11: Crafting Reports

Karly Nocera

Spring 2021

LESSON OBJECTIVES

1. Describe the purpose of using R Markdown as a communication and workflow tool
2. Incorporate Markdown syntax into documents
3. Communicate the process and findings of an analysis session in the style of a report

USE OF R STUDIO & R MARKDOWN SO FAR...

1. Write code
2. Document that code
3. Generate PDFs of code and its outputs
4. Integrate with Git/GitHub for version control

BASIC R MARKDOWN DOCUMENT STRUCTURE

1. **YAML Header** surrounded by — on top and bottom
   - YAML templates include options for html, pdf, word, markdown, and interactive
   - More information on formatting the YAML header can be found in the cheat sheet
2. **R Code Chunks** surrounded by "`on top and bottom + Create using`Cmd/Ctrl+Alt+I'
   - Can be named {r name} to facilitate navigation and autoreferencing
   - Chunk options allow for flexibility when the code runs and when the document is knitted
3. **Text** with formatting options for readability in knitted document

RESOURCES

Handy cheat sheets for R markdown can be found: here, and here.

There's also a quick reference available via the `Help→Markdown Quick Reference` menu.

Lastly, this website give a great & thorough overview.

THE KNITTING PROCESS

- The knitting sequence 

- Knitting commands in code chunks:
- `include = FALSE` - code is run, but neither code nor results appear in knitted file
- `echo = FALSE` - code not included in knitted file, but results are

- `eval = FALSE` - code is not run in the knitted file
- `message = FALSE` - messages do not appear in knitted file
- `warning = FALSE` - warnings do not appear...
- `fig.cap = "..."` - adds a caption to graphical results

WHAT ELSE CAN R MARKDOWN DO?

See: https://rmarkdown.rstudio.com and class recording. * Languages other than R... * Various outputs...

WHY R MARKDOWN?

<Fill in our discussion below with bullet points. Use italics and bold for emphasis (hint: use the cheat sheets or `Help` →`Markdown Quick Reference` to figure out how to make bold and italic text).>

- R markdown is limited to ~~one language~~ **many languages**
- Easier for organizting thoughts and separating context from script
- Tidier way to share code and results with others
- Easier user interface (*mostly*)
- No need to constantly **re-run** code from the beginning
- *Helpful* To create a line break, add new line space in between
- *Helpful* To find the guide: Help » Markdown Quick Guide (file/edit/code bar, not right hand side Help tab)

TEXT EDITING CHALLENGE

Create a table below that details the example datasets we have been using in class. The first column should contain the names of the datasets and the second column should include some relevant information about the datasets. (Hint: use the cheat sheets to figure out how to make a table in Rmd)

Dataset Names	Information
EPAair_O3_NC2018_raw	2018 EPA air samples of ozone levels in North Carolina
NTL-LTER_Lake_Carbon_Raw	Water data on dissolved organic and inorganic carbon, particulate organic matter, CO2, etc.
NWIS_SiteFlowData_NE_Raw	Water gage heights at different sample sites

R CHUNK EDITING CHALLENGE

Installing packages

Create an R chunk below that installs the package `knitr`. Instead of commenting out the code, customize the chunk options such that the code is not evaluated (i.e., not run).

```
install.packages("knitr")
```

Setup

Create an R chunk below called "setup" that checks your working directory, loads the packages `tidyverse`, `lubridate`, and `knitr`, and sets a ggplot theme. Remember that you need to disable R throwing a message, which contains a check mark that cannot be knitted.

Load the NTL-LTER_Lake_Nutrients_Raw dataset, display the head of the dataset, and set the date

column to a date format.

