Statistical analyses - simulation outputs

J. Guibert & A. Dupaix

20/11/2022 - update 23/05/2024 - 17/12/2024

Contents

Required data sets

```
WD <- ".."
PATH_OUTPUT <- file.path(WD, "Outputs_sim_smallarea")
PATH_FUNC <- file.path(WD, "Functions")
source(file.path(PATH_FUNC, "stepAIC_gam.R"))

NLOG_VE <- read.csv(file.path(PATH_OUTPUT, "NLOG_VE.csv"), head = T)

NLOG_VE$logNLOG <- log(NLOG_VE$nlogmean)
NLOG_VE$chlacr <- scale(NLOG_VE$chlamean)
NLOG_VE$slacr <- scale(NLOG_VE$slamean)
NLOG_VE$SSCIcr <- scale(NLOG_VE$SSCImean)
NLOG_VE$FSLEcr <- scale(NLOG_VE$FSLEmean)
NLOG_VE$MNcr <- scale(NLOG_VE$MNmean)

NLOG_VE_sup_zero_Moz <- NLOG_VE %>% dplyr::filter(Zone == 'MOZ', nlogmean > 0)
NLOG_VE_sup_zero_North <- NLOG_VE %>% dplyr::filter(Zone == 'WIO', nlogmean > 0)

# NLOG_VE_zero <- read.csv(file.path(PATH_OUTPUT, "NLOG_VE_zero.csv"), head = T)
#
# NLOG_VE_zero_Moz <- read.csv(file.path(PATH_OUTPUT, "NLOG_VE_zero_Moz.csv"), head = T)
# NLOG_VE_zero_North <- read.csv(file.path(PATH_OUTPUT, "NLOG_VE_zero_North.csv"), head = T)
#
# dfMN_epi<-read.csv(file.path(PATH_OUTPUT, "MN_epi_mean.csv"), header = T)
# dfMN_u<-read.csv(file.path(PATH_OUTPUT, "MN_umeso_mean.csv"), header = T)
# dfMN_mu<-read.csv(file.path(PATH_OUTPUT, "MN_mumeso_mean.csv"), header = T)
# dfMN_ml<-read.csv(file.path(PATH_OUTPUT, "MN_mlmeso_mean.csv"), header = T)
# dfMN_hml<-read.csv(file.path(PATH_OUTPUT, "MN_hmlmeso_mean.csv"), header = T)

# df_eff <- read.csv(file.path(PATH_OUTPUT, "df_eff.csv"), head = T)
# df_eff_new <- read.csv(file.path(PATH_OUTPUT, "df_eff_new_100.csv"), head = T)
# df_eff_new_50 <- read.csv(file.path(PATH_OUTPUT, "df_eff_new_50.csv"), head = T)
# df_eff_new_150 <- read.csv(file.path(PATH_OUTPUT, "df_eff_new_150.csv"), head = T)

# For the Sensitivity analysis (T = 10)

# NLOG_VE_zero_Moz_10 <- NLOG_VE_zero_Moz[NLOG_VE_zero_Moz$NumOBS>=10,]
# NLOG_VE_zero_North_10 <- NLOG_VE_zero_North[NLOG_VE_zero_North$NumOBS>=10,]
# NLOG_VE_sup_zero_Moz_10 <- NLOG_VE_sup_zero_Moz[NLOG_VE_sup_zero_Moz$NumOBS>=10,]
# NLOG_VE_sup_zero_North_10 <- NLOG_VE_sup_zero_North[NLOG_VE_sup_zero_North$NumOBS>=10,]
```

I : Correlations between environmental variables (Kendall tests)

