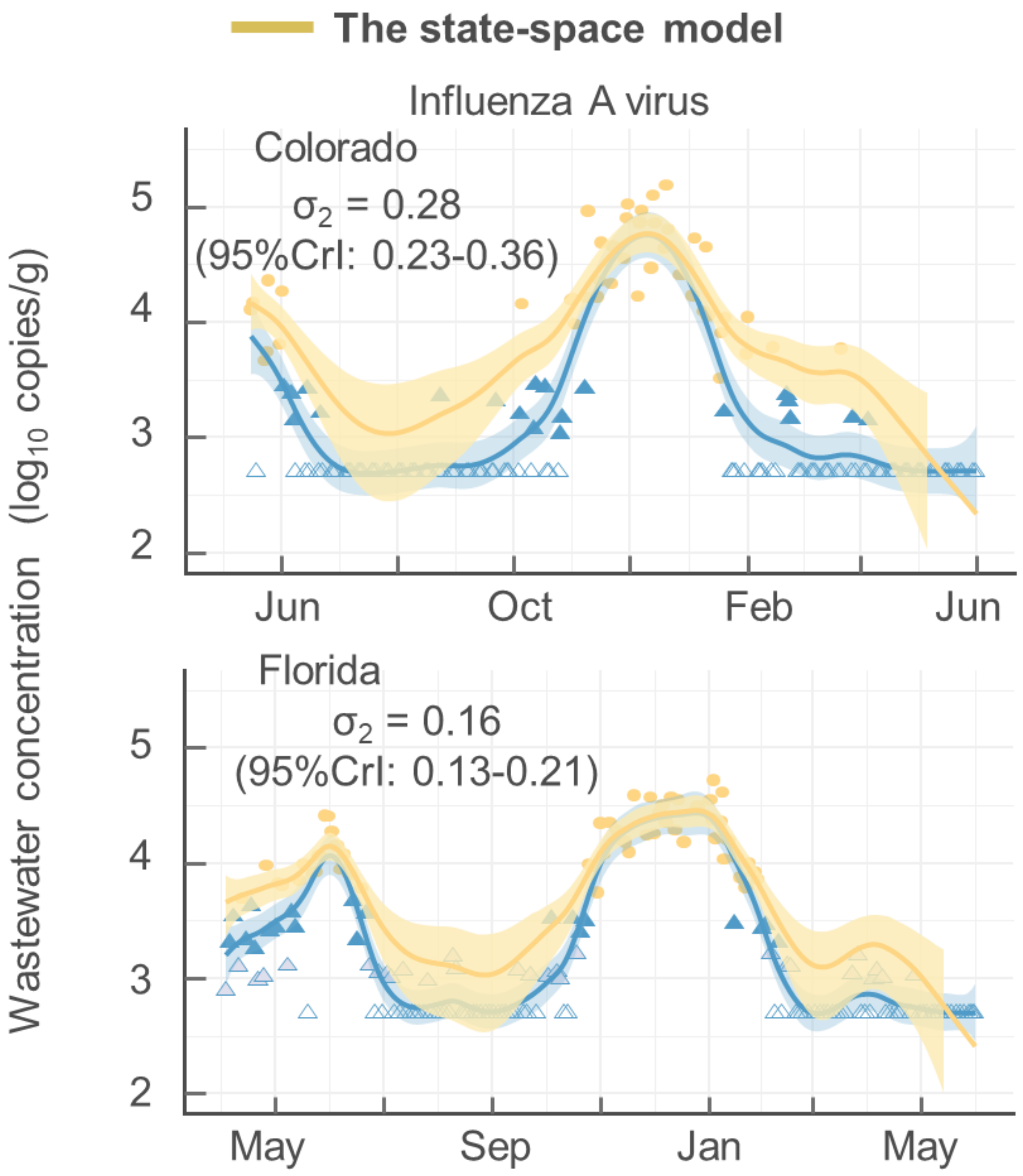
Brief guide for using the state-space model



Content

Name of used R and stan files
(state_space_model_with_logistic.R)
(state_space_model_with_logistic.stan)

- Install and download (slide: 2 ~ 3)
- Import wastewater-based data (slide: 4 ~ 9)
- Set prior distribution (slide 10)
- Implement analysis (slide 11 ~ 13)

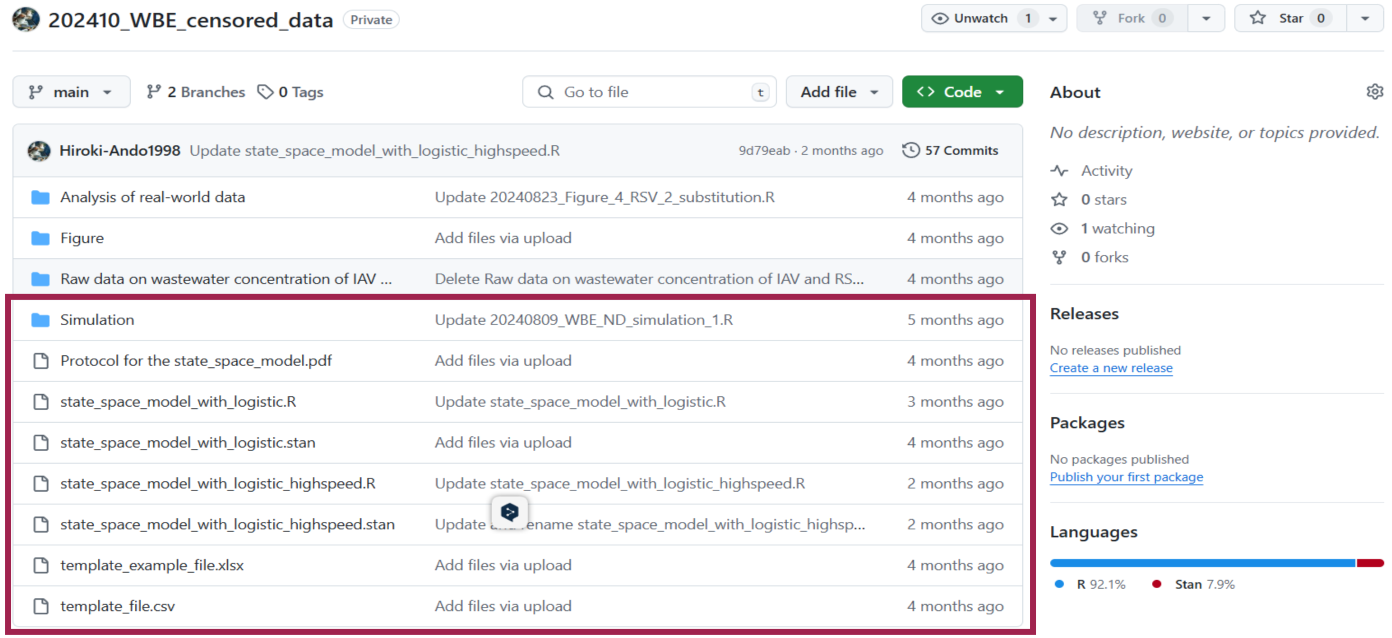
Name of used R and stan files
(state_space_model_with_logistic_highspeed.R)
(state_space_model_with_logistic_highspeed.stan)

