# Analysis Chlorophyll RFU 2022

Emanuel Mauch

2023-07-20

## Introduction

This is a report of what I've done so far regarding the analysis of Chlorophyll RFU from the Greenland Sondes data 2022. The analysis includes a t-test and a linear mixed model.

## Step 0: set up R-script

```
rm(list= ls())

setwd("~/ZIVI_EAWAG/project_22")

source("~/ZIVI_EAWAG/project_22/Moritz_Luehrig_paper_stuff/methods_packages.R")

# Mixed model packages
library(lme4)
library(nlme)

# Assisting packages
library(car)
library(pastecs)
library(ggpubr)
library(GGally)
library(effects)
library(arm)
library(MuMIn)

sonde_key = fread( "~/ZIVI_EAWAG/project_22/ponds_sonde_key.txt", header=T)
# Treatments:
# S: fish from single population (L26)
# D: fish from 2 populations (double, L26+ERL33)
# NF: no fish

# daily pond-wise means (want to account for daily fluctuations by using the daily
# means of the individual ponds, simultaneously having enough datapoints to fit the model
# -> need more datapoints than levels of the random effect Sonde later)
dat = fread("~/ZIVI_EAWAG/project_22/data/ponds_sonde_data_daily_avg.txt")
```

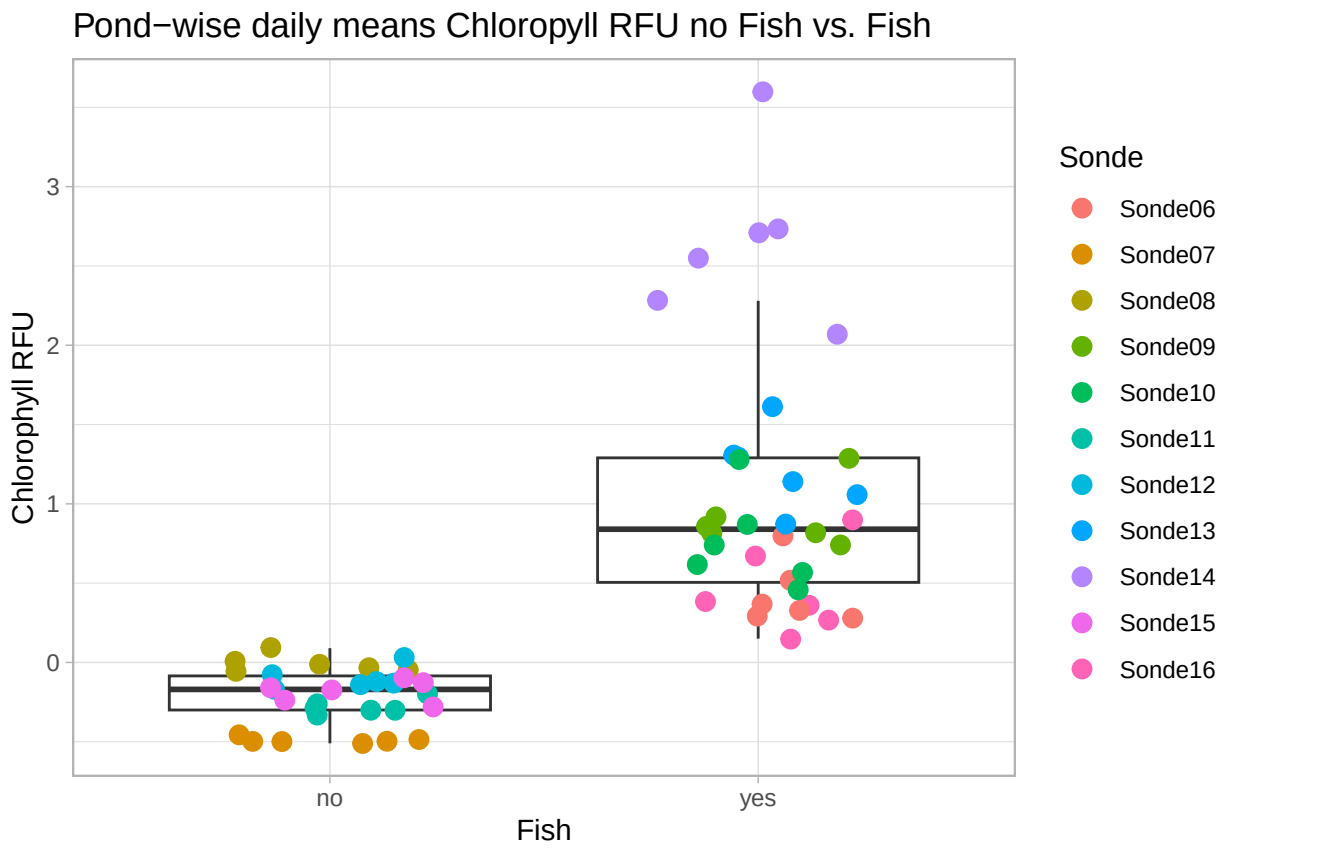
## Step 1: format dataset

```
dat = dat[, c("Pond", "Treatment", "Sonde", "Chlorophyll_RFU")]

# Summarize S (fish from single population) and D (fish from two populations) into F (Fish)
dat$Fish <- recode(dat$Treatment, "c('S', 'D') = 'yes'; c('NF') = 'no'")

# Convert characters to factors
dat$Fish <- factor(dat$Fish)
dat$Sonde <- factor(dat$Sonde)
```

## Step 2: EDA

### Step 2.1: Visualize research question

```
# Graphically
ggplot(dat, aes(Fish, Chlorophyll_RFU)) +
  geom_boxplot(outlier.shape = NA) +
  geom_jitter(aes(col = Sonde), width = 0.25, alpha=1, size = 3) +
  labs(y = "Chlorophyll RFU", title = "Pond-wise daily means Chloropyll RFU no Fish vs. Fish") +
  theme_light()
```



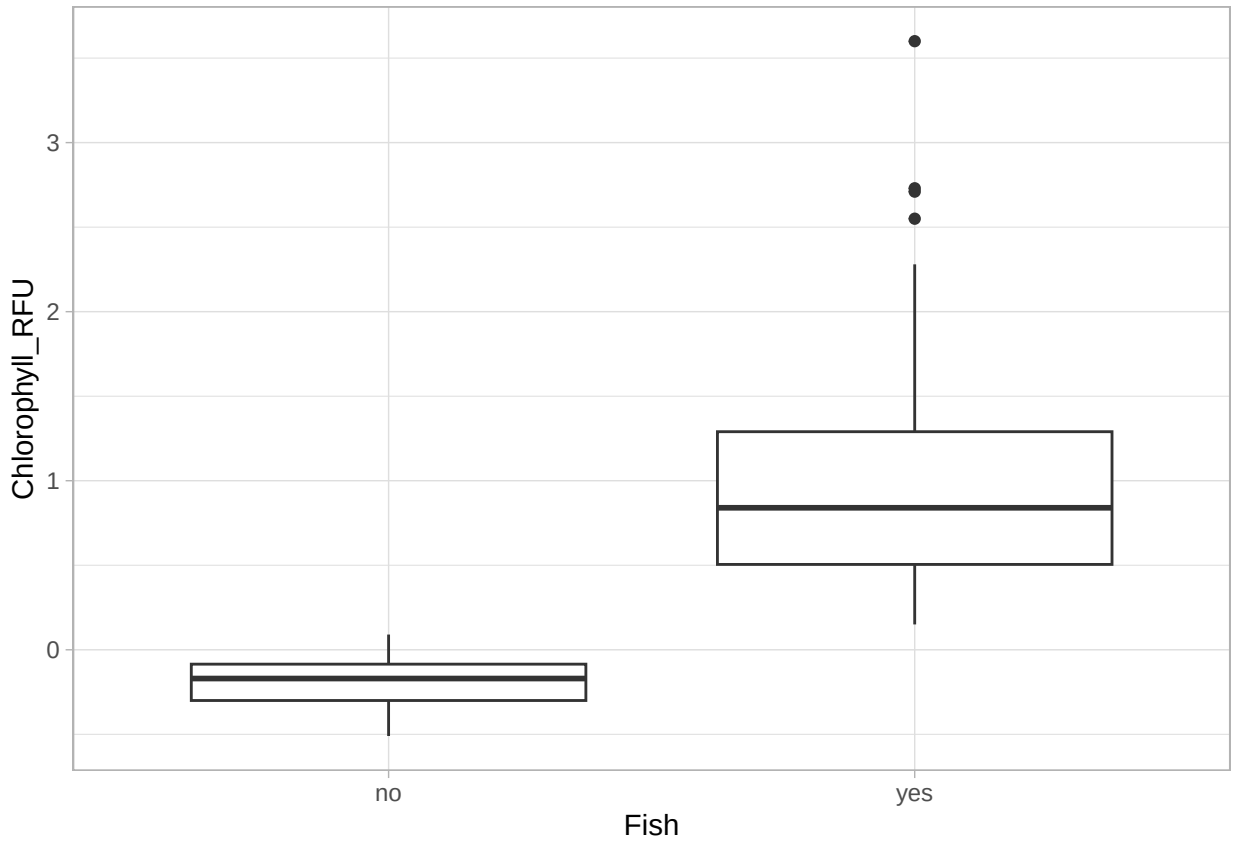
```
# Different sondes seem to have different intercepts, even though the treatment
# is the same
```

### Step 2.2: Independence of Chlorophyll_RFU

Although Chlorophyll_RFU is strongly time-dependent, we sort of accounted for this dependence by taking the overall or daily means of the individual ponds.

### Step 2.3. homogeneity of Chlorophyll_RFU

```
ggplot(dat, aes(Fish, Chlorophyll_RFU)) +
  geom_boxplot() +
  theme_light()
```




```
# Variance in ponds with fish much higher
```

## Step 3: model fitting

Approach:

1. T-test

- Since the assumption of homogeneity of variance is violated (ponds with Fish seem to vary much more in their diurnal Chlorophyll pattern, which is also part of the research question, but can be addressed at a later stage), we can just do a two-sample t-test allowing for unequal variances.
- A concern is not accounting for the fact that the sondes have an impact on the measurements as well (as well as the solar radiation).
- I addition, I'm not sure what consequences follow if one uses aggregated data (comparing pond-wise means) to do a t-test. What I am sure is that one looses information about the distribution of chloro in the different ponds, as well power to detect a relationship.

2. Mixed model

- I was also thinking about a mixed model with fish as fixed effects and sondes as a random effect with random intercepts. For that, I will be using the daily means, to sort of adjust for the daily, time-dependent fluctuations in Chlorophyll RFU

Hypotheses:

- H0: no associaton between the presence/absence of fish and chlorophyll RFU
- H1: association between presence/absence of fish and chlorophyll RFU

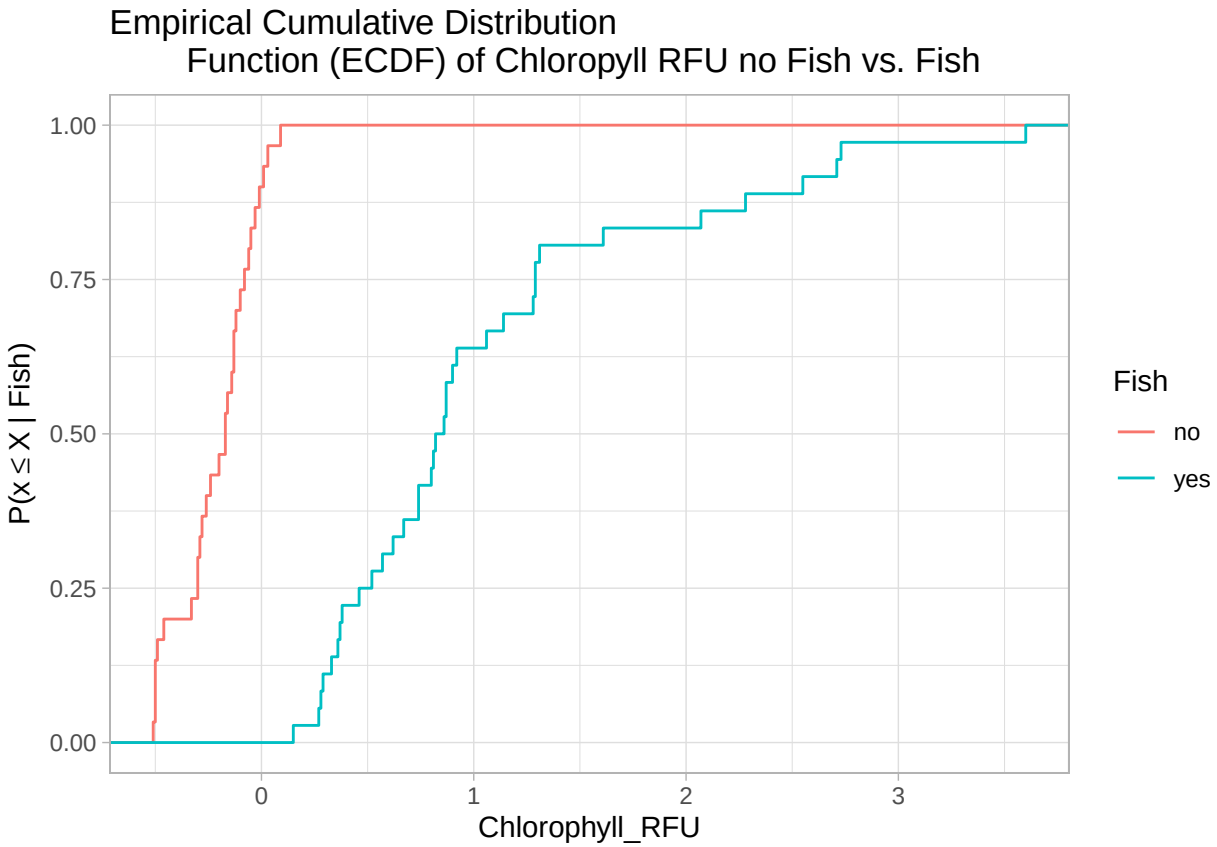
### Step 3.1: T-test

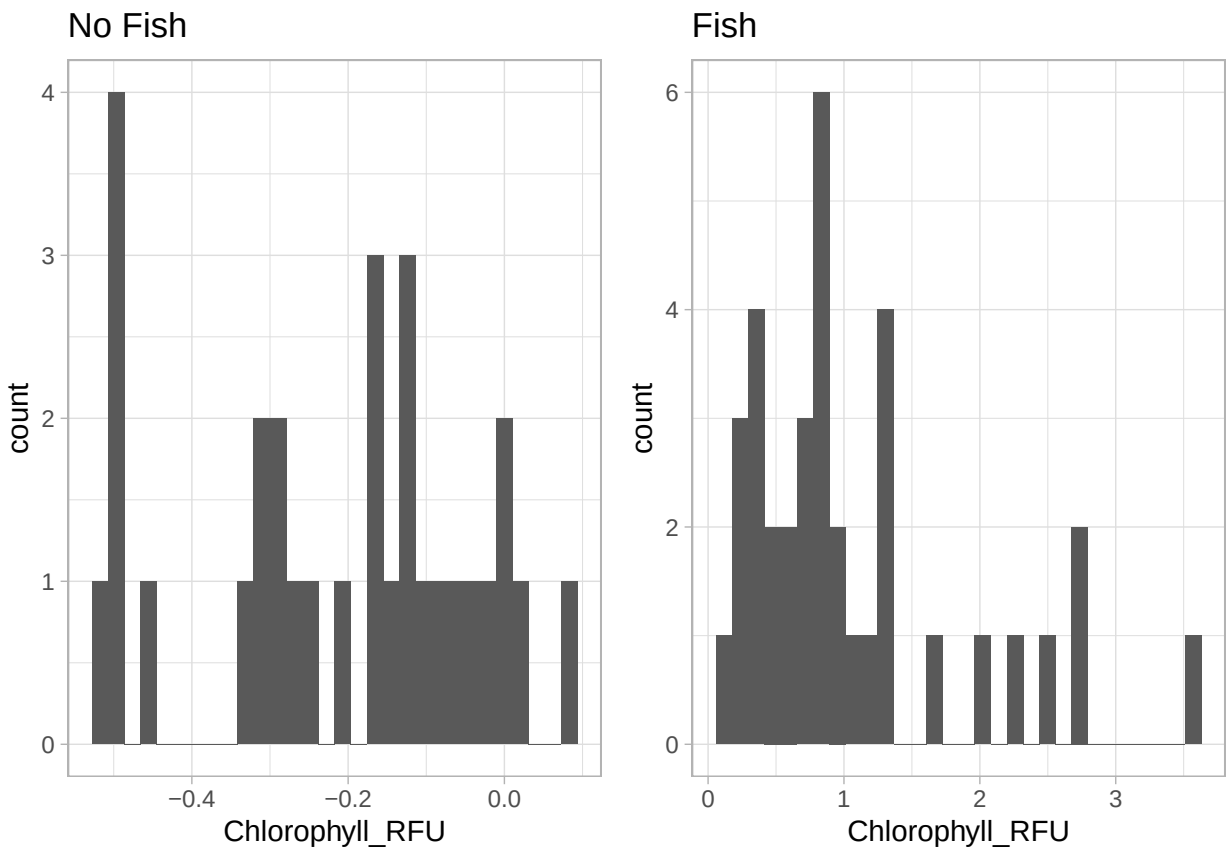
```
# Welch two-sample t-test
(t1 <- t.test(formula = dat$Chlorophyll_RFU ~ dat$Fish, var.equal = FALSE, conf.level = 0.95))
```

```
##
##  Welch Two Sample t-test
##
## data:  dat$Chlorophyll_RFU by dat$Fish
## t = -9.1193, df = 38.85, p-value = 3.376e-11
## alternative hypothesis: true difference in means between group no and group yes is not equal to 0
## 95 percent confidence interval:
##  -1.5672001 -0.9981332
## sample estimates:
##  mean in group no mean in group yes
##        -0.2126667         1.0700000
```

```
# est. mean for ponds no Fish: -0.21
# est. mean for ponds Fish: 1.07
# Esimated difference in means no Fish - Fish: -1.28
# 95% Wald-CI: from -1.57 to -1.00
# p < 0.0001
# There is strong evidence for an association between the presence/absence of fish in the
# ponds and the mean Chlorophyll RFU measured.
```

**Step 3.2: Mixed effects model: EDA**






- Odd distribution of chloro for ponds without fish
- Overdispesion in ponds with fish ("long tail")
- In gerneral, Chlorphyll_RFU does not seem to be normally distributed.