Customize the chunk options such that the code is run but is not displayed in the final document.

```
## [1] "Z:/Environmental_Data_Analytics_2021"
```

```
##   lakeid  lakename year4 daynum sampledate depth_id depth tn_ug tp_ug nh34 no23
## 1      L Paul Lake  1991    140    5/20/91        1  0.00   538    25   NA   NA
## 2      L Paul Lake  1991    140    5/20/91        2  0.85   285    14   NA   NA
## 3      L Paul Lake  1991    140    5/20/91        3  1.75   399    14   NA   NA
## 4      L Paul Lake  1991    140    5/20/91        4  3.00   453    14   NA   NA
## 5      L Paul Lake  1991    140    5/20/91        5  4.00   363    13   NA   NA
## 6      L Paul Lake  1991    140    5/20/91        6  6.00   583    37   NA   NA
##   po4 comments
## 1  NA
## 2  NA
## 3  NA
## 4  NA
## 5  NA
## 6  NA
```

Data Exploration, Wrangling, and Visualization

Create an R chunk below to create a processed dataset do the following operations:

- Include all columns except lakeid, depth_id, and comments
- Include only surface samples (depth = 0 m)
- Drop rows with missing data

Create a second R chunk to create a summary dataset with the mean, minimum, maximum, and standard deviation of total nitrogen concentrations for each lake. Create a second summary dataset that is identical except that it evaluates total phosphorus. Customize the chunk options such that the code is run but not displayed in the final document.

Create a third R chunk that uses the function `kable` in the knitr package to display two tables: one for the summary dataframe for total N and one for the summary dataframe of total P. Use the `caption = " "` code within that function to title your tables. Customize the chunk options such that the final table is displayed but not the code used to generate the table.

Create a fourth and fifth R chunk that generates two plots (one in each chunk): one for total N over time with different colors for each lake, and one with the same setup but for total P. Decide which geom option will be appropriate for your purpose, and select a color palette that is visually pleasing and accessible. Customize the chunk options such that the final figures are displayed but not the code used to generate the figures. In addition, customize the chunk options such that the figures are aligned on the left side of the page. Lastly, add a fig.cap chunk option to add a caption (title) to your plot that will display underneath the figure.

```
nutrients_processed <-
    nutrients %>%
    select(lakename:sampledate, depth:po4) %>%
    filter(depth == 0) %>%
    drop_na()
```