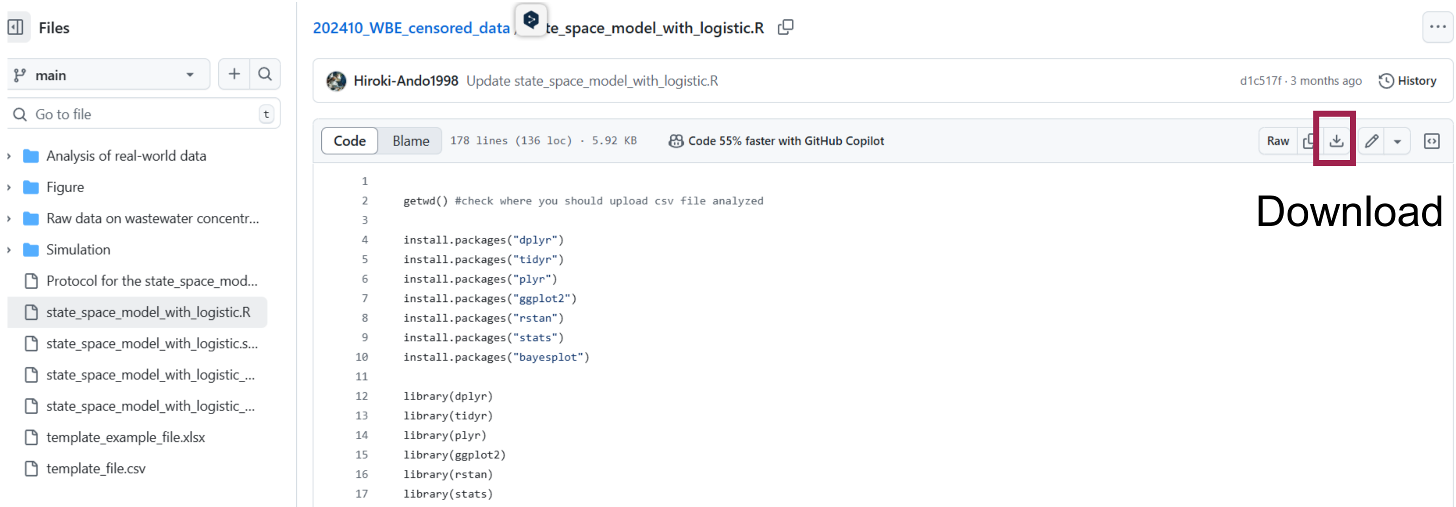
- High speed analysis

Contact: hirokiando@arizona.edu

- Install "R" and "Rstudio"
https://rstudio-education.github.io/hopr/starting.html

- Download R file, csv file, and stan file from the github.





- Open the CSV file downloaded in your laptop
- Input wastewater-based data in the template CSV file

	A	B	C	D	E
	date	number_of_tested_sa	number_of_positive_samples	concentration	
	2022/1/1				
	2022/1/2				
	2022/1/3	2	2	30955.34439	
	2022/1/4	3	3	24419.80553	
	2022/1/5				
	2022/1/6	NA	NA	NA	
	2022/1/7	NA	NA	NA	
	2022/1/8	2	1	1334.963692	
	2022/1/9	5	3	1842.572202	
	2022/1/10	5	3	2632.292098	
	2022/1/11	5	3	3171.828039	
	2022/1/12	4	3	2149.002467	
	2022/1/13	5	2	4058.253557	
	2022/1/14	4	1	13460.97495	

Data in non sampling day should be blank or “NA”.