Related to Figure A1 (values not directly displayed in the paper nor in the Appendix)

```
# cor.test(NLOG_VE$chlamean, NLOG_VE$sstmean, method = "kendall")
# cor.test(NLOG_VE$chlamean, NLOG_VE$slamean, method = "kendall")
# cor.test(NLOG_VE$chlamean, NLOG_VE$SSCImean, method = "kendall")
# cor.test(NLOG_VE$chlamean, NLOG_VE$FSLEmean, method = "kendall")
# cor.test(NLOG_VE$chlamean, NLOG_VE$MNmean, method = "kendall")
# cor.test(NLOG_VE$sstmean, NLOG_VE$slamean, method = "kendall")
# cor.test(NLOG_VE$sstmean, NLOG_VE$SSCImean, method = "kendall")
```

```
# cor.test(NLOG_VE$sstmean, NLOG_VE$FSLEmean, method = "kendall")
# cor.test(NLOG_VE$sstmean, NLOG_VE$MNmean, method = "kendall")
# cor.test(NLOG_VE$slamean, NLOG_VE$SSCImean, method = "kendall")
# cor.test(NLOG_VE$slamean, NLOG_VE$FSLEmean, method = "kendall")
# cor.test(NLOG_VE$slamean, NLOG_VE$MNmean, method = "kendall")
# cor.test(NLOG_VE$SSCImean, NLOG_VE$MNmean, method = "kendall")
# cor.test(NLOG_VE$FSLEmean, NLOG_VE$SSCImean, method = "kendall")
# cor.test(NLOG_VE$FSLEmean, NLOG_VE$MNmean, method = "kendall")
```

III : Correlations between NLOG and environmental variables (Kendall tests)

Related to Figure 4 and Table A2

```
cor.test(NLOG_VE_sup_zero_Moz$nlogmean, NLOG_VE_sup_zero_Moz$chlamean, method = "kendall")
```

```
##
##  Kendall's rank correlation tau
##
## data:  NLOG_VE_sup_zero_Moz$nlogmean and NLOG_VE_sup_zero_Moz$chlamean
## z = -7.8816, p-value = 3.231e-15
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##         tau
## -0.03506033
```

```
cor.test(NLOG_VE_sup_zero_North$nlogmean, NLOG_VE_sup_zero_North$chlamean, method = "kendall")
```

```
##
##  Kendall's rank correlation tau
##
## data:  NLOG_VE_sup_zero_North$nlogmean and NLOG_VE_sup_zero_North$chlamean
## z = 25.805, p-value < 2.2e-16
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##       tau
## 0.03984424
```

```
cor.test(NLOG_VE_sup_zero_Moz$nlogmean, NLOG_VE_sup_zero_Moz$slamean, method = "kendall")
```

```
##
##  Kendall's rank correlation tau
##
## data:  NLOG_VE_sup_zero_Moz$nlogmean and NLOG_VE_sup_zero_Moz$slamean
## z = -20.578, p-value < 2.2e-16
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##         tau
## -0.09153637
```

```
cor.test(NLOG_VE_sup_zero_North$nlogmean, NLOG_VE_sup_zero_North$slamean, method = "kendall")
```

```
##
##  Kendall's rank correlation tau
##
```

```
## data:  NLOG_VE_sup_zero_North$nlogmean and NLOG_VE_sup_zero_North$slamean
## z = 35.724, p-value < 2.2e-16
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##       tau
