Homework 3

Kinga Bihari

2024-05-28

Table of contents

Kinga's forked repo link: (https://github.com/biharikinga/bihari-kinga_homework-03.git)

Read in packages/data & do relevant analysis for plotting

Note: Data exploration, model diagnostics, model selection not included!

```
# general use
library(tidyverse)
library(janitor)
library(readxl)
library(here)

# model predictions
library(ggeffects)

# model tables
library(flextable)
library(modelsummary)

# read in data
drought_exp <- read_xlsx(path = here("data",
                                     "Valliere_etal_EcoApps_Data.xlsx"),
                         sheet = "First Harvest")
```

```r
# clean data
drought_exp_clean <- drought_exp %>%
  clean_names() %>% # nicer column names
  mutate(species_name = case_when( # adding column with species scientific names
    species == "ENCCAL" ~ "Encelia californica", # bush sunflower
    species == "ESCCAL" ~ "Eschscholzia californica", # California poppy
    species == "PENCEN" ~ "Penstemon centranthifolius", # Scarlet bugler
    species == "GRICAM" ~ "Grindelia camporum", # great valley gumweed
    species == "SALLEU" ~ "Salvia leucophylla", # Purple sage
    species == "STIPUL" ~ "Nasella pulchra", # Purple needlegrass
    species == "LOTSCO" ~ "Acmispon glaber" # deerweed
  )) %>%
  relocate(species_name, .after = species) %>% # moving species_name column after species
  mutate(water_treatment = case_when( # adding column with full treatment names
    water == "WW" ~ "Well watered",
    water == "DS" ~ "Drought stressed"
  )) %>%
  relocate(water_treatment, .after = water) # moving water_treatment column after water

# model creation
model4 <- lm(total_g ~ water_treatment + species_name,
             data = drought_exp_clean)
# look at model
summary(model4)
```

```
Call:
lm(formula = total_g ~ water_treatment + species_name, data = drought_exp_clean)

Residuals:
      Min        1Q    Median        3Q       Max
-0.157087 -0.046953 -0.003733  0.041244  0.192657

Coefficients:
                                      Estimate Std. Error t value Pr(>|t|)
(Intercept)                            0.05455    0.02451   2.225  0.02973 *
water_treatmentWell watered            0.11695    0.01733   6.746 5.90e-09 ***
species_nameEncelia californica        0.21774    0.03243   6.714 6.70e-09 ***
species_nameEschscholzia californica   0.23164    0.03243   7.143 1.22e-09 ***
species_nameGrindelia camporum         0.31335    0.03243   9.662 5.53e-14 ***
species_nameNasella pulchra            0.22881    0.03243   7.055 1.72e-09 ***
```

```
species_namePenstemon centranthifolius  0.05003    0.03243   1.543  0.12799
species_nameSalvia leucophylla          0.12020    0.03243   3.706  0.00045 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.07252 on 62 degrees of freedom
Multiple R-squared:  0.7535,    Adjusted R-squared:  0.7257
F-statistic: 27.08 on 7 and 62 DF,  p-value: < 2.2e-16
```

```
# create model predictions table
model_preds <- ggpredict(model4,
                         terms = c("water_treatment",
                                   "species_name"))
```

1: Multiple linear regression: model selection and construction

1.a. Table 1. Model selection for multiple linear regression of total biomass as a function of SLA, water treatment, and/or species. Five models are considered, with Model 0 as the null model (no predictors), Model 1 as the saturated model (all predictors), and Models 2-4 with some combination of the 3 predictors. The rows represent the different models, and columns represent the different aspects of each model. Rows are sorted by "best" (lowest AIC value) model to "worst" (highest AIC value) model.

```
# read in table of models
models <- read_csv("models.csv")

# use flextable to make nice table
flextable(models)
```

Model number	Model description	Predictors	AIC	AIC delta
4	water + species	water, species	-156.2	0.00
1	saturated model	SLA, water, species	-153.8	2.44
3	SLA + species	SLA, species	-124.1	32.12

Model number	Model description	Predictors	AIC	AIC delta
2	SLA + water	SLA, water	-95.8	60.37
0	null model	none	-75.0	81.22

1.b. To examine the influence of specific leaf area (SLA), water treatment (well watered or drought stressed), and species type on total biomass, I ran a multiple linear regression and chose the simplest linear model that predicted total biomass the best. Since there were 3 predictor variables, I ran 5 models (null, saturated, and 3 models with different combinations of the predictors). To determine the model that best predicted total biomass, I compared all 5 models and their Akaike Information Criterion (AIC) using model selection. Model 4, with water treatment and species predictors, had the lowest AIC and thus was the simplest model that fit the data best. To evaluate linear model assumptions, I ran diagnostics on all 5 models. I visually evaluated the linear relationship between the response and predictors, determined independent errors, ensured homoscedastic residuals, normally distributed residuals, and lack of residual outliers.

