WESA GAM, ZIP, & ZINB models by station

Sarah Popov

2023-02-06

Data summary

Data summary

Dataset: one count record per station per survey date, 1387 records. 11.3% of the records are zeroes.

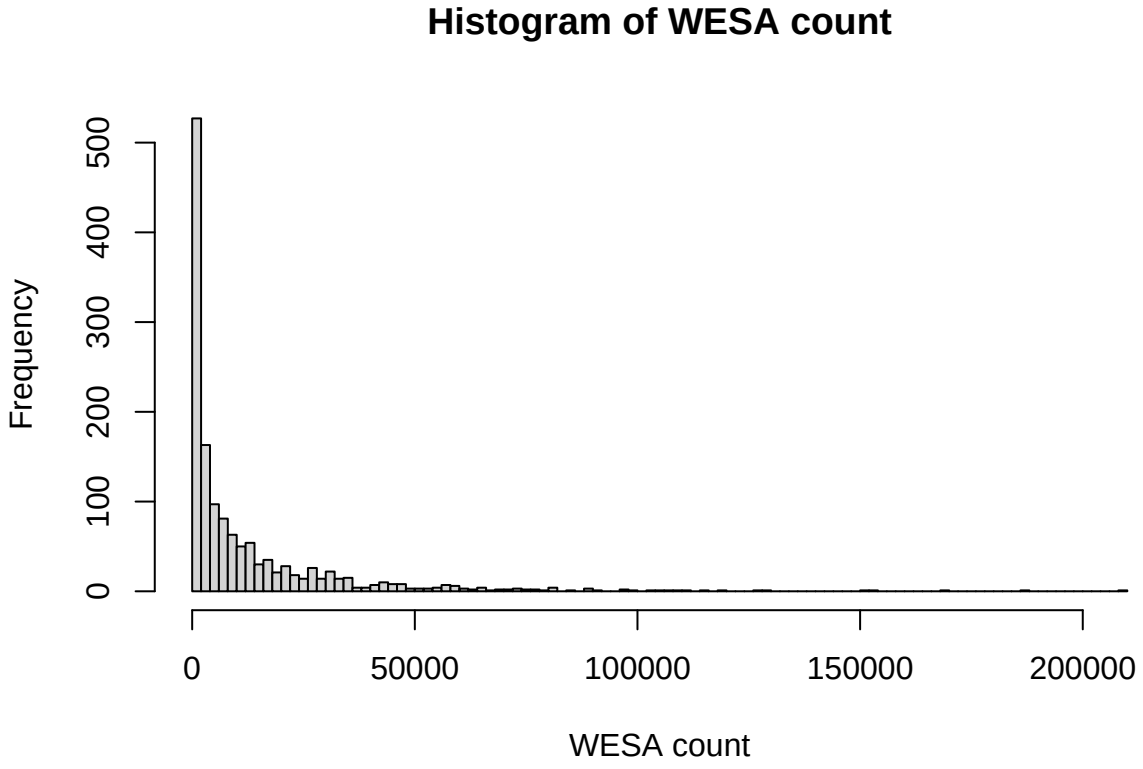
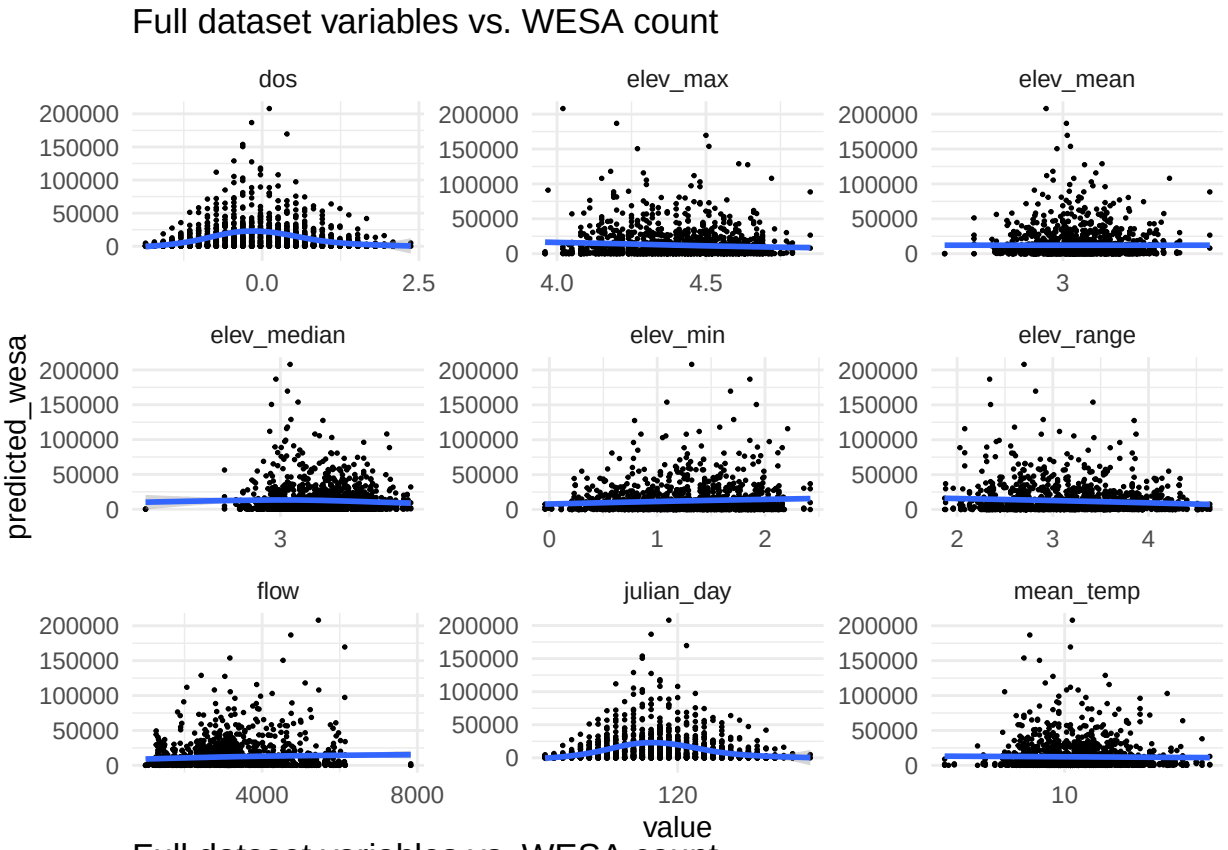
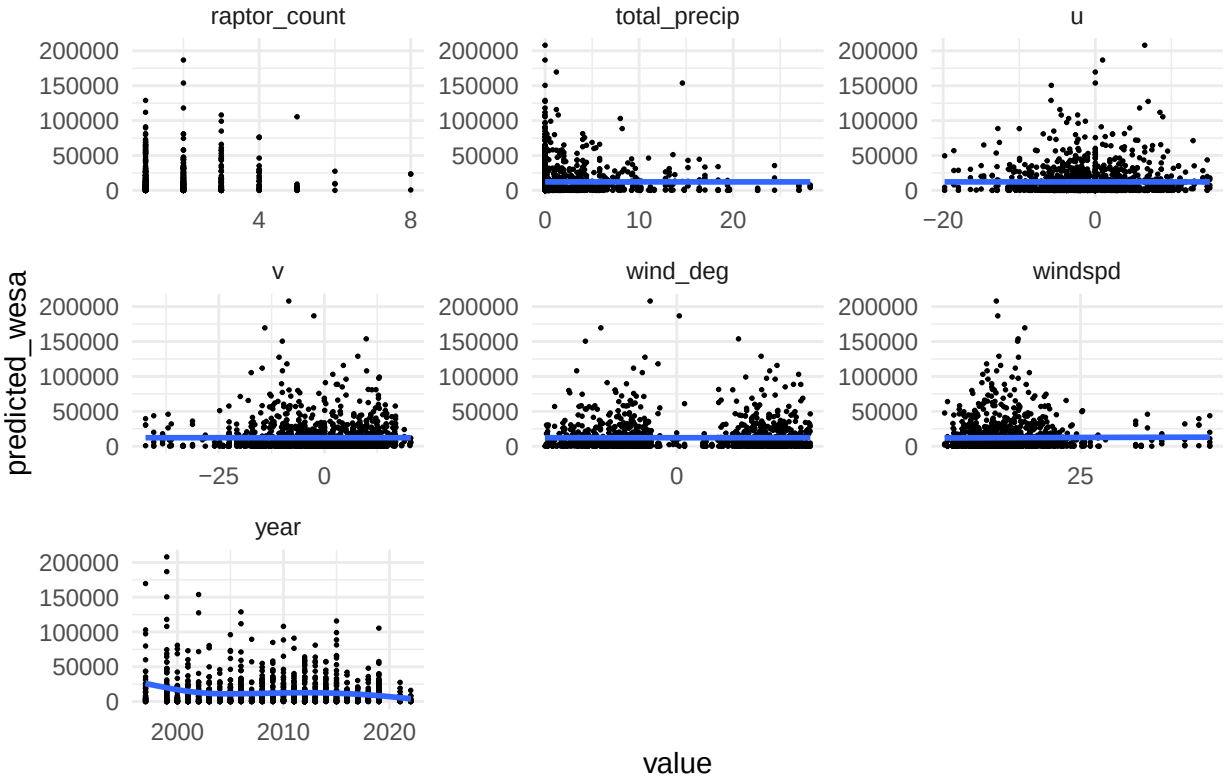
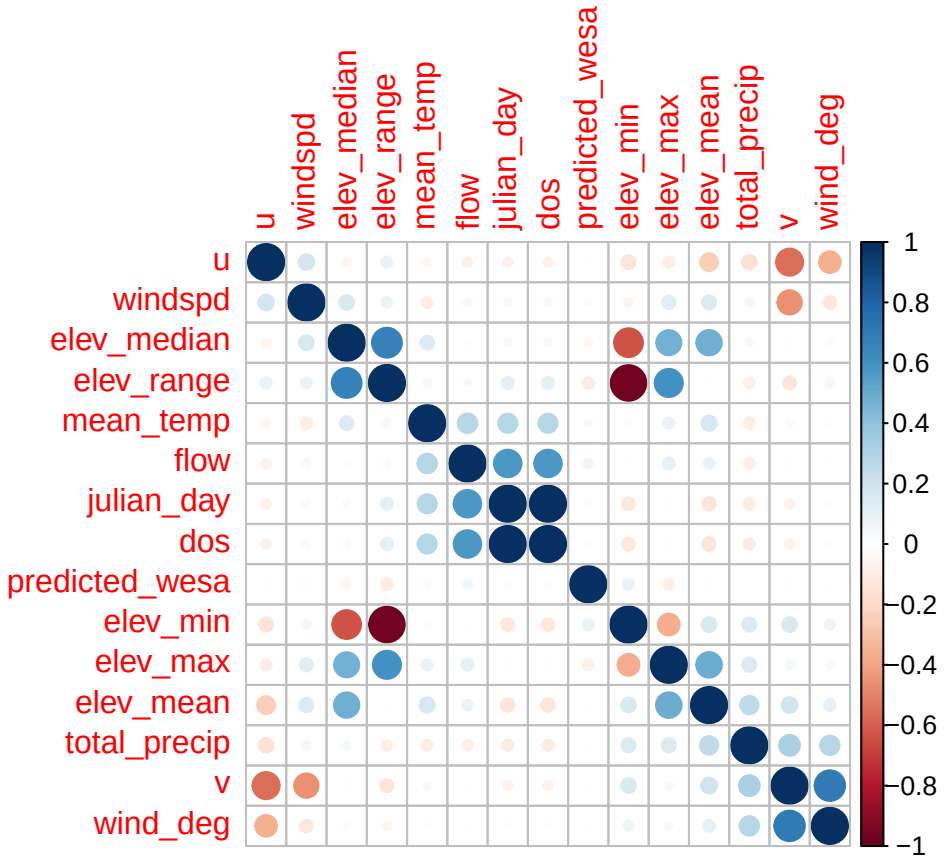