## 0.05515882
```

```
cor.test(NLOG_VE_sup_zero_Moz$nlogmean, NLOG_VE_sup_zero_Moz$SSCImean, method = "kendall")
```

```
##
##  Kendall's rank correlation tau
##
## data:  NLOG_VE_sup_zero_Moz$nlogmean and NLOG_VE_sup_zero_Moz$SSCImean
## z = -36.686, p-value < 2.2e-16
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##       tau
## -0.1631913
```

```
cor.test(NLOG_VE_sup_zero_North$nlogmean, NLOG_VE_sup_zero_North$SSCImean, method = "kendall")
```

```
##
##  Kendall's rank correlation tau
##
## data:  NLOG_VE_sup_zero_North$nlogmean and NLOG_VE_sup_zero_North$SSCImean
## z = 30.12, p-value < 2.2e-16
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##       tau
## 0.04650593
```

```
cor.test(NLOG_VE_sup_zero_Moz$nlogmean, NLOG_VE_sup_zero_Moz$FSLEmean, method = "kendall")
```

```
##
##  Kendall's rank correlation tau
##
## data:  NLOG_VE_sup_zero_Moz$nlogmean and NLOG_VE_sup_zero_Moz$FSLEmean
## z = -59.621, p-value < 2.2e-16
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##       tau
## -0.2652155
```

```
cor.test(NLOG_VE_sup_zero_North$nlogmean, NLOG_VE_sup_zero_North$FSLEmean, method = "kendall")
```

```
##
##  Kendall's rank correlation tau
##
## data:  NLOG_VE_sup_zero_North$nlogmean and NLOG_VE_sup_zero_North$FSLEmean
## z = -86.459, p-value < 2.2e-16
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##       tau
## -0.1334952
```

```
cor.test(NLOG_VE_sup_zero_Moz$nlogmean, NLOG_VE_sup_zero_Moz$MNmean, method = "kendall")
```

```
##
```

```
##  Kendall's rank correlation tau
##
## data:  NLOG_VE_sup_zero_Moz$nlogmean and NLOG_VE_sup_zero_Moz$MNmean
## z = 51.213, p-value < 2.2e-16
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##       tau
## 0.2278138
```

```
cor.test(NLOG_VE_sup_zero_North$nlogmean, NLOG_VE_sup_zero_North$MNmean, method = "kendall")
```

```
##
##  Kendall's rank correlation tau
##
## data:  NLOG_VE_sup_zero_North$nlogmean and NLOG_VE_sup_zero_North$MNmean
## z = 109.76, p-value < 2.2e-16
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##       tau
## 0.1694663
```

```
wilcox.test(NLOG_VE_sup_zero_Moz$nlogmean, NLOG_VE_sup_zero_North$nlogmean)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  NLOG_VE_sup_zero_Moz$nlogmean and NLOG_VE_sup_zero_North$nlogmean
## W = 3.279e+09, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

IV : Construction of models

GAM for the WIO zone