**Step 3.2.2: Fit Model**

```
# Identifying the random structure
m1 <- lmer(Chlorophyll_RFU ~ Fish + (1|Sonde), REML = TRUE, data = dat)
# I had this simple model in mind

m2 <- lmer(Chlorophyll_RFU ~ Fish + (1|Sonde) + (1|Pond), REML = TRUE, data = dat)
# Maybe there is also variability caused by the different ponds, independent
# of the treatment (different baseline Chlorophyll RFU for whatever reason)?
# Although pond and sonde will be highly correlated

anova(m1, m2)
```

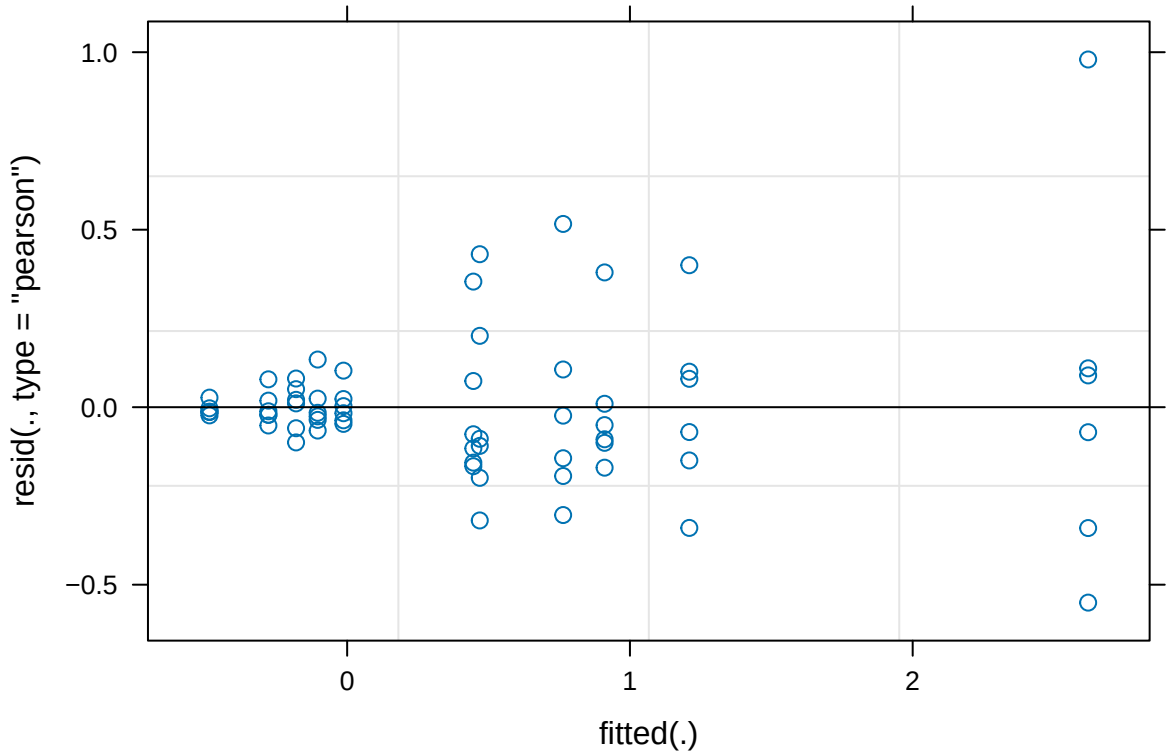
```
## refitting model(s) with ML (instead of REML)

## Data: dat
## Models:
## m1: Chlorophyll_RFU ~ Fish + (1 | Sonde)
## m2: Chlorophyll_RFU ~ Fish + (1 | Sonde) + (1 | Pond)
```

```
##    npar    AIC    BIC  logLik deviance Chisq Df Pr(>Chisq)
## m1    4 42.683 51.441 -17.341   34.683
## m2    5 44.683 55.631 -17.341   34.683     0  1          1
```

```
# Random intercept for pond doesn't lead to improvement
# take m1

# Rename
m <- m1
```

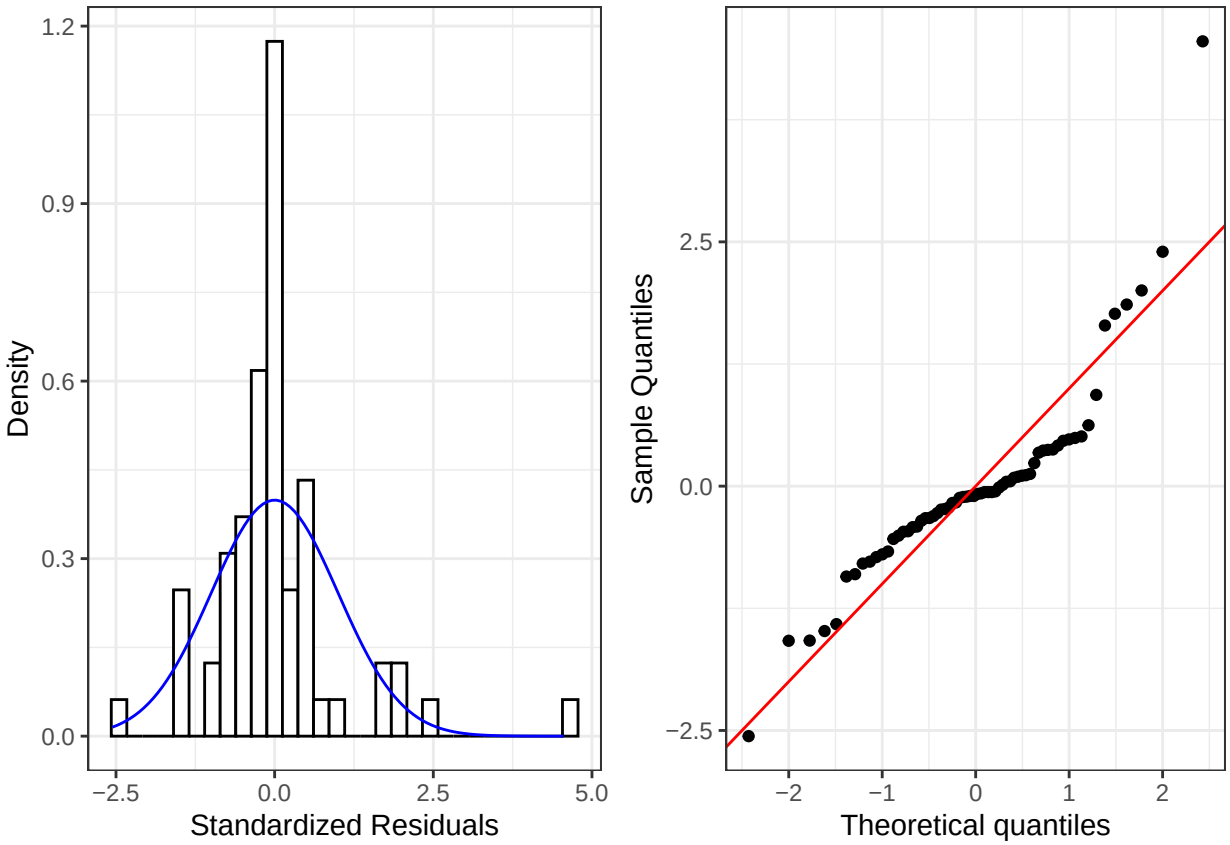
```
plot(m, warning = F, message = F)
```

**Step 3.2.3: Model diagnostics**

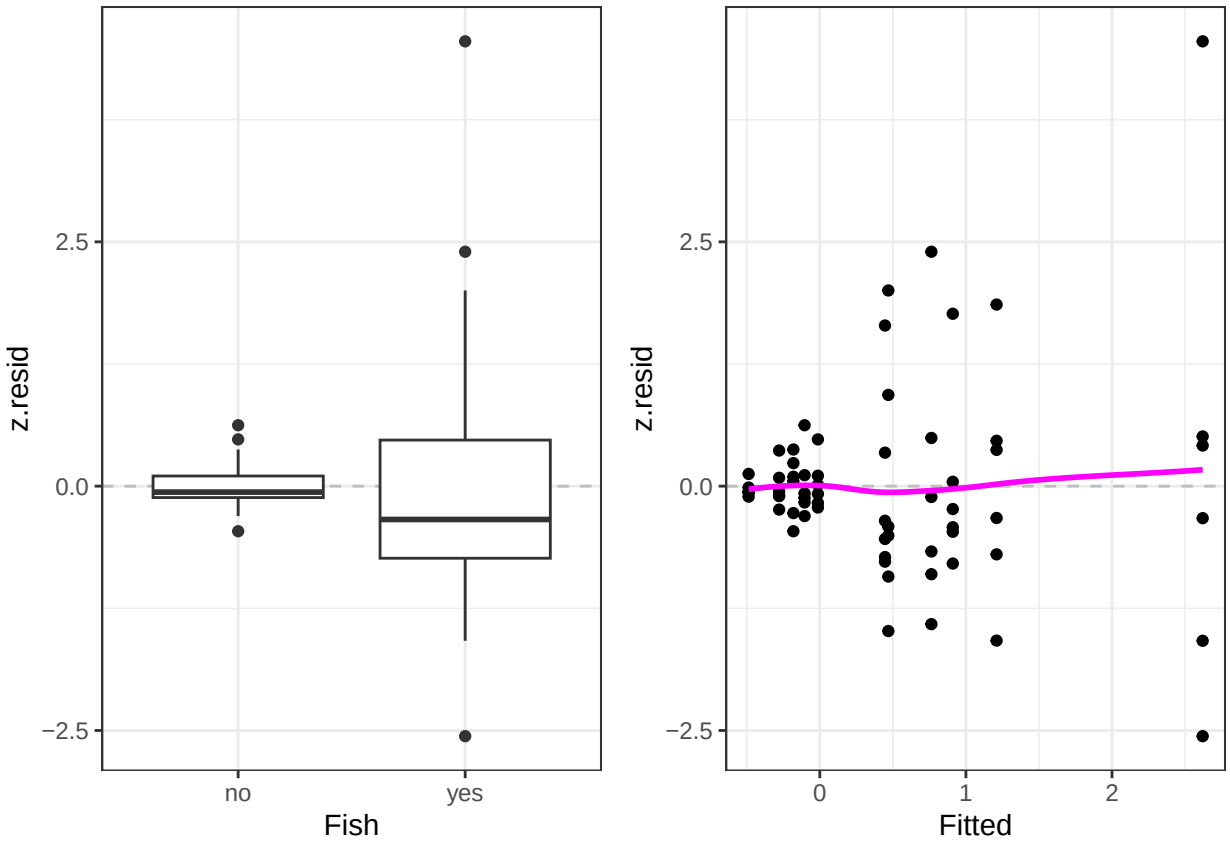



- Residual structure looks odd
- Residuals seem to increase, as the fitted values increase (heteroscedasticity)

```
# Diagnostics dataframe
df <- data.frame(z.resid= scale(residuals(m)),
                                 Fish= m@frame$Fish,
                                 Fitted= fitted.values(m))
```




- Deviations from the assumption of normally distributed residuals
- over/underdispersion
- This is probably largely driven through Sonde14, which had very high values of chlorophyll RFU, lead to this deviation from the normality assumption.




Heteroscedasticity in Fish

```
# Account for heteroscedasticity in Fish by applying a variance-function
# for the residuals:
# Constant variance for each level of Fish, but allowing for different variances
# in the 2 levels
vf1 <- varIdent(form= ~ 1|Fish)

# Refit our new model with package nlme
m.vf <- lme(Chlorophyll_RFU ~ Fish,
            random = (~1|Sonde),
            weights = vf1,
            na.action=na.exclude,
            data = dat)

# Refit previous model with package nlme
m.or <- lme(Chlorophyll_RFU ~ Fish,
            random = (~1|Sonde),
            na.action=na.exclude,
            data = dat)

# Compare to model without specified variance function
anova(m.or, m.vf)
```

```
##      Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## m.or     1  4 44.45521 53.09074 -18.227606
```

```
## m.vf     2  5 -8.86717  1.92725   9.433583 1 vs 2 55.32238  <.0001
```

```
# Model with specified variance function seems to be a better fit

# Rename
m <- m.vf
```

### Step 3.2.4: Model interpretation and predctions

```
Anova(m)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: Chlorophyll_RFU
##       Chisq Df Pr(>Chisq)
## Fish 11.688  1  0.0006292 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Fish contributes significantly to explain the variability
# observed in Chlorophyll_RFU
S(m)
```

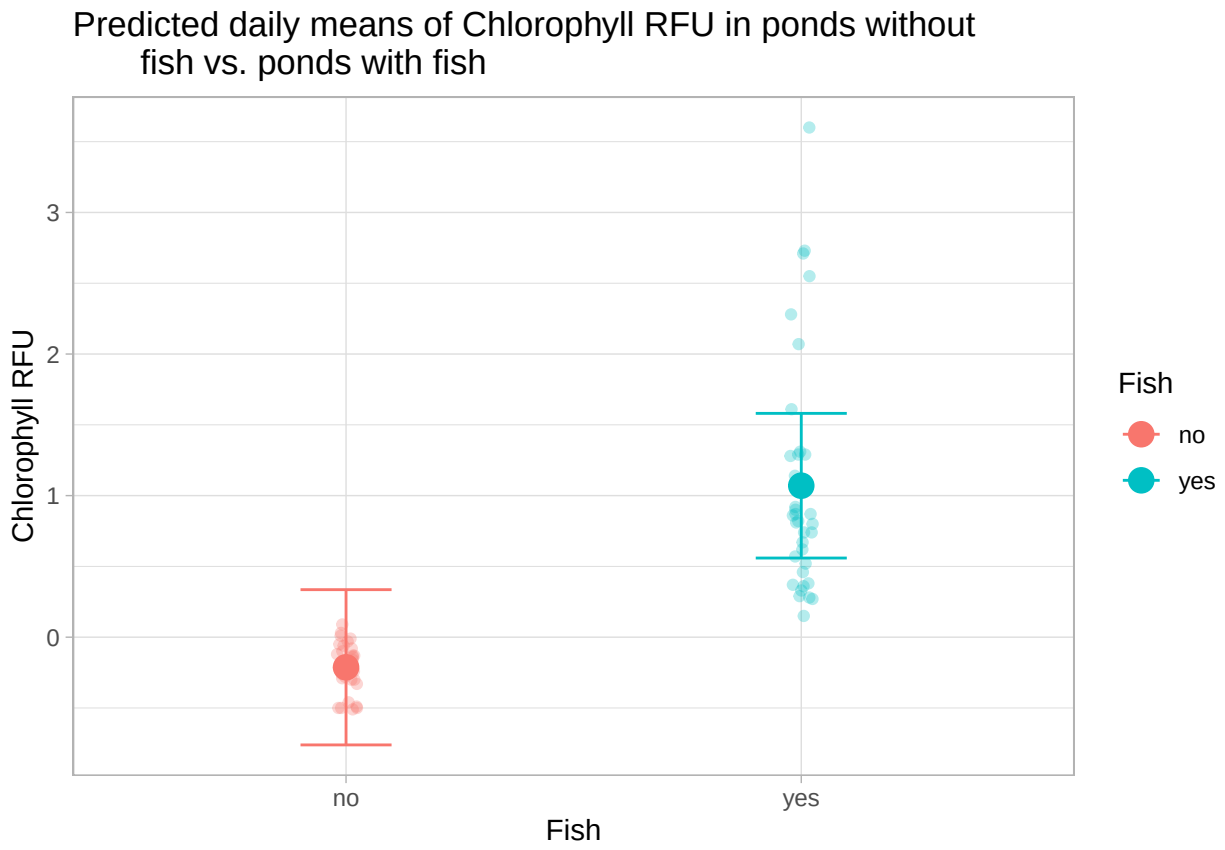
```
## Linear mixed model fit by REML,  Data: dat
##
## Fixed Effects:
##  Formula: Chlorophyll_RFU ~ Fish
##
##             Estimate Std.Error df t value Pr(>|t|)
## (Intercept)  -0.2127    0.2745 55  -0.775  0.44179
## Fishyes       1.2827    0.3752  9   3.419  0.00764 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Random effects:
##  Formula: ~1 | Sonde
##         (Intercept) Residual
## StdDev:      0.6134    0.313
##
## Variance function:
##  Structure: Different standard deviations per stratum
##  Formula: ~1 | Fish
##  Parameter estimates:
##       yes        no
## 1.0000000 0.1746085
##
## Number of Observations: 66
## Number of Groups: 11
##
## logLik     df    AIC    BIC
##   9.43      5  -8.87   1.93
```

```
# Fixed effects
# - The expected daily mean RFU level for ponds without Fish is -0.21
# - Compared to ponds without fish, ponds with fish are expected to have daily
# Chlorophyll RFU levels increased by 1.28 units (p = 0.008)

# Random effects
# The additional standard deviation in Chlorophyll RFU caused by the different Sondes
# is estimated to be 0.61

# Variance function
# Compared to ponds with fish, ponds without fish are expected to have a standard
# deviation reduced by a factor of 0.17 of Chlorophyll RFU-levels

# Pseudo-R-squared for mixed models
r.squaredGLMM(m)
```

```
##           R2m       R2c
## [1,] 0.4662197 0.8897032
```

```
# Proportion of variance explained by fixed effects: 0.47
# Proportion of variance explained by entire model: 0.89
```

**Step 3.2.5: Prediction plot**

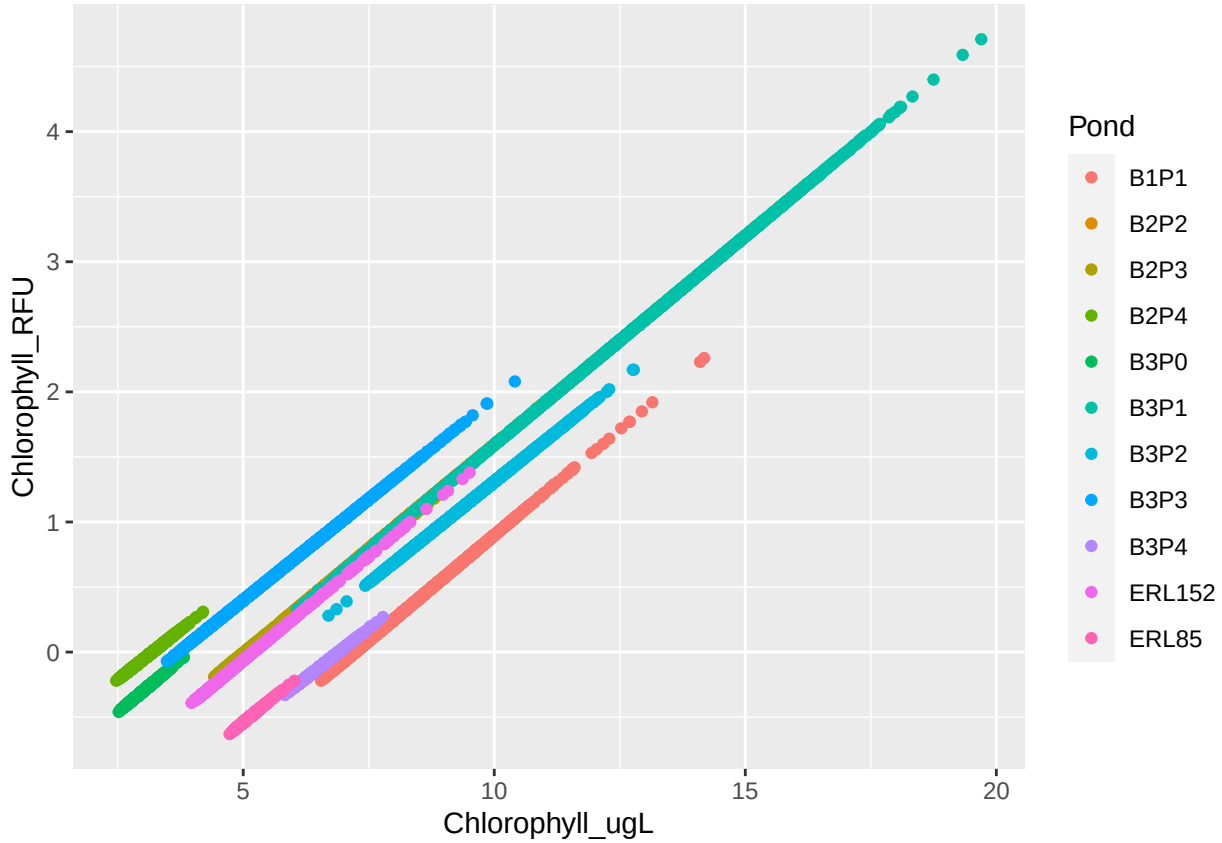


## Step 4: Relationship between Chlorophyll RFU and Chlorophyll ug/L

Now that we have an impression about what levels of Chlorophyll RFU to expect in these ponds, we can ask ourselves how this RFU values translate to ug/L?

### Step 4.1: Visualization of the question

```
# For that, we can plot Chlorophyll RFU vs. Chlorophyll ug/L of the time series
# data from 2022:
all <- fread("~/ZIVI_EAWAG/project_22/data/ponds_sonde_data_all.txt")

# Lets convert Pond to an unordered factor
all$Pond <- factor(all$Pond)

# plot
ggplot(data = all, aes(x = Chlorophyll_ugL, y = Chlorophyll_RFU, col = Pond)) +
  geom_point()
```




There seems to be a perfectly linear relationship between the 2 variables. Additionally it seems very much apparent that the different ponds seem to have different intercepts, but same slopes.

### Step 4.1: Model Chlorophyll RFU

```
# Lets do a simple mixed model that can estimate a global slope and intercept, yet acknowledges that
# observations are clustered in different sondes/ponds
m1 <- lmer(Chlorophyll_RFU ~ Chlorophyll_ugL + (1|Pond), data = all)
# Random intercept pond: Quantifies the additional variance around the global slope
# caused by the ponds.

summary(m1)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Chlorophyll_RFU ~ Chlorophyll_ugL + (1 | Pond)
##    Data: all
##
## REML criterion at convergence: -306659.1
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -5.8845 -0.6299 -0.0201  0.6039  7.5906
##
## Random effects:
##  Groups   Name        Variance  Std.Dev.
##  Pond     (Intercept) 0.1753014 0.418690
##  Residual             0.0000304 0.005513
## Number of obs: 40573, groups:  Pond, 11
##
## Fixed effects:
##                   Estimate Std. Error  t value
## (Intercept)     -1.672e+00  1.262e-01   -13.24
## Chlorophyll_ugL  3.181e-01  3.002e-05 10598.43
##
## Correlation of Fixed Effects:
##             (Intr)
## Chlrphyll_L -0.002
```

```
# (Intercept): global intercept across all ponds
# If there is no Chlorophyll at all in the water, the Chlorophyll RFU measured
# by the sonde is expected to be -1.67 units.

# Chlorophyll_ugL: global slope of Chlorophyll_ugL
# For 1 ug/L increase in Chlorophyll, Chlorophyll RFU is expected to increase
# by 0.318 units.
```

### Step 4.2: Formulate the fitted equation

The fitted equation for the fixed part is: Chlorophyll_RFU = -1.672 + 0.3181 * Chlorophyll_ugL

## Step 5: Bringing it all together

Interpretation of the mixed model describing the relationship between Fish- Treatment and Chlorophyll_RFU in terms of Chlorophyll ug/L:

- We learned from Step 3.2 that the expected daily mean RFU level for ponds without fish is -0.21. This would correspond to a daily mean Chlorophyll ug/L-level of (-0.21 + 1.672)/0.3181 = 4.60, according to our model.
- Additionally, according to the model from Step 3.2, the expected daily mean RFU level for ponds with Fish is -0.21 + 1.28 = 1.07. This would correspond to a daily mean Chlorophyll ug/L-level of (1.07 + 1.672)/0.3181 = 8.62, according to our model.

# 6: Appendix

### Version and packages used to generate this report:

```
## 2023-08-09 13:54:02.860588 Europe/Zurich


## R version 4.3.1 (2023-06-16 ucrt)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 19045)
##
## Matrix products: default
##
##
## locale:
## [1] LC_COLLATE=German_Switzerland.utf8  LC_CTYPE=German_Switzerland.utf8
## [3] LC_MONETARY=German_Switzerland.utf8 LC_NUMERIC=C
## [5] LC_TIME=German_Switzerland.utf8
##
## time zone: Europe/Zurich
## tzcode source: internal
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] MuMIn_1.47.5      arm_1.13-1        MASS_7.3-60       effects_4.2-2
##  [5] GGally_2.1.2      ggpubr_0.6.0      pastecs_1.3.21    car_3.1-2
##  [9] carData_3.0-5     nlme_3.1-162      lme4_1.1-34       Matrix_1.5-4.1
## [13] zoo_1.8-12        viridis_0.6.3     viridisLite_0.4.2 forcats_1.0.0
## [17] stringr_1.5.0     dplyr_1.1.2       purrr_1.0.1       readr_2.1.4
## [21] tidyr_1.3.0       tibble_3.2.1      ggplot2_3.4.2     tidyverse_2.0.0
## [25] lubridate_1.9.2   data.table_1.14.8 cowplot_1.1.1     bit64_4.0.5
## [29] bit_4.0.5         pacman_0.5.1
##
## loaded via a namespace (and not attached):
##  [1] tidyselect_1.2.0   farver_2.1.1       fastmap_1.1.1      reshape_0.8.9
##  [5] digest_0.6.32      estimability_1.4.1 timechange_0.2.0   lifecycle_1.0.3
##  [9] survival_3.5-5     magrittr_2.0.3     compiler_4.3.1     rlang_1.1.1
## [13] tools_4.3.1        utf8_1.2.3         yaml_2.3.7         knitr_1.43
## [17] ggsignif_0.6.4     labeling_0.4.2     plyr_1.8.8         RColorBrewer_1.1-3
## [21] abind_1.4-5        withr_2.5.0        nnet_7.3-19        grid_4.3.1
## [25] stats4_4.3.1       fansi_1.0.4        colorspace_2.1-0   scales_1.2.1
## [29] insight_0.19.3     cli_3.6.1          survey_4.2-1       rmarkdown_2.23
## [33] generics_0.1.3     rstudioapi_0.14    tzdb_0.4.0         minqa_1.2.5
```

```
## [37] DBI_1.1.3          splines_4.3.1      mitools_2.4        vctrs_0.6.3
## [41] boot_1.3-28.1      hms_1.1.3          rstatix_0.7.2      glue_1.6.2
## [45] nloptr_2.0.3       stringi_1.7.12     gtable_0.3.3       munsell_0.5.0
## [49] pillar_1.9.0       htmltools_0.5.5    R6_2.5.1           evaluate_0.21
## [53] lattice_0.21-8     highr_0.10         backports_1.4.1    broom_1.0.5
## [57] Rcpp_1.0.10        coda_0.19-4        gridExtra_2.3      mgcv_1.8-42
## [61] xfun_0.39          pkgconfig_2.0.3
```

## 7: Code used to generate the plots

```
# ECDF no Fish vs. Fish daily means
ggplot(dat, aes(Chlorophyll_RFU, col = Fish)) +
   stat_ecdf() +
   labs(y = expression("P(x"<="X | Fish)"), title = "Empirical Cumulative Distribution
        Function (ECDF) of Chloropyll RFU no Fish vs. Fish") +
   theme_light()

# Histogram Chlorophyll RFU no Fish vs. Fish
ggarrange(
  ggplot(subset(dat, Fish=="no")) +
     geom_histogram(aes(x=Chlorophyll_RFU)) +
     theme_light() +
     labs(title = "No Fish"),
  ggplot(subset(dat, Fish=="yes")) +
     geom_histogram(aes(x=Chlorophyll_RFU)) +
     theme_light() +
     labs(title = "Fish")
  )

# Distribution of Chlorophyll_RFU
ggarrange(
  ggplot(df, aes(z.resid)) +
    geom_histogram(aes(y= ..density..), fill= "white", col= "black") +
    stat_function(fun= dnorm,
                  args= list(mean(df$z.resid),
                             sd(df$z.resid)),
                  n= 1e2,
                  col= "blue") +
    labs(y= "Density", x= "Standardized Residuals") +
    theme_bw(),
  ggplot(df, aes(sample= z.resid)) +
    stat_qq() +
    geom_abline(intercept= 0, slope= 1, col= "red") +
    labs(x= "Theoretical quantiles", y= "Sample Quantiles") +
    theme_bw(),
  ncol= 2)

# Residual plots
ggarrange(
  ggplot(df, aes(Fish, z.resid)) +
    geom_hline(yintercept= 0, lty= 2, col= "grey") +
```

```
    geom_boxplot() +
    theme_bw(),
  ggplot(df, aes(Fitted, z.resid)) +
    geom_hline(yintercept= 0, lty= 2, col= "grey") +
    geom_point() +
    geom_smooth(col= "magenta", se= F) +
    theme_bw(),
  ncol= 2)

# Prediction plot
pred.m<- data.frame(Effect("Fish", m))
pred.m$Fish<- factor(pred.m$Fish)

ggplot(pred.m, aes(Fish, fit, col= Fish)) +
  geom_errorbar(aes(ymin= lower, ymax= upper), width= .2, position= "dodge") +
  geom_point(size= 4, position= position_dodge(width= .2)) +
  geom_point(data= m$data, aes(y= Chlorophyll_RFU), alpha= .3,
             position= position_jitterdodge(dodge.width= .2,
                                            jitter.width= .1,
                                            jitter.height= 0)) +
  labs(y= "Chlorophyll RFU", title = "Predicted daily means of Chlorophyll RFU in ponds without
       fish vs. ponds with fish") +
  theme_light()
```