Figure 1: Histogram of WESA count per station per survey date. Plenty of zeroes...


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, two 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 ~ station_n + year_c + s(dos) + s(year,
    bs = "re"), data = dat, family = poisson, method = "ML")
pgam1 <- gam(predicted_wesa ~ station_n + s(flow) + year_c +
    s(dos) + s(year, bs = "re"), data = dat, family = poisson,
    method = "ML")
pgam2 <- gam(predicted_wesa ~ station_n + s(flow) + station_n:flow +
    year_c + s(dos) + s(year, bs = "re"), data = dat, family = poisson,
    method = "ML")

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

AIC(pgam0, pgam1, pgam2)
```

```
##             df      AIC
## pgam0 38.00000 13327684
## pgam1 47.00000 13095451
## pgam2 51.99883 12651156
```

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

```
##     df      AIC
## pm0  9 13767483
## pm1 10 13673799
## pm2 15 13274866
```

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

```
    year_c + s(dos) + s(year, bs = "re"), data = dat, family = nb,
    method = "ML")

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

AIC(nbgam0, nbgam1, nbgam2)
```

```
##              df      AIC
## nbgam0 31.00055 26240.13
## nbgam1 34.73274 26237.07
## nbgam2 39.06017 26232.10
```

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

```
##      df      AIC
## nbm0 10 26275.99
## nbm1 11 26274.30
## nbm2 16 26271.95
```

```
## Zero-inflated Poisson mgcv's ziplss can only fit using
## REML
zipgam0 <- gam(list(predicted_wesa ~ station_n + year_c + s(dos) +
    s(year, bs = "re"), ~station_n), data = dat, family = ziplss,
    method = "REML")
zipgam1 <- gam(list(predicted_wesa ~ station_n + s(flow) + year_c +
    s(dos) + s(year, bs = "re"), ~station_n), data = dat, family = ziplss,
    method = "REML")
zipgam2 <- gam(list(predicted_wesa ~ station_n + s(flow) + station_n:flow +
    year_c + s(dos) + s(year, bs = "re"), ~station_n + flow),
    data = dat, family = ziplss, method = "REML")
zipgam3 <- gam(list(predicted_wesa ~ station_n + year_c + s(dos) +
    s(year, bs = "re"), ~station_n * flow), data = dat, 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 ~ station_n + year_c + I(dos^2) +
    (1 | year), zi = ~station_n, data = dat, family = poisson)
zipm1 <- glmmTMB(predicted_wesa ~ station_n + scale(flow) + year_c +
    I(dos^2) + (1 | year), zi = ~station_n, data = dat, family = poisson)
zipm2 <- glmmTMB(predicted_wesa ~ station_n + scale(flow) + year_c +
    I(dos^2) + (1 | year), zi = ~station_n + flow, data = dat,
    family = poisson)
zipm3 <- glmmTMB(predicted_wesa ~ station_n * scale(flow) + year_c +
    I(dos^2) + (1 | year), zi = ~station_n * flow, data = dat,
```

```
    family = poisson)

# nbinom2 better fit than nbinom1 in all 3
zinb0 <- glmmTMB(predicted_wesa ~ station_n + year_c + I(dos^2) +
    (1 | year), zi = ~station_n, data = dat, family = nbinom2)
zinb1 <- glmmTMB(predicted_wesa ~ station_n + scale(flow) + year_c +
    I(dos^2) + (1 | year), zi = ~station_n + flow, data = dat,
    family = nbinom2)
zinb2 <- glmmTMB(predicted_wesa ~ station_n * scale(flow) + year_c +
    I(dos^2) + (1 | year), zi = ~station_n + flow, data = dat,
    family = nbinom2)

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

```
##               df      AIC
## zipgam0 44.00000 11172183
## zipgam1 53.00000 10942997
## zipgam2 60.43551 10652362
## zipgam3 50.00000 11172151
```

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

```
##       df         AIC
## zipm0 15 11666470.11
## zipm1 16 11577970.86
## zipm2 17 11577969.07
## zipm3 27          NA
## zinb0 16    25619.28
## zinb1 18    25614.86
## zinb2 23    25618.58
```