```
GAM_North <- mgcv::gam(logNLOG ~ s(chlacr, k = 6) + s(slacr, k = 6) +
                         s(SSCIcr, k = 6) + s(FSLEcr, k = 6) + s(MNcr, k = 6),
                       data = NLOG_VE_sup_zero_North)
summary(GAM_North)
```

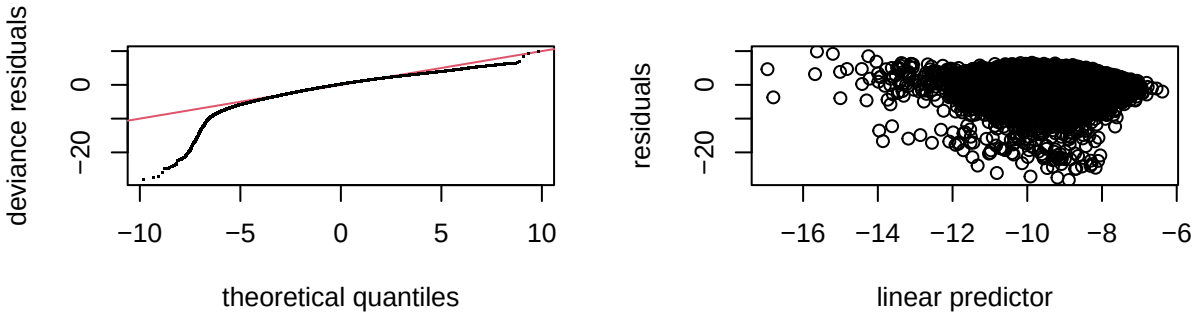
```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## logNLOG ~ s(chlacr, k = 6) + s(slacr, k = 6) + s(SSCIcr, k = 6) +
##     s(FSLEcr, k = 6) + s(MNcr, k = 6)
##
## Parametric coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -9.564231   0.004999   -1913   <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
```

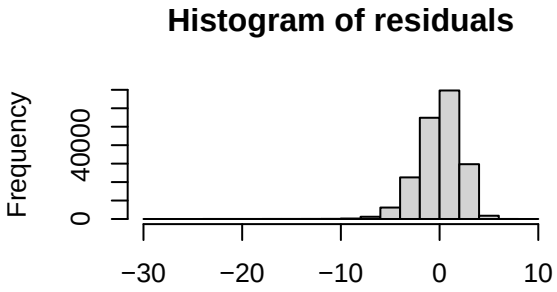
```
## s(chlacr) 4.982  5.000  354.0  <2e-16 ***
## s(slacr)  4.981  5.000  338.8  <2e-16 ***
## s(SSCIcr) 4.270  4.768  175.4  <2e-16 ***
## s(FSLEcr) 4.982  5.000  572.8  <2e-16 ***
## s(MNcr)   4.954  4.999 1901.6  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.104   Deviance explained = 10.5%
## GCV = 4.6592  Scale est. = 4.6586    n = 186429
```

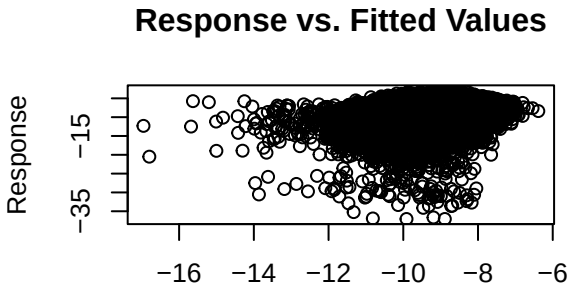
```
GAM_North2 <- stepAIC.gam(GAM_North)
summary(GAM_North2)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## logNLOG ~ s(chlacr, k = 6) + s(slacr, k = 6) + s(SSCIcr, k = 6) +
##     s(FSLEcr, k = 6) + s(MNcr, k = 6)
##
## Parametric coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -9.564231   0.004999   -1913   <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(chlacr) 4.982  5.000  354.0  <2e-16 ***
## s(slacr)  4.981  5.000  338.8  <2e-16 ***
## s(SSCIcr) 4.270  4.768  175.4  <2e-16 ***
## s(FSLEcr) 4.982  5.000  572.8  <2e-16 ***
## s(MNcr)   4.954  4.999 1901.6  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.104   Deviance explained = 10.5%
## GCV = 4.6592  Scale est. = 4.6586    n = 186429
```