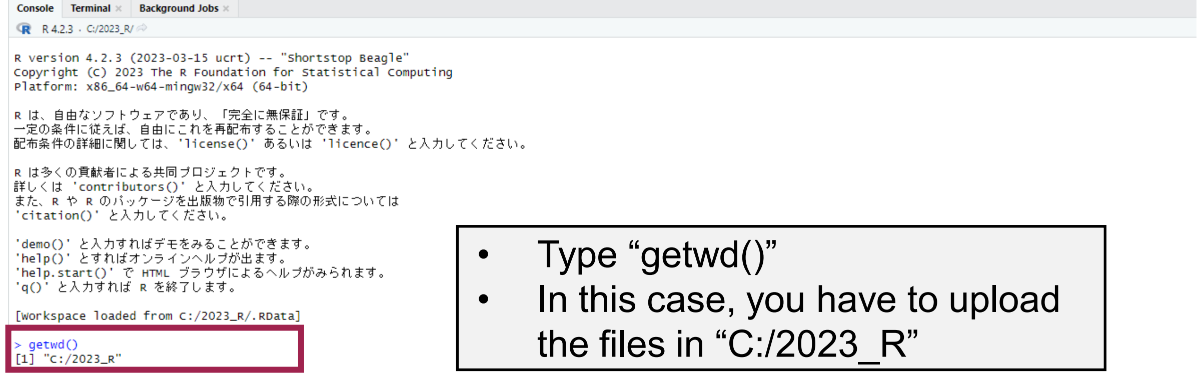
Integer

Integer

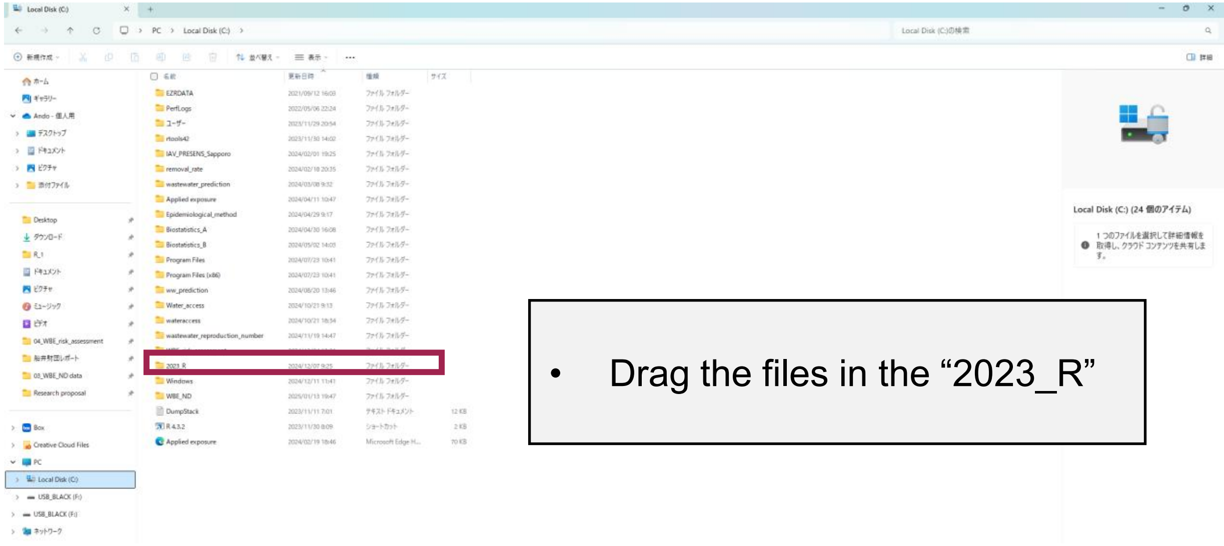
Copies/L

- Open “Rstudio” in your desktop

- Check where you have to upload CSV file, R and Stan file.