Table 2: Total Nitrogen Summary Statistics

Mean	Min	Max	Standard Deviation
610.9982	45.67	2870.302	333.8124

Table 3: Total Phosphorous Summary Statistics

Mean	Min	Max	Standard Deviation
18.96872	0	101.05	12.72575

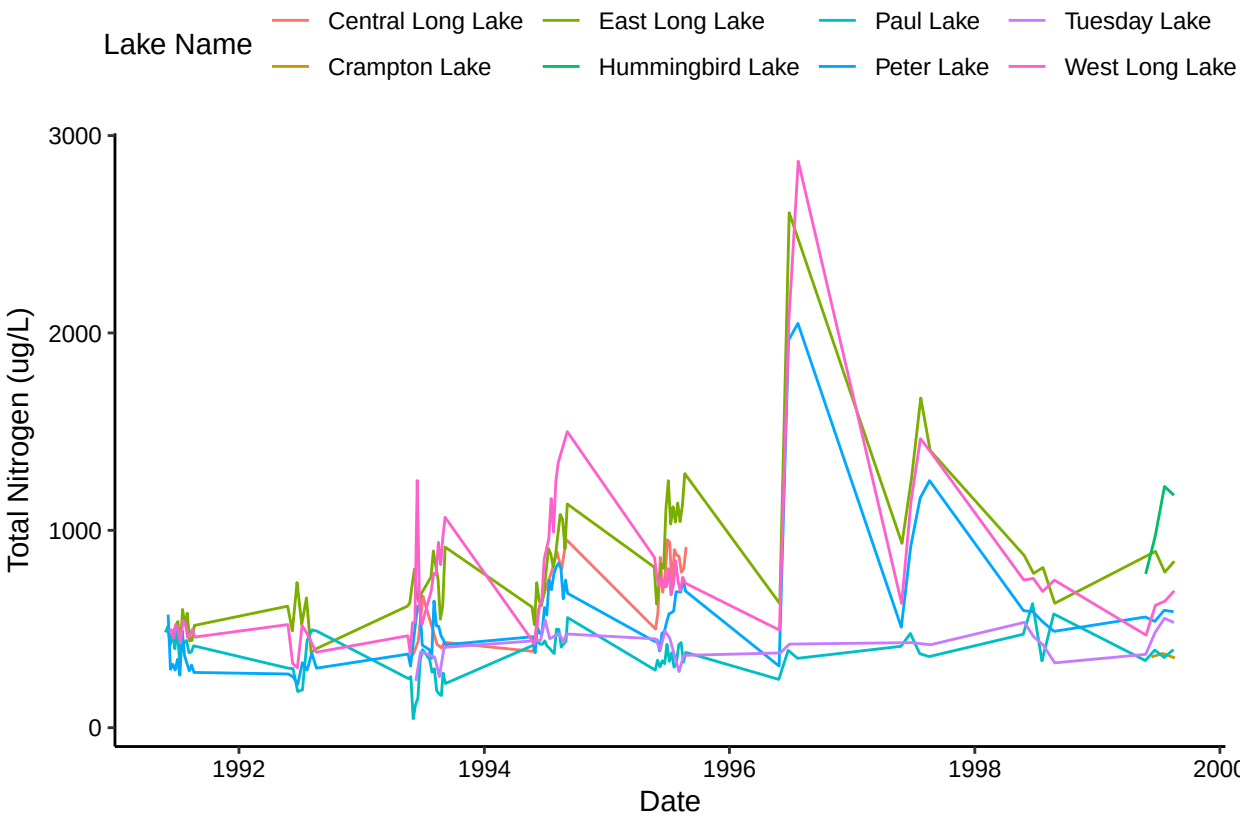

Figure 1: Total Nitrogen by Lake Over Time

Communicating results

Write a paragraph describing your findings from the R coding challenge above. This should be geared toward an educated audience but one that is not necessarily familiar with the dataset. Then insert a horizontal rule below the paragraph. Below the horizontal rule, write another paragraph describing the next steps you might take in analyzing this dataset. What questions might you be able to answer, and what analyses would you conduct to answer those questions?

> The first set of basic time series graphs illustrates that total nitrogen seems to be less variable over time than total phosphorous. This pattern is further emphasized in the facet grid graphs where it visualizes the pattern for each lake by year: with few exceptions (such as 1996 and West Long Lake) nitrogen levels remain fairly constant throughout the year with slight elevations near the end of summer; in contrast, phosphorous levels show a variety of temporal trends depending on their year and lake, however it does appear that its levels increase tend to increase with warmer temperatures (later in the summer). Lastly, the facet wrap graphs demonstrate that East Long, West Long, and Peter Lakes experience significantly more variability in nitrogren and phosphorous over the years than the others, although it is worth noting that there is insufficient data samples