```
par(mfrow = c(2,2))
gam.check(GAM_North2)
```

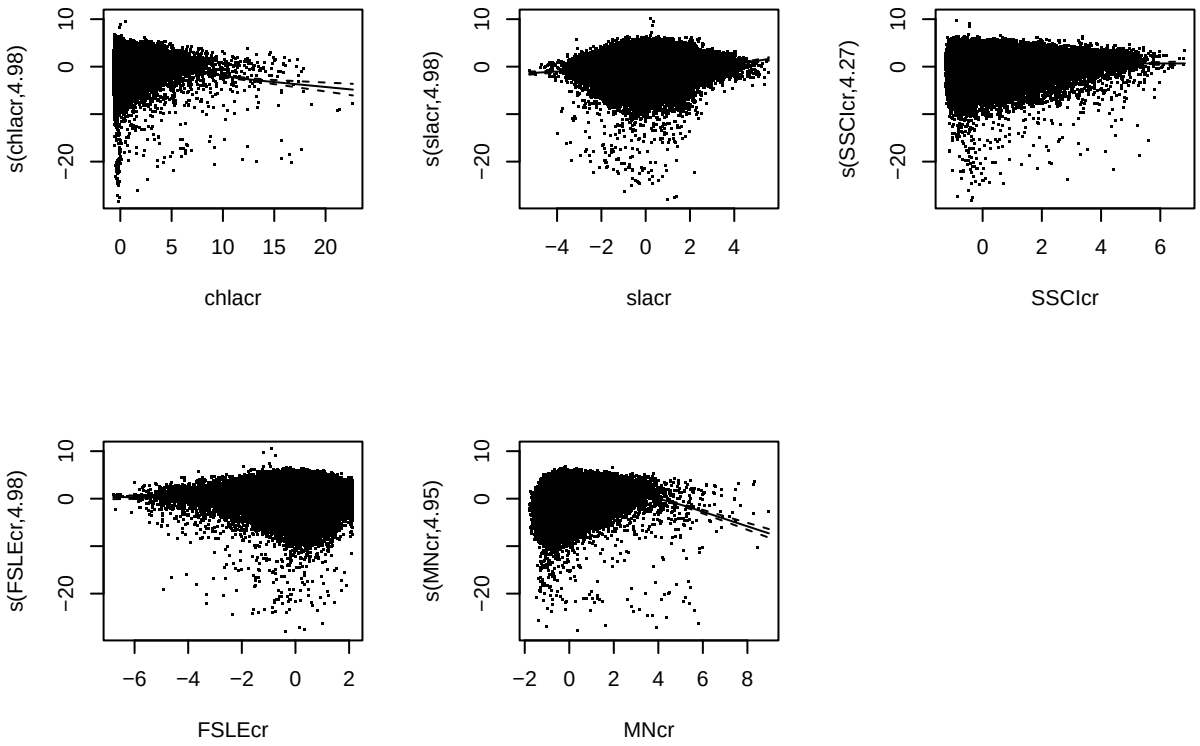