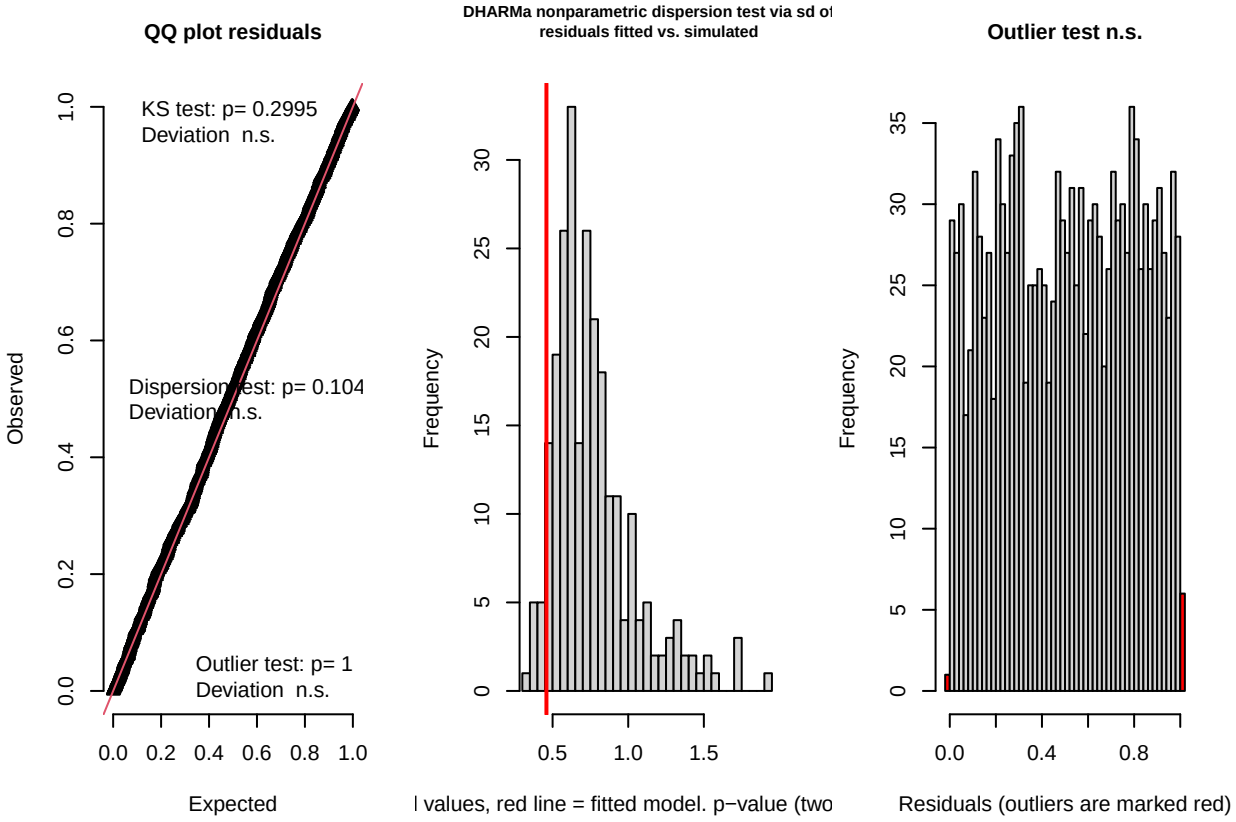
```
# Compare them all
bbmle::AICtab(pgam0, pgam1, pgam2, pm0, pm1, pm2, nbgam0, nbgam1,
    nbgam2, nbm0, nbm1, nbm2, zipgam0, zipgam1, zipgam2, zipm0,
    zipm1, zipm2, zinb0, zinb1, zinb2)
```

```
##         dAIC       df
## zinb1          0.0 18
## zinb2          3.7 23
## zinb0          4.4 16
## nbgam2       617.2 39.1
## nbgam1       622.2 34.7
## nbgam0       625.3 31
## nbm2         657.1 16
## nbm1         659.4 11
## nbm0         661.1 10
## zipgam2 10626746.7 60.4
## zipgam1 10917382.0 53
## zipgam0 11146567.6 44
## zipm2   11552354.2 17
## zipm1   11552356.0 16
## zipm0   11640855.3 15
```

```
## pgam2   12625540.9 52
## pgam1   13069835.9 47
## pm2     13249251.6 15
## pgam0   13302069.5 38
## pm1     13648184.1 10
## pm0     13741868.1 9
```

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 = 0.02614, p-value = 0.2995
## alternative hypothesis: two-sided
##
##
## $dispersion
##
##  DHARMa nonparametric dispersion test via sd of residuals fitted vs.
##  simulated
##
## data:  simulationOutput
```

```
## dispersion = 0.59606, p-value = 0.104
## alternative hypothesis: two.sided
##
##
## $outliers
##
##  DHARMa outlier test based on exact binomial test with approximate
##  expectations
##
## data:  simulationOutput
## outliers at both margin(s) = 11, observations = 1387, p-value = 1
## alternative hypothesis: true probability of success is not equal to 0.007968127
## 95 percent confidence interval:
##  0.003965475 0.014145965
## sample estimates:
## frequency of outliers (expected: 0.00796812749003984 )
##                                             0.007930786

## $uniformity
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  simulationOutput$scaledResiduals
## D = 0.02614, p-value = 0.2995
## alternative hypothesis: two-sided
##
##
## $dispersion
##
##  DHARMa nonparametric dispersion test via sd of residuals fitted vs.
##  simulated
##
## data:  simulationOutput
## dispersion = 0.59606, p-value = 0.104
## alternative hypothesis: two.sided
##
##
## $outliers
##
##  DHARMa outlier test based on exact binomial test with approximate
##  expectations
##
## data:  simulationOutput
## outliers at both margin(s) = 11, observations = 1387, p-value = 1
## alternative hypothesis: true probability of success is not equal to 0.007968127
## 95 percent confidence interval:
##  0.003965475 0.014145965
## sample estimates:
## frequency of outliers (expected: 0.00796812749003984 )
##                                             0.007930786
```

Test for zero inflation

```
##
```

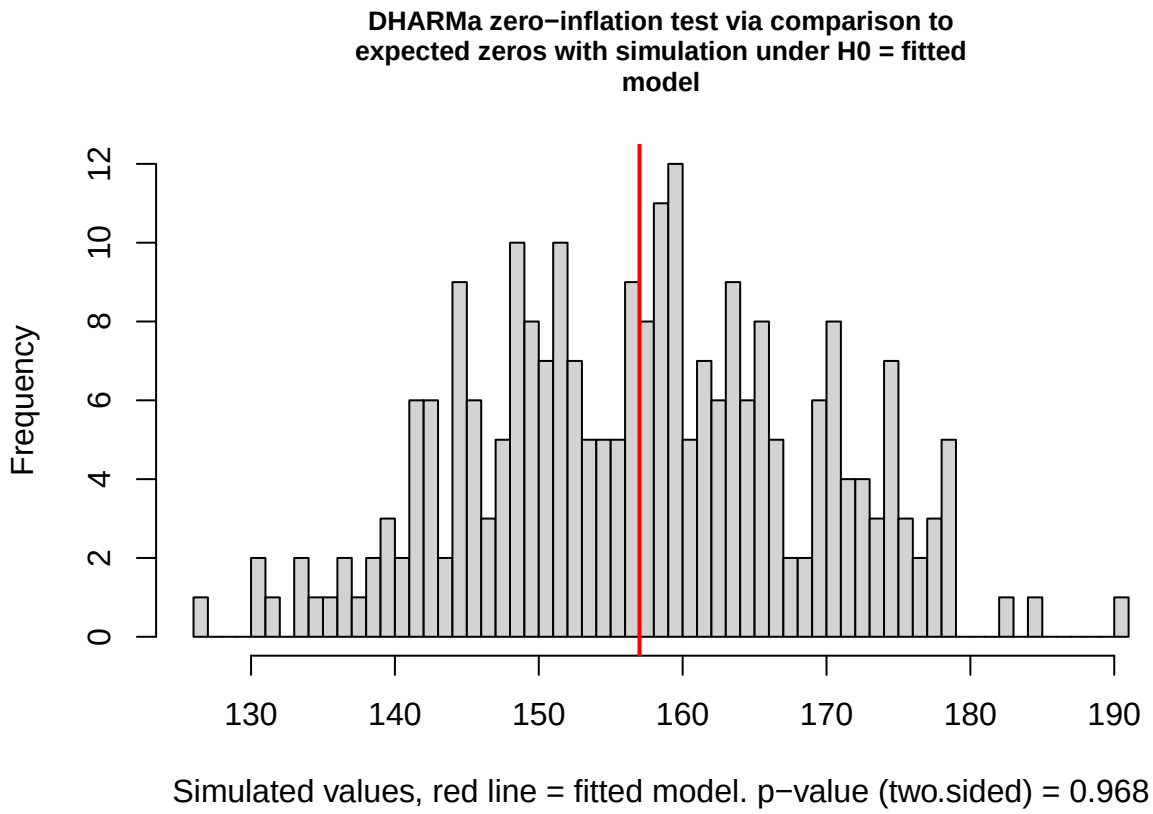


Figure 2: The zero-inflation test indicates we're fitting the zeros very well with the base model.

```
##  DHARMa zero-inflation test via comparison to expected zeros with
##  simulation under H0 = fitted model
##
## data:  simulationOutput
## ratioObsSim = 0.99556, p-value = 0.968
## alternative hypothesis: two.sided
```

Full model

```
##  Family: nbinom2  ( log )
## Formula:
## predicted_wesa ~ station_n * scale(flow) + year_c + scale(mean_temp) +
##     scale(elev_range) + tide + scale(total_precip) + scale(u) +
##     I(dos^2) + (1 | year)
## Zero inflation:                   ~.
## Data: dat
##
##      AIC      BIC   logLik deviance df.resid
##  25458.2  25672.8 -12688.1  25376.2     1346
##
## Random effects:
##
## Conditional model:
##  Groups Name        Variance Std.Dev.
##  year   (Intercept) 0.1574   0.3968
## Number of obs: 1387, groups:  year, 24
##
## Zero-inflation model:
##  Groups Name        Variance Std.Dev.
##  year   (Intercept) 0.7673   0.876
## Number of obs: 1387, groups:  year, 24
##
## Dispersion parameter for nbinom2 family (): 0.887
##
## Conditional model:
##                                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)                         10.499228   0.118747   88.42   <2e-16 ***
## station_nView corner                -0.153823   0.129524   -1.19   0.2350
## station_nPilings                    -1.010683   0.099284  -10.18   <2e-16 ***
## station_nBend                       -1.832555   0.143827  -12.74   <2e-16 ***
## station_n34th St pullout            -1.402858   0.121980  -11.50   <2e-16 ***
## station_nCoal Port                  -1.008139   0.098812  -10.20   <2e-16 ***
## scale(flow)                         -0.006661   0.078038   -0.09   0.9320
## year_c                              -0.180841   0.087564   -2.07   0.0389 *
## scale(mean_temp)                    -0.009402   0.036256   -0.26   0.7954
## scale(elev_range)                   -0.065923   0.038363   -1.72   0.0857 .
## tiderising                           0.168633   0.078030    2.16   0.0307 *
## scale(total_precip)                  0.030669   0.032968    0.93   0.3522
## scale(u)                             0.011724   0.033612    0.35   0.7272
## I(dos^2)                            -0.836862   0.030653  -27.30   <2e-16 ***
## station_nView corner:scale(flow)    -0.030466   0.129237   -0.24   0.8136
## station_nPilings:scale(flow)        -0.156612   0.101124   -1.55   0.1215
## station_nBend:scale(flow)           -0.023932   0.143571   -0.17   0.8676
```

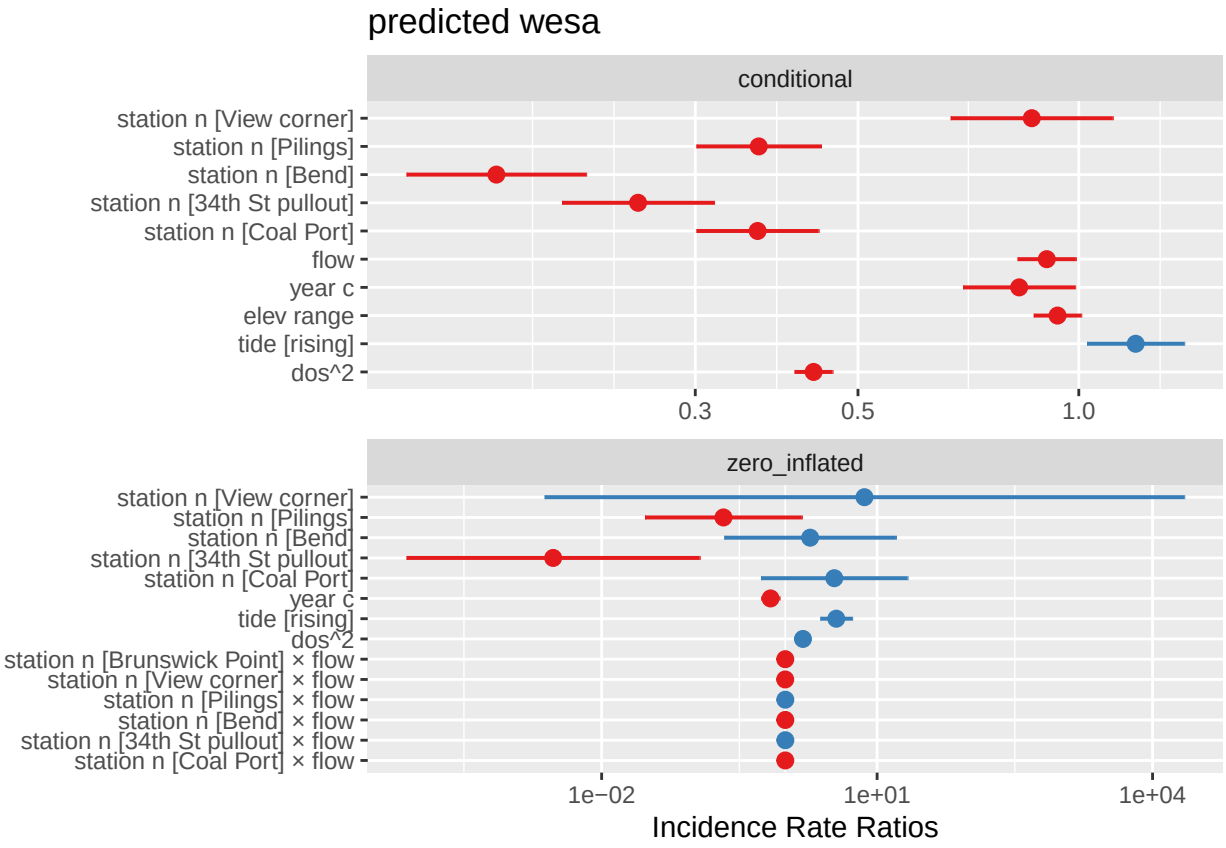
```
## station_n34th St pullout:scale(flow) -0.220996   0.119776   -1.85   0.0650 .
## station_nCoal Port:scale(flow)       -0.081179   0.103589   -0.78   0.4332
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Zero-inflation model:
##                                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)                         -5.31096    0.47644 -11.147  < 2e-16 ***
## station_nView corner                -3.04408    3.22833  -0.943  0.34572
## station_nPilings                     1.74292    0.41308   4.219 2.45e-05 ***
## station_nBend                        2.17631    0.47731   4.560 5.13e-06 ***
## station_n34th St pullout             0.99957    0.64251   1.556  0.11978
## station_nCoal Port                   2.38157    0.39384   6.047 1.48e-09 ***
## scale(flow)                         -0.50076    0.34725  -1.442  0.14928
## year_c                              -0.39474    0.21860  -1.806  0.07096 .
## scale(mean_temp)                    -0.27744    0.12419  -2.234  0.02549 *
## scale(elev_range)                   -0.13957    0.11809  -1.182  0.23724
## tiderising                           1.09014    0.24368   4.474 7.69e-06 ***
## scale(total_precip)                 -0.19198    0.11804  -1.626  0.10387
## scale(u)                             0.03478    0.11229   0.310  0.75674
## I(dos^2)                             0.53286    0.09778   5.450 5.04e-08 ***
## station_nView corner:scale(flow)    -1.67855    2.53179  -0.663  0.50734
## station_nPilings:scale(flow)         1.13032    0.37977   2.976  0.00292 **
## station_nBend:scale(flow)            0.42895    0.41877   1.024  0.30568
## station_n34th St pullout:scale(flow)  2.11627    0.51867   4.080 4.50e-05 ***
## station_nCoal Port:scale(flow)       0.30585    0.37004   0.827  0.40849
## ---
## 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: nbinom2  ( log )
## Formula:
## predicted_wesa ~ station_n + scale(flow) + year_c + scale(elev_range) +
##     tide + I(dos^2) + (1 | year)
## Zero inflation:
## ~station_n + year_c + tide + I(dos^2) + station_n:flow
## Data: dat
##
##      AIC      BIC   logLik deviance df.resid
##  25479.3  25625.9 -12711.7  25423.3     1359
##
## Random effects:
##
## Conditional model:
##  Groups Name        Variance Std.Dev.
##  year   (Intercept) 0.1685   0.4105
## Number of obs: 1387, groups:  year, 24
##
## Dispersion parameter for nbinom2 family (): 0.881
```

```
##
## Conditional model:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)                10.49033    0.12066   86.94   <2e-16 ***
## station_nView corner       -0.14755    0.12982   -1.14   0.2557
## station_nPilings           -1.00465    0.09967  -10.08   <2e-16 ***
## station_nBend              -1.82833    0.14381  -12.71   <2e-16 ***
## station_n34th St pullout -1.38356    0.12172  -11.37   <2e-16 ***
## station_nCoal Port         -1.00826    0.09779  -10.31   <2e-16 ***
## scale(flow)                -0.10049    0.04665   -2.15   0.0312 *
## year_c                     -0.18713    0.08971   -2.09   0.0370 *
## scale(elev_range)          -0.06675    0.03803   -1.75   0.0793 .
## tiderising                  0.17856    0.07781    2.29   0.0217 *
## I(dos^2)                   -0.83248    0.03036  -27.42   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Zero-inflation model:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   -3.194e+00  8.634e-01  -3.700 0.000216 ***
## station_nView corner           1.986e+00  4.092e+00   0.485 0.627473
## station_nPilings              -1.551e+00  1.003e+00  -1.545 0.122256
## station_nBend                  6.235e-01  1.097e+00   0.568 0.569932
## station_n34th St pullout      -5.821e+00  1.875e+00  -3.104 0.001908 **
## station_nCoal Port             1.229e+00  9.373e-01   1.311 0.189763
## year_c                        -3.696e-01  1.162e-01  -3.181 0.001466 **
## tiderising                     1.279e+00  2.029e-01   6.304 2.90e-10 ***
## I(dos^2)                       4.435e-01  8.763e-02   5.061 4.17e-07 ***
## station_nBrunswick Point:flow -5.290e-04  2.847e-04  -1.858 0.063141 .
## station_nView corner:flow     -2.134e-03  2.119e-03  -1.007 0.313860
## station_nPilings:flow          4.371e-04  1.391e-04   3.143 0.001674 **
## station_nBend:flow            -1.464e-05  1.989e-04  -0.074 0.941327
## station_n34th St pullout:flow  1.432e-03  3.494e-04   4.098 4.16e-05 ***
## station_nCoal Port:flow       -2.251e-04  1.218e-04  -1.848 0.064642 .
## ---
## 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 = 0.030382, p-value = 0.1544
## alternative hypothesis: two-sided
##
##
## $dispersion
##
##  DHARMa nonparametric dispersion test via sd of residuals fitted vs.
##  simulated
##
## data:  simulationOutput
## dispersion = 0.57669, p-value = 0.08
## alternative hypothesis: two.sided
##
##
## $outliers
##
##  DHARMa outlier test based on exact binomial test with approximate
##  expectations
##
```

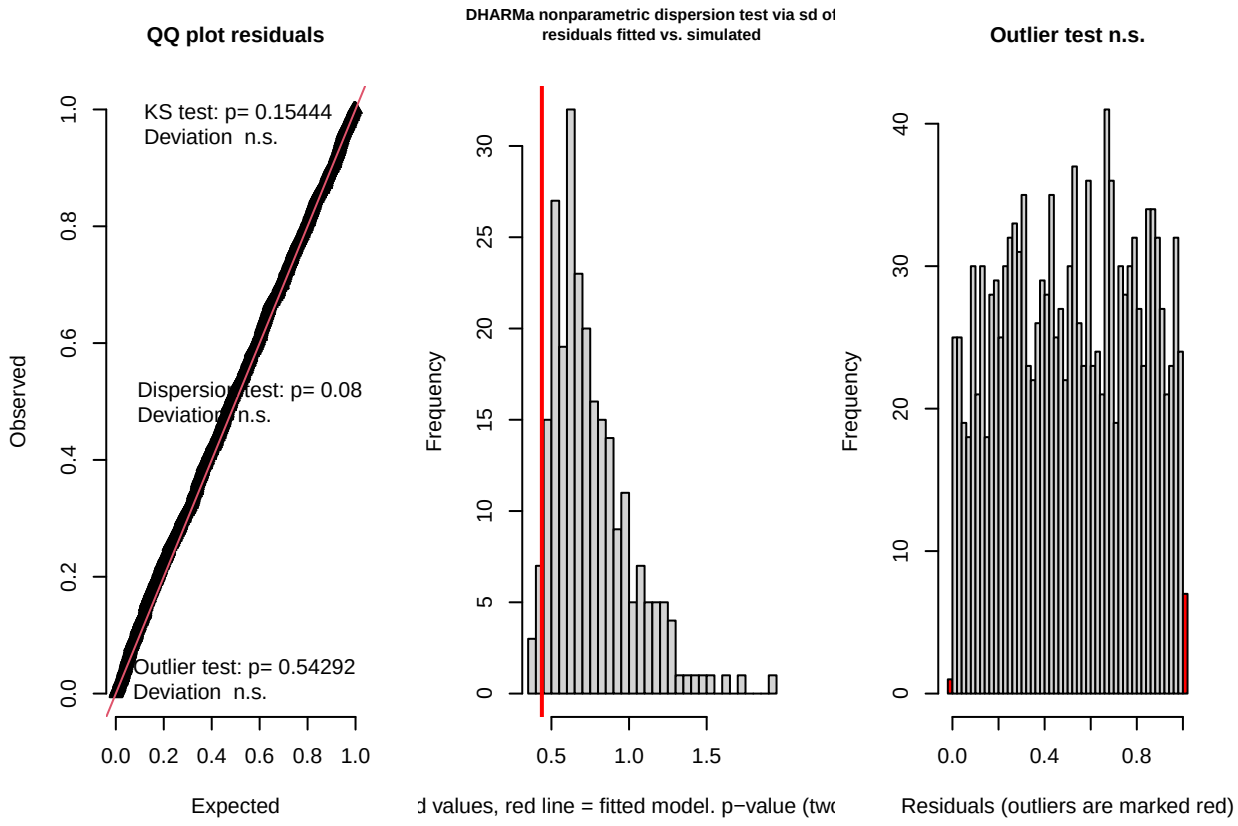


Figure 3: Residual diagnostics.

```
## data:  simulationOutput
## outliers at both margin(s) = 13, observations = 1387, p-value = 0.5429
## alternative hypothesis: true probability of success is not equal to 0.007968127
## 95 percent confidence interval:
##  0.004999762 0.015974353
## sample estimates:
## frequency of outliers (expected: 0.00796812749003984 )
##                                             0.009372747

## $uniformity
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  simulationOutput$scaledResiduals
## D = 0.030382, p-value = 0.1544
## alternative hypothesis: two-sided
##
##
## $dispersion
##
##  DHARMa nonparametric dispersion test via sd of residuals fitted vs.
##  simulated
##
## data:  simulationOutput
## dispersion = 0.57669, p-value = 0.08
## alternative hypothesis: two.sided
##
##
## $outliers
##
##  DHARMa outlier test based on exact binomial test with approximate
##  expectations
##
## data:  simulationOutput
## outliers at both margin(s) = 13, observations = 1387, p-value = 0.5429
## alternative hypothesis: true probability of success is not equal to 0.007968127
## 95 percent confidence interval:
##  0.004999762 0.015974353
## sample estimates:
## frequency of outliers (expected: 0.00796812749003984 )
##                                             0.009372747


##
##  DHARMa zero-inflation test via comparison to expected zeros with
##  simulation under H0 = fitted model
##
## data:  simulationOutput
## ratioObsSim = 1.0007, p-value = 0.976
## alternative hypothesis: two.sided
```

Residuals vs. predicted

```
##
```

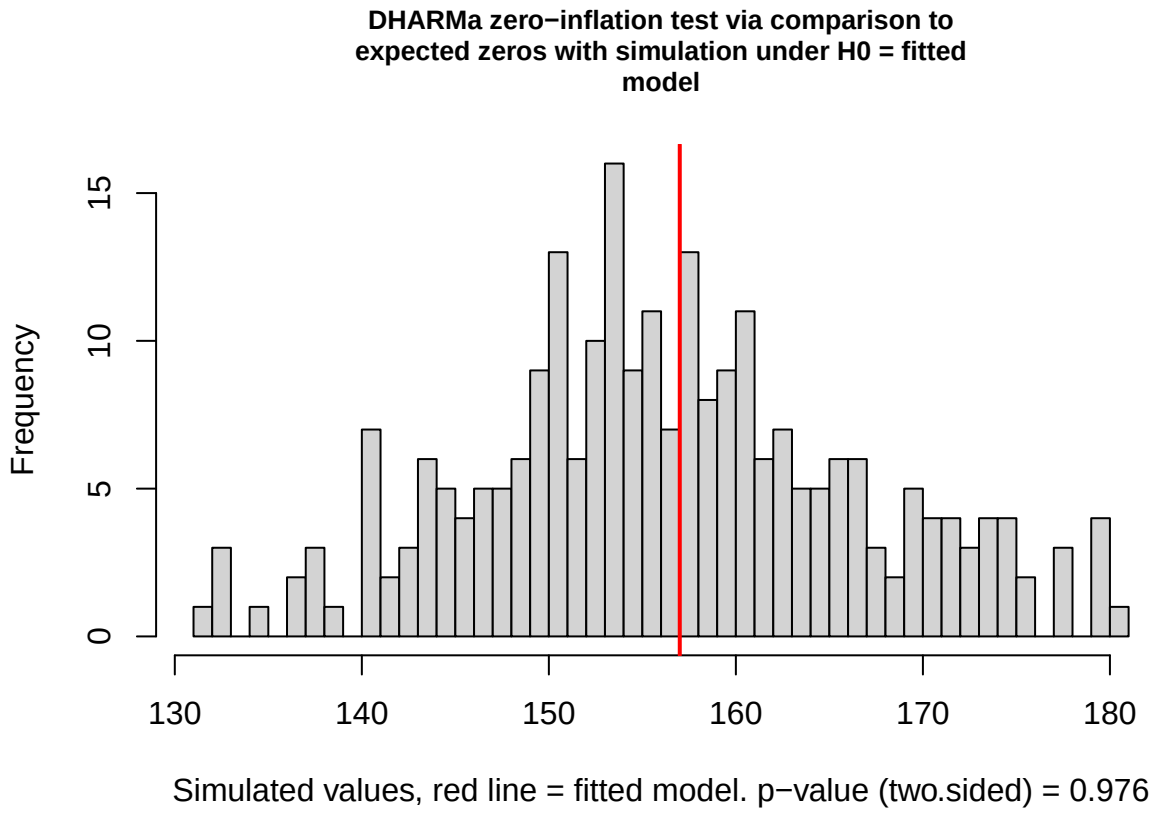
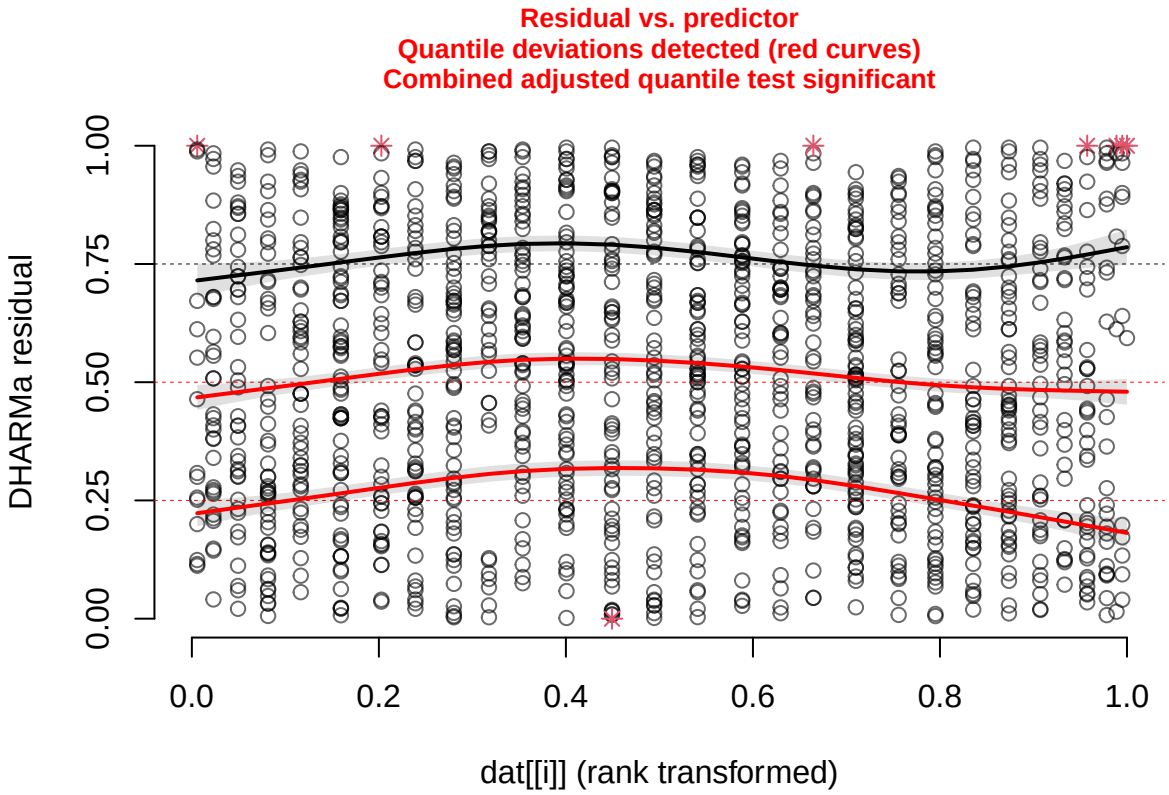
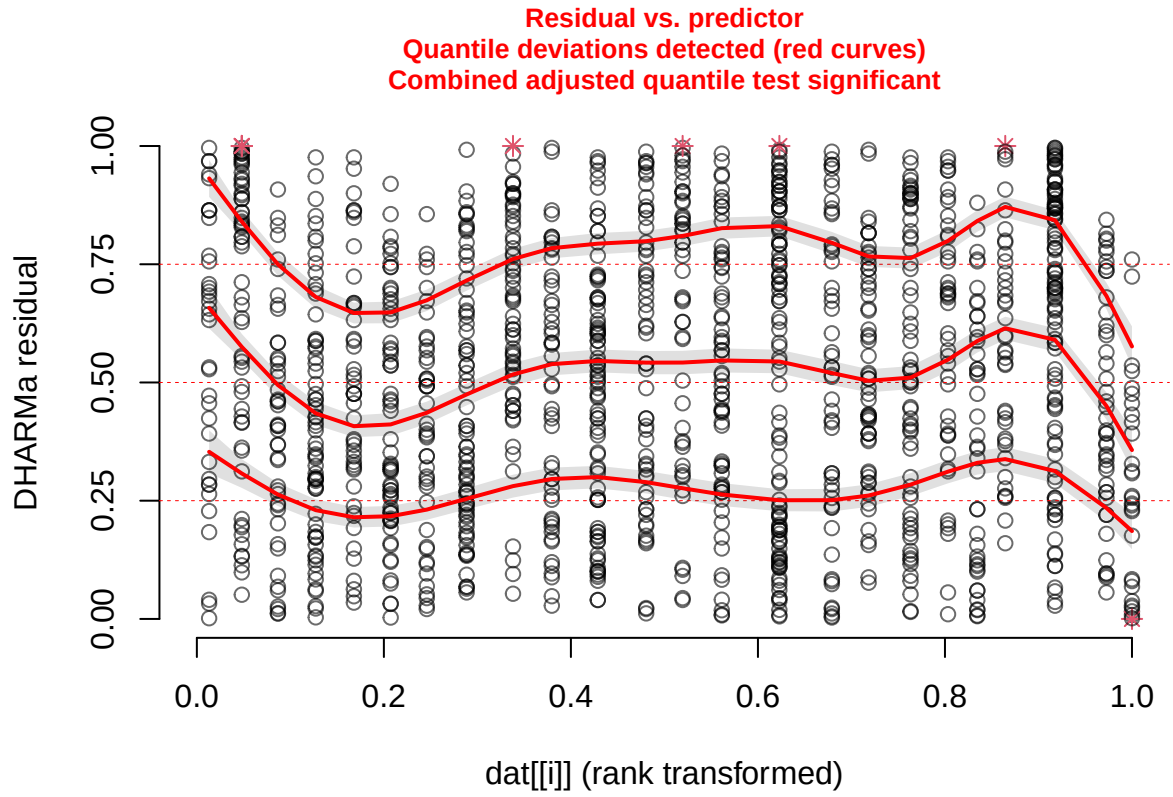


Figure 4: Testing for overdispersion. Still not quite predicting the number of zeroes exactly correctly but better than before.

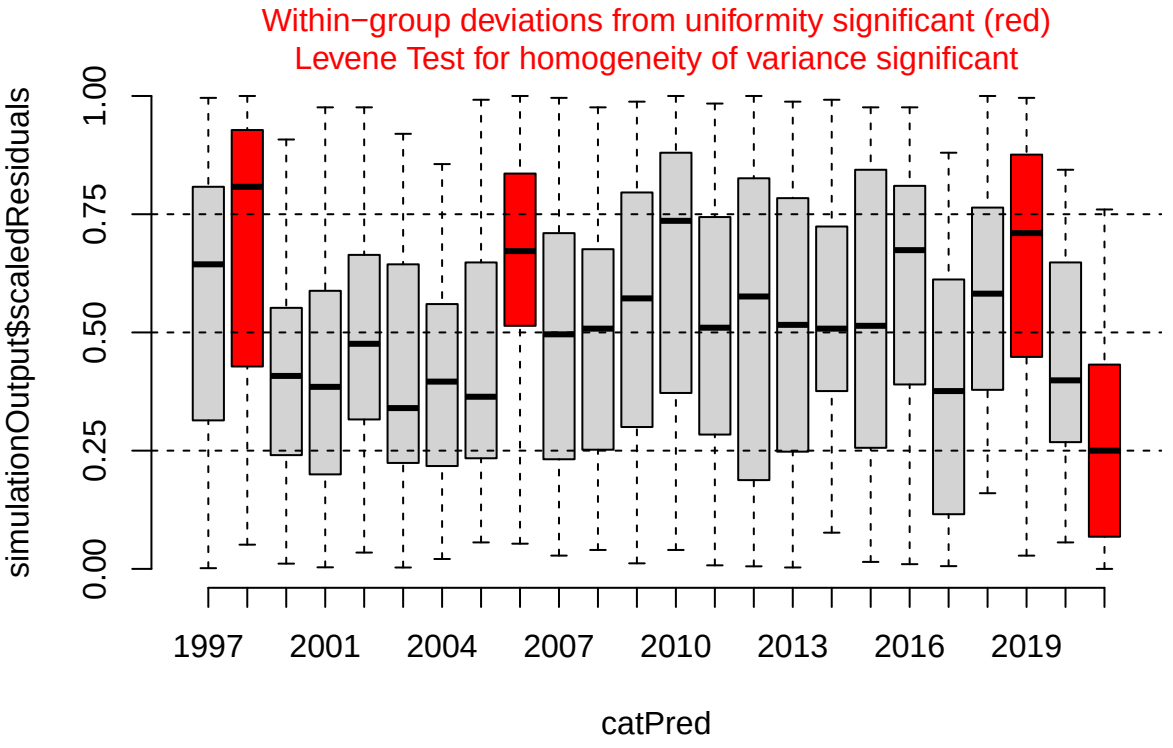
```
##  dos
```


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
dat[[i]] (rank transformed)

```
##
##  year_c
```

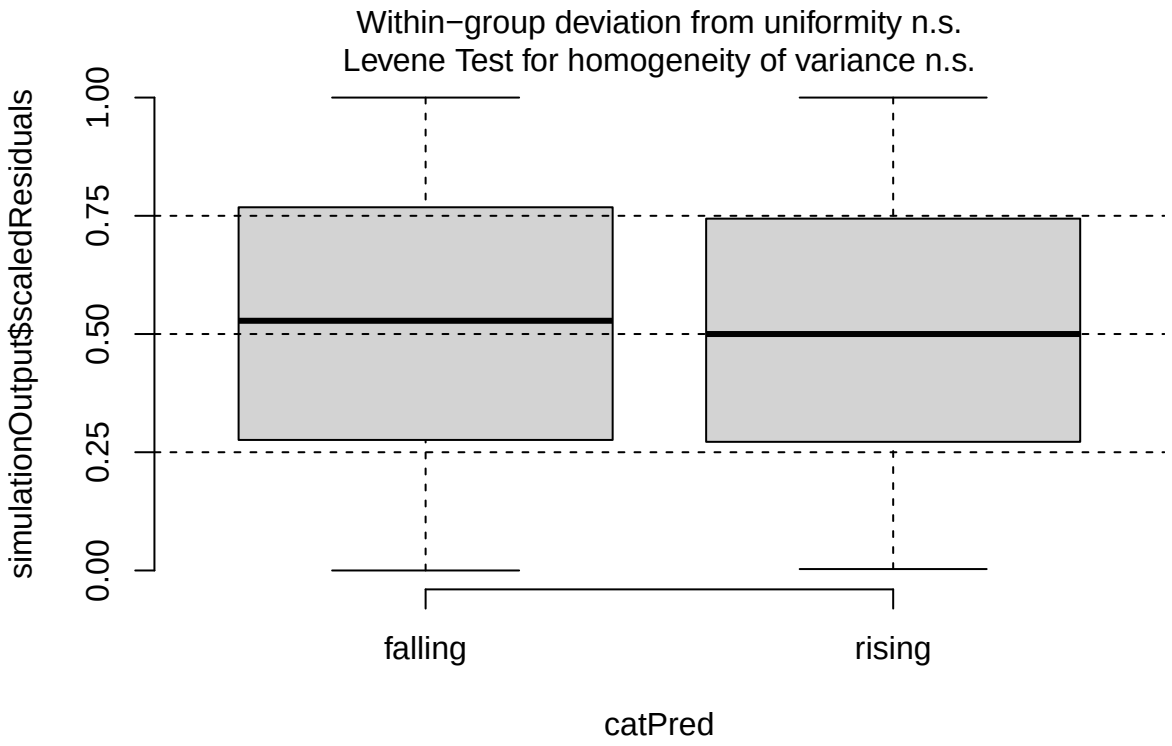

Residual vs. predictor
Quantile deviations detected (red curves)
Combined adjusted quantile test significant
DHARMa residual
0.00
0.25
0.50
0.75
1.00
0.0
0.2
0.4
0.6
0.8
1.0
dat[[i]] (rank transformed)

```
##
##   year
```

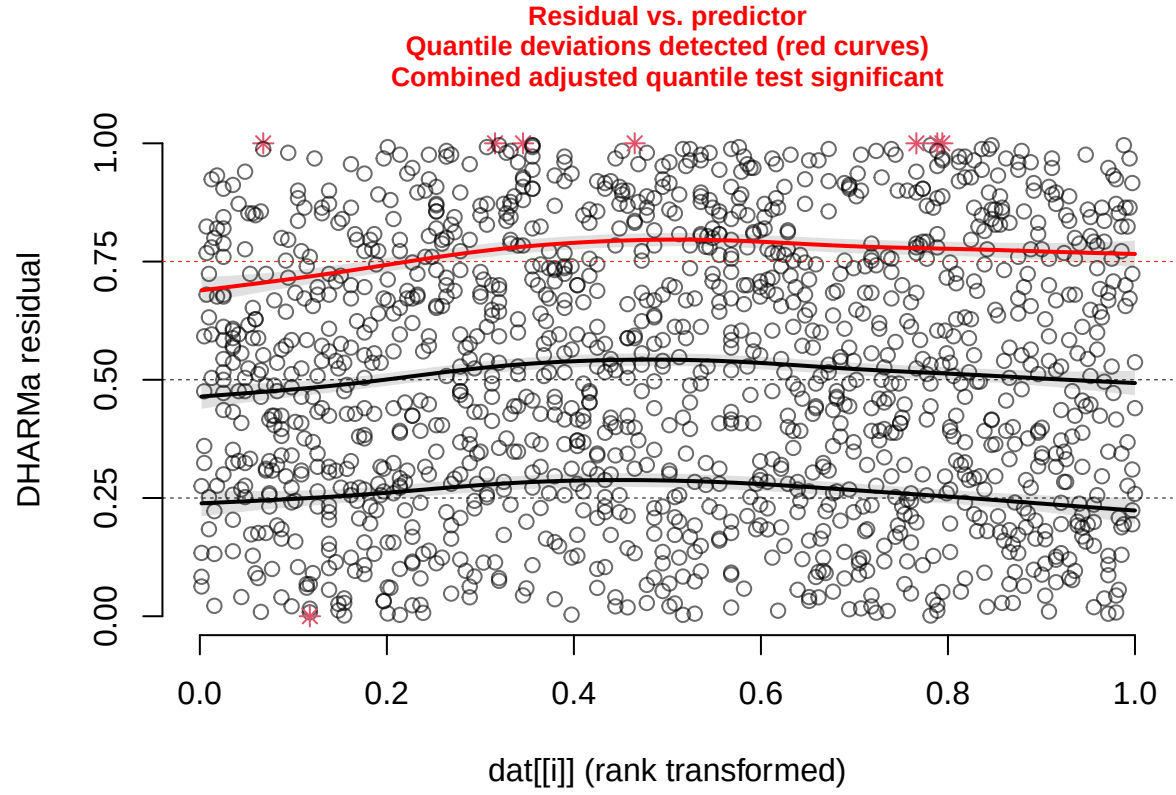

Within−group deviations from uniformity significant (red)
Levene Test for homogeneity of variance significant
simulationOutput$scaledResiduals
0.00
0.25
0.50
0.75
1.00
1997
2001
2004
2007
2010
2013
2016
2019
catPred

