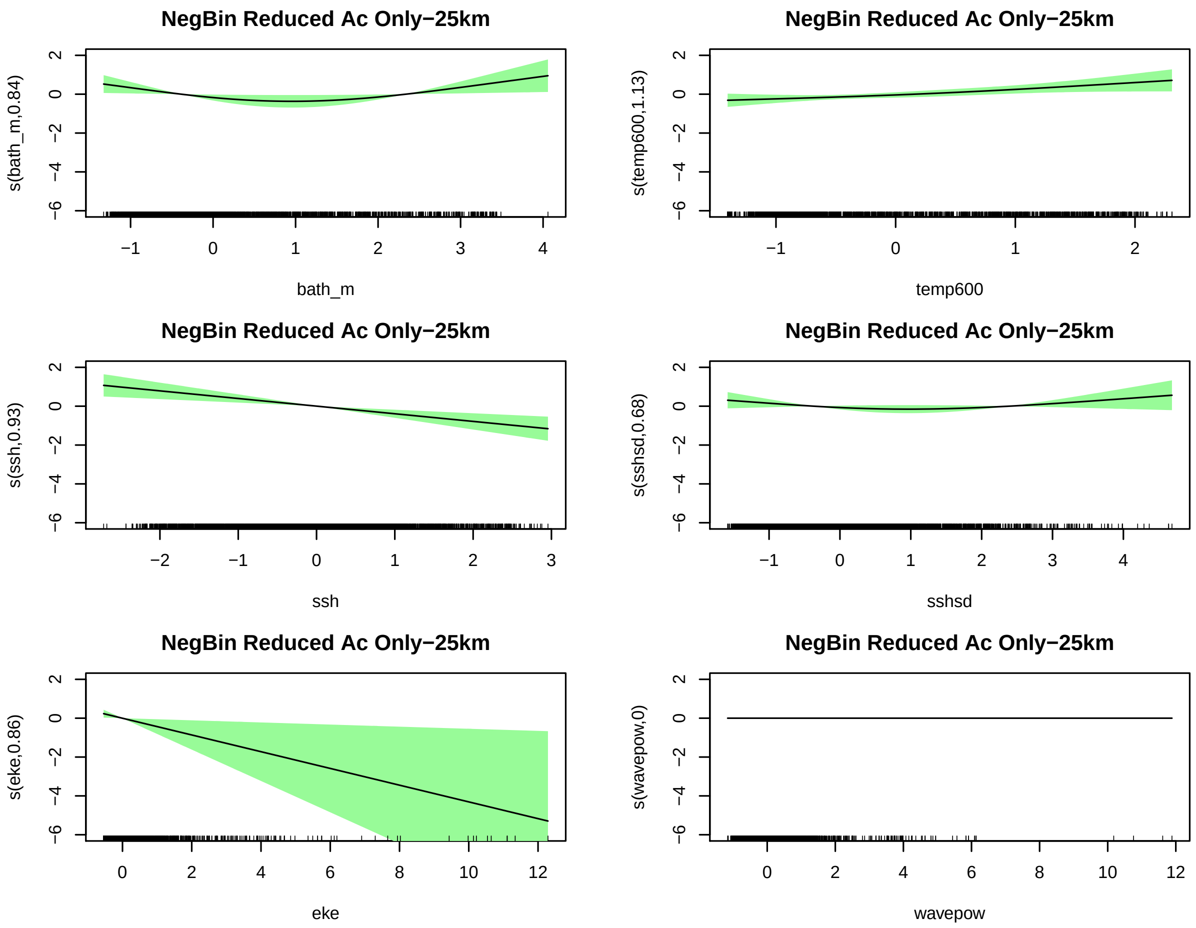
```
[1] 3784.442
```

Reduced Models

- Negative Binomial: higher deviance explained, lower AIC than Tweedie
- Removed non-significant variables:
  - distance to land
  - distance to seamount
  - sst
  - chla
  - eke

```r
# * Does NOT include sighted acoustic encounters

nbS999b <- gam(pa ~ s(bath_m, 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 = trainS999, family = nb, link = "log",
    select = TRUE, method = "REML")
summary(nbS999b)
```

```
Family: Negative Binomial(0.394)
Link function: log

Formula:
pa ~ s(bath_m, 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 z value Pr(>|z|)
(Intercept) -22.8779     0.1156  -197.9   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                 edf Ref.df Chi.sq  p-value
s(bath_m)  8.371e-01      2  5.172 0.012348 *
s(temp600) 1.130e+00      2  7.478 0.003994 **
s(ssh)     9.340e-01      2 13.976 0.000108 ***
s(sshsd)   6.783e-01      2  2.151 0.074580 .
s(eke)     8.574e-01      2  5.250 0.013274 *
s(wavepow) 1.919e-05      2  0.000 0.710937
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.0362   Deviance explained = 8.93%
-REML = 425.63  Scale est. = 1         n = 3659
```

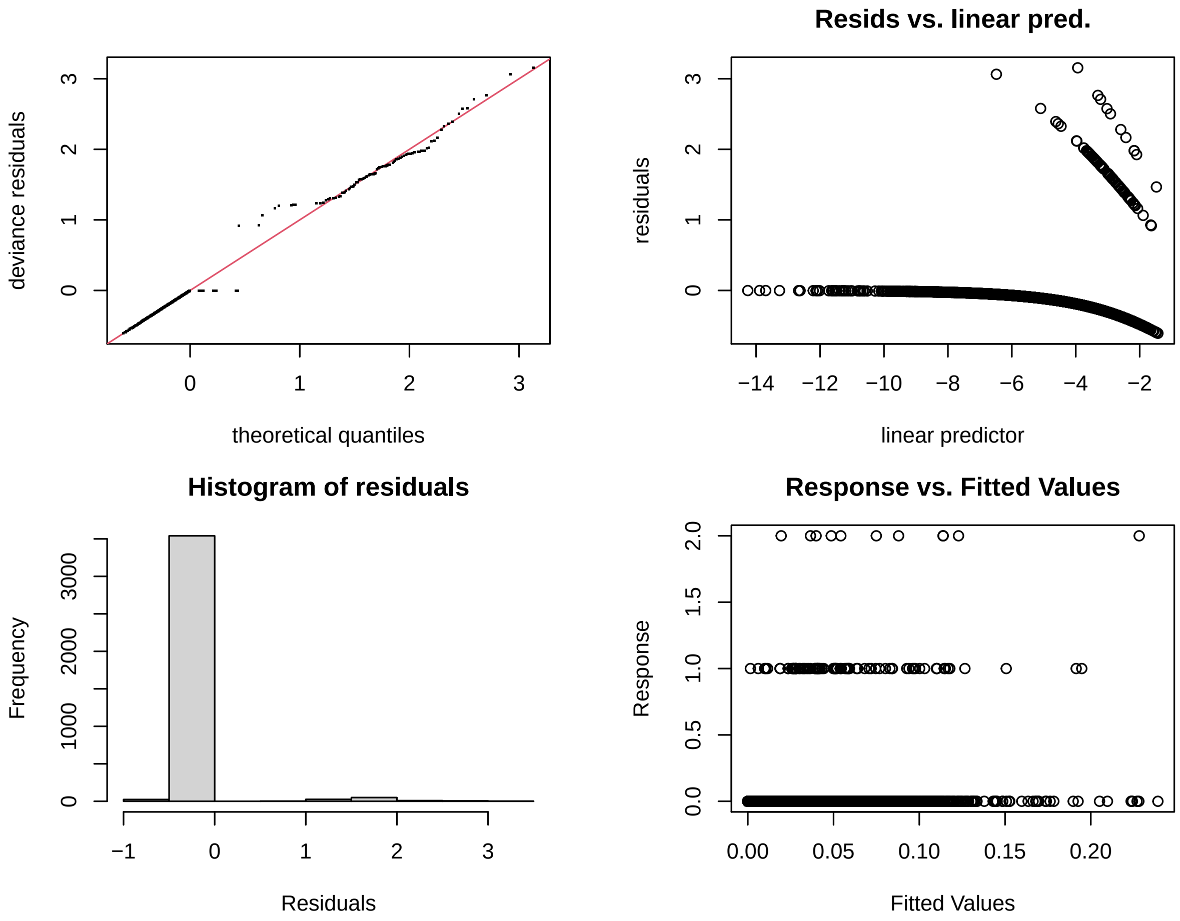


- Remove wave power

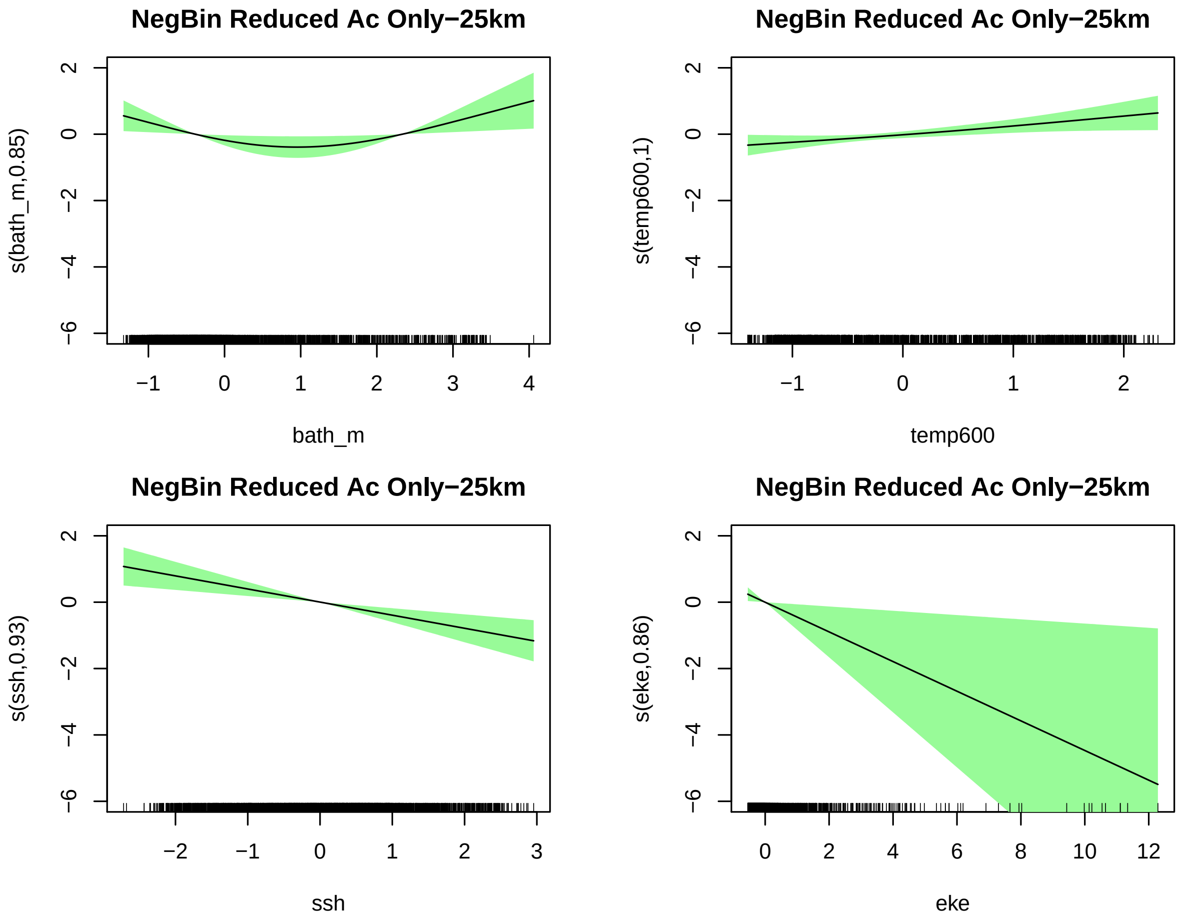
- Remove sshsd

```
# * Does NOT include sighted acoustic encounters

nbS999c <- gam(pa ~ s(bath_m, k = 3) + s(temp600, k = 3) + s(ssh,
    k = 3) + s(eke, k = 3) + offset(log.effort), data = trainS999,
    family = nb, link = "log", select = TRUE, method = "REML")
summary(nbS999c)
```

```
Family: Negative Binomial(0.368)
Link function: log

Formula:
pa ~ s(bath_m, k = 3) + s(temp600, k = 3) + s(ssh, k = 3) + s(eke,
    k = 3) + offset(log.effort)

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -22.8702     0.1157  -197.6   <2e-16 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Approximate significance of smooth terms:
              edf Ref.df Chi.sq  p-value
s(bath_m)  0.8509      2  5.754 0.008932 **
s(temp600) 0.9993      2  6.817 0.005356 **
s(ssh)     0.9343      2 14.059 0.000103 ***
s(eke)     0.8620      2  5.471 0.011726 *
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

R-sq.(adj) =  0.0323   Deviance explained = 8.41%
-REML = 426.13  Scale est. = 1         n = 3659
```



```
Method: REML   Optimizer: outer newton
full convergence after 12 iterations.
Gradient range [-8.70497e-05,5.662692e-06]