```
##
## Method: GCV   Optimizer: magic
## Smoothing parameter selection converged after 13 iterations.
## The RMS GCV score gradient at convergence was 3.14294e-07 .
## The Hessian was positive definite.
## Model rank =  26 / 26
##
## Basis dimension (k) checking results. Low p-value (k-index<1) may
## indicate that k is too low, especially if edf is close to k'.
##
##              k'  edf k-index p-value
## s(chlacr) 5.00 4.98    1.00    0.41
## s(slacr)  5.00 4.98    1.01    0.69
## s(SSCIcr) 5.00 4.27    1.00    0.41
## s(FSLEcr) 5.00 4.98    1.02    0.95
## s(MNcr)   5.00 4.95    0.99    0.23
```

```
plot(GAM_North2,residuals=T,pages=1)
```



Linear model for the MOZ zone

```
LM_Moz <- lm(logNLOG ~ chlacr + slacr + SSCIcr + FSLEcr + MNcr, data = NLOG_VE_sup_zero_Moz)
LM_Moz2 <- stepAIC(LM_Moz)
```

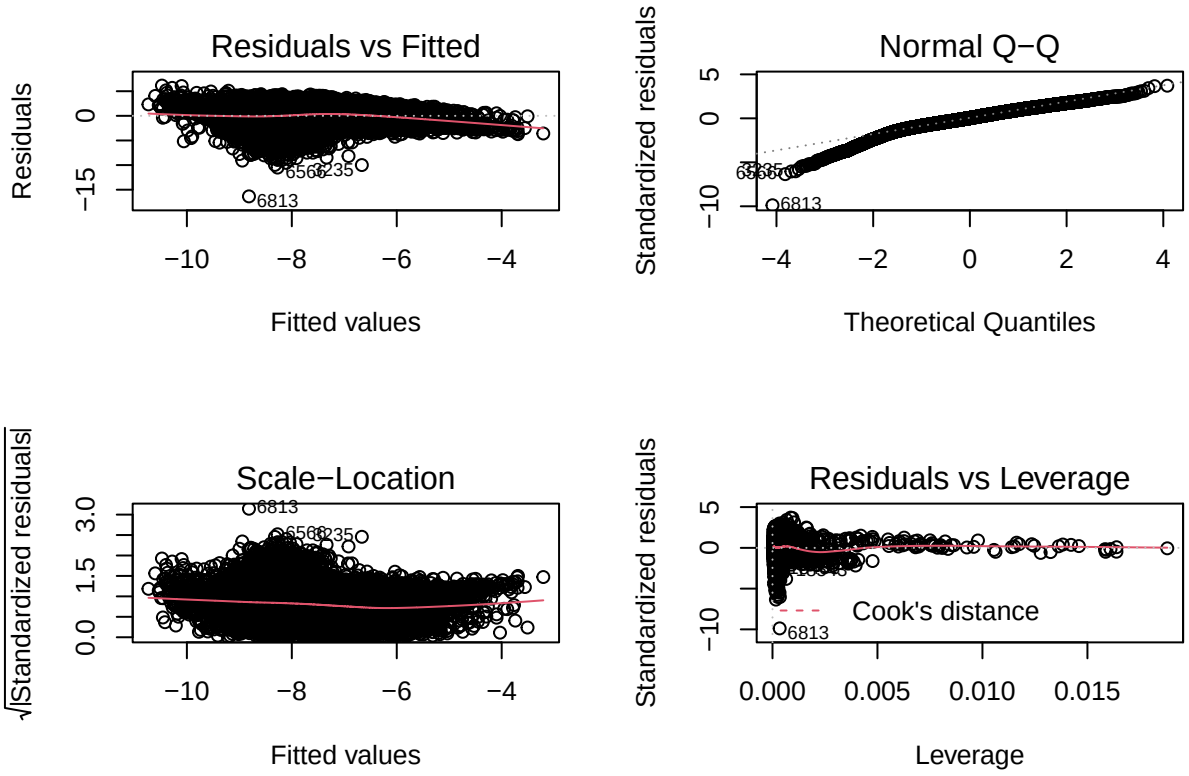
```
## Start:  AIC=22657.02
## logNLOG ~ chlacr + slacr + SSCIcr + FSLEcr + MNcr
##
##          Df Sum of Sq   RSS   AIC
## <none>                61558 22657
## - slacr   1     378.3 61936 22793
## - chlacr  1     620.1 62178 22880
## - MNcr    1    2227.5 63785 23454
## - SSCIcr  1    3368.0 64926 23852
## - FSLEcr  1   10325.0 71883 26138
```

```
summary(LM_Moz2)
```

```
##
## Call:
## lm(formula = logNLOG ~ chlacr + slacr + SSCIcr + FSLEcr + MNcr,
##     data = NLOG_VE_sup_zero_Moz)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.3536  -0.9314   0.0611   1.1320   6.1403
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -8.036415   0.019427 -413.68   <2e-16 ***
```

```
## chlacr       0.215837   0.014350   15.04   <2e-16 ***
## slacr       -0.126574   0.010775  -11.75   <2e-16 ***
## SSCIcr      -0.369376   0.010538  -35.05   <2e-16 ***
## FSLEcr      -0.613276   0.009992  -61.38   <2e-16 ***
## MNcr         0.205751   0.007217   28.51   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.656 on 22458 degrees of freedom
## Multiple R-squared:  0.2539, Adjusted R-squared:  0.2538
## F-statistic:  1529 on 5 and 22458 DF,  p-value: < 2.2e-16
```

