STAT 414 - Class Project

Part 3

Kyle Nessen, Justin Mai, Aiden Kelly, Arneh Begi

2024-11-18

Introduction

Our project focuses on the habitat characteristics of overwintering monarch butterflies. Specifically, we are investigating whether there are distinct patterns within a grove related to temperature, humidity, and light that monarch butterflies are selecting. Due to the structure of our data, we cannot directly predict monarch presence based on these climatic conditions. Instead, we are reversing the question and asking if monarch presence is a significant predictor of the following measures:

whether

- Average temperature (C)
- Minimum temperature (C)
- Maximum temperature (C)
- Average humidity (%)
- Average light (lux)
- Standard Deviation of light (lux)

We are analyzing daily summaries from weather stations positioned within eight monarch groves. Each grove is geographically distinct from each other, and within each grove are five weather stations, or "arrays." One array is placed at the location of overwintering butterflies ("Cluster") and four other arrays are positioned both within the grove (SW, NE) and outside the canopy (SE, NW). We created a new variable, butterfly_present, where Cluster arrays = 1, and all others = 0, which we use to assess if monarchs help predict climatic variables. We also account for time (seasonDay) by counting days since the beginning of the monitoring period (December 1st, 2018). We include seasonDay as both a fixed and random effect to account for both the overall seasonal temperature pattern and allow each grove to have its own unique seasonal trajectory, since groves may warm or cool at different rates due to their distinct physical characteristics (e.g., elevation, canopy cover, proximity to coast). Finally, we treat groves as random effects, as they are a sample from a larger pool of other potential groves that monarchs can overwinter at.

whether

effects

xx

where

Data

```
df <- read_csv('allgr_array_KianaRawdat.csv')
df <- df |>
  mutate(
    # Parse the month and day components
    month = as.integer(substr(month.day, 1, 2)),
    day = as.integer(substr(month.day, 4, 5)),
```

I will need to see some graphs -

```
  # Assign year based on month
  year = ifelse(month >= 10, 2022, 2023),

  # Create a Date column
  Date = as.Date(paste(year, month, day, sep = "-"), format = "%Y-%m-%d"),

  # Calculate seasonDay and seasonWeight
  seasonDay = as.numeric(difftime(Date, as.Date("2022-12-01"), units = "days"))
 )

colSums(is.na(df)) # lots of cols with missing vals so log like wont worksince lmer will drop nas

     ...1      array  month.day   temp.avg    hum.avg dew.pt.avg  light.avg
        0          0          0         31        618        618        109
 light.min  light.max   temp.min   temp.max   temp.std  light.std dew.pt.std
       109        109         31         31         31        109        618
   hum.std    hum.max    hum.min     dp.max     dp.min     daynum      grove
       618        618        618        662        662          0          0
     month        day       year       Date  seasonDay
         0          0          0          0          0

df$array <- as.factor(df$array)
df$array <- relevel(df$array, ref = "NE")
df$butterfly_present <- ifelse(df$array == "Cluster", 1, 0)

# Center variables
df$temp.avg_centered <- df$temp.avg - mean(df$temp.avg, na.rm = TRUE)
df$temp.min_centered <- df$temp.min - mean(df$temp.min, na.rm = TRUE)
df$temp.max_centered <- df$temp.max - mean(df$temp.max, na.rm = TRUE)
df$hum.avg_centered <- df$hum.avg - mean(df$hum.avg, na.rm = TRUE)
df$light.avg_centered <- df$light.avg - mean(df$light.avg, na.rm = TRUE)
df$light.std_centered <- df$light.std - mean(df$light.std, na.rm = TRUE)
```

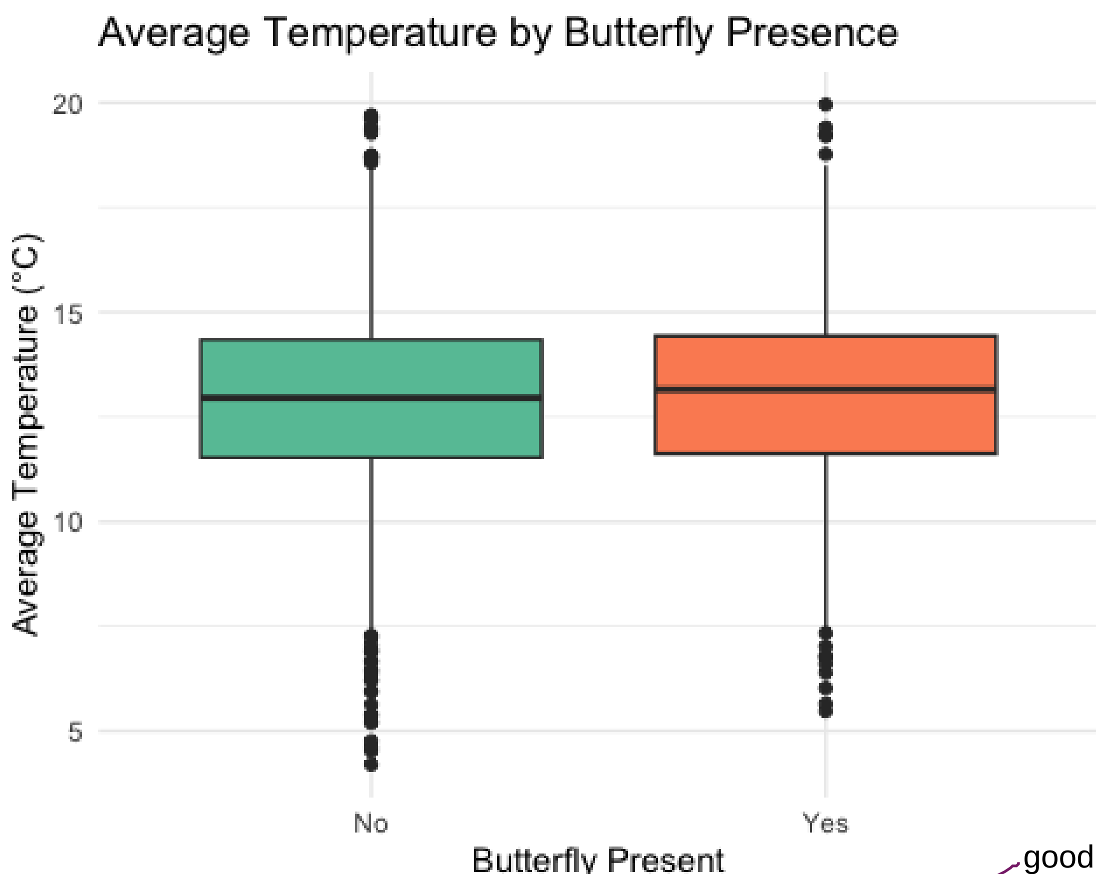
Temperature

Average Temperature (Daily)

```
temp.avg.model1 <- lmer(temp.avg_centered ~ seasonDay + (1 + seasonDay | grove), data=df)
temp.avg.model2 <- lmer(temp.avg_centered ~ seasonDay + butterfly_present + (1 + seasonDay | grove),
data=df)
anova(temp.avg.model1, temp.avg.model2)

Data: df
Models:
temp.avg.model1: temp.avg_centered ~ seasonDay + (1 + seasonDay | grove)
temp.avg.model2: temp.avg_centered ~ seasonDay + butterfly_present + (1 + seasonDay | grove)
                npar   AIC   BIC  logLik deviance  Chisq Df Pr(>Chisq)
temp.avg.model1    6 10332 10366 -5159.9    10320
temp.avg.model2    7 10333 10373 -5159.3    10319 1.2872  1     0.2566

performance::icc(temp.avg.model2)
```

Pick one of these to show the "null model"

```
# Intraclass Correlation Coefficient

  Adjusted ICC: 0.248
 Unadjusted ICC: 0.237

ggplot(df, aes(x=factor(butterfly_present), y=temp.avg)) +
 geom_boxplot(fill=c("#66c2a5", "#fc8d62")) +
 labs(x="Butterfly Present", y="Average Temperature (°C)",
     title="Average Temperature by Butterfly Presence") +
 scale_x_discrete(labels=c("No", "Yes")) +
 theme_minimal()
```



good

The intraclass correlation coefficient (ICC) of 0.248 indicates that about 25% of the variation in average temperature (after accounting for seasonal effects) can be attributed to differences between groves. When we added butterfly presence to the model, a likelihood ratio test showed no significant improvement in model fit (p = 0.2566), suggesting that locations selected by monarchs do not differ significantly in average temperature from other monitored locations within the groves.

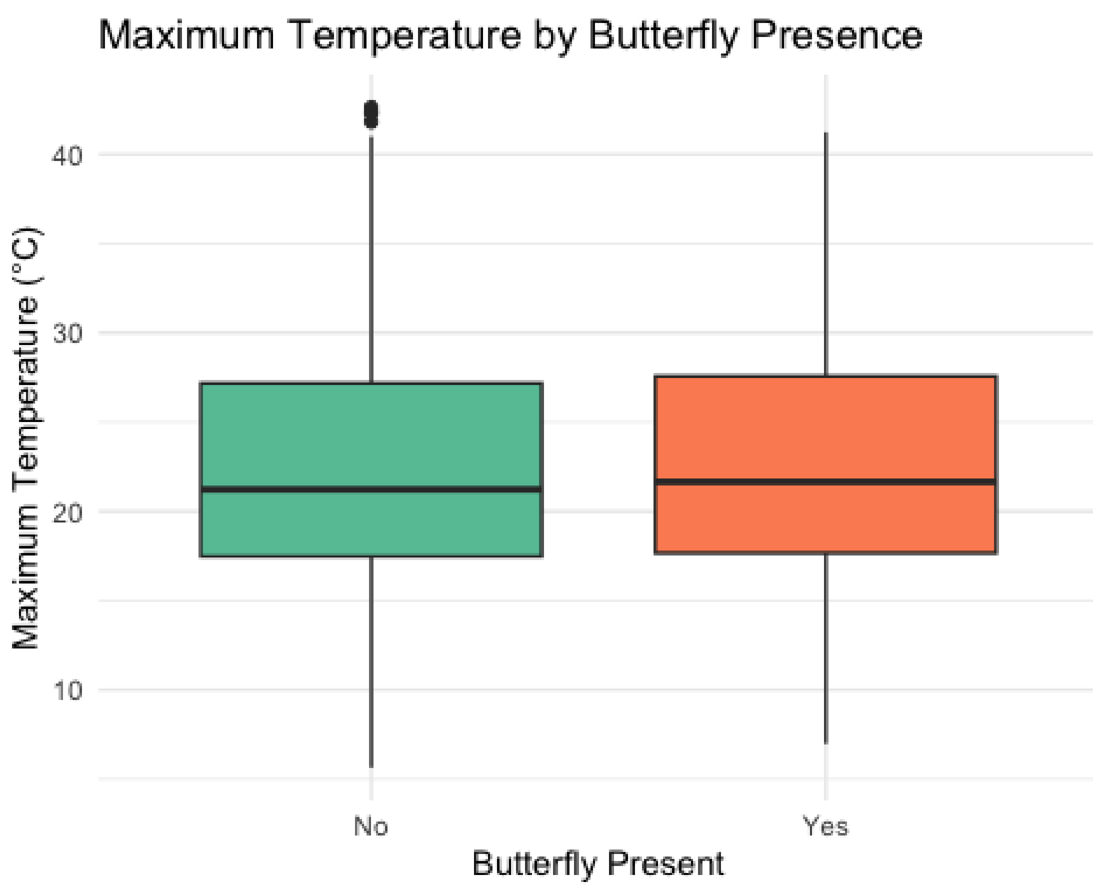
probably repeat here -after adjusting fo

Maximum Temperature

```
temp.max.model1 <- lmer(temp.max_centered ~ seasonDay + (1 + seasonDay | grove), data=df)
temp.max.model2 <- lmer(temp.max_centered ~ seasonDay + butterfly_present + (1 + seasonDay |
```

this is max temp in a day?

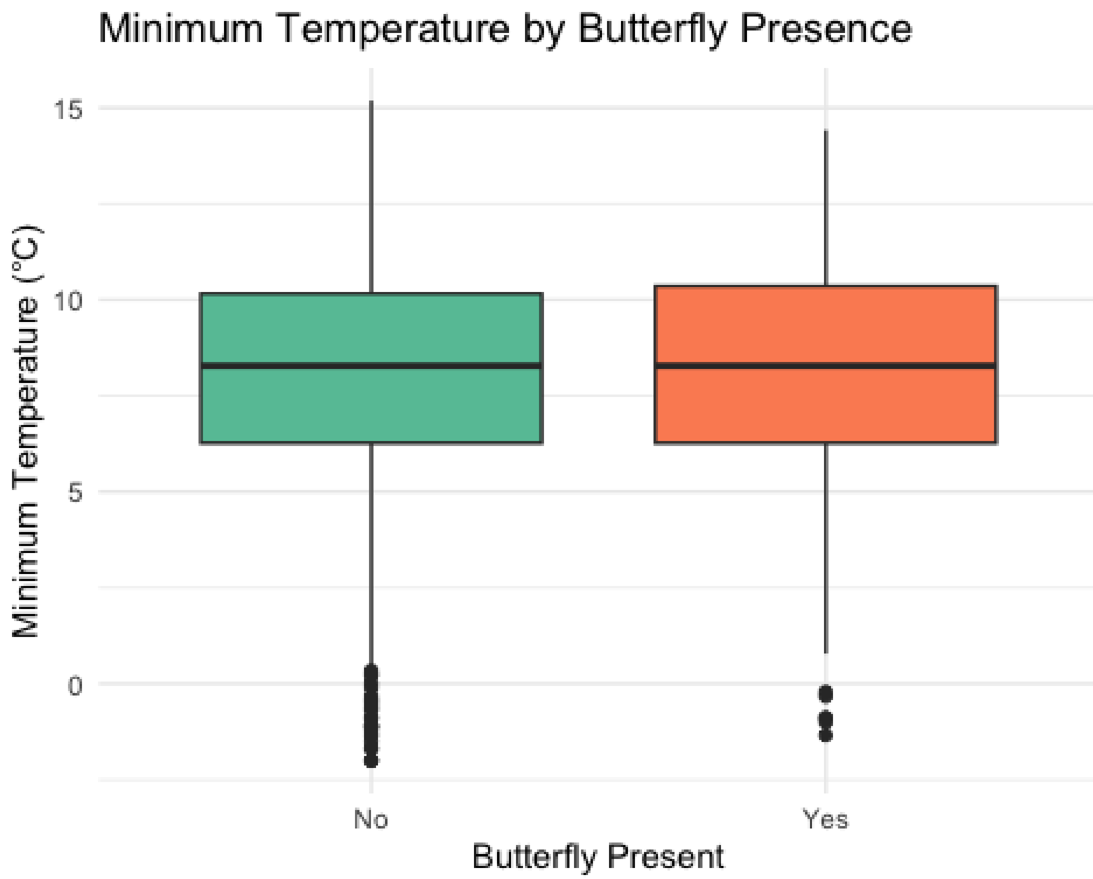
```
grove), data=df)
anova(temp.max.model1, temp.max.model2)

Data: df
Models:
temp.max.model1: temp.max_centered ~ seasonDay + (1 + seasonDay | grove)
temp.max.model2: temp.max_centered ~ seasonDay + butterfly_present + (1 + seasonDay | grove)
                npar   AIC   BIC  logLik deviance  Chisq Df Pr(>Chisq)
temp.max.model1    6 15126 15161 -7557.2    15114
temp.max.model2    7 15127 15167 -7556.3    15113 1.8249  1     0.1767

performance::icc(temp.max.model2)

# Intraclass Correlation Coefficient

    Adjusted ICC: 0.507
  Unadjusted ICC: 0.500

ggplot(df, aes(x=factor(butterfly_present), y=temp.max)) +
  geom_boxplot(fill=c("#66c2a5", "#fc8d62")) +
  labs(x="Butterfly Present", y="Maximum Temperature (°C)",
       title="Maximum Temperature by Butterfly Presence") +
  scale_x_discrete(labels=c("No", "Yes")) +
  theme_minimal()
```



The ICC of 0.507 indicates that about 51% of the variation in maximum temperature (after accounting for seasonal effects) can be attributed to differences between groves. When we added butterfly presence to the model, a likelihood ratio test showed no significant improvement in model fit (p = 0.1767), suggesting that locations selected by monarchs do not differ significantly in maximum temperature from other monitored locations within the groves.

Minimum Temperature

```
temp.min.model1 <- lmer(temp.min_centered ~ seasonDay + (1 + seasonDay | grove), data=df)
temp.min.model2 <- lmer(temp.min_centered ~ seasonDay + butterfly_present + (1 + seasonDay | grove), data=df)
anova(temp.min.model1, temp.min.model2)

Data: df
Models:
temp.min.model1: temp.min_centered ~ seasonDay + (1 + seasonDay | grove)
temp.min.model2: temp.min_centered ~ seasonDay + butterfly_present + (1 + seasonDay | grove)
                npar   AIC   BIC  logLik deviance Chisq Df Pr(>Chisq)
temp.min.model1    6 11416 11450 -5701.8    11404
temp.min.model2    7 11417 11457 -5701.5    11403 0.566  1     0.4518

performance::icc(temp.min.model2)

# Intraclass Correlation Coefficient

    Adjusted ICC: 0.473
  Unadjusted ICC: 0.468

ggplot(df, aes(x=factor(butterfly_present), y=temp.min)) +
  geom_boxplot(fill=c("#66c2a5", "#fc8d62")) +
  labs(x="Butterfly Present", y="Minimum Temperature (°C)",
       title="Minimum Temperature by Butterfly Presence") +
  scale_x_discrete(labels=c("No", "Yes")) +
  theme_minimal()
```

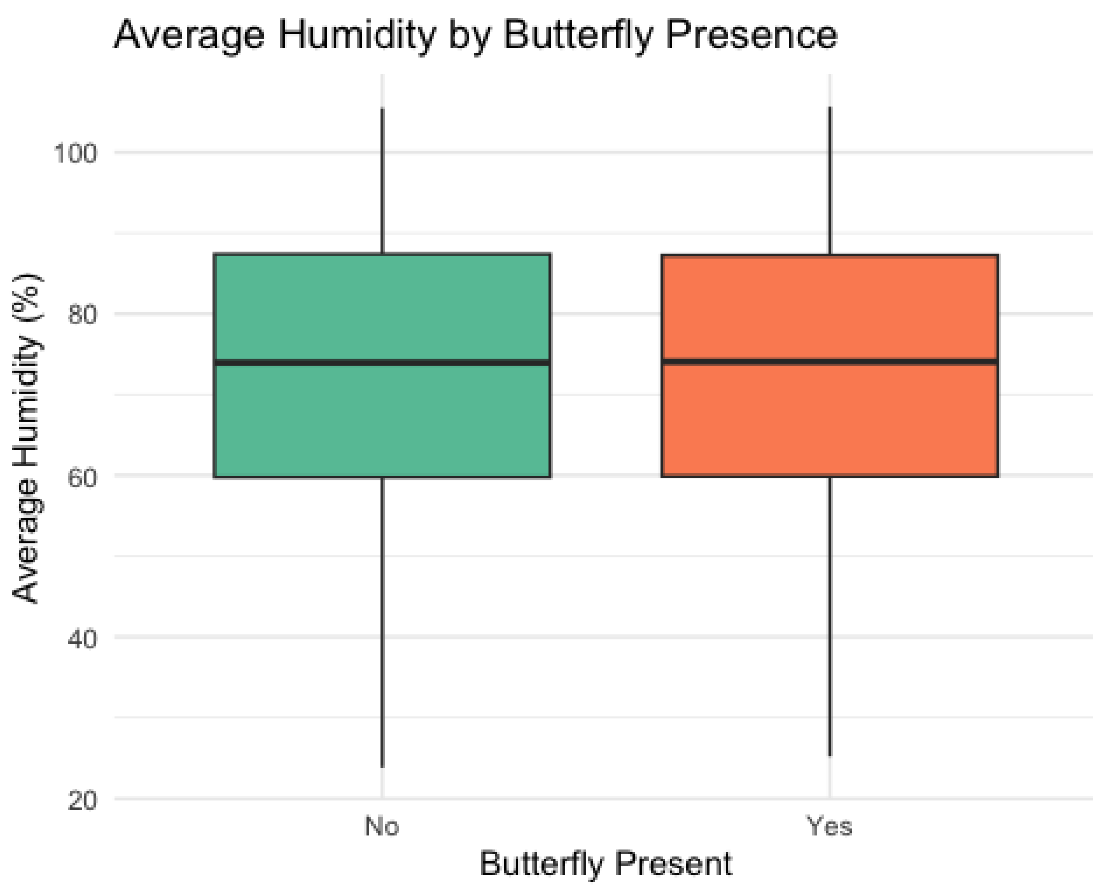


The ICC of 0.452 indicates that about 45% of the variation in minimum temperature (after accounting for seasonal effects) can be attributed to differences between groves. When we added butterfly presence to the model, a likelihood ratio test showed no significant improvement in model fit (p = 0.4518), suggesting that locations selected by monarchs do not differ significantly in minimum temperature from other monitored locations within the groves.

Humidity

```
hum.avg.model1 <- lmer(hum.avg_centered ~ seasonDay + (1 + seasonDay | grove), data=df)
hum.avg.model2 <- lmer(hum.avg_centered ~ seasonDay + butterfly_present + (1 + seasonDay | grove), data=df)
anova(hum.avg.model1, hum.avg.model2)

Data: df
Models:
hum.avg.model1: hum.avg_centered ~ seasonDay + (1 + seasonDay | grove)
hum.avg.model2: hum.avg_centered ~ seasonDay + butterfly_present + (1 + seasonDay | grove)
               npar   AIC   BIC  logLik deviance Chisq Df Pr(>Chisq)
hum.avg.model1    6 15261 15294 -7624.6    15249
hum.avg.model2    7 15262 15301 -7624.2    15248 0.744  1     0.3884

performance::icc(hum.avg.model2)
```

```
# Intraclass Correlation Coefficient

   Adjusted ICC: 0.110
 Unadjusted ICC: 0.105

ggplot(df, aes(x=factor(butterfly_present), y=hum.avg)) +
  geom_boxplot(fill=c("#66c2a5", "#fc8d62")) +
  labs(x="Butterfly Present", y="Average Humidity (%)",
       title="Average Humidity by Butterfly Presence") +
  scale_x_discrete(labels=c("No", "Yes")) +
  theme_minimal()
```

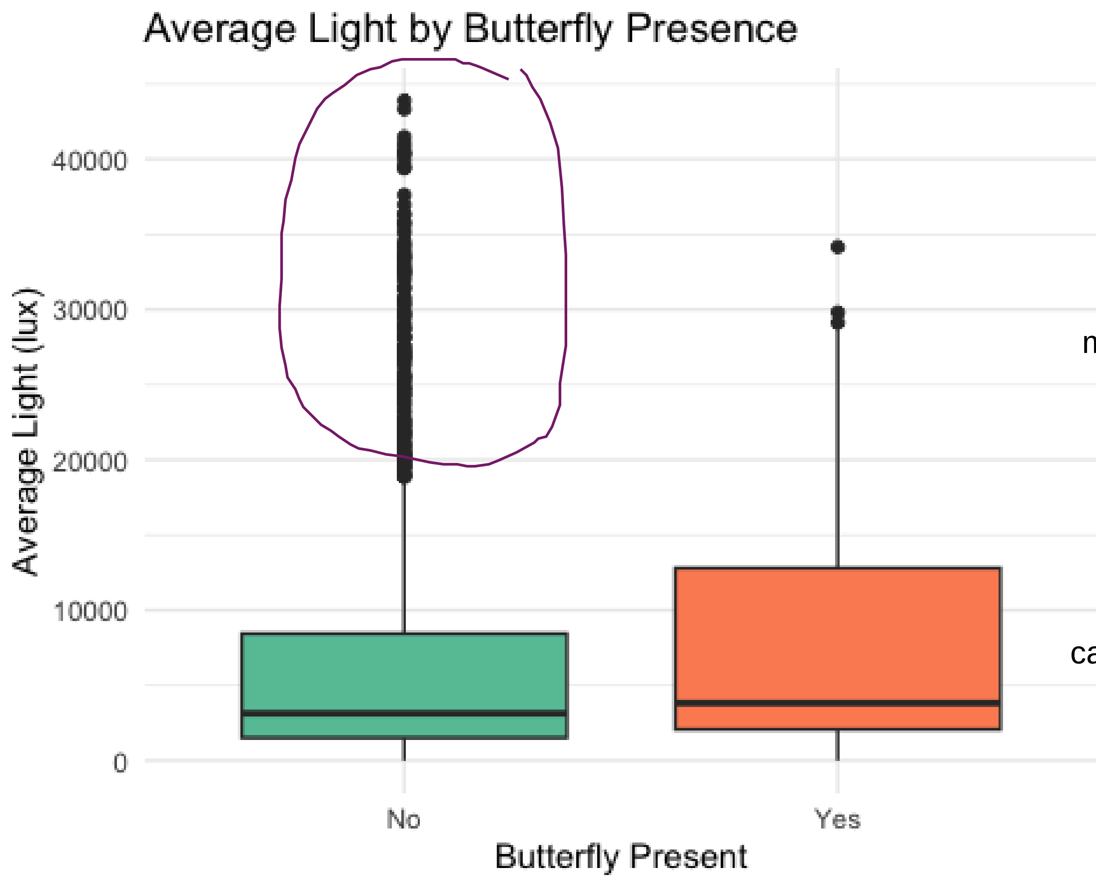


The ICC of 0.110 indicates that about 11% of the variation in average humidity (after accounting for seasonal effects) can be attributed to differences between groves. When we added butterfly presence to the model, a likelihood ratio test showed no significant improvement in model fit ($p = 0.3884$), suggesting that locations selected by monarchs do not differ significantly in average temperature from other monitored locations within the groves.

Light

Average light

```
# Just random intercepts, no random slopes
light.avg.model1 <- lmer(light.avg_centered ~ seasonDay + (1 | grove), data=df)
light.avg.model2 <- lmer(light.avg_centered ~ seasonDay + butterfly_present + (1 | grove), data=df)
anova(light.avg.model1, light.avg.model2)

Data: df
Models:
light.avg.model1: light.avg_centered ~ seasonDay + (1 | grove)
light.avg.model2: light.avg_centered ~ seasonDay + butterfly_present + (1 | grove)
                 npar   AIC   BIC logLik deviance  Chisq Df Pr(>Chisq)
light.avg.model1    4 47454 47477 -23723    47446
light.avg.model2    5 47449 47478 -23720    47439 6.8248  1    0.00899 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

performance::icc(light.avg.model2)

# Intraclass Correlation Coefficient

    Adjusted ICC: 0.135
  Unadjusted ICC: 0.134

ggplot(df, aes(x=factor(butterfly_present), y=light.avg)) +
  geom_boxplot(fill=c("#66c2a5", "#fc8d62")) +
  labs(x="Butterfly Present", y="Average Light (lux)",
       title="Average Light by Butterfly Presence") +
  scale_x_discrete(labels=c("No", "Yes")) +
  theme_minimal()
```



might be more going on

can you show some of th

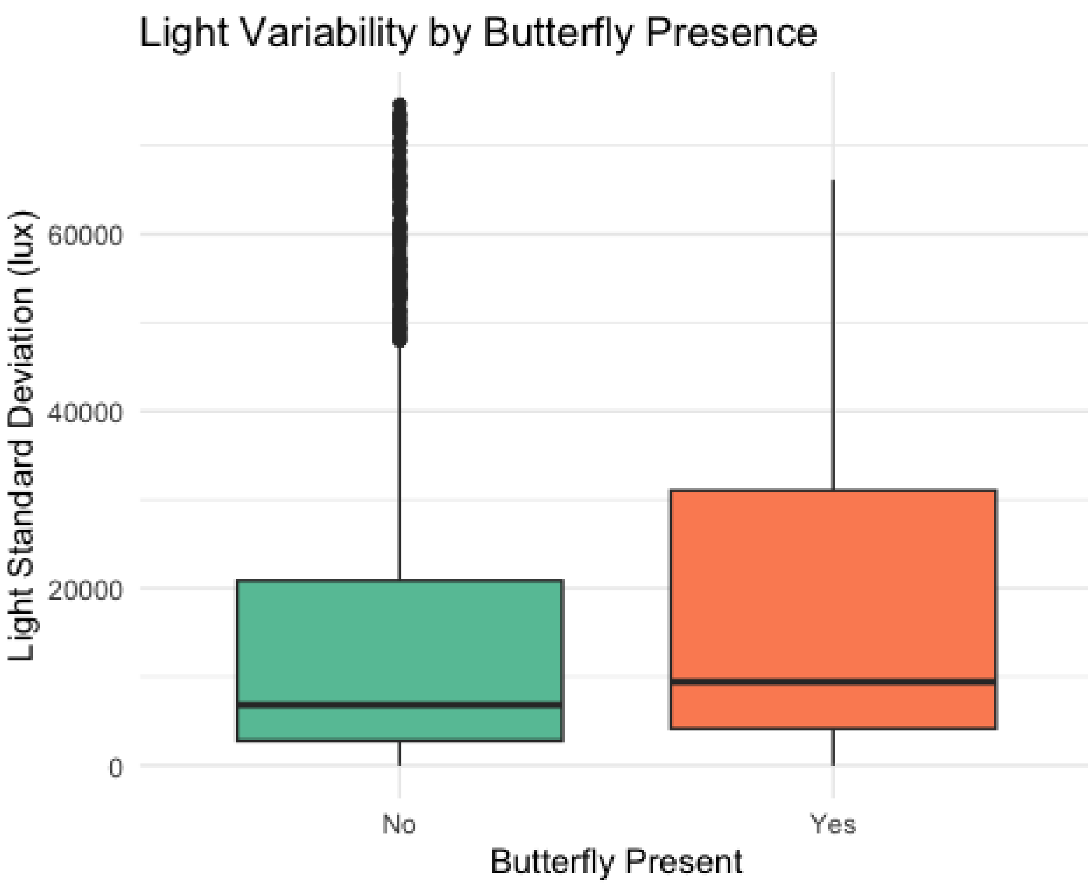
The ICC of 0.135 indicates that about 13.5% of the variation in average light can be attributed to differences between groves. When we added butterfly presence to the model, a likelihood ratio test showed a significant improvement in model fit (p = 0.00899), suggesting that locations selected by monarchs differ significantly in average light levels from other monitored locations within the groves. Note that due to model convergence issues, we had to simplify the random effects structure to only include random intercepts for groves.

simplify compared to what?

Standard deviation of light

```
light.std.model1 <- lmer(light.std_centered ~ seasonDay + (1 | grove), data=df)
light.std.model2 <- lmer(light.std_centered ~ seasonDay + butterfly_present + (1 | grove), data=df)
anova(light.std.model1, light.std.model2)

Data: df
Models:
light.std.model1: light.std_centered ~ seasonDay + (1 | grove)
light.std.model2: light.std_centered ~ seasonDay + butterfly_present + (1 | grove)
                 npar   AIC   BIC logLik deviance  Chisq Df Pr(>Chisq)    
light.std.model1    4 50856 50879 -25424    50848                         
light.std.model2    5 50843 50872 -25417    50833 14.589  1  0.0001337 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
performance::icc(light.std.model2)

# Intraclass Correlation Coefficient

    Adjusted ICC: 0.135
  Unadjusted ICC: 0.134

ggplot(df, aes(x=factor(butterfly_present), y=light.std)) +
  geom_boxplot(fill=c("#66c2a5", "#fc8d62")) +
  labs(x="Butterfly Present", y="Light Standard Deviation (lux)",
       title="Light Variability by Butterfly Presence") +
  scale_x_discrete(labels=c("No", "Yes")) +
  theme_minimal()
```



The ICC of 0.135 indicates that about 13.5% of the variation in light variability can be attributed to differences between groves. When we added butterfly presence to the model, a likelihood ratio test showed a highly significant improvement in model fit (p = 0.0001337), suggesting that locations selected by monarchs differ significantly in light variability from other monitored locations within the groves. The boxplots indicate that butterfly-present locations have higher light variability compared to other monitored locations.

So first need a Level 2 variable, like inland vs. coastal, across your groves. Then add

```
# Cross-level interaction model
cross_level_model <- lmer(light.avg_centered ~ seasonDay * butterfly_present + (1 | grove), data = df)

# Summary of the model
summary(cross_level_model)

Linear mixed model fit by REML ['lmerMod']
Formula: light.avg_centered ~ seasonDay * butterfly_present + (1 | grove)
   Data: df

REML criterion at convergence: 47385.9

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.5470 -0.6194 -0.2599  0.2045  4.7908

Random effects:
 Groups   Name        Variance Std.Dev.
 grove    (Intercept)  8407508 2900
 Residual             54623148 7391
Number of obs: 2295, groups:  grove, 9

Fixed effects:
                             Estimate Std. Error t value
(Intercept)                    54.815   1042.840   0.053
seasonDay                     -10.066      9.916  -1.015
butterfly_present           -1290.628    832.242  -1.551
seasonDay:butterfly_present    66.180     21.217   3.119

Correlation of Fixed Effects:
            (Intr) sesnDy bttrf_
seasonDay   -0.334
bttrfly_prs -0.172  0.402
ssnDy:bttr_  0.149 -0.435 -0.886
```

To further investigate the relationship between butterfly presence and light conditions over time, we fit a model including an interaction between seasonDay and butterfly presence. The model revealed a significant interaction between seasonDay and butterfly presence (Est. = 66.180, t = 3.119). This suggests that not only do monarchs select locations with different light conditions, but this relationship changes throughout the overwintering season. Note that this model also used the simplified random effects structure with only random intercepts for groves due to convergence issues with the more complex random effects structure.

graph!

So for purposes of this project, may need to play around with a second level 1 variable and try