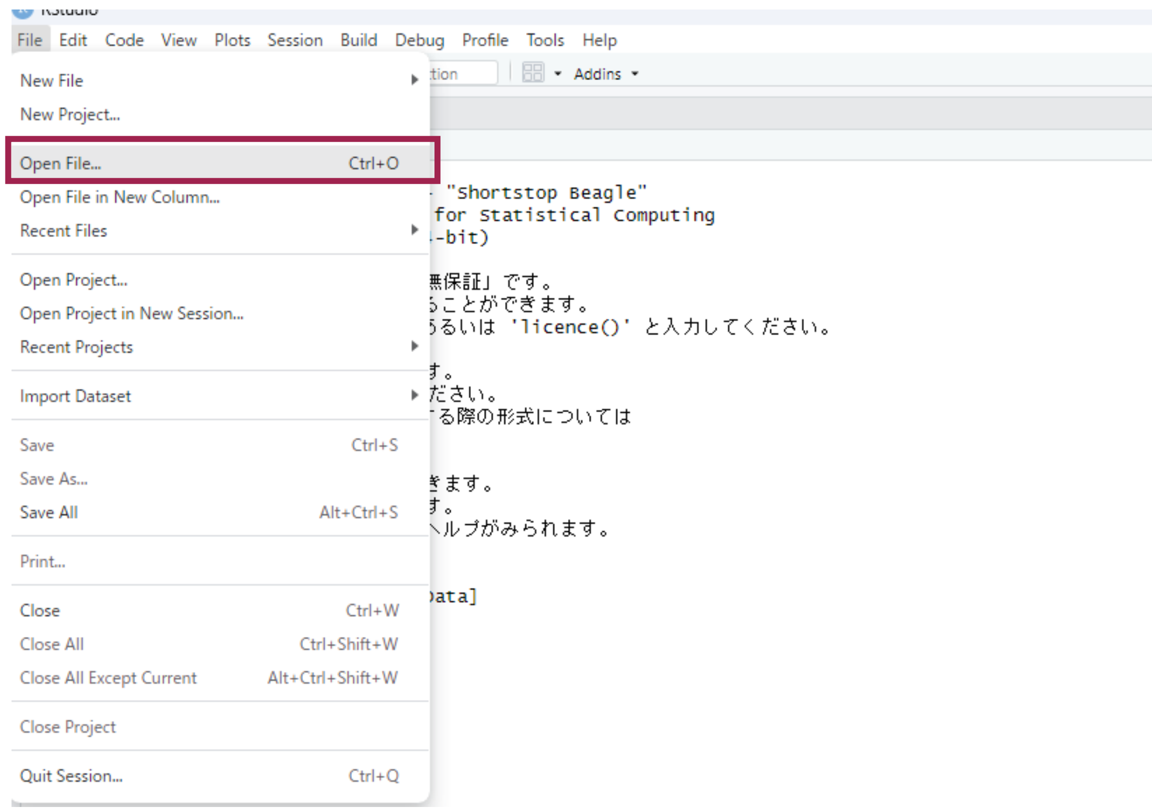


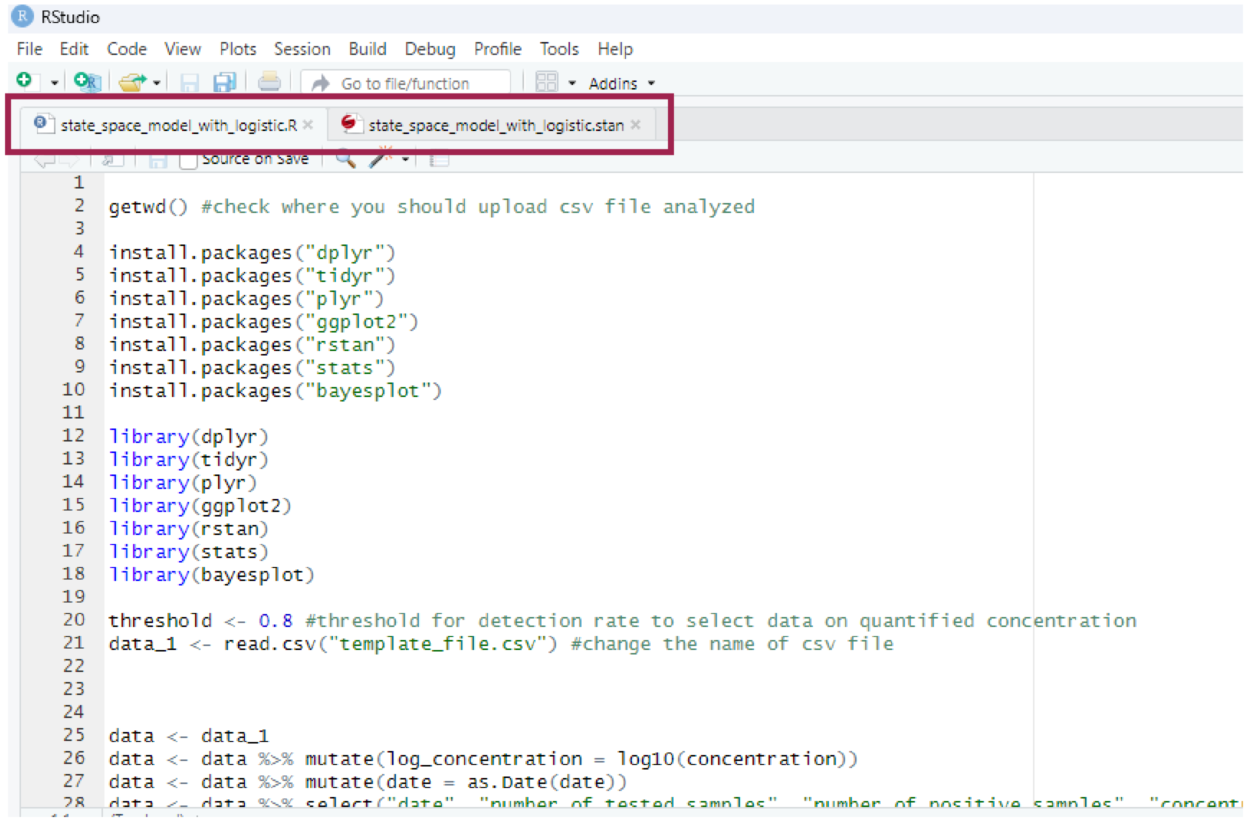
- Type “getwd()”
- In this case, you have to upload the files in “C:/2023_R”



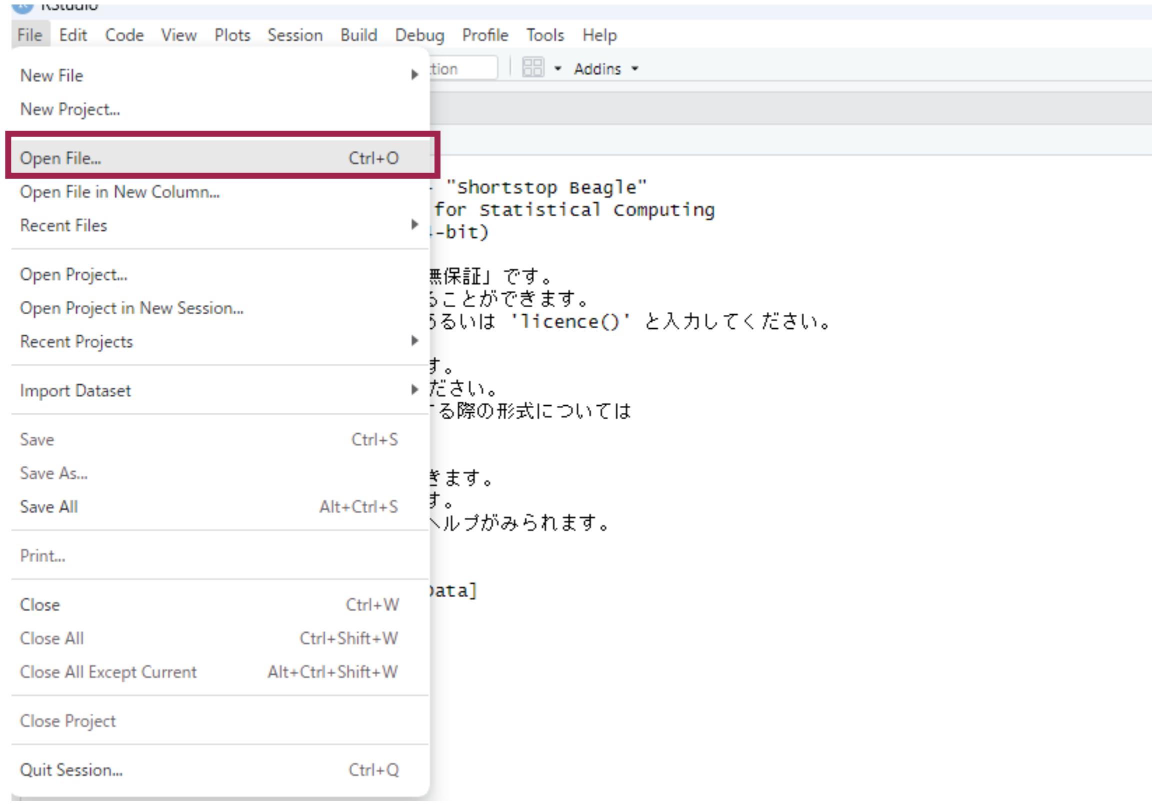
- Drag the files in the “2023_R”

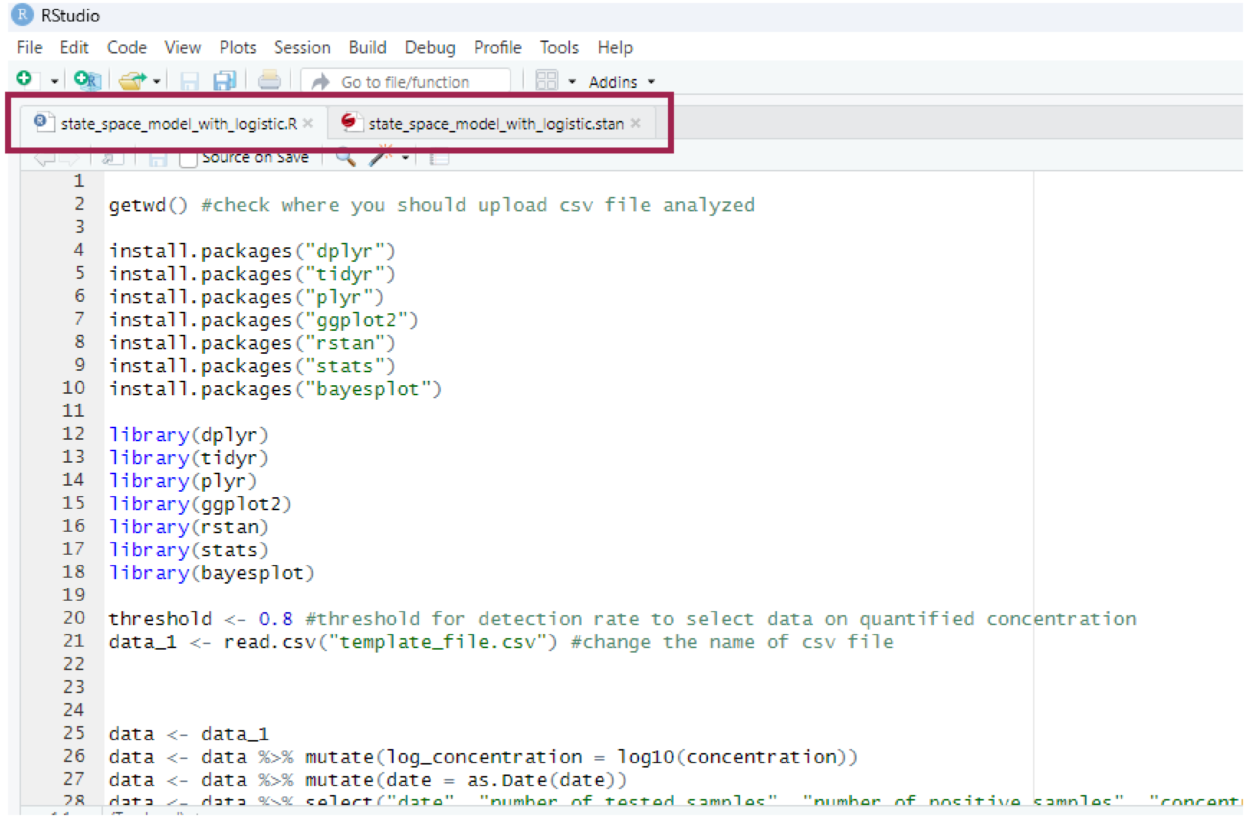
- Open R file and Stan file in Rstudio



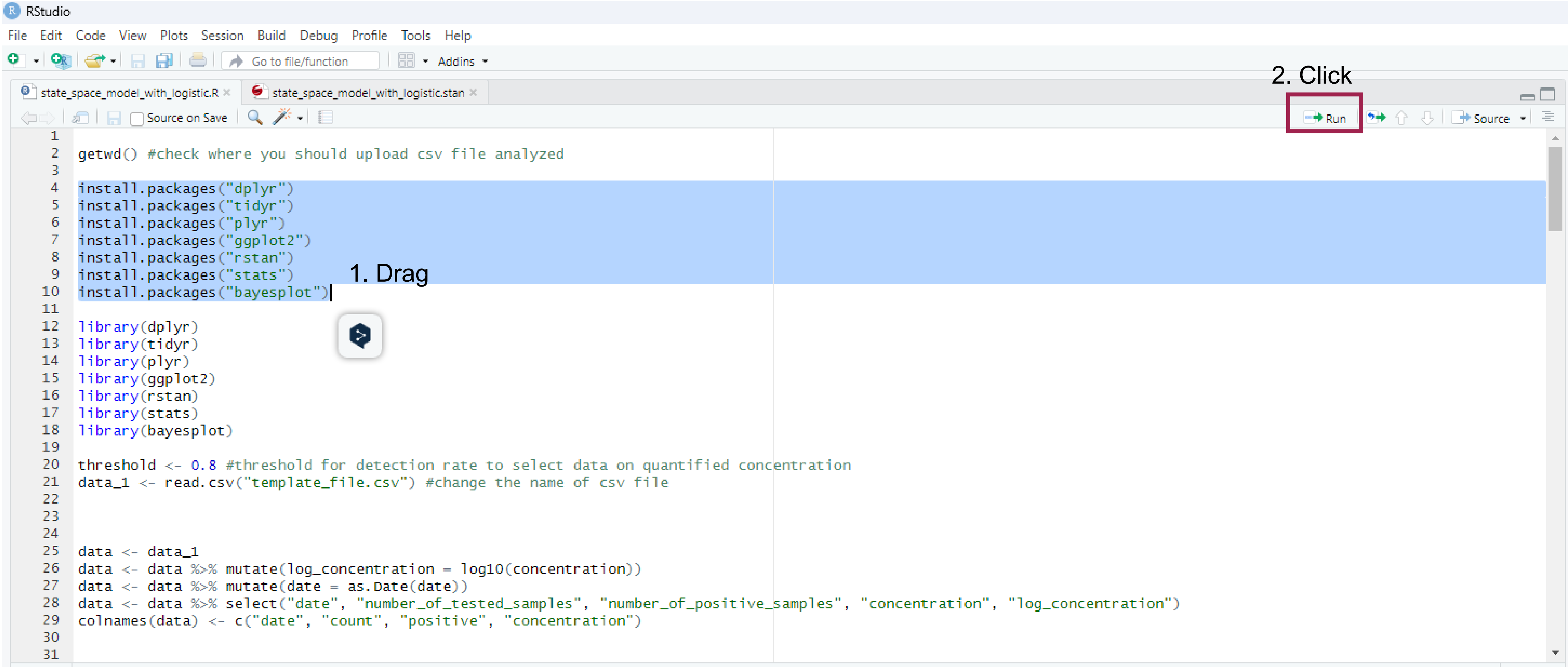


- Open R file and Stan file in Rstudio



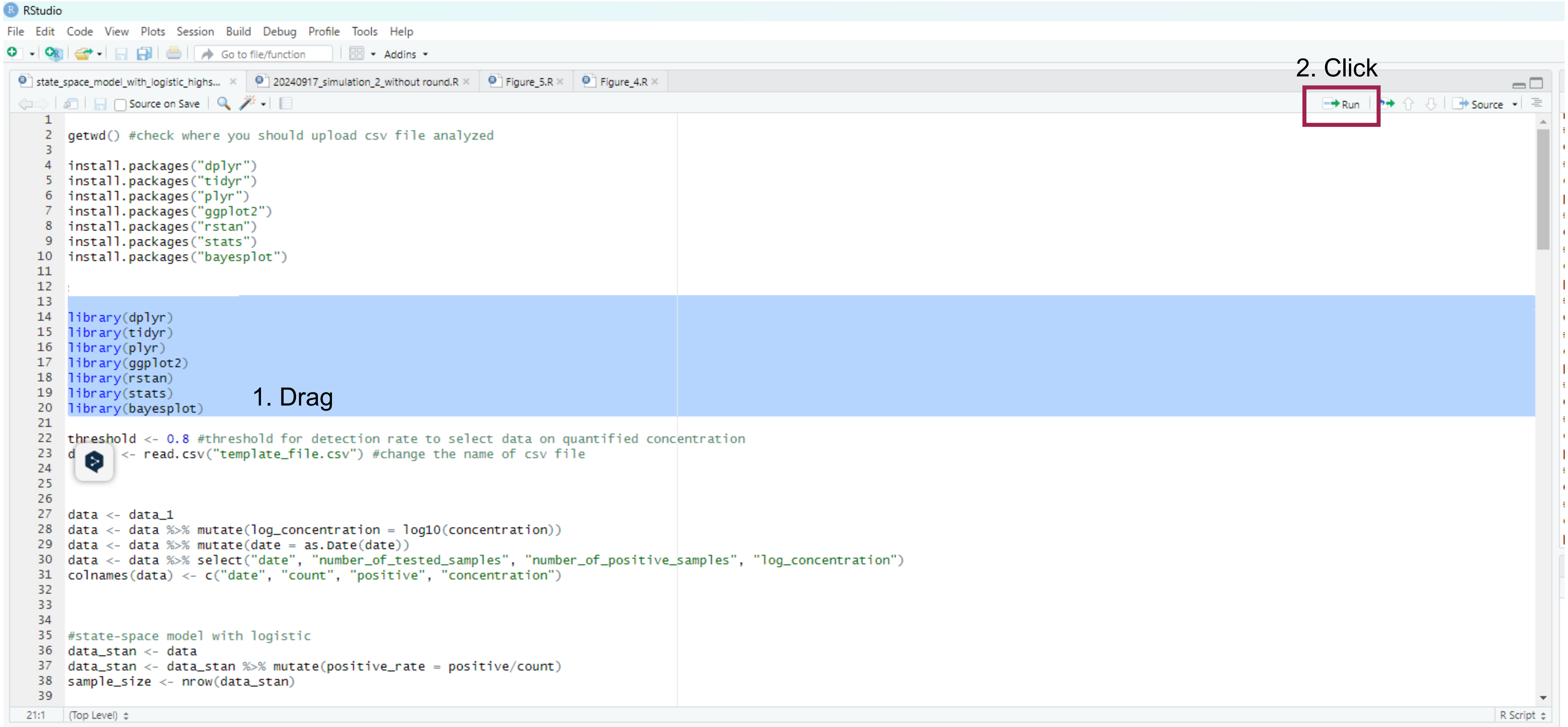


- Install R packages used in the analysis



Note: You no longer need to repeat this process once the packages are installed.