```
par(mfrow = c(2, 2))
plot(LM_Moz2)
```



GAM for the MOZ zone

```
GAM_Moz <- mgcv::gam(logNLOG ~ s(chlacr, k = 6) + s(slacr, k = 6) +
                       s(SSCIcr, k = 6) + s(FSLEcr, k = 6) + s(MNcr, k = 6),
                     data = NLOG_VE_sup_zero_Moz)
summary(GAM_Moz)
```

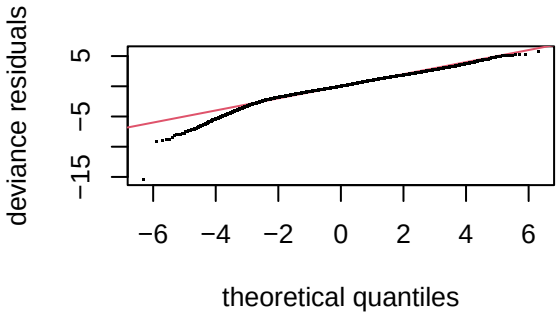
```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## logNLOG ~ s(chlacr, k = 6) + s(slacr, k = 6) + s(SSCIcr, k = 6) +
##     s(FSLEcr, k = 6) + s(MNcr, k = 6)
```

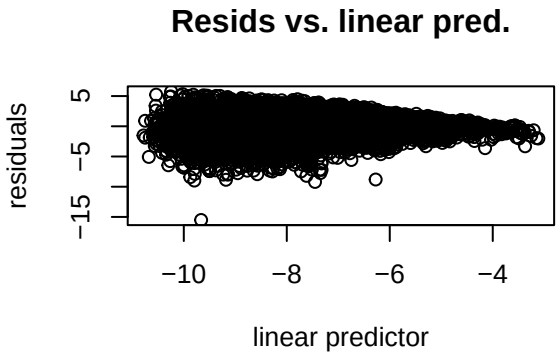
```
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -7.38057    0.01031  -715.8   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##             edf Ref.df      F p-value
## s(chlacr) 4.956  4.999 170.76  <2e-16 ***
## s(slacr)  4.438  4.865  79.24  <2e-16 ***
## s(SSCIcr) 4.575  4.916 202.12  <2e-16 ***
## s(FSLEcr) 4.989  5.000 713.88  <2e-16 ***
## s(MNcr)   4.990  5.000 358.40  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =   0.35   Deviance explained =   35%
## GCV = 2.3911  Scale est. = 2.3884    n = 22464
```

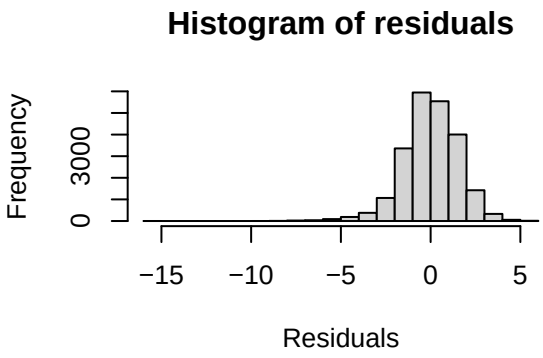
```
GAM_Moz2 <- stepAIC.gam(GAM_Moz)
summary(GAM_Moz2)
```

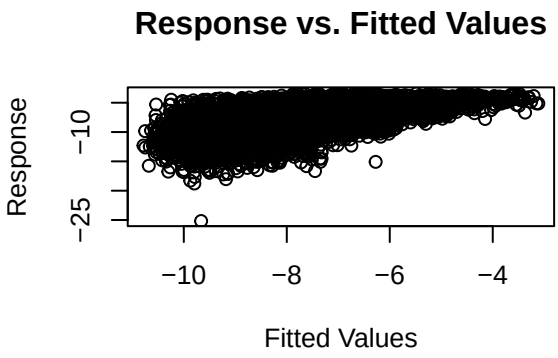
```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## logNLOG ~ s(chlacr, k = 6) + s(slacr, k = 6) + s(SSCIcr, k = 6) +
##     s(FSLEcr, k = 6) + s(MNcr, k = 6)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -7.38057    0.01031  -715.8   <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(chlacr) 4.956  4.999 170.76  <2e-16 ***
## s(slacr)  4.438  4.865  79.24  <2e-16 ***
## s(SSCIcr) 4.575  4.916 202.12  <2e-16 ***
## s(FSLEcr) 4.989  5.000 713.88  <2e-16 ***
## s(MNcr)   4.990  5.000 358.40  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =   0.35   Deviance explained =   35%
## GCV = 2.3911  Scale est. = 2.3884    n = 22464
```

```
par(mfrow = c(2,2))
gam.check(GAM_Moz2)
```



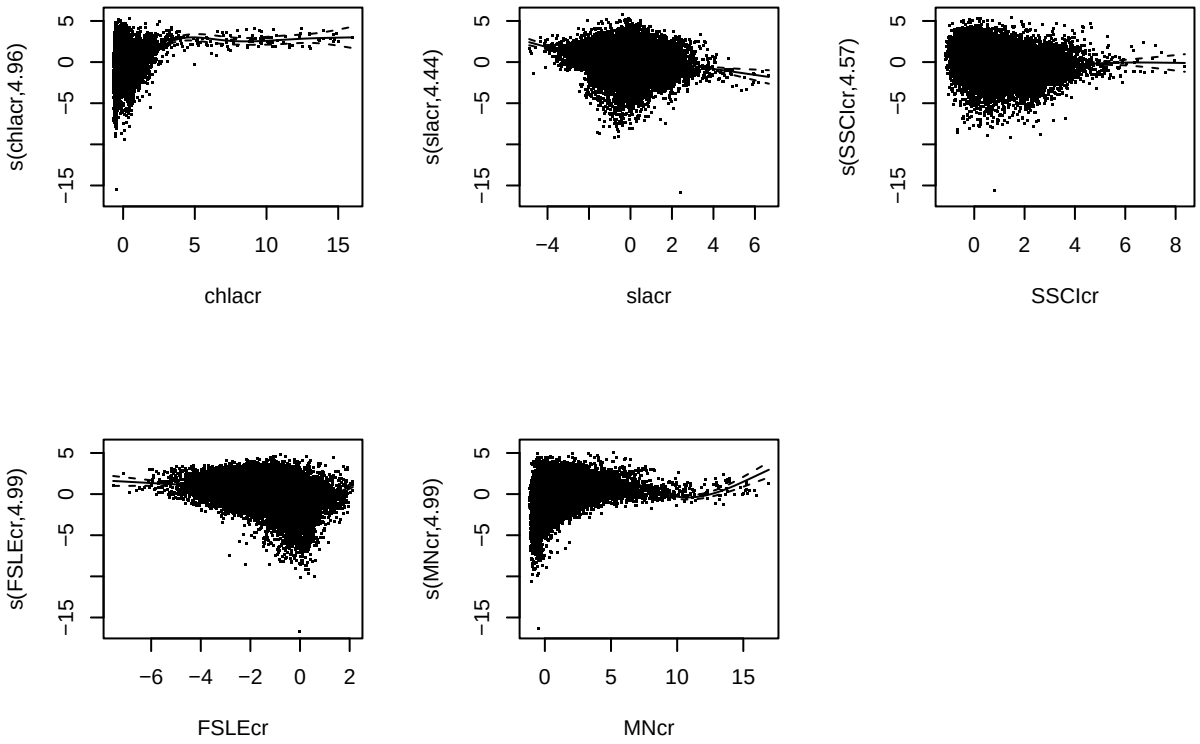







```
##
## Method: GCV   Optimizer: magic
## Smoothing parameter selection converged after 14 iterations.
## The RMS GCV score gradient at convergence was 2.668483e-07 .
## The Hessian was positive definite.
## Model rank =  26 / 26
##
## Basis dimension (k) checking results. Low p-value (k-index<1) may
## indicate that k is too low, especially if edf is close to k'.
##
##             k'  edf k-index p-value
## s(chlacr) 5.00 4.96    0.98   0.095 .
## s(slacr)  5.00 4.44    0.98   0.075 .
## s(SSCIcr) 5.00 4.57    1.00   0.525
## s(FSLEcr) 5.00 4.99    1.00   0.520
## s(MNcr)   5.00 4.99    0.99   0.130
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
plot(GAM_Moz2,residuals=T,pages=1)
```


s(chlacr,4.96)
chlacr
s(slacr,4.44)
slacr
s(SSCIcr,4.57)
SSCIcr
s(FSLEcr,4.99)
FSLEcr
s(MNcr,4.99)
MNcr