(score 426.1267 & scale 1).
Hessian positive definite, eigenvalue range [1.799498e-05,4.765316].
Model rank =  9 / 9

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)   2.000 0.851    0.81    0.04 *
s(temp600) 2.000 0.999    0.72  <2e-16 ***
s(ssh)      2.000 0.934    0.83    0.23
s(eke)      2.000 0.862    0.83    0.18
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```



SET 2

Full Models: Includes s(Longitude,Latitude)

Includes 2D Lat-Lon smoother to account for spatial structure in the data and fit the spatial variation not explained by the other predictors
* Notice that the temperature at 600m is no longer significant compared to the previous models
* Chlorophyll and SSHsd remain significant
+ Does this indicate that they aren't spatially structured and are independent of location?

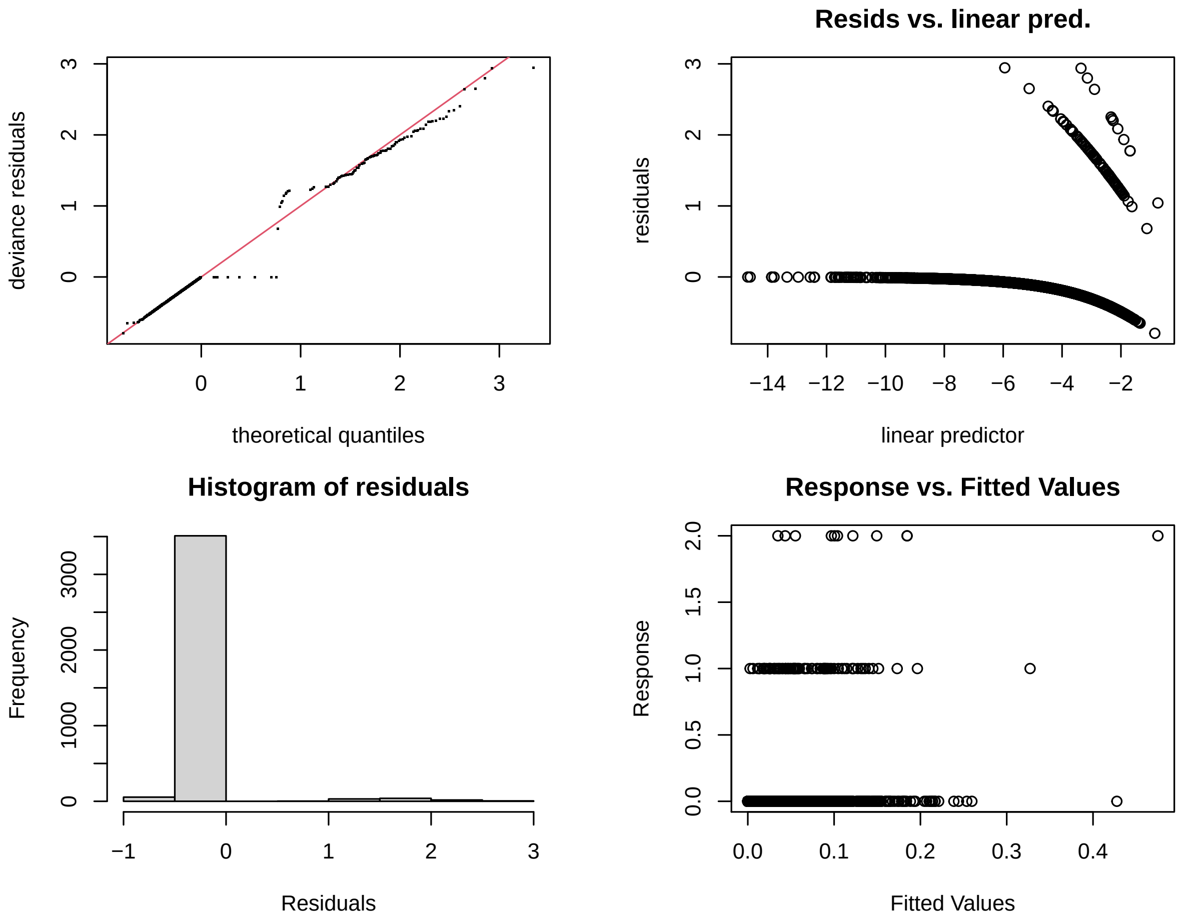
```
nbS999LL <- gam(pa ~ s(Longitude, Latitude) + s(bath_m, k = 3) +
    s(dist, k = 3) + s(d2smt, k = 3) + s(sst, k = 3) + s(chla,
    k = 3) + s(temp600, k = 3) + s(ssh, k = 3) + s(sshsd, k = 3) +
    s(eke, k = 3) + s(wavepow, k = 3) + offset(log.effort), data = trainS999,
    family = nb, link = "log", select = TRUE, method = "REML")
summary(nbS999LL)
```

```
Family: Negative Binomial(0.536)
Link function: log

Formula:
```

```
pa ~ s(Longitude, Latitude) + s(bath_m, k = 3) + s(dist, 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 z value Pr(>|z|)
(Intercept) -23.0247     0.1273  -180.9   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                            edf Ref.df Chi.sq  p-value
s(Longitude,Latitude) 8.124e+00     29 31.315 3.91e-06 ***
s(bath_m)             6.465e-01      2  1.851   0.0843 .
s(dist)               5.540e-01      2  1.041   0.1123
s(d2smt)              5.972e-05      2  0.000   0.3966
s(sst)                5.774e-01      2  1.252   0.1305
s(chla)               7.317e-01      2  2.473   0.0571 .
s(temp600)            4.397e-05      2  0.000   0.6480
s(ssh)                8.204e-01      2  4.476   0.0156 *
s(sshsd)              8.090e-01      2  4.226   0.0213 *
s(eke)                8.457e-01      2  4.737   0.0170 *
s(wavepow)            2.467e-05      2  0.000   0.5642
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.0542   Deviance explained = 15.6%
-REML = 418.76  Scale est. = 1         n = 3659
```

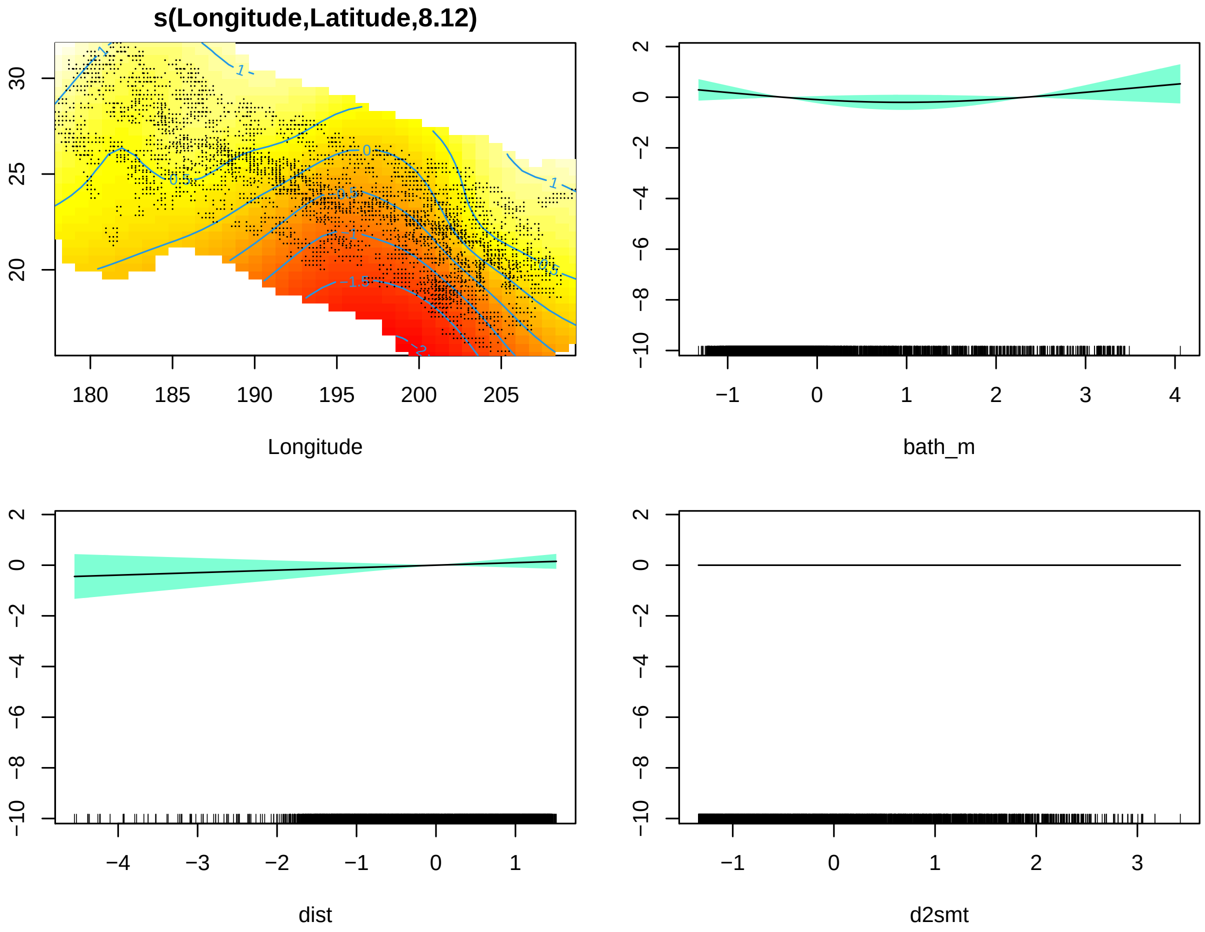
```
AIC(nbS999LL)
```

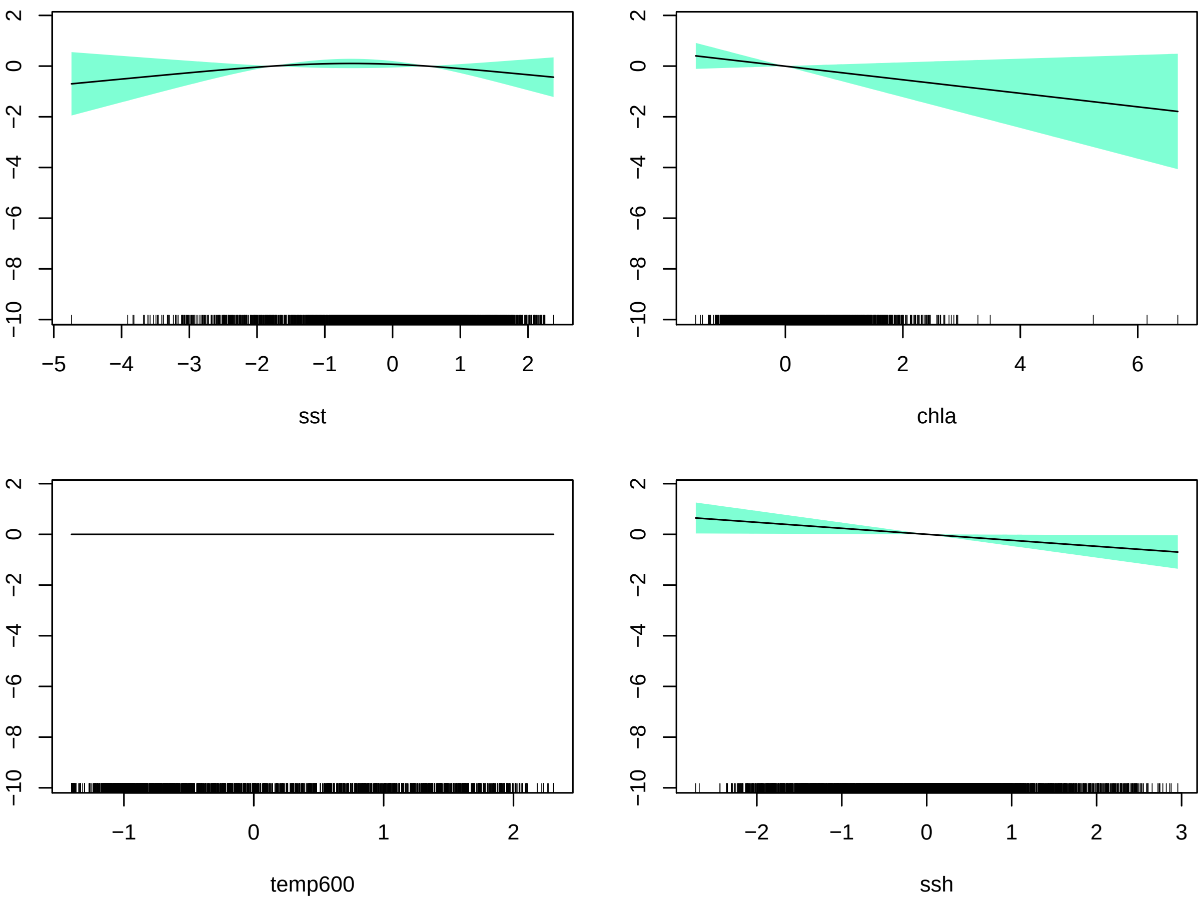
```
[1] 834.7093
```

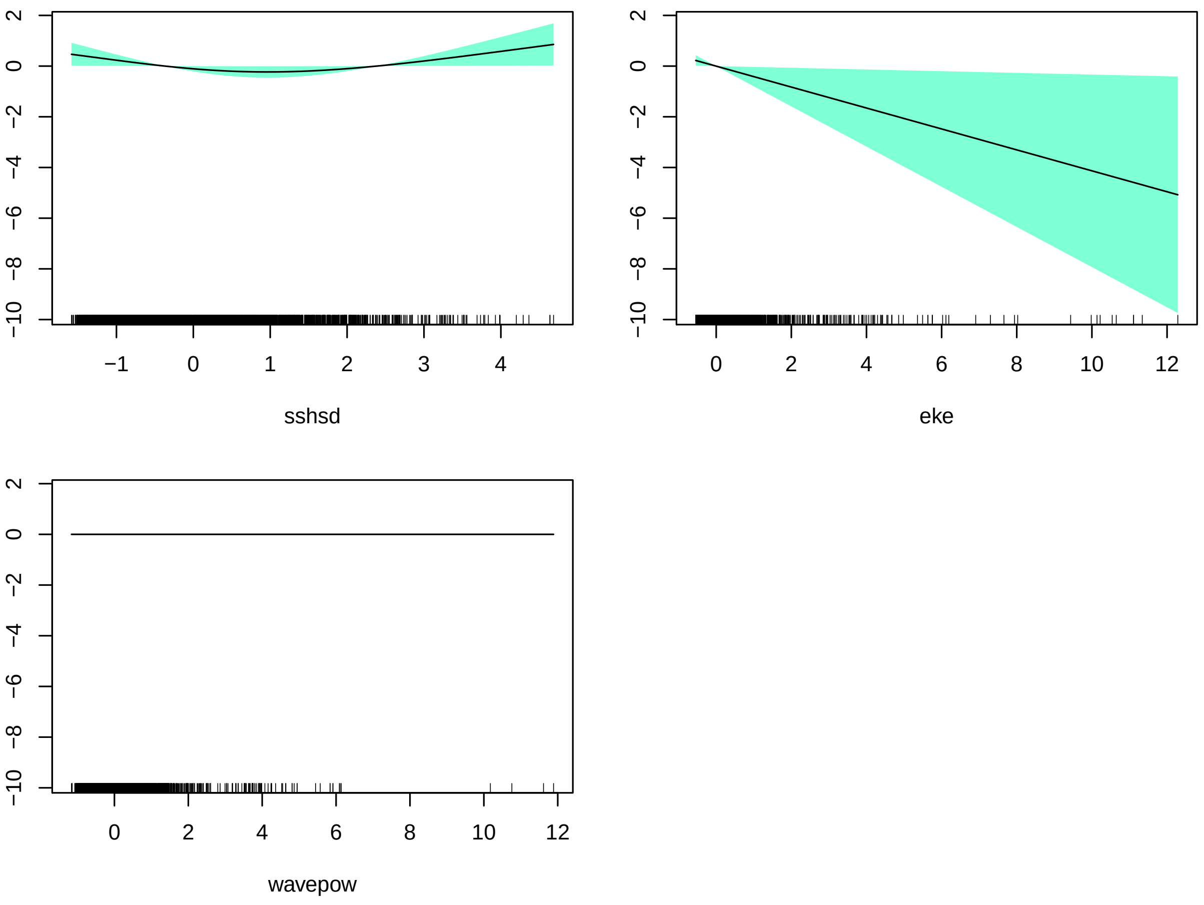


```
Method: REML   Optimizer: outer newton
full convergence after 14 iterations.
Gradient range [-0.0004623193,0.0001049906]
(score 418.7596 & scale 1).
Hessian positive definite, eigenvalue range [2.190574e-06,3.536584].
Model rank =  50 / 50

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(Longitude,Latitude) 2.90e+01 8.12e+00    0.80  <2e-16 ***
s(bath_m)             2.00e+00 6.46e-01    0.83    0.10
s(dist)               2.00e+00 5.54e-01    0.86    0.66
s(d2smt)              2.00e+00 5.97e-05    0.86    0.67
s(sst)                2.00e+00 5.77e-01    0.82    0.03 *
s(chla)               2.00e+00 7.32e-01    0.85    0.26
s(temp600)            2.00e+00 4.40e-05    0.74  <2e-16 ***
s(ssh)                2.00e+00 8.20e-01    0.84    0.29
s(sshsd)              2.00e+00 8.09e-01    0.82    0.02 *
s(eke)                2.00e+00 8.46e-01    0.84    0.20
s(wavepow)            2.00e+00 2.47e-05    0.79  <2e-16 ***
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```




2
0
−2
−4
−6
−8
−10
−5
−4
−3
−2
−1
0
1
2
sst
chla
0
2
4
6
temp600
−1
0
1
2
ssh
−2
−1
0
1
2
3



Checking Tweedie for comparison

```r
twS999LL <- gam(pa ~ s(Longitude, Latitude) + s(bath_m, k = 3) +
    s(dist, 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 = trainS999,
    family = tw, link = "log", select = TRUE, method = "REML")
summary(twS999LL)
```

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

Formula:
pa ~ s(Longitude, Latitude) + s(bath_m, k = 3) + s(dist, 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|)
```

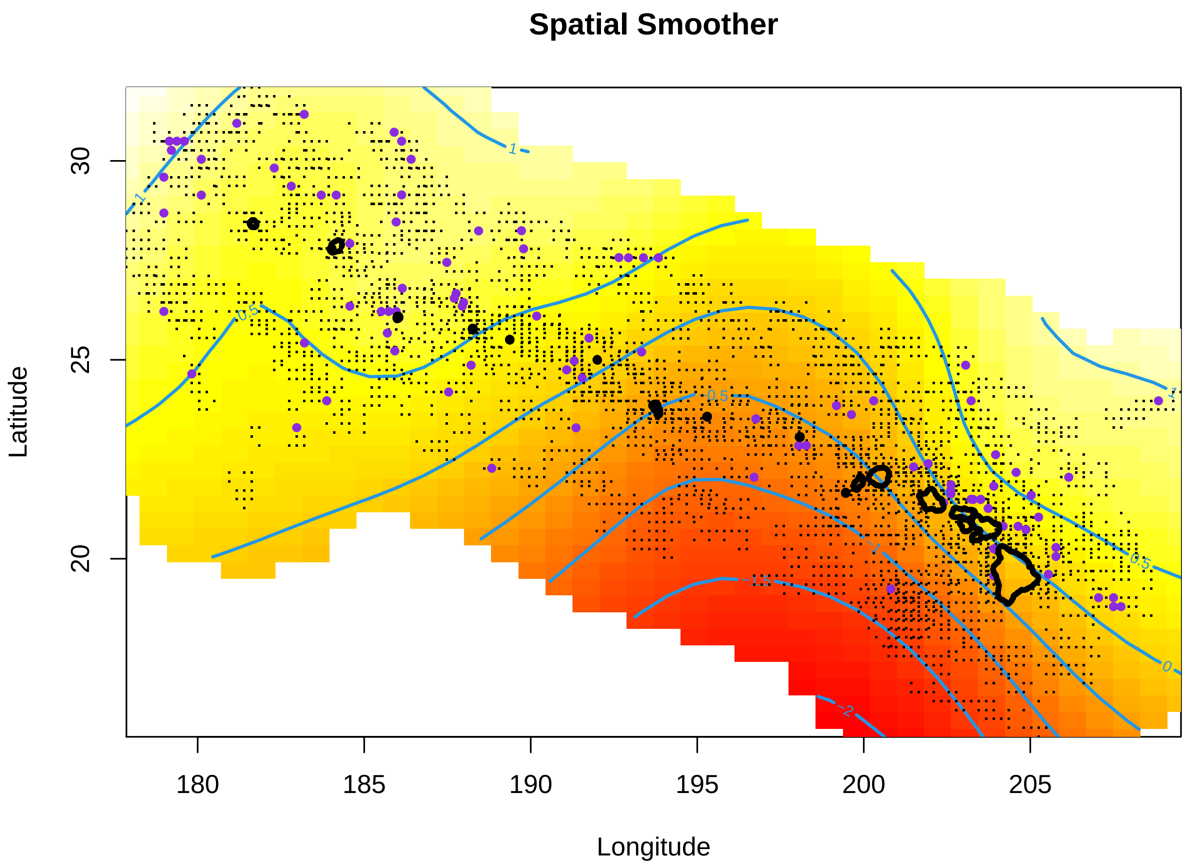


Figure 1: Purple dots represent acoustically detected encounters. Black dots are all data points(grid centroids)

```
(Intercept) -23.0393     0.1255  -183.5   <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(Longitude,Latitude) 9.041e+00     29 1.200 1.62e-06 ***
s(bath_m)             7.025e-01      2 1.182  0.06115 .  
s(dist)               5.511e-01      2 0.510  0.10537    
s(d2smt)              1.023e-01      2 0.058  0.26724    
s(sst)                6.318e-01      2 0.796  0.10296    
s(chla)               7.613e-01      2 1.476  0.04146 *  
s(temp600)            9.725e-05      2 0.000  0.62060    
s(ssh)                8.220e-01      2 2.298  0.01453 *  
s(sshsd)              8.495e-01      2 2.882  0.00865 ** 
s(eke)                8.548e-01      2 2.584  0.01319 *  
s(wavepow)            2.367e-04      2 0.000  0.42850    
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

R-sq.(adj) =  0.0577   Deviance explained = 14.4%
-REML = 297.21  Scale est. = 1.0388    n = 3659
```

```r
AIC(twS999LL)
```

```
[1] 3635.925
```

Reduced Models

- Negative Binomial: higher explained deviance, lower AIC than Tweedie
- Keep:
  - Lon, Lat
  - SSH

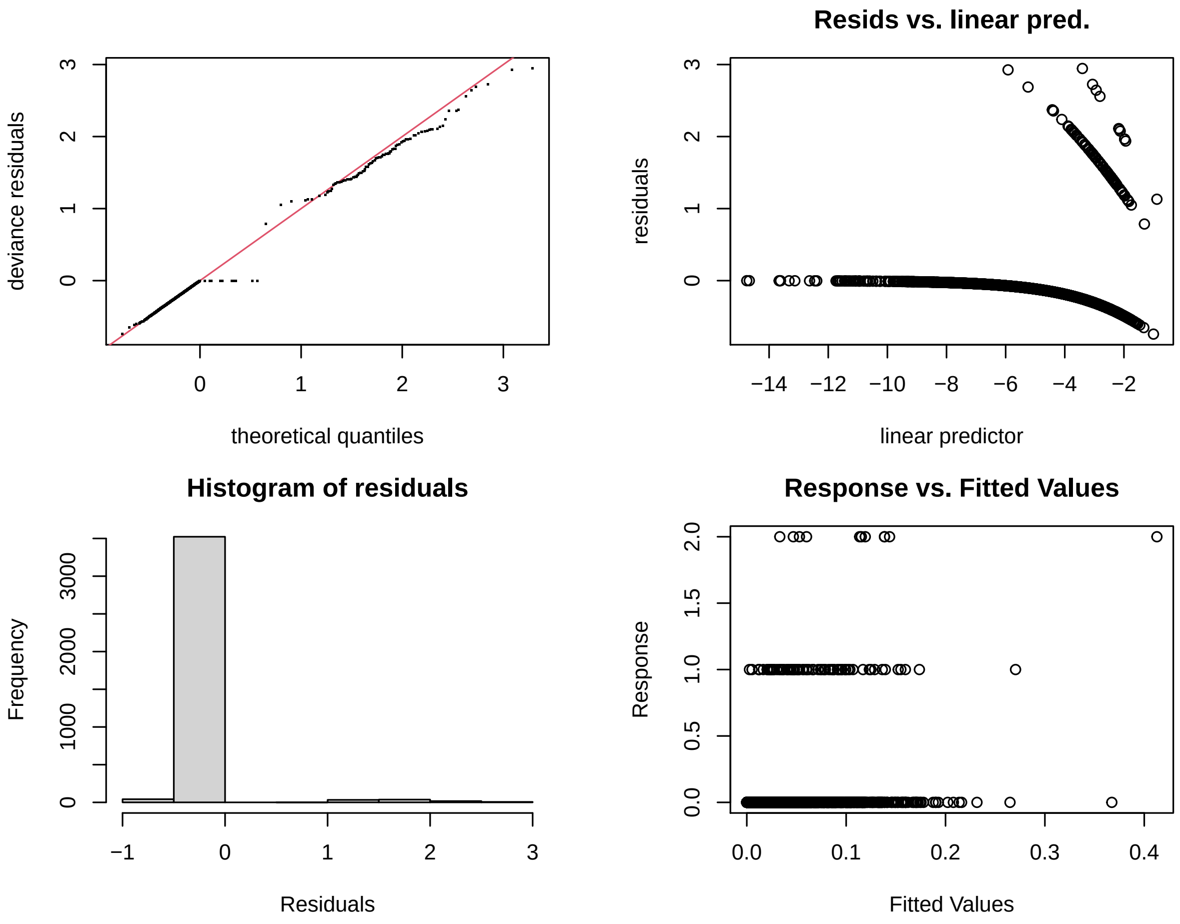
  - SSHsd

  - EKE

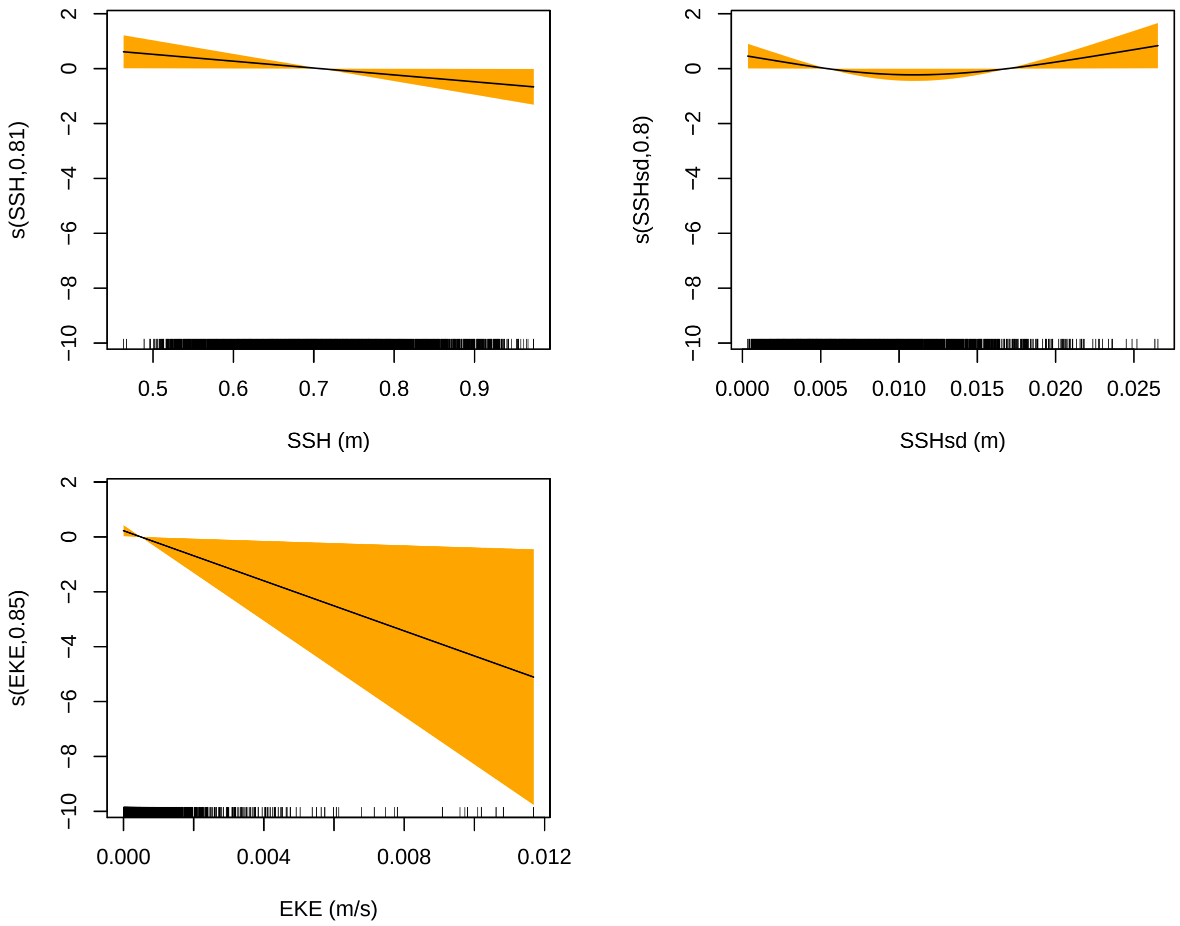
```r
colnames(trainS999)[38] <- "SSH"
colnames(trainS999)[40] <- "EKE"
colnames(trainS999)[39] <- "SSHsd"
nbS999LLb2 <- gam(pa ~ s(Longitude, Latitude) + s(SSH, k = 3) +
    s(SSHsd, k = 3) + s(EKE, k = 3) + offset(log.effort), data = trainS999,
    family = nb, link = "log", select = TRUE, method = "REML")
summary(nbS999LLb2)
```

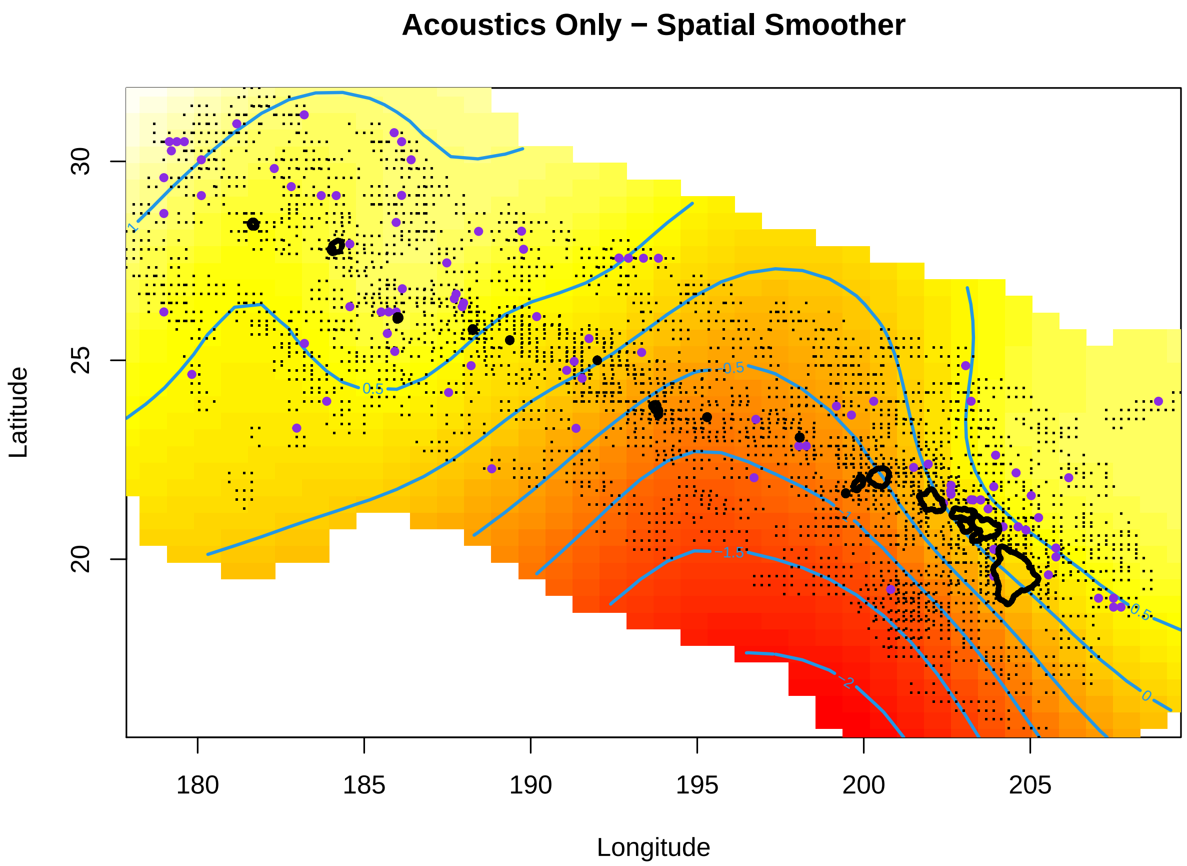
```
Family: Negative Binomial(0.502) 
Link function: log 

Formula:
pa ~ s(Longitude, Latitude) + s(SSH, k = 3) + s(SSHsd, k = 3) + 
    s(EKE, k = 3) + offset(log.effort)

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -23.0007     0.1257    -183   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                         edf Ref.df Chi.sq  p-value
s(Longitude,Latitude) 9.3310     29 38.954 3.24e-07 ***
s(SSH)                0.8082      2  4.178   0.0188 *
s(SSHsd)              0.8044      2  4.127   0.0226 *
s(EKE)                0.8465      2  4.826   0.0162 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.0499   Deviance explained = 14.7%
-REML =  419.8  Scale est. = 1         n = 3659
```

```r
# model diagnostics
par(mar = c(4, 4, 3, 3), mfrow = c(2, 2))
gam.check(nbS999LLb2)
```



```
Method: REML   Optimizer: outer newton
full convergence after 13 iterations.
Gradient range [-0.0001047822,5.191454e-06]
(score 419.802 & scale 1).
Hessian positive definite, eigenvalue range [8.055455e-06,3.76634].
Model rank =  36 / 36

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(Longitude,Latitude) 29.000  9.331    0.80  <2e-16 ***
s(SSH)                 2.000  0.808    0.84    0.21
s(SSHsd)               2.000  0.804    0.82    0.02 *
s(EKE)                 2.000  0.847    0.84    0.25
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```





Acoustics Only – Spatial Smoother

Predict Test Data

```
require(magrittr)
require(dplyr)

#### For twS999c, no spatial smoother ####
nbTrainFinal <- trainS999 %>% mutate(resid = resid(nbS999c),
    predict = predict(nbS999c))
predTrain <- predict.gam(nbS999c, type = "response")  #calculate MSE for these to compare with test set
nbTrainFinal$fit <- predTrain

# using scale of 0,1,2 makes this hard to interpret
nbMSEtrain <- mean((nbTrainFinal$pa - nbTrainFinal$fit)^2)  #MSE
# mean(abs((nbTrainFinal$pa - nbTrainFinal$fit))) #Mean
# absolute error Calculate MSE AFTER transforming the
# predictions back to the same scale as the observed data

nbPred <- predict.gam(nbS999c, newdata = testS999, type = "response",
    se.fit = TRUE)
nbTestFinal <- data.frame(testS999, fit = nbPred$fit, se.fit = nbPred$se.fit)
nbMSEtest <- mean((nbTestFinal$pa - nbTestFinal$fit)^2)  #MSE
```

```
# mean(abs((testFinal$pa - testFinal$fit))) #Mean absolute
# error


#### For nbS999LLb2, with spatial smoother #### pulling the
#### prediction and residual data from the model
nbTrainLL <- trainS999 %>% mutate(resid = resid(nbS999LLb2),
    predict = predict(nbS999LLb2))
predTrainLL <- predict.gam(nbS999LLb2, type = "response")  #calculate MSE for these to compare with tes
nbTrainLL$fit <- predTrainLL

# using scale of 0,1,2 makes this hard to interpret
nbMSEtrainLL <- mean((nbTrainLL$pa - nbTrainLL$fit)^2)  #MSE
# mean(abs((nbTrainFinal$pa - nbTrainFinal$fit))) #Mean
# absolute error Calculate MSE AFTER transforming the
# predictions back to the same scale as the observed data
colnames(testS999)[40] <- "EKE"
colnames(testS999)[38] <- "SSH"
colnames(testS999)[39] <- "SSHsd"
nbPredLL <- predict.gam(nbS999LLb2, newdata = testS999, type = "response",
    se.fit = TRUE)
nbTestLL <- data.frame(testS999, fit = nbPredLL$fit, se.fit = nbPredLL$se.fit)
nbMSEtestLL <- mean((nbTestLL$pa - nbTestLL$fit)^2)  #MSE

# mean(abs((testFinal$pa - testFinal$fit))) #Mean absolute
# error

# AIC
nbAIC <- AIC(nbS999c)
nbAICLL <- AIC(nbS999LLb2)

# Explained Deviance
nbExpDev = round(((nbS999c$null.deviance - nbS999c$deviance)/nbS999c$null.deviance) *
    100, 2)
nbExpDevLL = round(((nbS999LLb2$null.deviance - nbS999LLb2$deviance)/nbS999LLb2$null.deviance) *
    100, 2)
```

```
# make summary table of metrics

table = matrix(NA, nrow = 2, ncol = 5)
colnames(table) = c("Best Models", "ExpDev", "AIC", "MSEtrain",
    "MSEtest")

# enter info by row

table[1, ] <- c("nbS999c", paste0(nbExpDev, "%"), round(nbAIC,
    2), round(nbMSEtrain, 3), round(nbMSEtest, 3))

table[2, ] <- c("nbS999LLc", paste0(nbExpDevLL, "%"), round(nbAICLL,
    2), round(nbMSEtrainLL, 3), round(nbMSEtestLL, 3))
require(knitr)
kable(table, caption = "Negative Binomial Model Summary Metrics")
```

Table 1: Negative Binomial Model Summary Metrics

Best Models	ExpDev	AIC	MSEtrain	MSEtest
nbS999c	8.41%	849.45	0.033	0.031
nbS999LLc	14.74%	834.36	0.032	0.03

Conclusions

The best-fit models built using the 'Acoustics Only' data set included the negative binomial distribution as they resulted to higher explained deviance and lower AIC values when compared to the Tweedie distribution. The models with the spatial smoother (Set 2) also performed better overall, yielding the spatial smoother, SSH, SSHsd and EKE as the significant predictor variables.

For the Set 2 models, there is a negative relationship between sperm whale encounters and SSH and EKE. More sperm whale encounters occurred at 0.5 m and gradually declined as SSH increased to 0.9 m. Lower EKE values (0-0.001 m/s) related to higher encounter rates and then showed a steep decline as EKE increased. THe variation in SSH (SSHsd) shows a decrease in sperm whale encounters as it approaches 0.012 m, with more sperm whale encounters occurring at SSHsd of 0 m and 0.025 m. This suggests that the acoustically-detected sperm whale groups are potentially more likely to occur in areas of downwelling (as indicated by the lower SSH) and with weaker current velocities (negative EKE relationship). The relationship with SSHsd is interesting, as the slight dip in the plot indicates some amount of variation in SSH that is not as ideal as no variation or higher variation.

The negative relationships with SSH and EKE suggest that sperm whale groups occur in potentially calmer areas near downwelling zones, or zones with relatively lower SSH at the instance the whales are detected. Keeping in mind the 8 km spatial resolution of the monthly averaged data, it's hard to fully explain whether the significant variables suggest that the acoustically-detected whales are in more productive or less productive regions.