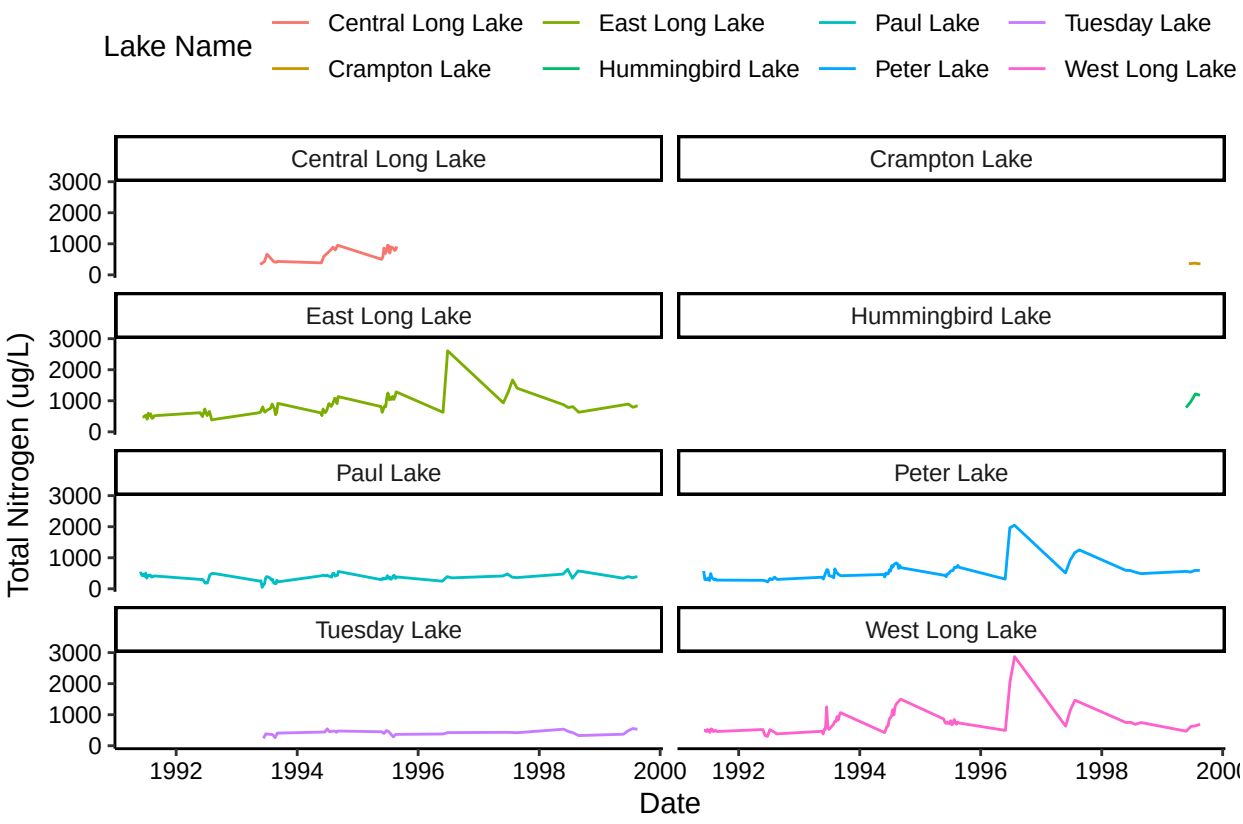

Figure 2: Total Nitrogen by Lake Over Time

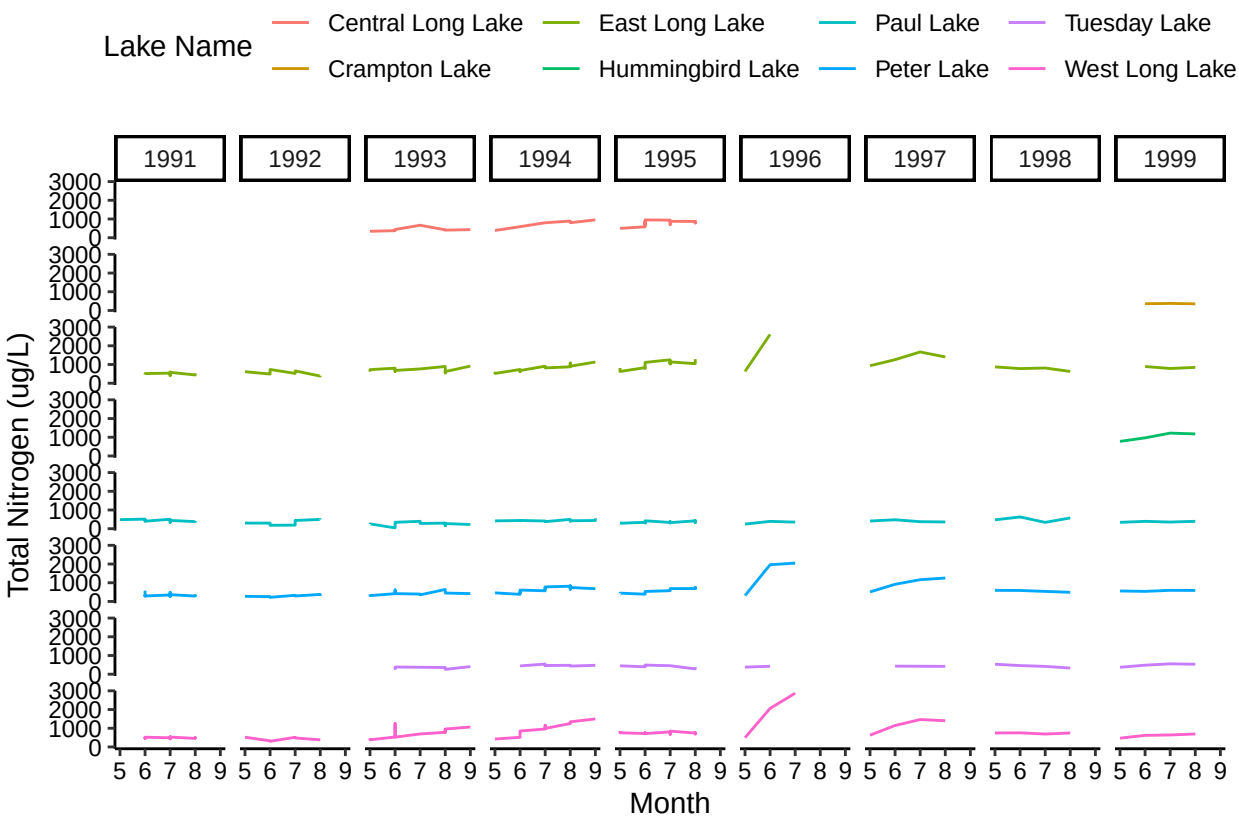

Figure 3: Total Nitrogen by Lake Over Time

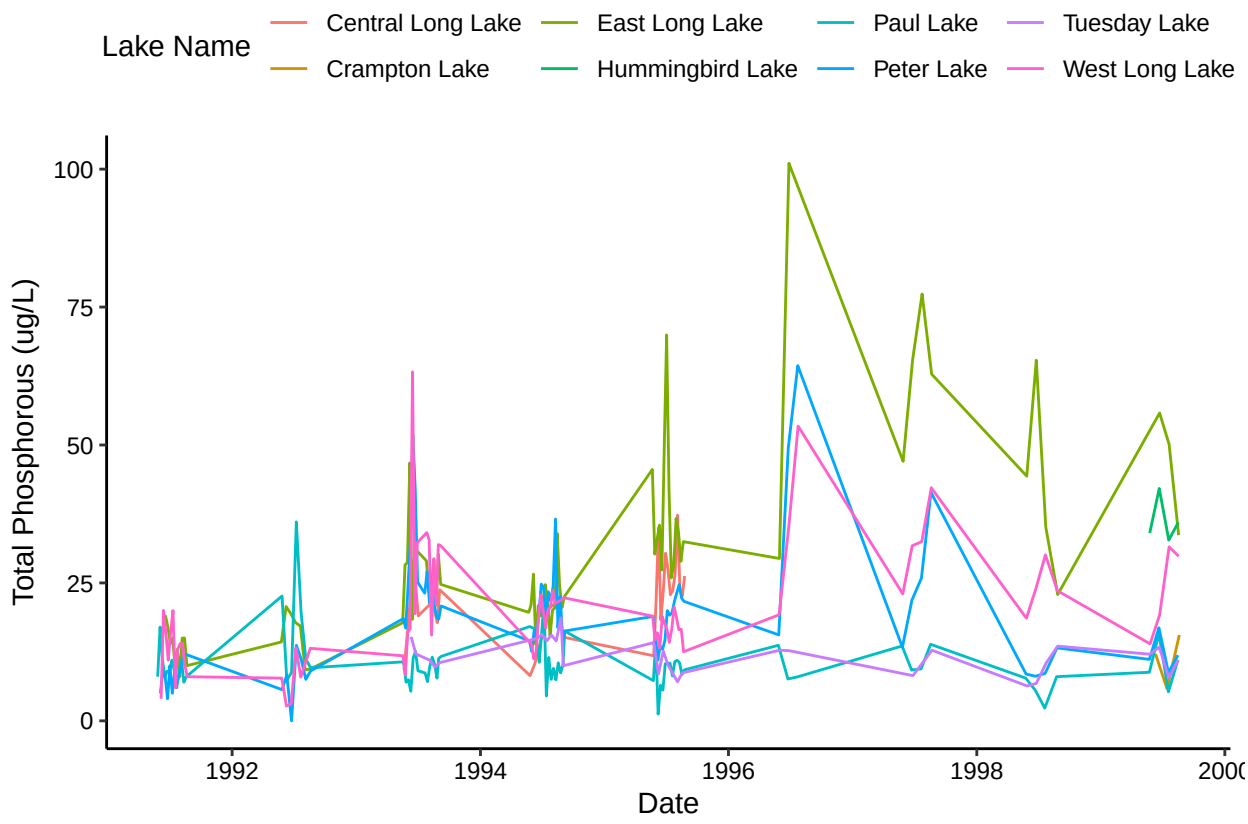

Figure 4: Total Phosphorous by Lake Over Time

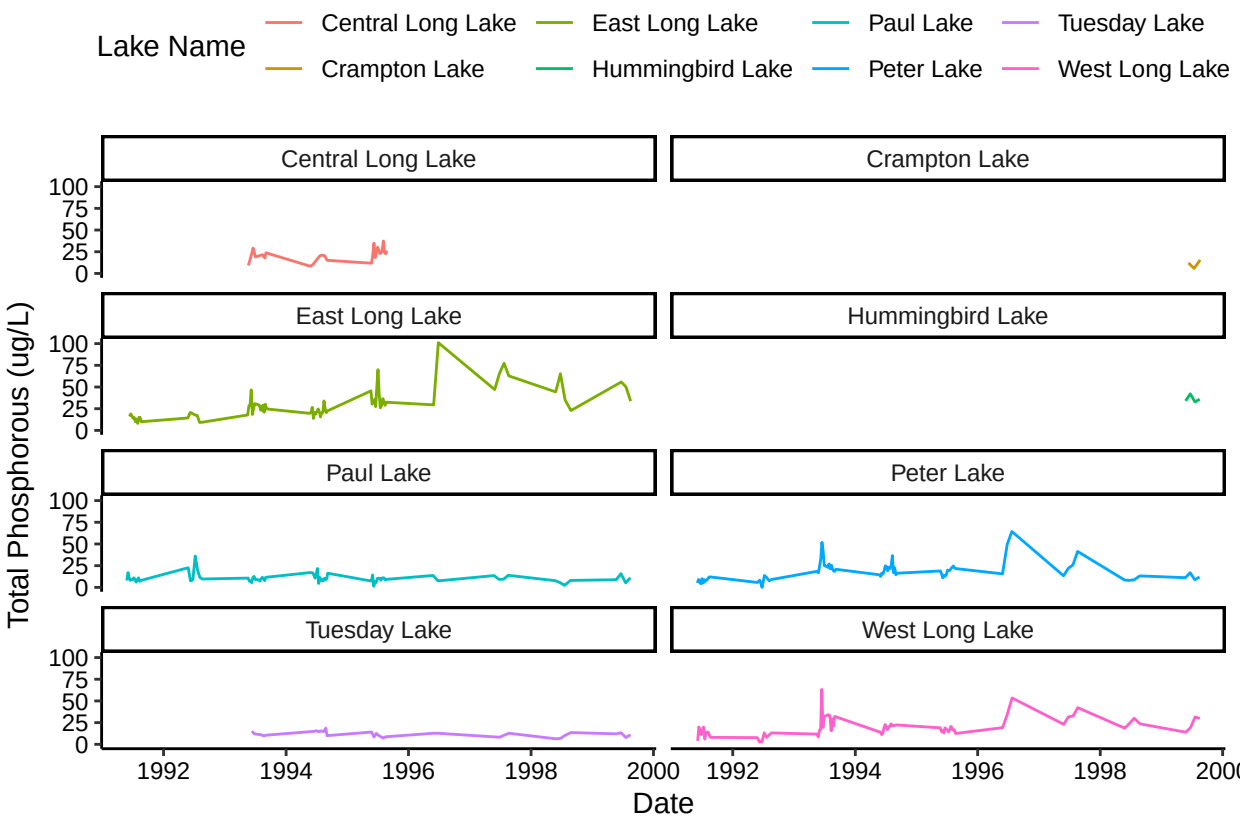

Figure 5: Total Phosphorous by Lake Over Time

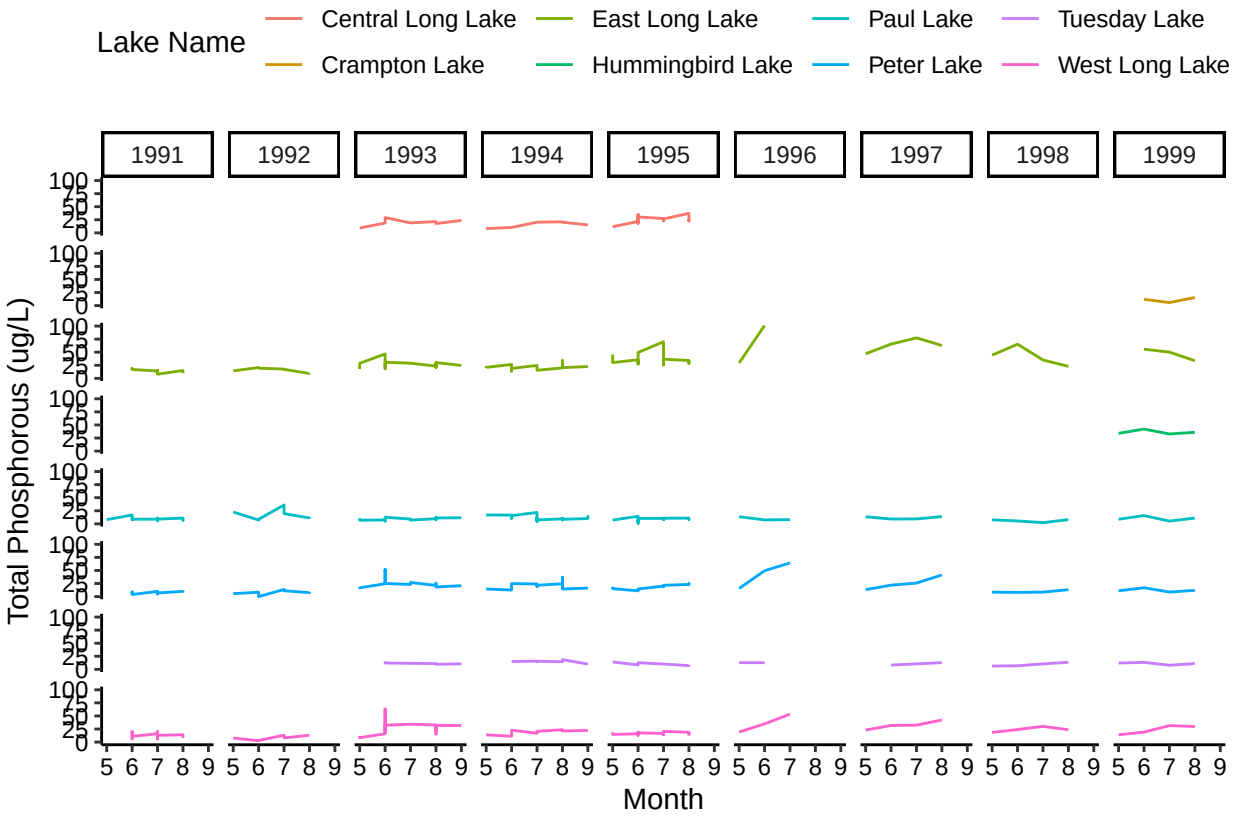

Figure 6: Total Phosphorous by Lake Over Time

to make this assessment of Crampton and Hummingbird Lakes.
********** The next steps in analyzing this dataset should include corresponding temperature and gage heights to answer whether weather and storm events can explain some of these trends. Specifically, running regressions to identify correlations, establish an ordinary least squares best fit line, and test for joint correlations can be conducted to answer these questions. This output will also indicate the strength of our model and indicate whether there are omitted variables that we have not considered.

KNIT YOUR PDF

When you have completed the above steps, try knitting your PDF to see if all of the formatting options you specified turned out as planned. This may take some troubleshooting.

OTHER R MARKDOWN CUSTOMIZATION OPTIONS

We have covered the basics in class today, but R Markdown offers many customization options. A word of caution: customizing templates will often require more interaction with LaTeX and installations on your computer, so be ready to troubleshoot issues.

Customization options for pdf output include:

- Table of contents
- Number sections
- Control default size of figures
- Citations
- Template (more info here)

pdf_document:
toc: true
number_sections: true
fig_height: 3
fig_width: 4
citation_package: natbib
template: