WESA GAM, ZIP, & ZINB models by N/S

Sarah Popov

2023-02-06

Data summary

Dataset: one count record per N/S region per survey date, 820 records. 8.4% of the records are zeroes.

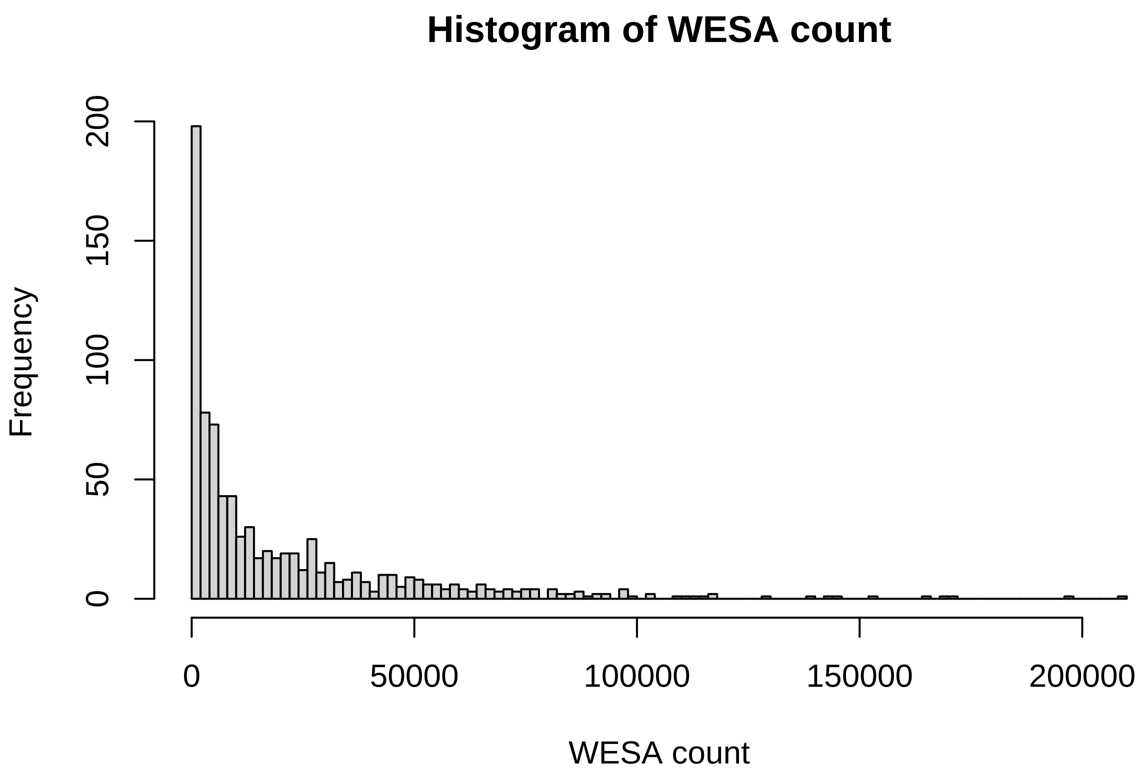
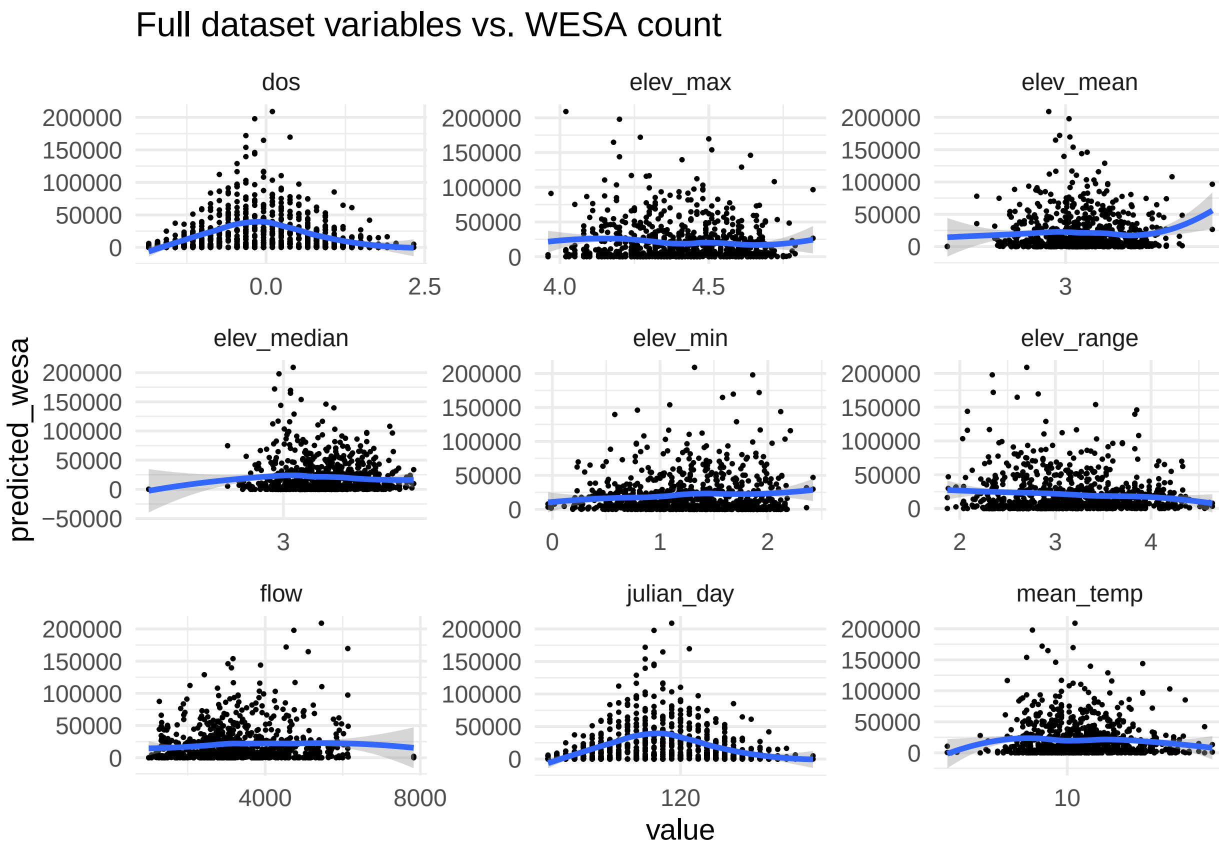
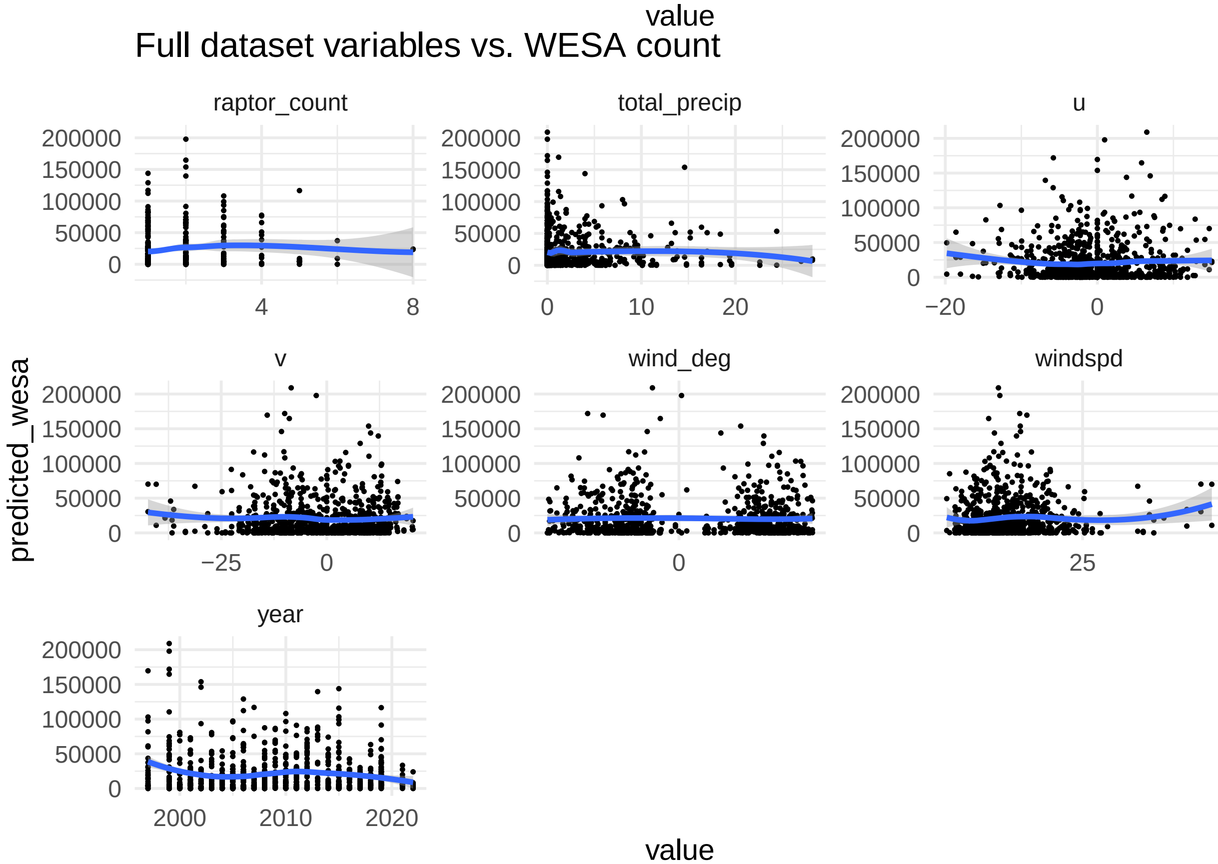
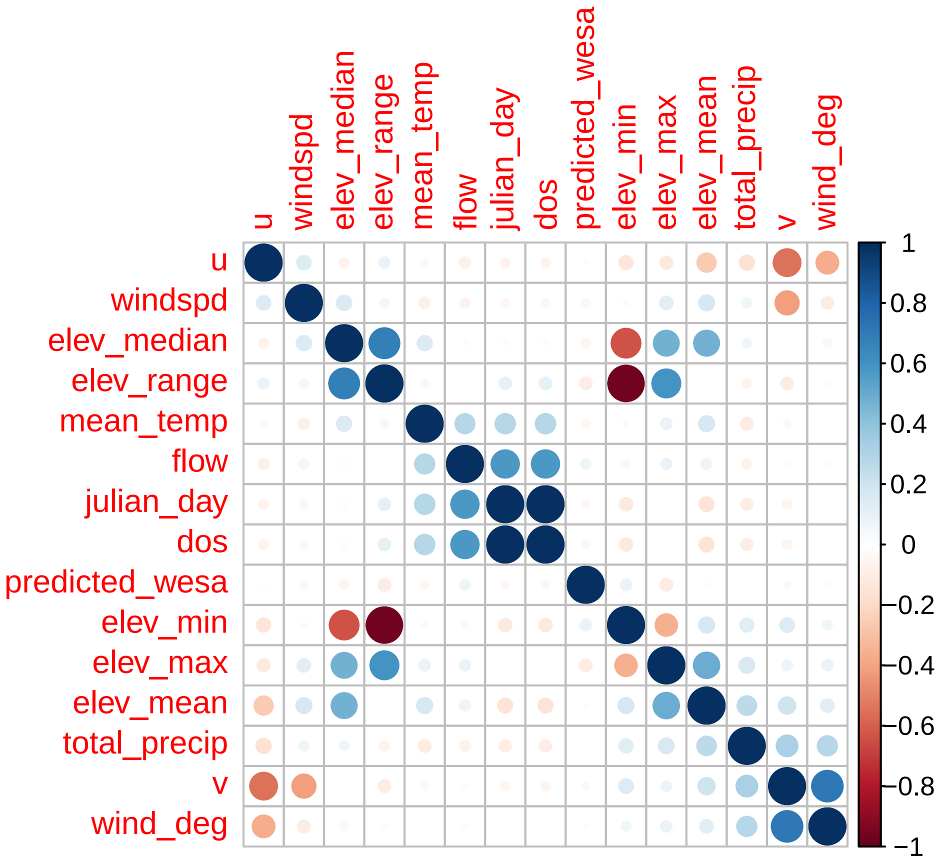