1.c. & 1.d.

```
# plot model predictions
ggplot(model_preds,              # df
       aes(x = reorder(group,    # reordering x-axis by mass
                       -predicted),
           y = predicted)) +     # x/y axes

# model predictions
  geom_point(aes(color = x)) +  # color by water treatment

# facet by species
  facet_wrap("group") +

# add underlying data
  geom_point(drought_exp_clean,              # df
    mapping = aes(reorder(x = species_name,
                          -total_g),
                  y = total_g,               # x/y axes
                  color = water_treatment), # color by water treatment
      alpha = 0.2) +                         # translucent points

# plot 95% CI
```

```
  geom_errorbar(aes(ymin = conf.low,
                    ymax = conf.high,
                    color = x)) +           # color by water treatment

# labels (x/y axes, caption)
  labs(x = "Species",
       y = "Total biomass weight (g)",
       caption = "Figure 1. Species and water treatment predict total biomass weight in
       local plants. Colors represent water treatment: blue points are well watered,
       orange are drought stressed. Translucent points represent all recorded weights
       of plants. Opaque points represent model predictions for that species, and error
       bars are the 95% confidence interval. Species names are shortened to just genus
       on x-axis for readability. Dataset from Valliere et al. (2019) on Dryad.") +

# clean background
  theme_bw() +

# custom colors
  scale_color_manual(values = c("skyblue2", "orange2")) +

# other formatting stuff
  theme(legend.position = "none",           # no legend
    axis.text.x = element_text(angle = 60,
                               hjust = 1,
                               vjust = 1),  # angle x axis labels
    plot.caption = element_text(hjust = 0,  # caption left adjusted
                                size = 12), # font size
    plot.caption.position = "plot") +       # caption fill entire plot
  scale_x_discrete(labels = c("Grindelia",
                              "Eschscholzia",
                              "Nasella",
                              "Encelia",
                              "Salvia",
                              "Penstemon",
                              "Acmispon"))    # shorter x axis labels
```

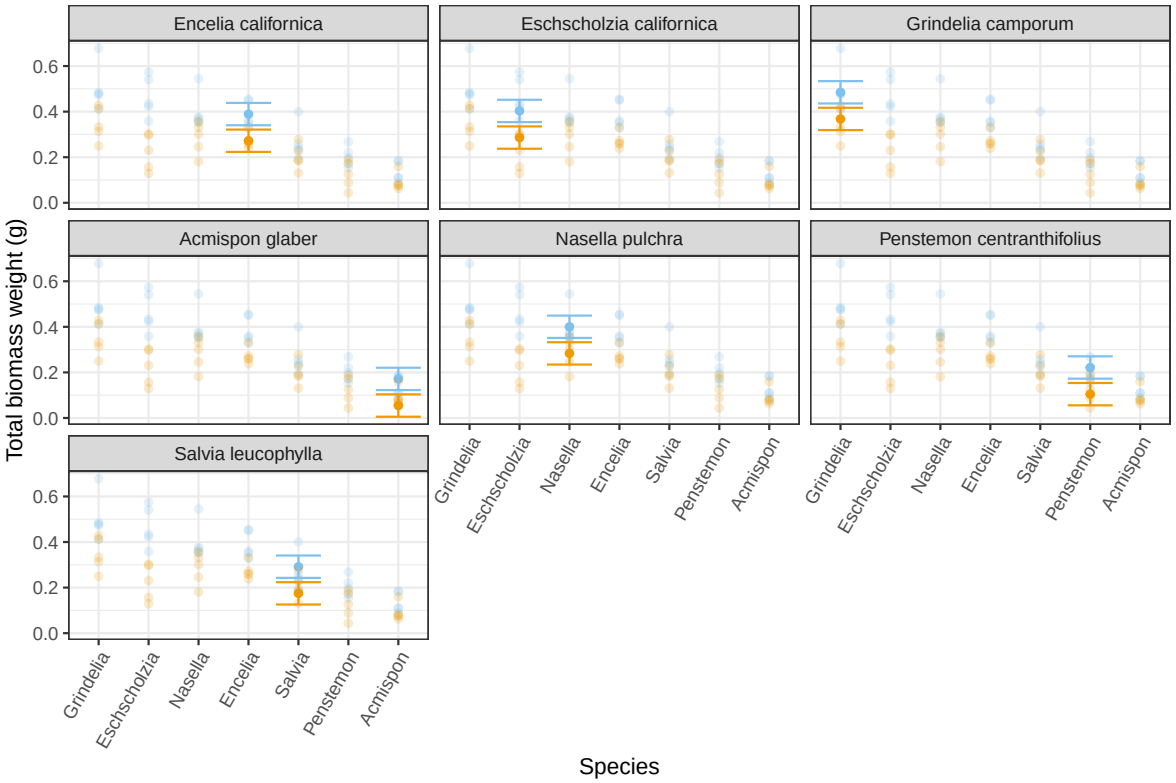


Figure 1. Species and water treatment predict total biomass weight in local plants. Colors represent water treatment: blue points are well watered, orange are drought stressed. Translucent points represent all recorded weights of plants. Opaque points represent model predictions for that species, and error bars are the 95% confidence interval. Species names are shortened to just genus on x−axis for readability. Dataset from Valliere et al. (2019) on Dryad.

1.e. Comparing the models, I found that Model 4 (AIC = -156.2), which used species and water treatment as predictors, best predicted total plant biomass (multiple linear regression, $F(7, 26) = 27.08$, $p < 0.001$, $= 0.05$, $R^2 = 0.73$). I found that on average, the well watered treatment predicted higher biomass across all species by about 0.2 g when compared to drought stressed treatment. I found that *Grindelia camporum* had the highest biomass, while *Acmispon glaber* had the lowest biomass, with a difference of about 0.3 g.

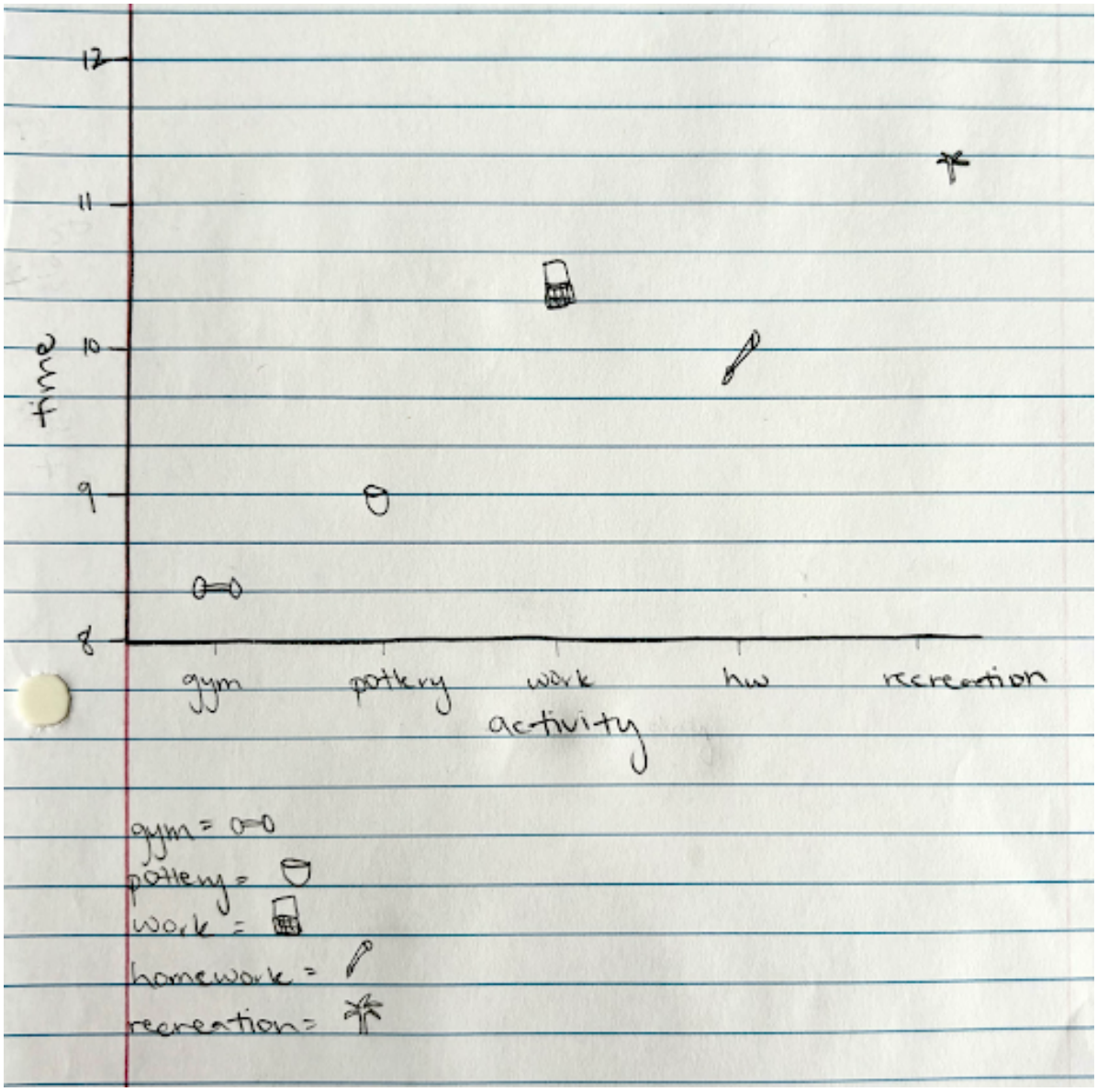
2: Affective visualization

Personal data exploration:

```
# read in personal data for #3 exploratory plot
breakfast <- read_csv("stats brekkie data.csv") |>
  clean_names() |>    # clean names
  slice(1:38)         # select dates I have data for
```

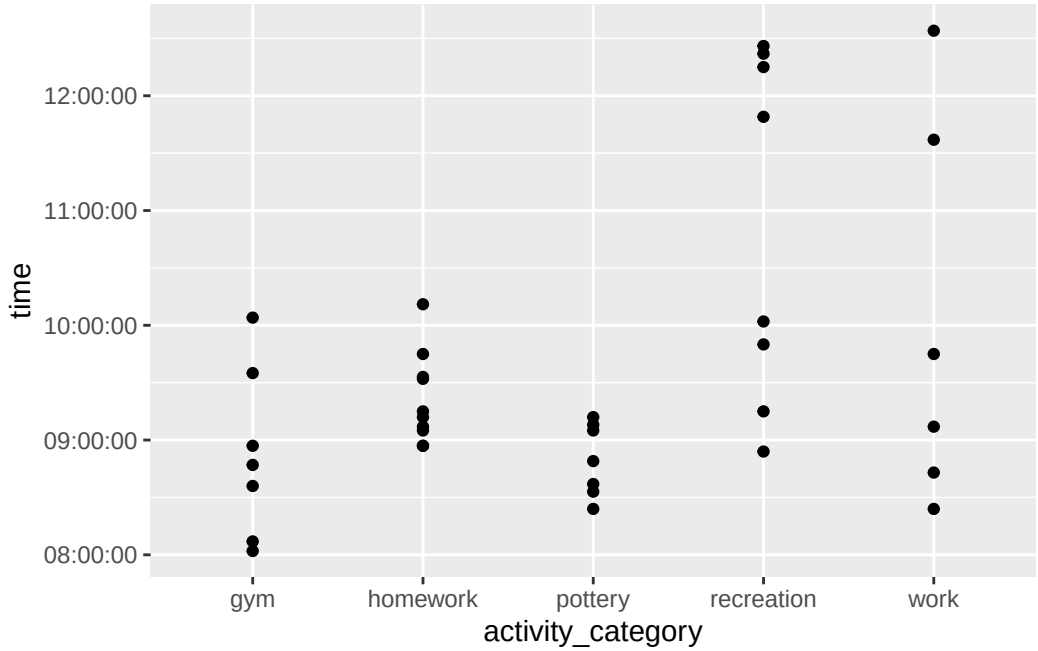
2.a. My personal data is focused on when/what I eat for breakfast based on what I'm doing after. Since the "theme" is breakfast and activities, I could use these as my visualizations. I'm thinking of using emojis of breakfast foods or activities to be the points of a scatterplot. Or, I could scale the size of the emojis to represent how many times I've eaten that food or done that activity.

2.b.




```
# exploratory plot
ggplot(data = breakfast,
```

```
       aes(x = activity_category,
           y = time)) +  # dataframe & axes
# scatter plot!
  geom_point()
```



2.c.

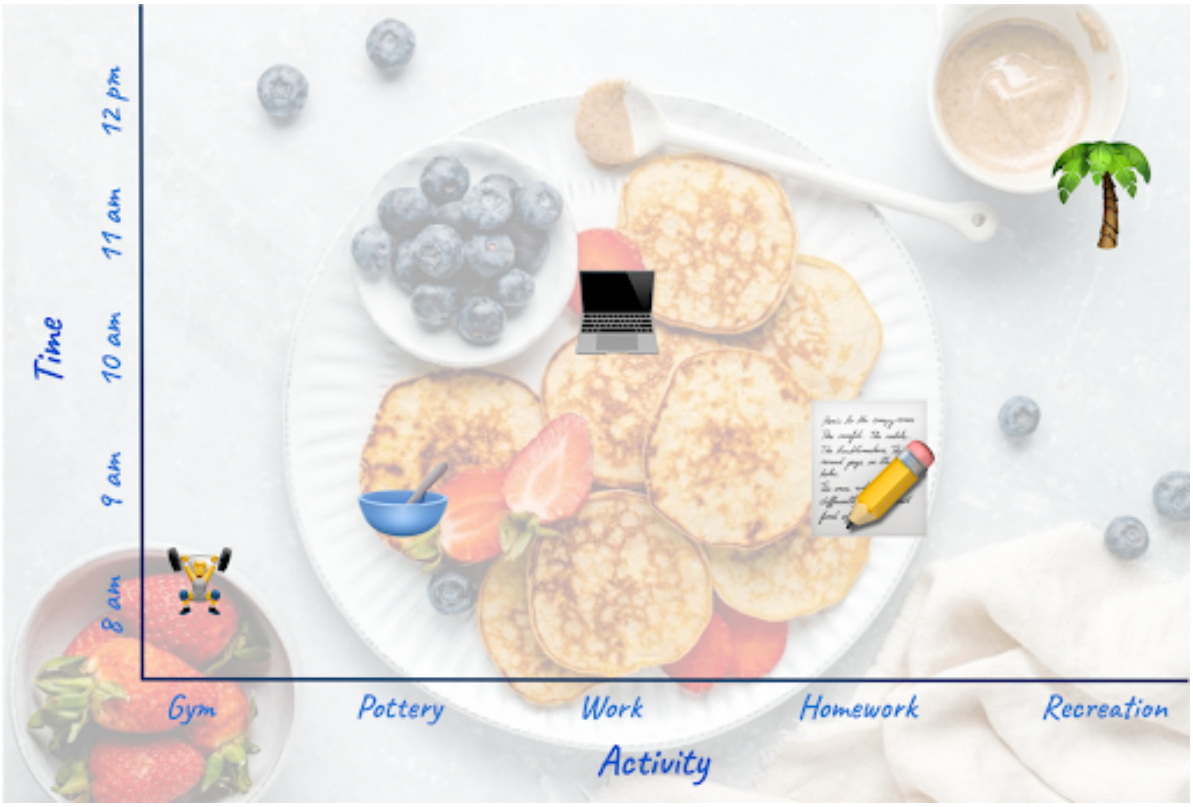
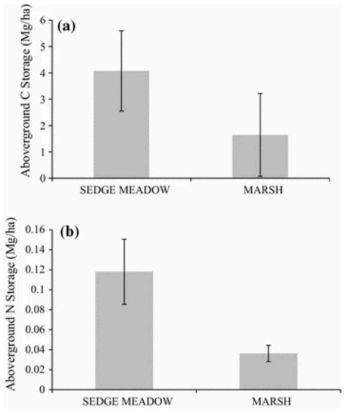


Figure 2. When I have breakfast based on what activity I'm doing after. The icons represent the activity type and the mean time of day (gym = 8:17 am, pottery = 8:49 am, work = 10:25 am, homework = 9:02 am, recreation = 11:34 am) I eat breakfast before doing the corresponding activity. The icons' size corresponds to how many times I have recorded doing that activity (gym = 7, pottery = 8, work = 8, homework = 11, recreation = 9). The background is an example of what I usually eat for breakfast.

2.d. This is an artistic rendition of how the activity I do after breakfast influences what time I eat breakfast, with icons representing the type of activity, and the size of the icon representing how many times I recorded that activity (recreation the most, gym the least). I was influenced by artists who use emojis, which simply and easily convey a lot more information than just words, to communicate information in infographics on various social media news forums (such as the accounts @feminist and @thenewsmovement). The form is "digital collage", since I used various images, icons, and text to create my graph. I used Adobe Photoshop to create this piece, as the creative freedom to move all elements wherever I wanted to was crucial in being able to bring together everything into a cohesive message.

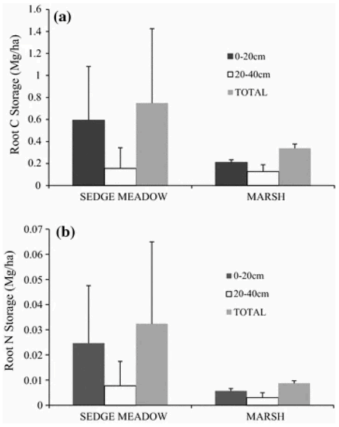
3: Statistical critique

3.a. Their main objective is: to quantify C and N storage of a restored wetland at the Emiquon Preserve (a restored wetland) and examine it for additional C sequestration by comparing its C storage with C storage of reference natural wetlands. They used a one-way ANOVA to determine how restored wetland soil depth influenced OC storage and TN storage of root biomass and soils. They also did a nested three-way ANOVA test to determine how the reference wetland type, soil depth, and site (nested within wetland type) influenced OC storage and TN storage of above and below ground plant biomass and soil.



Average storage (Mg ha^{-1}) of organic carbon (**a**) and total nitrogen (**b**) in aboveground plant biomass in two reference sedge meadow sites and two reference marsh sites. The average was calculated based on ten samples. *Bars* represent one standard deviation

Fig. 2: Main message is that sedge meadows (restored wetlands) have significantly higher aboveground N and C storage than the reference marshes (natural wetlands). x-axis is aboveground C or N storage, y-axis is the wetland type (sedge or marsh).



Average storage (Mg ha^{-1}) of organic carbon (**a**) and total nitrogen (**b**) of root biomass of two reference sedge meadow sites and two reference marsh sites at 0–20, 20–40, and total 0–40 cm. The average was calculated based on ten composited samples. *Bars* represent one SD

Fig. 3: Main message is that sedge meadows (restored wetlands) have significantly higher root N and C storage than the reference marshes (natural wetlands). Both have lower root N and C storage in the upper 20 cm of soil than 20-40 cm deep. x-axis is root C or N storage, y-axis is the wetland type (sedge or marsh).

3.b. They represented their statistics pretty clearly in the figures. However, they didn't use color to differentiate between the sedge and marsh, and the use of white as a color in Fig. 3 is confusing because at first glance it just looks like a gap. The x and y axes are logical and easy to read, with correct units. They show standard deviation on top of all of the data, but no other summary statistics. The SD bars do help to show the large variance in data, but SE + mean could've been a better way to visualize that. They don't have any applicable model predictions.

3.c. I think they handled visual clutter very well. The data:ink ratio is very high - all of the ink is being used to describe the data in a meaningful way. There are no gridlines, the plots aren't cluttered with irrelevant data, the labels are also informative but not too detailed, and have units where appropriate. The a) and b) plots are also labeled clearly and explained in the caption instead of cluttering the plot with unnecessary titles/descriptions. The legend in figure 5 is simple but necessary as well.

3.d. My first recommendation is to use color instead of gray gradient (color blind friendly colors), especially replacing the white with a color so it's clear that's also data, not just blank space. I would also color code sedge and marsh in fig. 3 to make it more visually pleasing, and do different shades of those colors for the 3 depths in fig. 5. I would increase the axes text size a bit to make it easier to read. I would maybe replace standard deviation with mean/SE and/or add the underlying data as well. I would keep everything else the same because overall it's a pretty good graph.