```
##
```

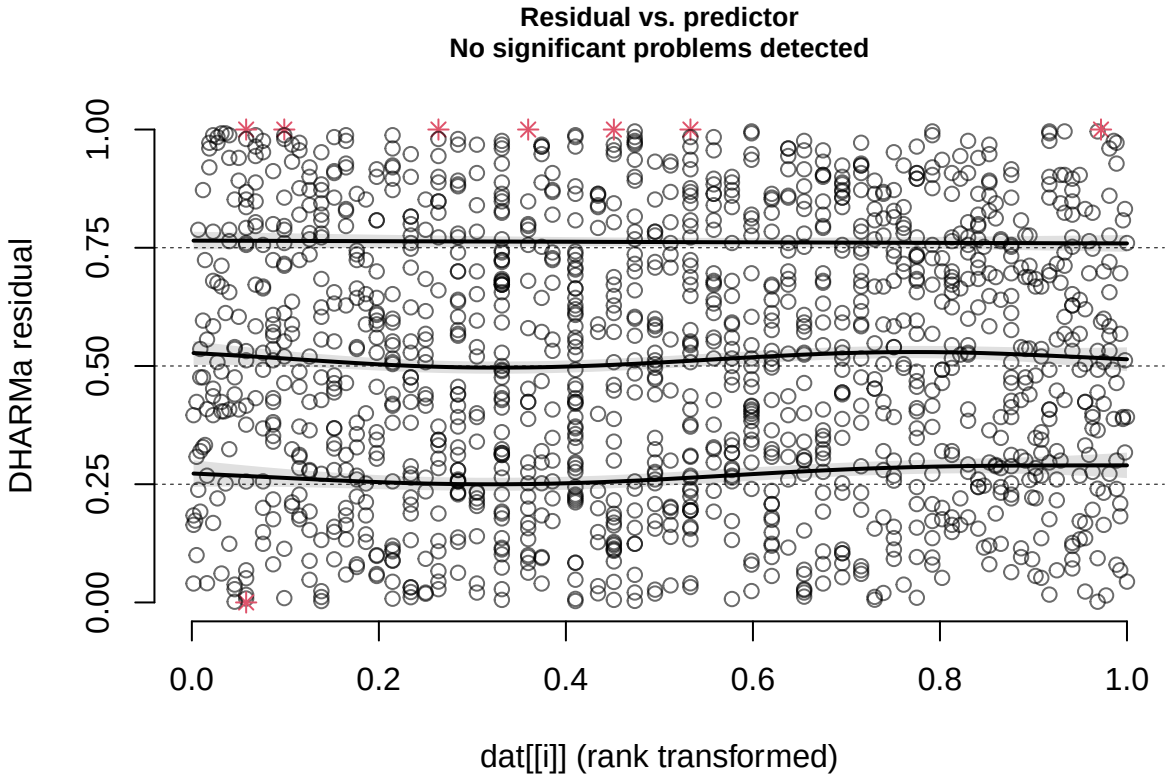
```
##  tide
```


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

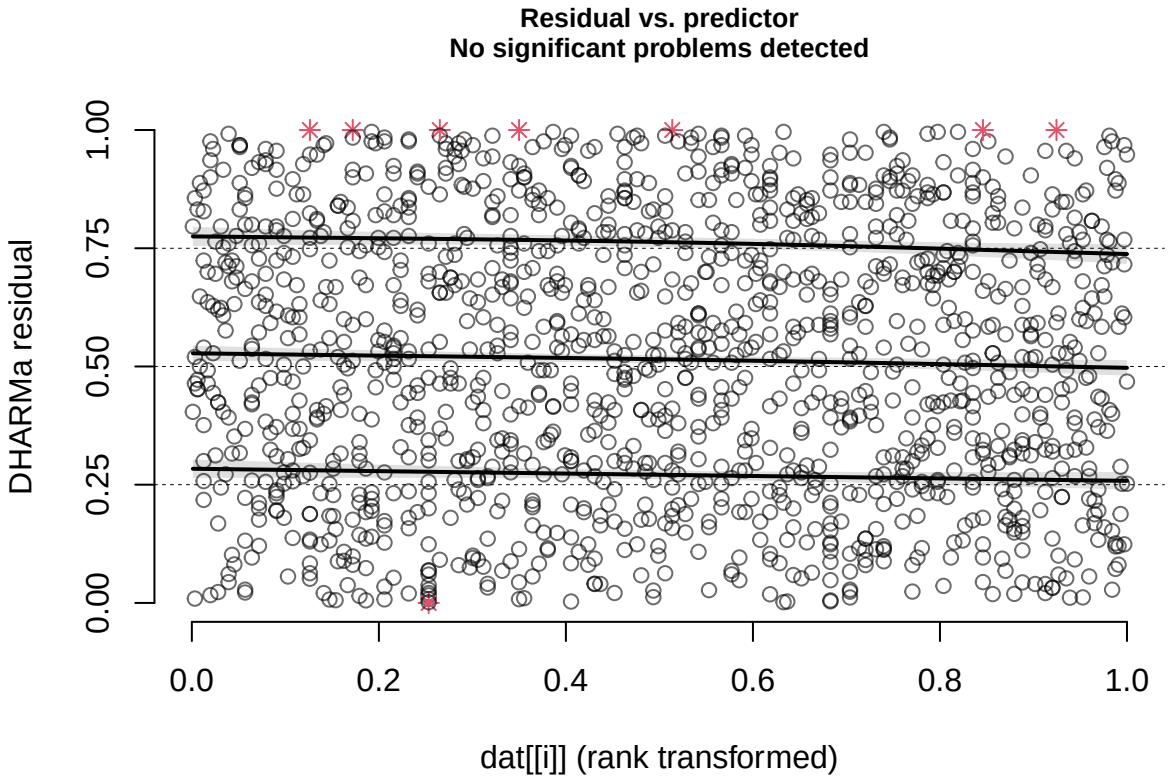
```
##
##  flow
```


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
dat[[i]] (rank transformed)

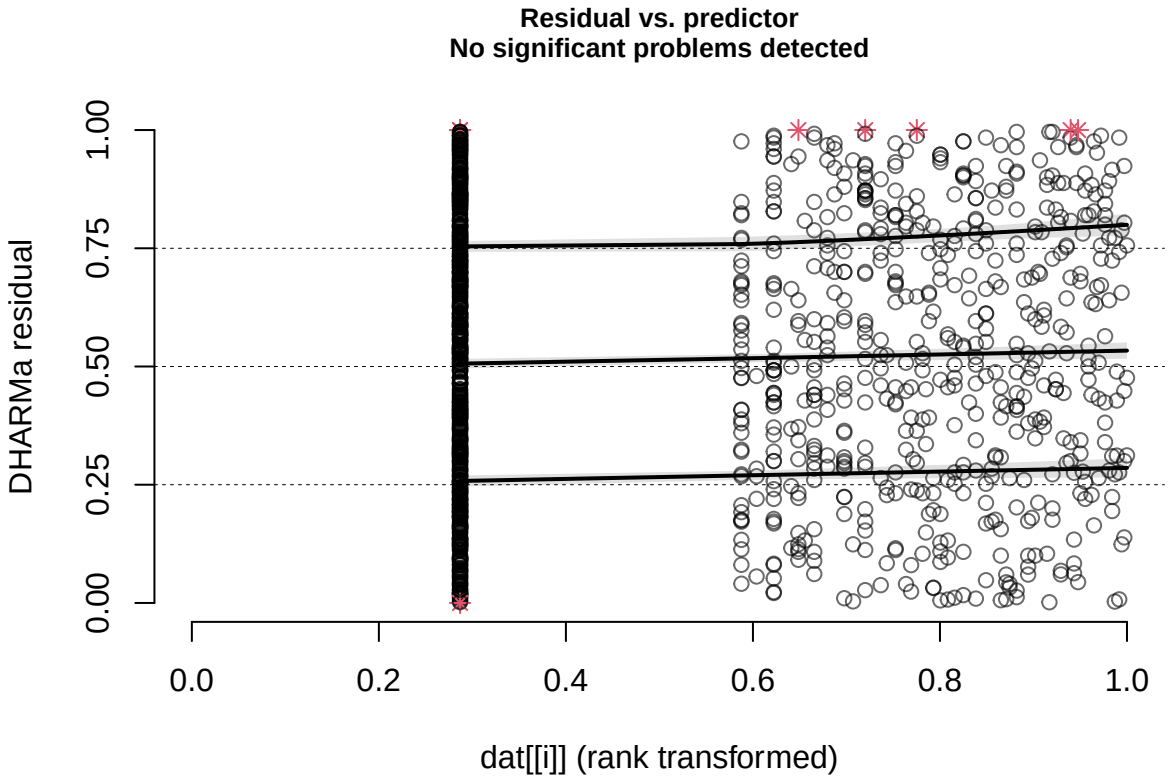
```
##
##  mean_temp
```



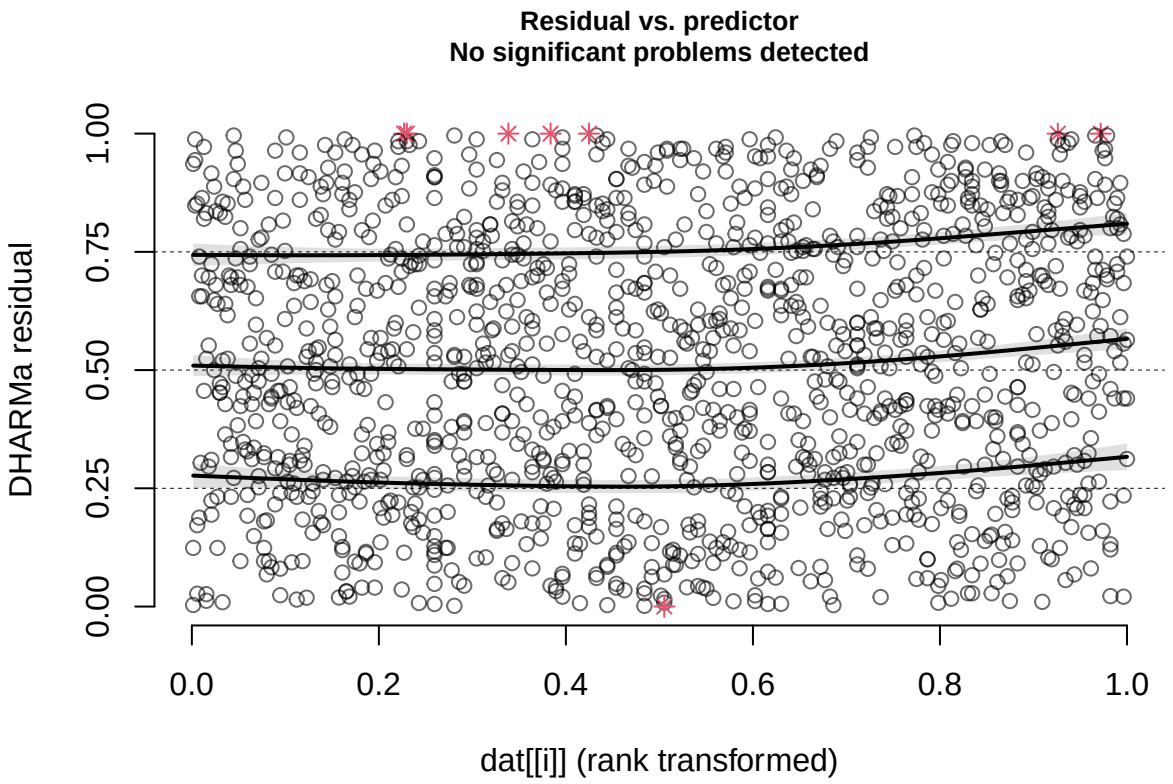

```
##
##  elev_range
```




```
##
##  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
dat[[i]] (rank transformed)