Figure 1: Histogram of WESA count per N/S group per survey date. Plenty of zeroes...


Full dataset variables vs. WESA count
dos
elev_max
elev_mean
elev_median
elev_min
elev_range
flow
julian_day
mean_temp
predicted_wesa
value


Full dataset variables vs. WESA count
raptor_count
total_precip
u
v
wind_deg
windspd
year
predicted_wesa
value



u
windspd
elev_median
elev_range
mean_temp
flow
julian_day
dos
predicted_wesa
elev_min
elev_max
elev_mean
total_precip
v
wind_deg
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1

Models

From the initial `glmmTMB` explorations, three things jumped out:

1. The negative binomial distribution fits the data best.
2. A simplified random effects structure eliminates all model convergence issues.
3. A non-linear approach (GAM) potentially might fit the data better.

```
# Base script by Gavin Simpson
# https://fromthebottomoftheheap.net/2017/05/04/compare-mgcv-with-glmmtmb/
# https://gist.github.com/gavinsimpson/8a0f0e072b095295cf5f7af2762e05a7

library("mgcv")
library("glmmTMB")

## Poisson Models
pgam0 <- gam(predicted_wesa ~ n_s + year_c + s(dos) + s(year,
    bs = "re"), data = dat3, family = poisson, method = "ML")
pgam1 <- gam(predicted_wesa ~ n_s + s(flow) + year_c + s(dos) +
    s(year, bs = "re"), data = dat3, family = poisson, method = "ML")
pgam2 <- gam(predicted_wesa ~ n_s + s(flow) + n_s:flow + year_c +
    s(dos) + s(year, bs = "re"), data = dat3, family = poisson,
    method = "ML")

pm0 <- glmmTMB(predicted_wesa ~ n_s + year_c + I(dos^2) + (1 |
    year), data = dat3, family = poisson)
pm1 <- glmmTMB(predicted_wesa ~ n_s + scale(flow) + year_c +
    I(dos^2) + (1 | year), data = dat3, family = poisson)
pm2 <- glmmTMB(predicted_wesa ~ n_s * scale(flow) + year_c +
    I(dos^2) + (1 | year), data = dat3, family = poisson)

AIC(pgam0, pgam1, pgam2)
```

```
##       df     AIC
## pgam0 34 8003056
## pgam1 43 7782737
## pgam2 44 7720742
```

```
AIC(pm0, pm1, pm2)
```

```
##     df     AIC
## pm0  5 8450215
## pm1  6 8304914
## pm2  7 8240747
```

```
## Negative binomial models
nbgam0 <- gam(predicted_wesa ~ n_s + year_c + s(dos) + s(year,
    bs = "re"), data = dat3, family = nb, method = "ML")
nbgam1 <- gam(predicted_wesa ~ n_s + s(flow) + year_c + s(dos) +
    s(year, bs = "re"), data = dat3, family = nb, method = "ML")
nbgam2 <- gam(predicted_wesa ~ n_s + s(flow) + n_s:flow + year_c +
    s(dos) + s(year, bs = "re"), data = dat3, family = nb, method = "ML")
```

```r
nbm0 <- glmmTMB(predicted_wesa ~ n_s + year_c + I(dos^2) + (1 |
    year), data = dat3, family = nbinom2)
nbm1 <- glmmTMB(predicted_wesa ~ n_s + scale(flow) + year_c +
    I(dos^2) + (1 | year), data = dat3, family = nbinom2)
nbm2 <- glmmTMB(predicted_wesa ~ n_s * scale(flow) + year_c +
    I(dos^2) + (1 | year), data = dat3, family = nbinom2)

AIC(nbgam0, nbgam1, nbgam2)
```

```
##               df      AIC
## nbgam0 26.54186 16670.59
## nbgam1 27.79454 16666.51
## nbgam2 28.94616 16668.14
```

```r
AIC(nbm0, nbm1, nbm2)
```

```
##      df      AIC
## nbm0  6 16700.31
## nbm1  7 16699.52
## nbm2  8 16700.83
```

```r
## Zero-inflated Poisson mgcv's ziplss can only fit using
## REML
zipgam0 <- gam(list(predicted_wesa ~ n_s + year_c + s(dos) +
    s(year, bs = "re"), ~n_s), data = dat3, family = ziplss,
    method = "REML")
zipgam1 <- gam(list(predicted_wesa ~ n_s + s(flow) + year_c +
    s(dos) + s(year, bs = "re"), ~n_s), data = dat3, family = ziplss,
    method = "REML")
zipgam2 <- gam(list(predicted_wesa ~ n_s + s(flow) + n_s:flow +
    year_c + s(dos) + s(year, bs = "re"), ~n_s + flow), data = dat3,
    family = ziplss, method = "REML")
zipgam3 <- gam(list(predicted_wesa ~ n_s + year_c + s(dos) +
    s(year, bs = "re"), ~n_s * flow), data = dat3, family = ziplss,
    method = "REML")
## check the things converged zipgam0$outer.info ## full
## convergence zipgam1$outer.info ## full convergence
## zipgam2$outer.info ## full convergence
## zipgam3$outer.info ## full convergence

zipm0 <- glmmTMB(predicted_wesa ~ n_s + year_c + I(dos^2) + (1 |
    year), zi = ~n_s, data = dat3, family = poisson)
zipm1 <- glmmTMB(predicted_wesa ~ n_s + scale(flow) + year_c +
    I(dos^2) + (1 | year), zi = ~n_s, data = dat3, family = poisson)
zipm2 <- glmmTMB(predicted_wesa ~ n_s + scale(flow) + year_c +
    I(dos^2) + (1 | year), zi = ~n_s + flow, data = dat3, family = poisson)
zipm3 <- glmmTMB(predicted_wesa ~ n_s * scale(flow) + year_c +
    I(dos^2) + (1 | year), zi = ~n_s * flow, data = dat3, family = poisson)

zinb0 <- glmmTMB(predicted_wesa ~ n_s + year_c + I(dos^2) + (1 |
    year), zi = ~n_s, data = dat3, family = nbinom1)
zinb1 <- glmmTMB(predicted_wesa ~ n_s + scale(flow) + year_c +
```

```r
    I(dos^2) + (1 | year), zi = ~n_s + flow, data = dat3, family = nbinom1)
zinb2 <- glmmTMB(predicted_wesa ~ n_s * scale(flow) + year_c +
    I(dos^2) + (1 | year), zi = ~n_s + flow, data = dat3, family = nbinom1)

AIC(zipgam0, zipgam1, zipgam2, zipgam3)
```

```
##                df     AIC
## zipgam0 36.00000 6845442
## zipgam1 45.00000 6652772
## zipgam2 46.99877 6610734
## zipgam3 38.00000 6845443
```

```r
AIC(zipm0, zipm1, zipm2, zipm3, zinb0, zinb1, zinb2)
```

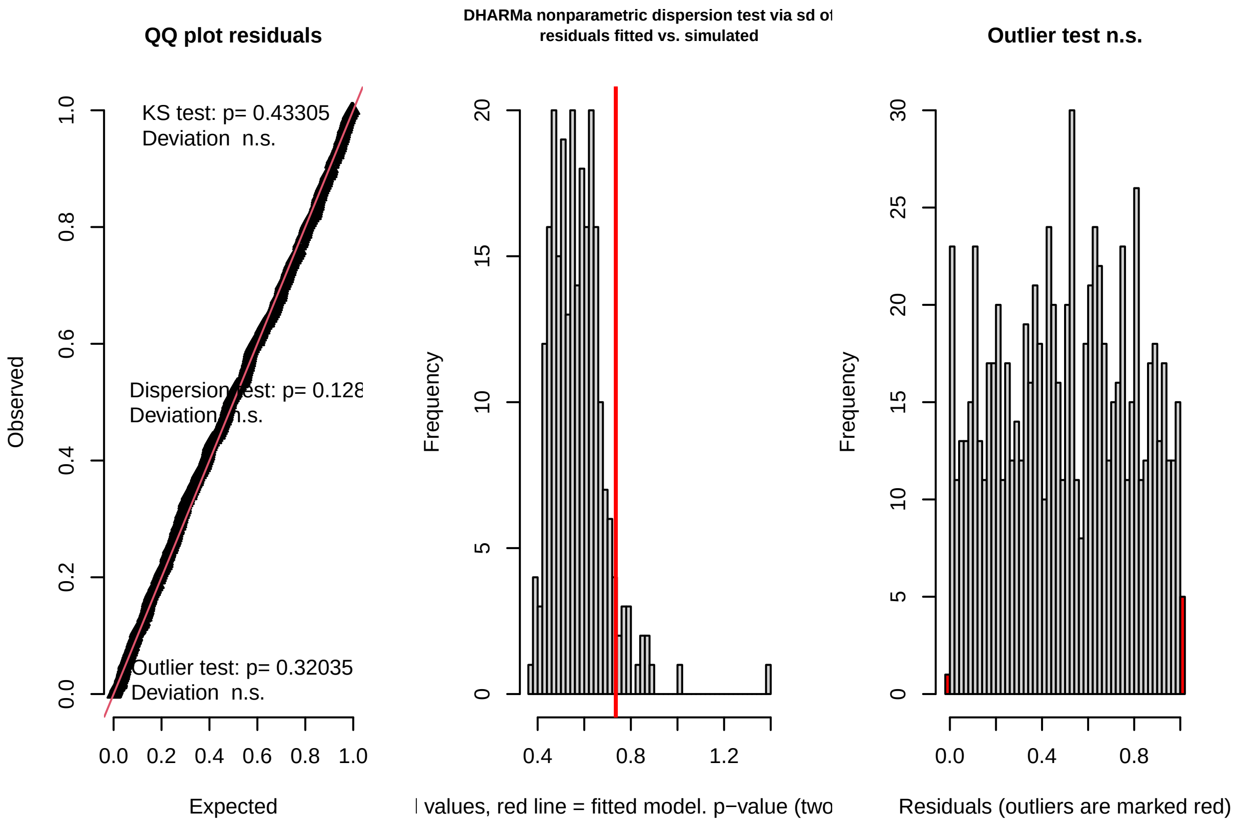
```
##       df        AIC
## zipm0  7 7337842.56
## zipm1  8 7180753.13
## zipm2  9 7180754.98
## zipm3 11 7135974.38
## zinb0  8   16180.70
## zinb1 10   16171.51
## zinb2 11   16167.57
```

```r
# Compare them all
bbmle::AICtab(pgam0, pgam1, pgam2, pm0, pm1, pm2, nbgam0, nbgam1,
    nbgam2, nbm0, nbm1, nbm2, zipgam0, zipgam1, zipgam2, zipm0,
    zipm1, zipm2, zinb0, zinb1, zinb2)
```

```
##         dAIC      df
## zinb2         0.0 11
## zinb1         3.9 10
## zinb0        13.1 8
## nbgam1      498.9 27.8
## nbgam2      500.6 28.9
## nbgam0      503.0 26.5
## nbm1        532.0 7
## nbm0        532.7 6
## nbm2        533.3 8
## zipgam2 6594566.2 47
## zipgam1 6636604.6 45
## zipgam0 6829274.4 36
## zipm1   7164585.6 8
## zipm2   7164587.4 9
## zipm0   7321675.0 7
## pgam2   7704574.7 44
## pgam1   7766569.1 43
## pgam0   7986888.1 34
## pm2     8224579.3 7
## pm1     8288746.9 6
## pm0     8434047.7 5
```

Best-fit diagnostics

Diagnostics indicate underdispersion in our data. Even though it's the best-fit model, it's underpredicting zeros.




```
## $uniformity
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  simulationOutput$scaledResiduals
## D = 0.030439, p-value = 0.4331
## alternative hypothesis: two-sided
##
##
## $dispersion
##
##  DHARMa nonparametric dispersion test via sd of residuals fitted vs.
##  simulated
##
## data:  simulationOutput
## dispersion = 1.2798, p-value = 0.128
## alternative hypothesis: two.sided
##
##
## $outliers
##
##  DHARMa outlier test based on exact binomial test with approximate
##  expectations
```

```
##
## data:  simulationOutput
## outliers at both margin(s) = 9, observations = 820, p-value = 0.3203
## alternative hypothesis: true probability of success is not equal to 0.007968127
## 95 percent confidence interval:
##  0.005030689 0.020732492
## sample estimates:
## frequency of outliers (expected: 0.00796812749003984 )
##                                              0.01097561
```

```
## $uniformity
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  simulationOutput$scaledResiduals
## D = 0.030439, p-value = 0.4331
## alternative hypothesis: two-sided
##
##
## $dispersion
##
##  DHARMa nonparametric dispersion test via sd of residuals fitted vs.
##  simulated
##
## data:  simulationOutput
## dispersion = 1.2798, p-value = 0.128
## alternative hypothesis: two.sided
##
##
## $outliers
##
##  DHARMa outlier test based on exact binomial test with approximate
##  expectations
##
## data:  simulationOutput
## outliers at both margin(s) = 9, observations = 820, p-value = 0.3203
## alternative hypothesis: true probability of success is not equal to 0.007968127
## 95 percent confidence interval:
##  0.005030689 0.020732492
## sample estimates:
## frequency of outliers (expected: 0.00796812749003984 )
##                                              0.01097561
```

Test for zero inflation

```
##
##  DHARMa zero-inflation test via comparison to expected zeros with
##  simulation under H0 = fitted model
##
## data:  simulationOutput
## ratioObsSim = 0.90433, p-value = 0.392
## alternative hypothesis: two.sided
```

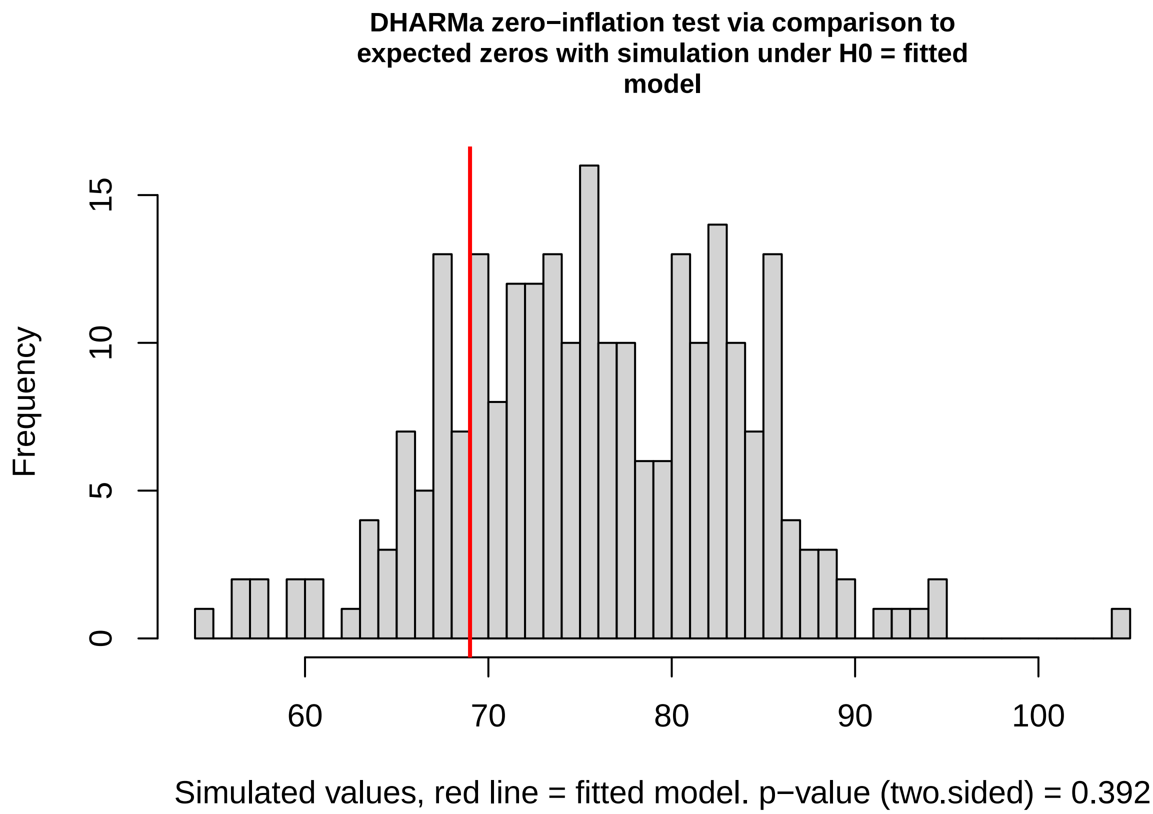


Figure 2: The zero-inflation test confirms we're underpredicting zeroes, despite this being the 'best-fit' model with the lowest AIC.

Full model

```
##  Family: nbinom1  ( log )
## Formula:
## predicted_wesa ~ n_s * scale(flow) + year_c + scale(mean_temp) +
##     scale(elev_range) + tide + scale(total_precip) + scale(u) +
##     I(dos^2) + (1 | year)
## Zero inflation:                  ~.
## Data: dat3
##
##      AIC      BIC   logLik deviance df.resid
##  16087.8  16205.5  -8018.9  16037.8      795
##
## Random effects:
##
## Conditional model:
##  Groups Name        Variance Std.Dev.
##  year   (Intercept) 0.1175   0.3428
## Number of obs: 820, groups:  year, 24
##
## Zero-inflation model:
##  Groups Name        Variance Std.Dev.
##  year   (Intercept) 0.2803   0.5295
## Number of obs: 820, groups:  year, 24
##
## Dispersion parameter for nbinom1 family (): 1.24e+04
##
## Conditional model:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         10.796120   0.082294  131.19  < 2e-16 ***
## n_sS                -1.091614   0.055440  -19.69  < 2e-16 ***
## scale(flow)         -0.097463   0.039034   -2.50  0.01253 *
## year_c              -0.124939   0.071961   -1.74  0.08253 .
## scale(mean_temp)     0.053709   0.029644    1.81  0.07002 .
## scale(elev_range)   -0.085658   0.029694   -2.88  0.00392 **
## tiderising           0.029601   0.059018    0.50  0.61598
## scale(total_precip)  0.004169   0.025360    0.16  0.86942
## scale(u)             0.042797   0.025816    1.66  0.09737 .
## I(dos^2)            -0.737516   0.035440  -20.81  < 2e-16 ***
## n_sS:scale(flow)    -0.148369   0.055654   -2.67  0.00768 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Zero-inflation model:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -8.4428     1.4317  -5.897 3.70e-09 ***
## n_sS                  4.6169     1.3413   3.442 0.000577 ***
## scale(flow)          -1.3735     1.0139  -1.355 0.175516
## year_c               -0.7836     0.2552  -3.071 0.002135 **
## scale(mean_temp)     -0.2737     0.2036  -1.344 0.178796
## scale(elev_range)    -0.5078     0.2111  -2.405 0.016156 *
## tiderising            1.9877     0.4794   4.146 3.38e-05 ***
## scale(total_precip)  -0.4051     0.2829  -1.432 0.152218
## scale(u)              0.4990     0.2014   2.478 0.013217 *
```

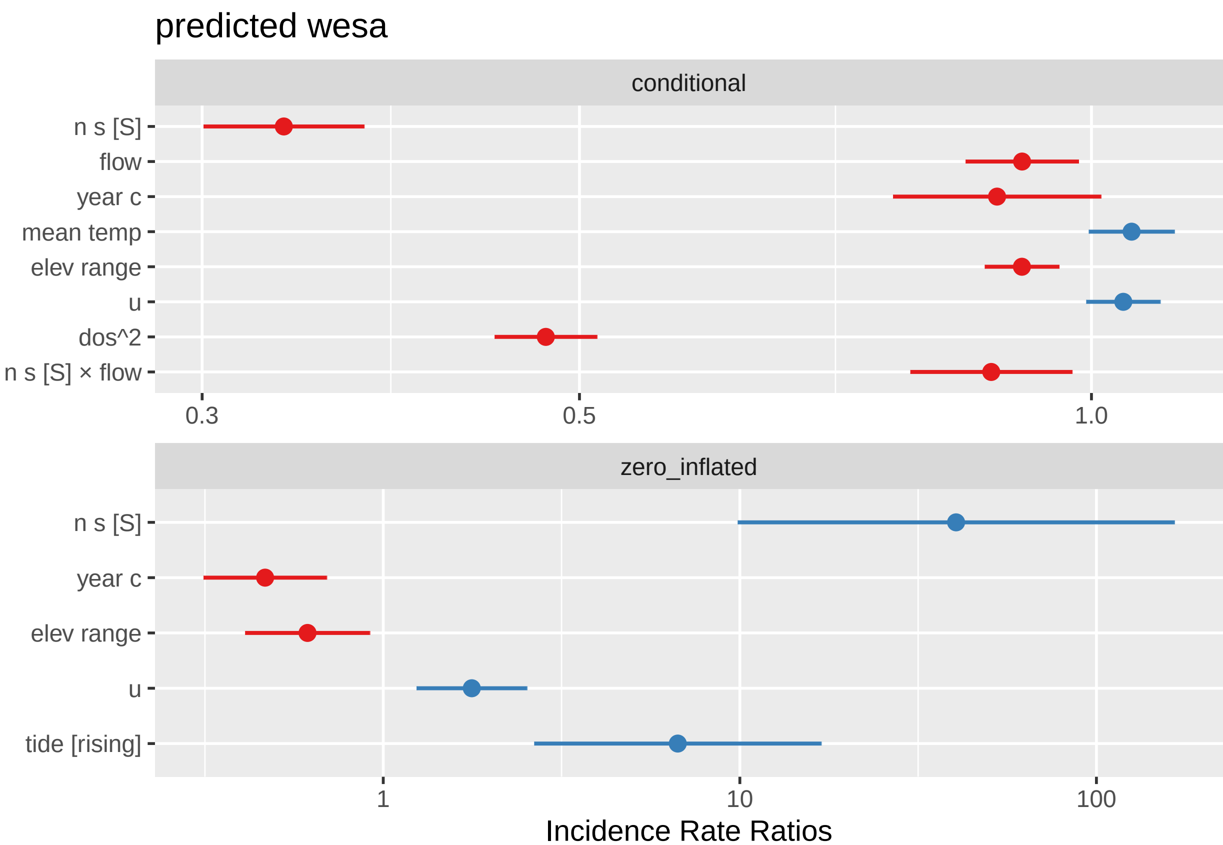
```
## I(dos^2)              0.1793     0.1909   0.939 0.347594
## n_sS:scale(flow)      1.1353     1.0171   1.116 0.264351
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Final model

Backwards stepwise selection; first removed insignificant terms from zi model, then subsequently removed insignificant terms from full model using AIC backwards selection (`drop1` command).

```
##  Family: nbinom1  ( log )
## Formula:
## predicted_wesa ~ n_s + scale(flow) + year_c + scale(mean_temp) +
##     scale(elev_range) + scale(u) + I(dos^2) + (1 | year) + n_s:scale(flow)
## Zero inflation:
## ~n_s + year_c + scale(elev_range) + tide + scale(u)
## Data: dat3
##
##      AIC      BIC   logLik deviance df.resid
##  16084.2  16164.2  -8025.1  16050.2      803
##
## Random effects:
##
## Conditional model:
##  Groups Name        Variance Std.Dev.
##  year   (Intercept) 0.1174   0.3427
## Number of obs: 820, groups:  year, 24
##
## Dispersion parameter for nbinom1 family (): 1.24e+04
##
## Conditional model:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       10.80899    0.07915  136.57  < 2e-16 ***
## n_sS              -1.09341    0.05547  -19.71  < 2e-16 ***
## scale(flow)       -0.09393    0.03896   -2.41 0.015902 *
## year_c            -0.12782    0.07177   -1.78 0.074928 .
## scale(mean_temp)   0.05436    0.02956    1.84 0.065862 .
## scale(elev_range) -0.09414    0.02562   -3.68 0.000238 ***
## scale(u)           0.04310    0.02554    1.69 0.091475 .
## I(dos^2)          -0.73873    0.03531  -20.92  < 2e-16 ***
## n_sS:scale(flow)  -0.13566    0.05585   -2.43 0.015141 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Zero-inflation model:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -7.1174     0.8219  -8.660  < 2e-16 ***
## n_sS                3.6990     0.7192   5.143 2.70e-07 ***
## year_c             -0.7629     0.2029  -3.760  0.00017 ***
## scale(elev_range)  -0.4890     0.2053  -2.383  0.01719 *
## tiderising          1.9017     0.4727   4.024 5.73e-05 ***
## scale(u)            0.5717     0.1816   3.148  0.00165 **
## ---
```

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



Final model diagnostics

```
## $uniformity
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  simulationOutput$scaledResiduals
## D = 0.041122, p-value = 0.1249
## alternative hypothesis: two-sided
##
##
## $dispersion
##
##  DHARMa nonparametric dispersion test via sd of residuals fitted vs.
##  simulated
##
## data:  simulationOutput
## dispersion = 1.2571, p-value = 0.184
## alternative hypothesis: two.sided
##
##
## $outliers
##
```

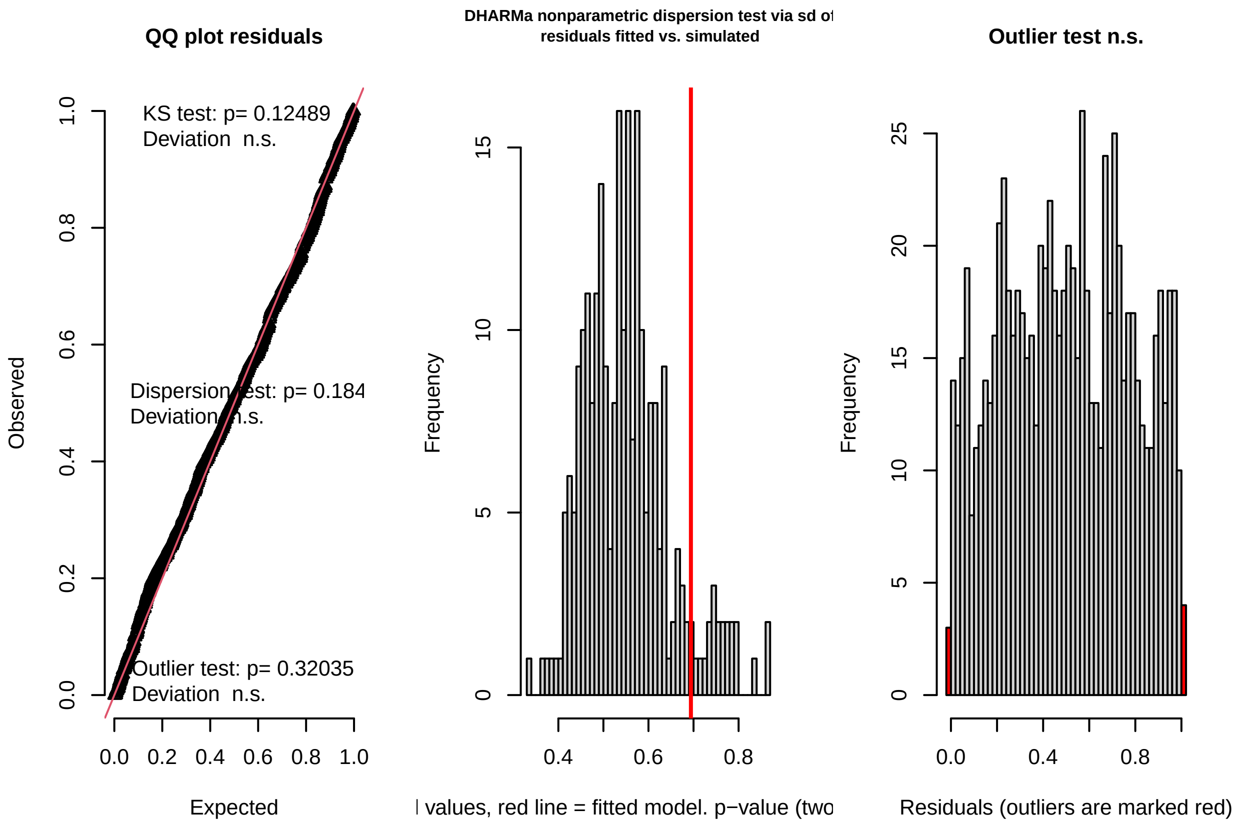


Figure 3: Residual diagnostics.

```
##  DHARMa outlier test based on exact binomial test with approximate
##  expectations
##
## data:  simulationOutput
## outliers at both margin(s) = 9, observations = 820, p-value = 0.3203
## alternative hypothesis: true probability of success is not equal to 0.007968127
## 95 percent confidence interval:
##  0.005030689 0.020732492
## sample estimates:
## frequency of outliers (expected: 0.00796812749003984 )
##                                              0.01097561
```

```
## $uniformity
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  simulationOutput$scaledResiduals
## D = 0.041122, p-value = 0.1249
## alternative hypothesis: two-sided
##
##
## $dispersion
##
##  DHARMa nonparametric dispersion test via sd of residuals fitted vs.
##  simulated
##
## data:  simulationOutput
## dispersion = 1.2571, p-value = 0.184
## alternative hypothesis: two.sided
##
##
## $outliers
##
##  DHARMa outlier test based on exact binomial test with approximate
##  expectations
##
## data:  simulationOutput
## outliers at both margin(s) = 9, observations = 820, p-value = 0.3203
## alternative hypothesis: true probability of success is not equal to 0.007968127
## 95 percent confidence interval:
##  0.005030689 0.020732492
## sample estimates:
## frequency of outliers (expected: 0.00796812749003984 )
##                                              0.01097561
```

```
##
##  DHARMa zero-inflation test via comparison to expected zeros with
##  simulation under H0 = fitted model
##
## data:  simulationOutput
## ratioObsSim = 0.87764, p-value = 0.224
## alternative hypothesis: two.sided
```

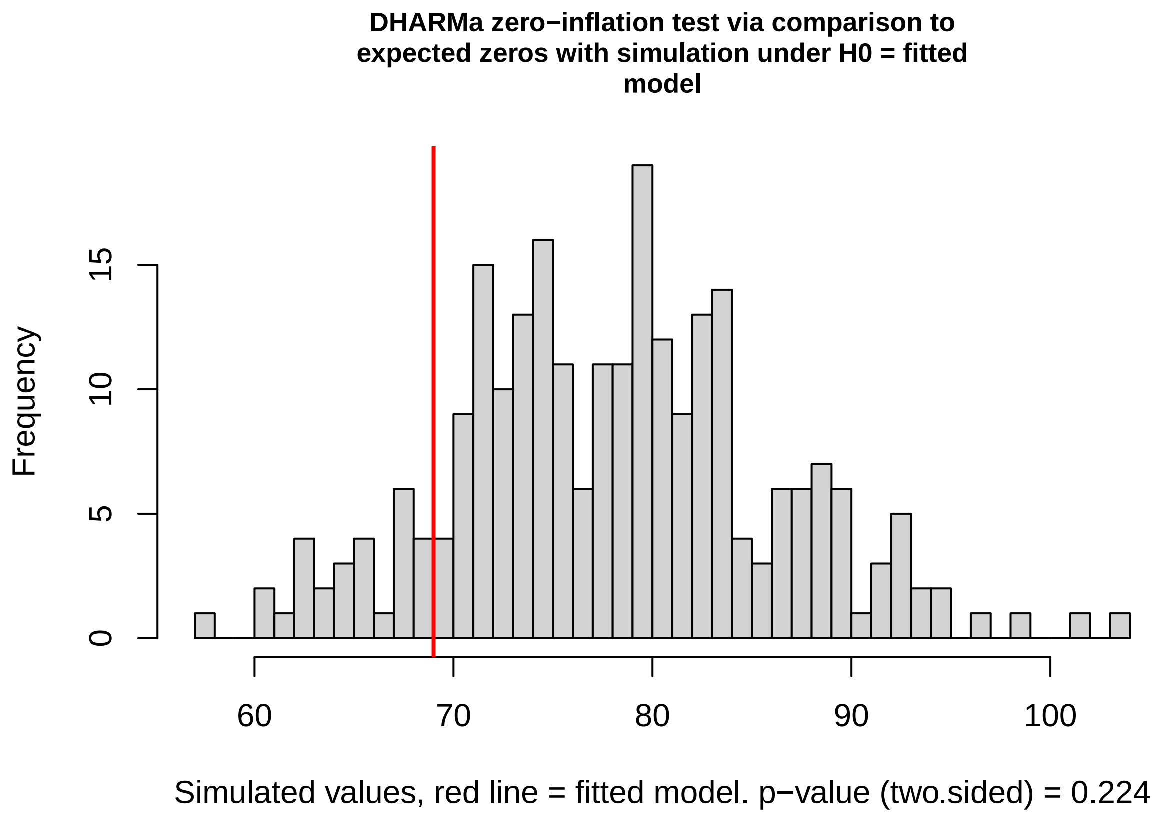
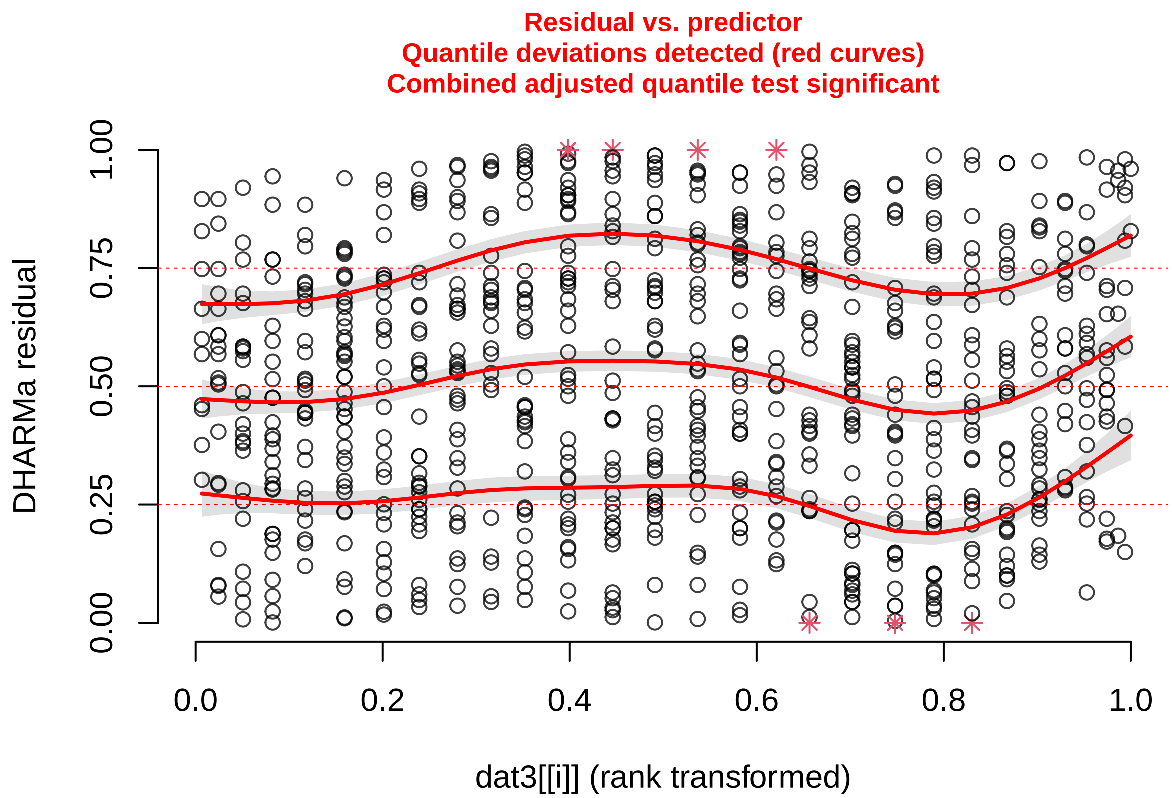


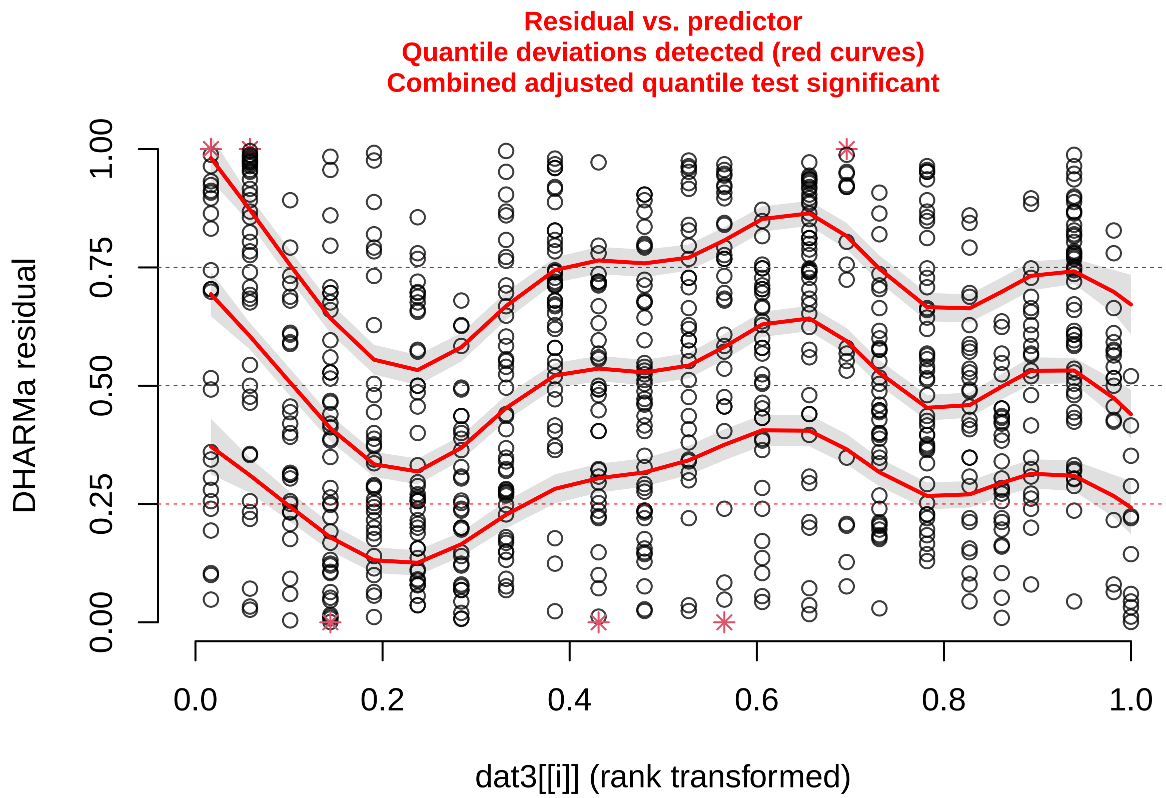
Figure 4: Testing for overdispersion. Still not quite predicting the number of zeroes exactly correctly but better than before.

Residuals vs. predicted

```
##
##   dos
```

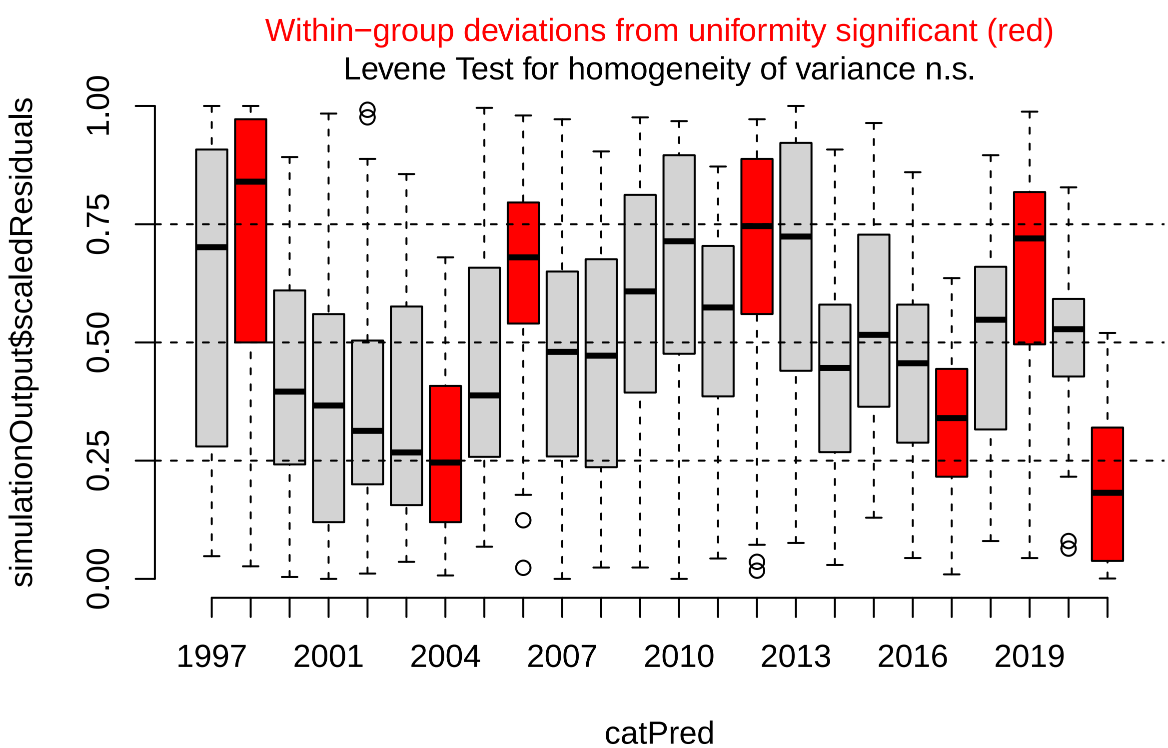



```
##
##   year_c
```



Residual vs. predictor
Quantile deviations detected (red curves)
Combined adjusted quantile test significant
DHARMa residual
0.00
0.25
0.50
0.75
1.00
0.0
0.2
0.4
0.6
0.8
1.0
dat3[[i]] (rank transformed)

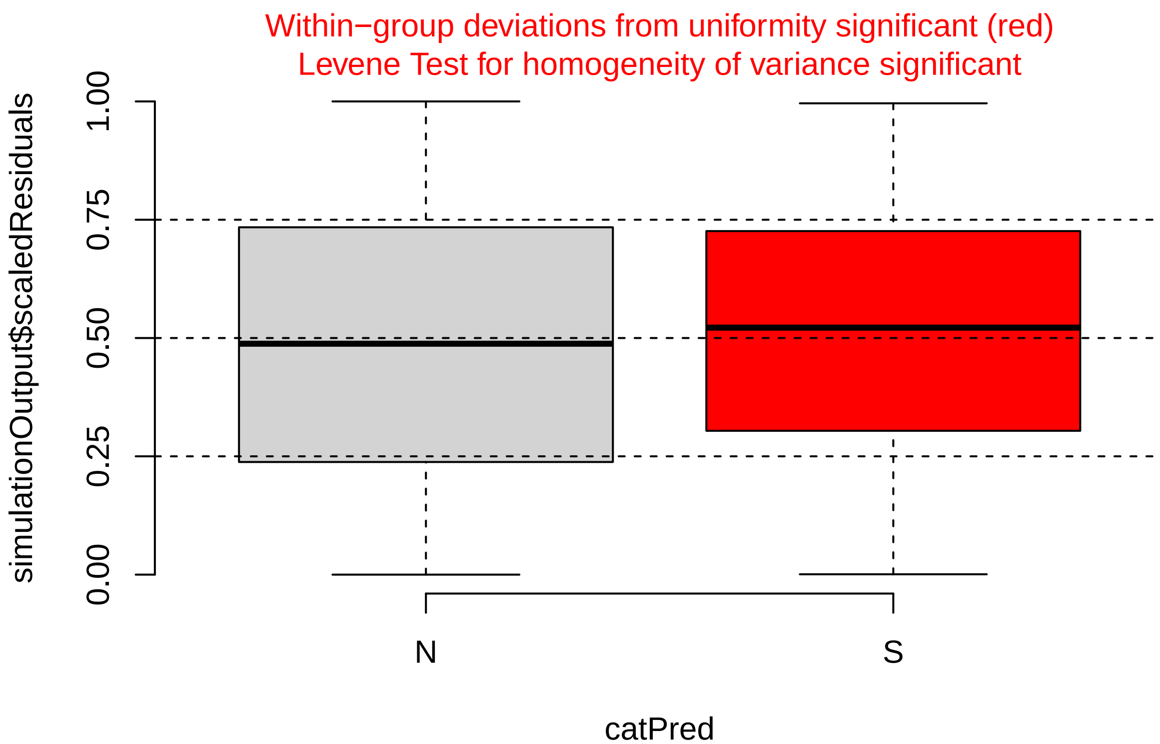

```
##
## year
```



Within−group deviations from uniformity significant (red)
Levene Test for homogeneity of variance n.s.
simulationOutput$scaledResiduals
0.00
0.25
0.50
0.75
1.00
1997
2001
2004
2007
2010
2013
2016
2019
catPred

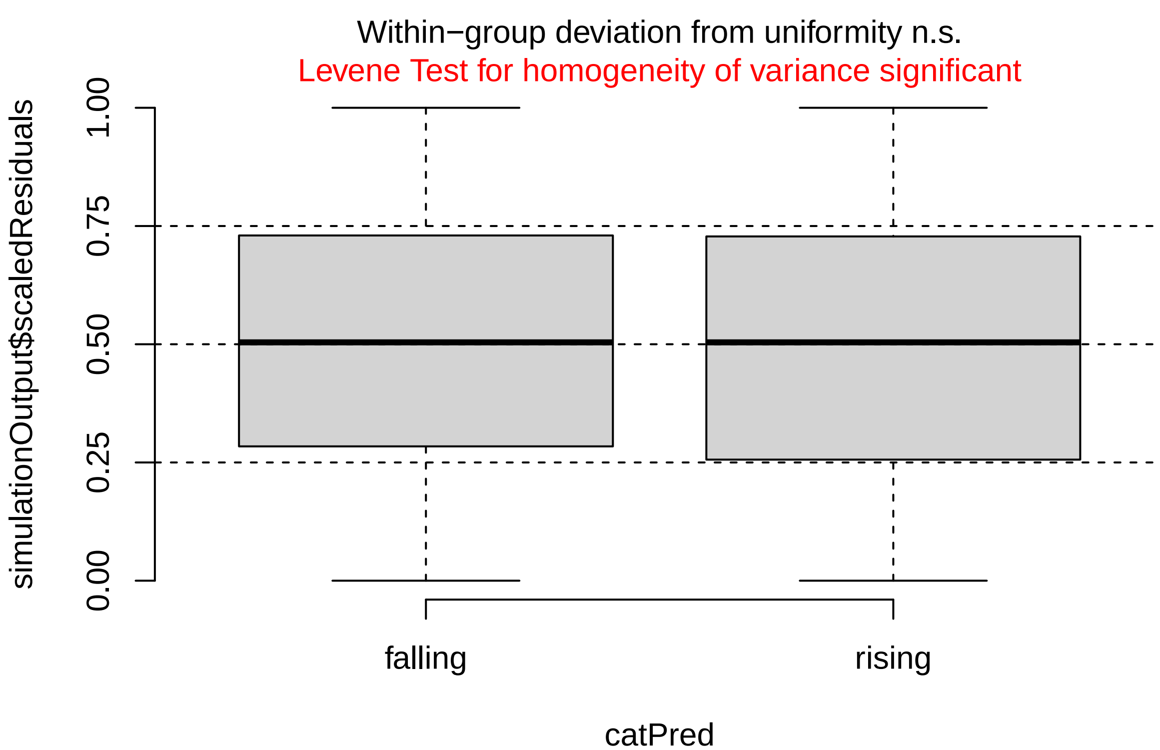

```
##
```

```
##  n_s
```



Within−group deviations from uniformity significant (red)
Levene Test for homogeneity of variance significant
simulationOutput$scaledResiduals
0.00
0.25
0.50
0.75
1.00
N
S
catPred

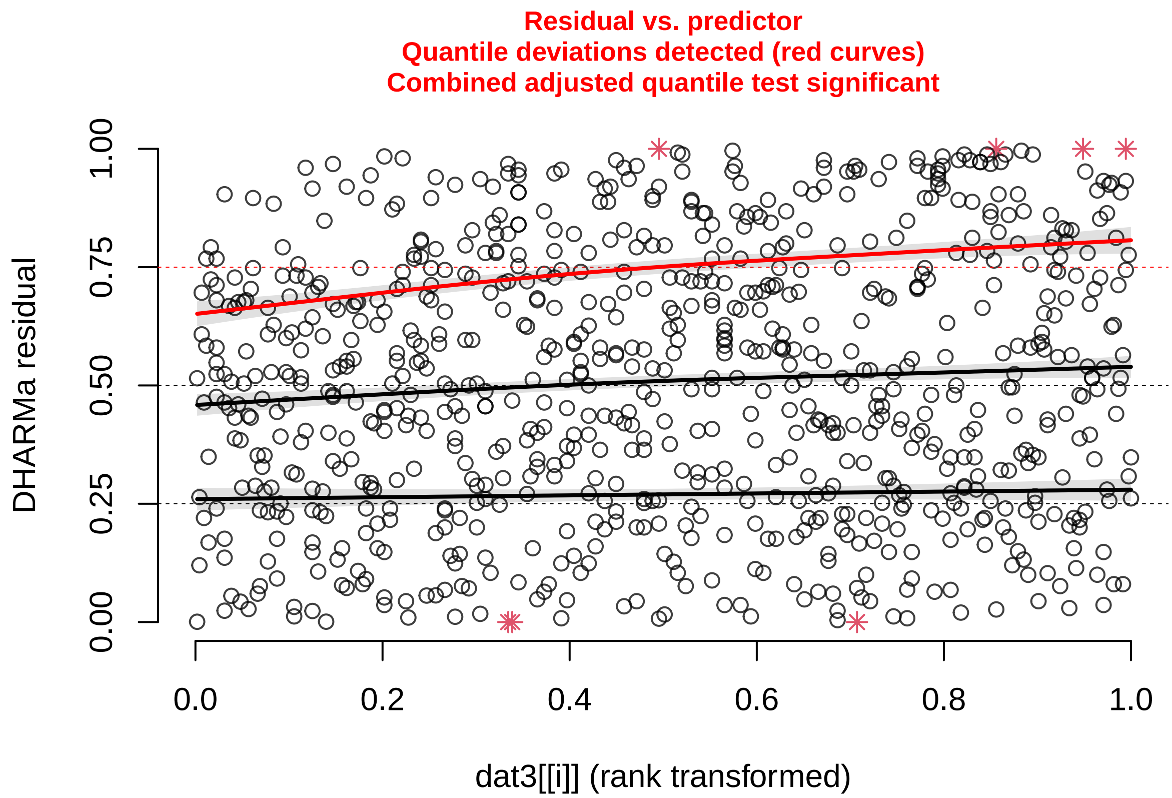

```
##
##  tide
```



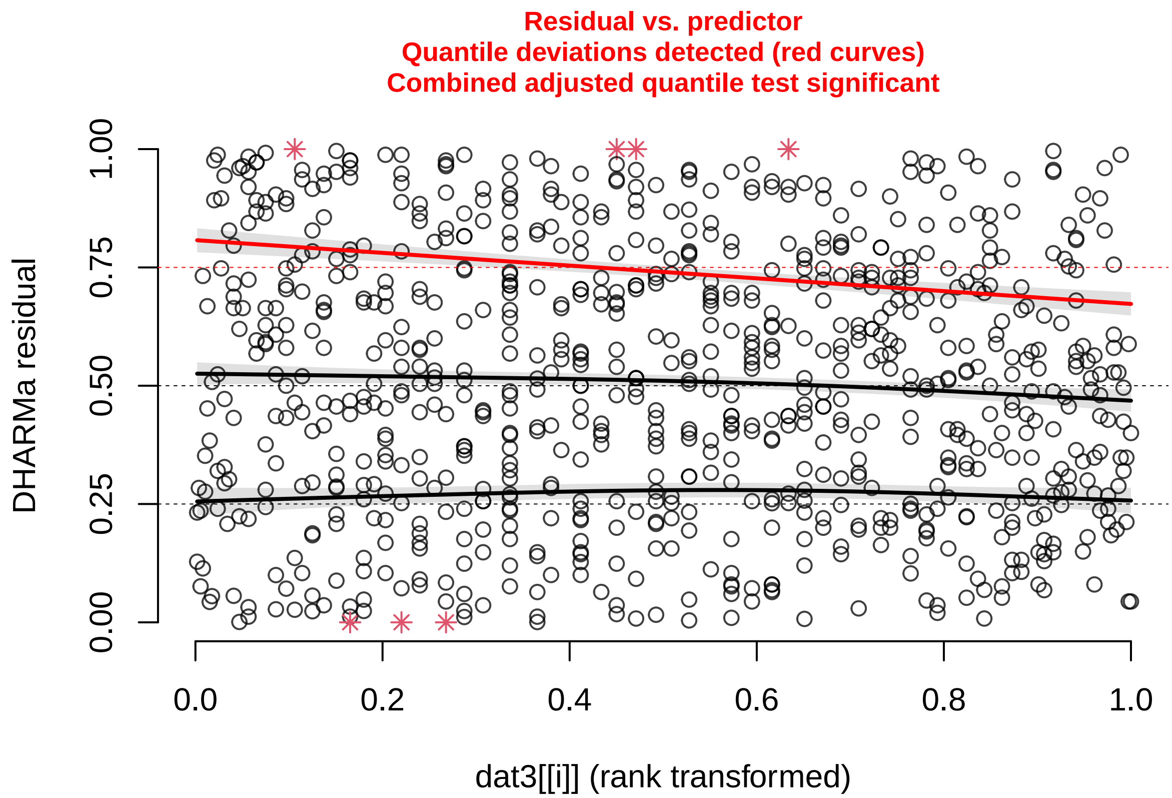
Within−group deviation from uniformity n.s.
Levene Test for homogeneity of variance significant
simulationOutput$scaledResiduals
0.00
0.25
0.50
0.75
1.00
falling
rising
catPred


```
##
```

```
## flow
```

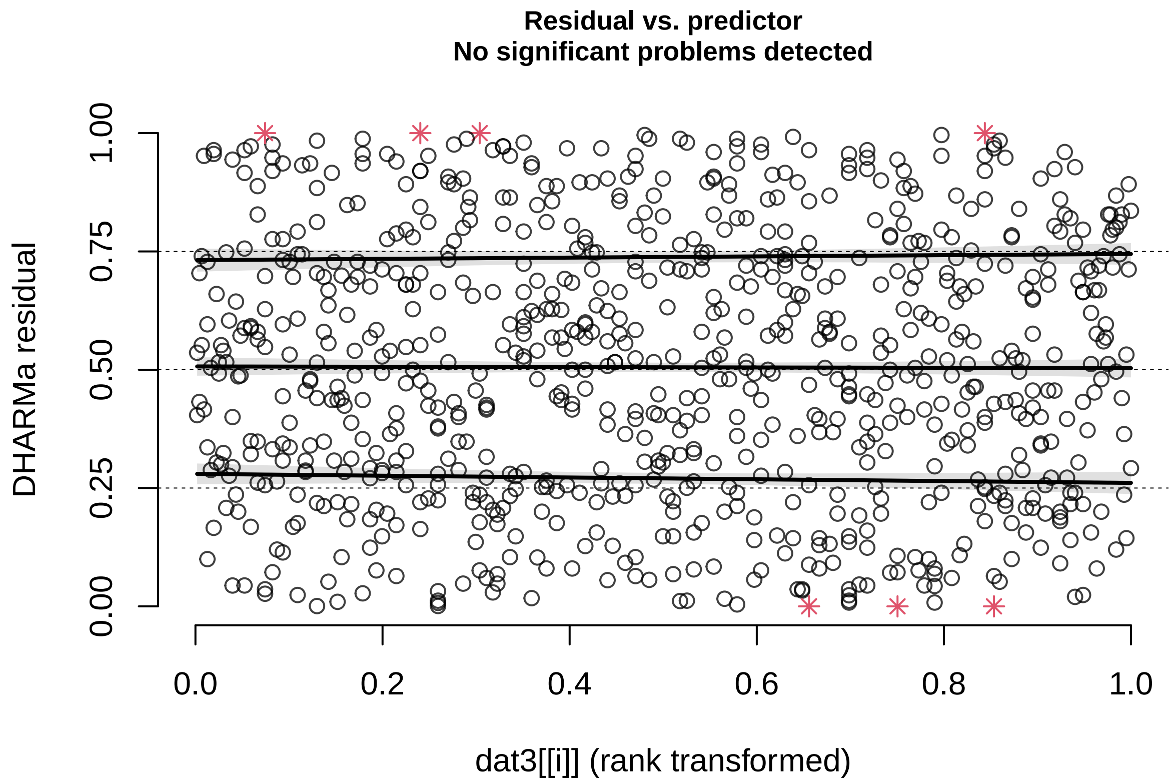



```
##
## mean_temp
```



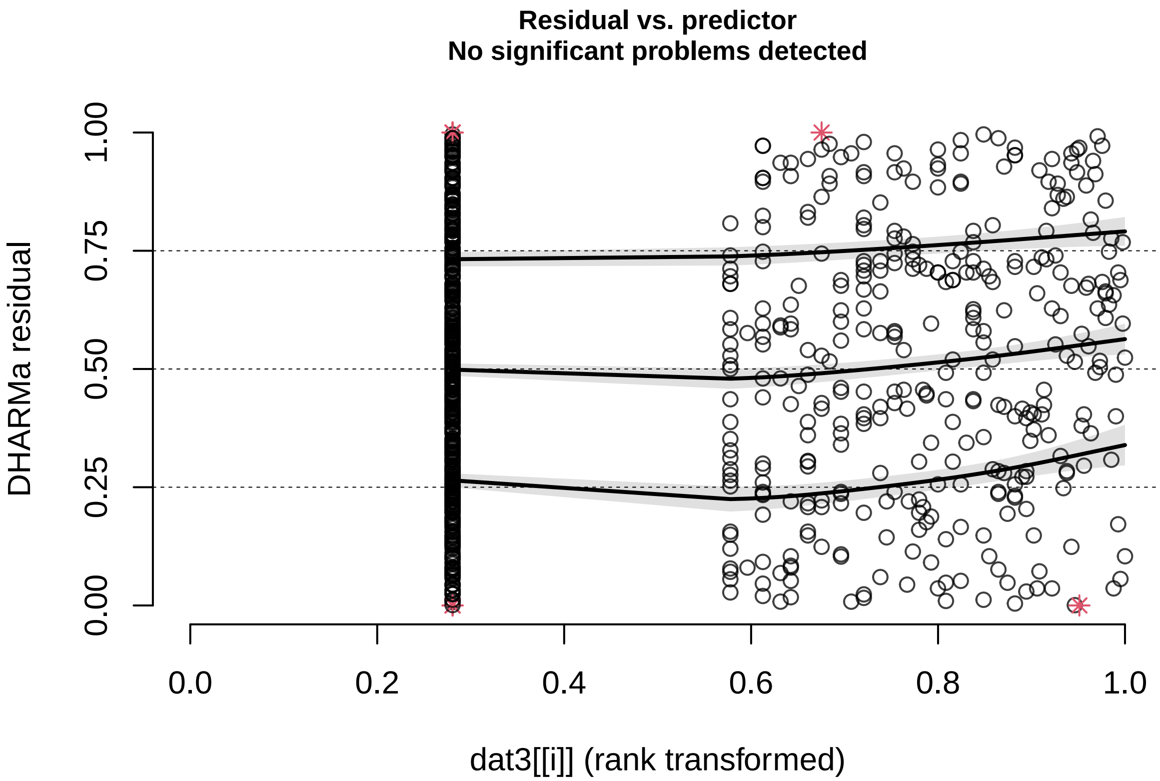
Residual vs. predictor
Quantile deviations detected (red curves)
Combined adjusted quantile test significant
DHARMa residual
0.00
0.25
0.50
0.75
1.00
0.0
0.2
0.4
0.6
0.8
1.0
dat3[[i]] (rank transformed)


```
##
##   elev_range
```

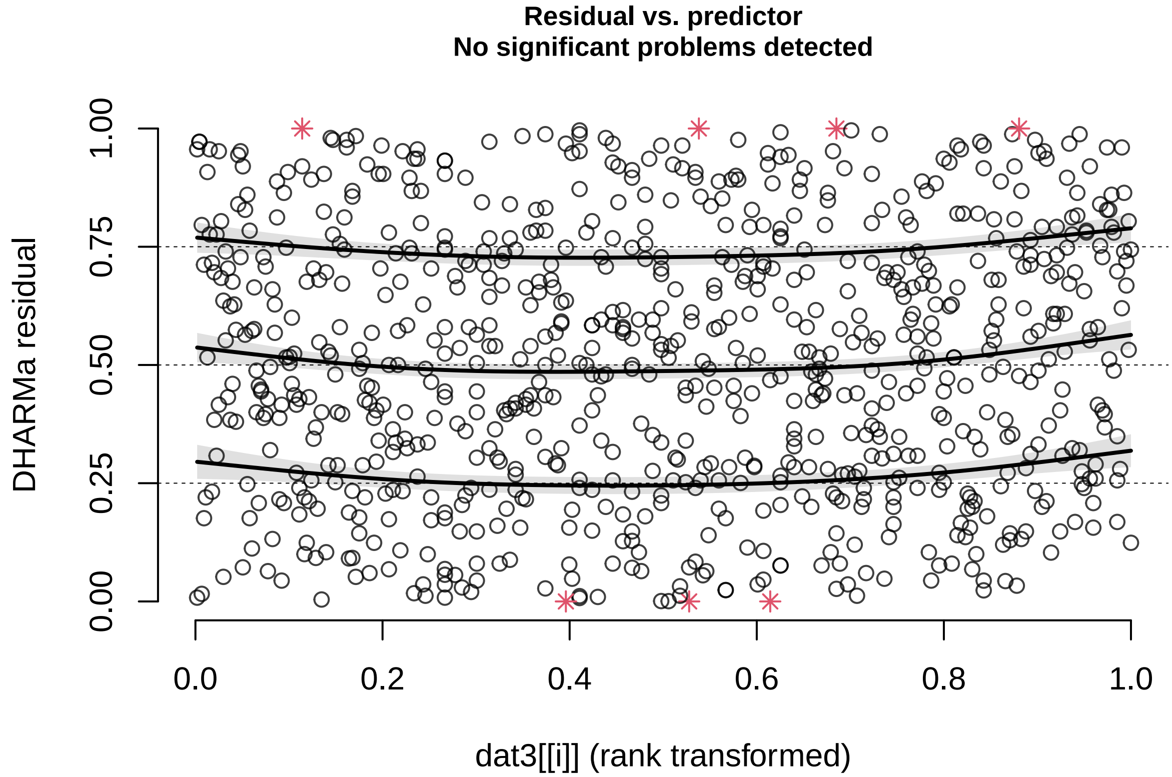


Residual vs. predictor
No significant problems detected
DHARMa residual
0.00
0.25
0.50
0.75
1.00
0.0
0.2
0.4
0.6
0.8
1.0
dat3[[i]] (rank transformed)

```
##
##  total_precip
```




```
##
##  u
```


Residual vs. predictor
No significant problems detected
DHARMa residual
0.00
0.25
0.50
0.75
1.00
0.0
0.2
0.4
0.6
0.8
1.0
dat3[[i]] (rank transformed)