- Upload library used in the analysis



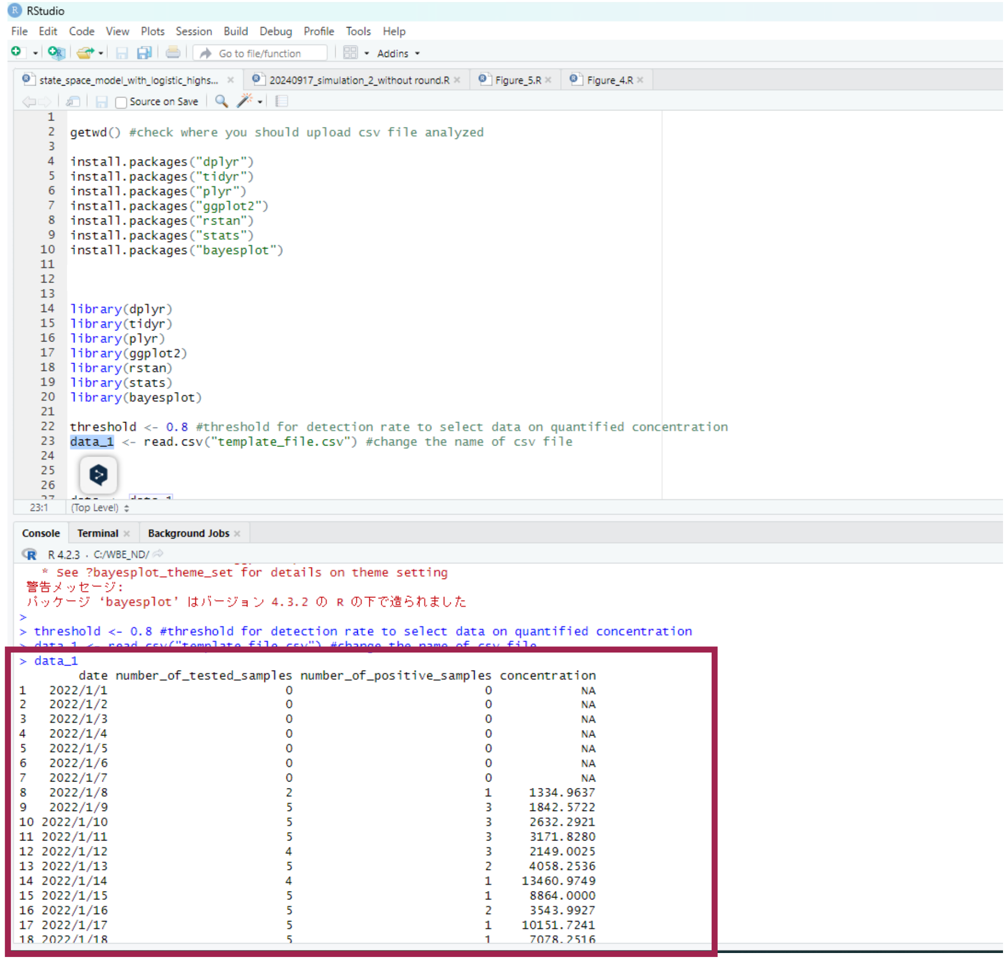
- Set arbitrary threshold (0.7 ~ 1.0)
- Import data

state_space_model_with_logistic.R | state_space_model_with_logistic.stan

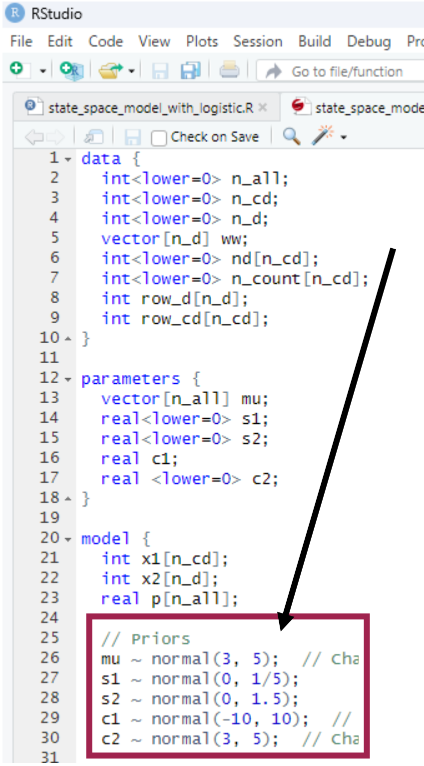
Source on Save

```r
getwd() #check where you should upload csv file analyzed

install.packages("dplyr")
install.packages("tidyr")
install.packages("plyr")
install.packages("ggplot2")
install.packages("rstan")
install.packages("stats")
install.packages("bayesplot")

library(dplyr)
library(tidyr)
library(plyr)
library(ggplot2)
library(rstan)
library(stats)
library(bayesplot)

threshold <- 0.8 #threshold for detection rate to select data on quantified concentration
data_1 <- read.csv("template_file.csv") #change the name of csv file
```

Check the name of CSV file you uploaded



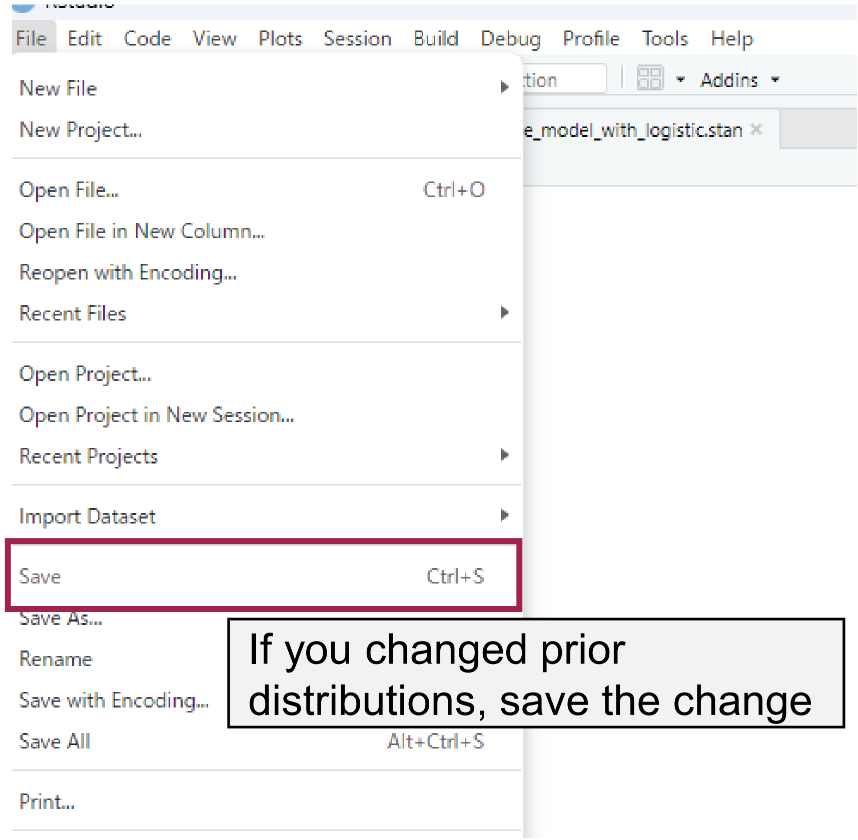
- Prior distributions are important for convergence of parameters
  (Prior distribution should be decided from available information)



mu: log-transformed wastewater concentration (log10 copies/L)

- In the case of “mu”, prior distribution was set as a normal distribution with mean of 3 and standard deviation of 5 to cover typical wastewater concentration range (0 ~ 8 log10 copies/L)

- If average of wastewater concentration of a virus is known to be around 7 log10 copies/L, mean of prior distribution should be 7.



If you changed prior distributions, save the change

- S_1 is a parameter of the state-formula.
- S_2 is a parameter of the observation formula (i.e., measurement error).
- C_1 and C_2 are parameters of logistic function.

- Drag the row from 25-87 and click Run

state_space_model_with_logistic.R × | state_space_model_with_logistic.stan ×

Source on Save | Run | Source

```r
25 data <- data_1
   data <- data %>% mutate(log_concentration = log10(concentration))
   data <- data %>% mutate(date = as.Date(date))
28 data <- data %>% select("date", "number_of_tested_samples", "number_of_positive_samples", "concentration", "log_concentration")
29 colnames(data) <- c("date", "count", "positive", "concentration")
30
31
32
33 #state-space model with logistic
34 data_stan <- data
35 data_stan <- data_stan %>% mutate(positive_rate = positive/count)
36 sample_size <- nrow(data_stan)
37
38 #vector of row number used for the analysis
39 #pick row numbers for censored data
40 data_row_D <- data.frame(true = which((data_stan$positive_rate >= threshold)))
41 sample_size_D <- nrow(data_row_D)
42 #pick row numbers for censored data
43 data_row_CD <- data.frame(true = which(data_stan$count > 0))
```

```r
60 mcmc <- stan(
61   file = "state_space_model_with_logistic.stan",
62   data = data_list_ww,
63   seed = 1,
64   chain = 4,
65   iter = 200000,
66   warmup = 100000,
67   thin = 4
68 )
```

MCMC condition can be changed according to your data.(https://mc-stan.org/rstan/reference/stan.html)

```r
70 print(mcmc, pars = c("c1", "c2", "s1", "s2"), probe = c(0.025, 0.50, 0.975))
71
72 #check traceplots if you want
73 #mcmc_combo(mcmc, pars = c("c1", "c2", "s1", "s2"))
74
75
76 #MCMC samples and 95% credible intervals
77 mcmc_sample <- rstan::extract(mcmc)
78 state_name <- "mu"
79 result <- data.frame(t(apply(
80   X = mcmc_sample[[state_name]],
81   MARGIN = 2,
82   FUN = quantile,
83   probs = c(0.025, 0.5, 0.975) #credible interval can be changed
84 )))
85
86 colnames(result) <- c("low", "median", "upr")
87 data_estimated_concentration <- cbind(data, result)
```

- Check the convergence of parameters in the state-space model

```
> print(mcmc, pars = c("c1", "c2", "s1", "s2"), probe = c(0.025, 0.50, 0.975))
Inference for Stan model: anon_model.
4 chains, each with iter=4000; warmup=2000; thin=4;
post-warmup draws per chain=500, total post-warmup draws=2000.

     mean se_mean   sd   2.5%    25%    50%    75%  97.5% n_eff Rhat
c1 -24.55    0.11 3.76 -32.08 -27.04 -24.37 -21.96 -17.60  1208 1.00
c2   8.04    0.03 1.18   5.82   7.21   8.01   8.82  10.39  1565 1.00
s1   0.04    0.00 0.01   0.02   0.03   0.04   0.05   0.07   272 1.02
s2   0.96    0.00 0.15   0.71   0.85   0.94   1.04   1.31  1805 1.00
```

- Rhat should be lower than 1.10 (https://mc-stan.org/misc/warnings.html#bulk-ess)
- N_eff should be higher than 400 (ideally, 1000)

To improve Rhat and n_eff

- Fix prior distribution
- Increase chain, iter (warmup), and thin
- Conduct re-parameterization (**Slide number 14 - 15**)

- Create figures and check the estimation of wastewater concentration

state_space_model_with_logistic.R | state_space_model_with_logistic.stan

Source on Save | Run | Source

```r
#figure
data_fig <- data_data_estimated_concentration
data_fig <- data_fig %>% mutate(posi = if_else(count < 1, "no", if_else(positive >= threshold*count, "yes", "no")))

plot <- ggplot(data_fig, aes(x = as.Date(date)))
plot <- plot + geom_point(aes(y = concentration, color = posi, shape = posi))
plot <- plot + scale_shape_manual(values = c(24, 16))
plot <- plot + scale_color_manual(values = c("#FEC44F","#0570B0"))
plot <- plot + geom_ribbon(aes(ymin = low, ymax = upr), fill = "#3690C0", alpha = 0.3)
plot <- plot + geom_line(aes(y = median), color = "#0570B0", size = 1)
plot <- plot + labs(x = "Date", y = "wastewater concentration")
plot <- plot + scale_x_date(limits = c(as.Date("2022-01-01"), as.Date("2023-06-01")), date_breaks = "2 months", date_labels = "%b") #change the date
plot <- plot + scale_y_continuous(limits = c(2.0, 5.5), breaks = seq(2.0, 5.5, 1)) #change the scale of y axis
plot <- plot + theme_bw()
plot <- plot + theme(
  axis.line = element_line(linewidth = 1.0, lineend = "square"),
  text = element_text(colour ="black", size = 14),
  legend.position = "none",
  axis.ticks = element_line(linewidth = 1.0),
  axis.ticks.length = unit(-2, "mm"))
plot
```

Change
- X axis (date)
- Y axis (log10 wastewater concentration)

https://ggplot2.tidyverse.org/reference/scale_continuous.html

- Export the result in CSV file

state_space_model_with_logistic.R* | state_space_model_with_logistic.stan

Source on Save | Run | Sour

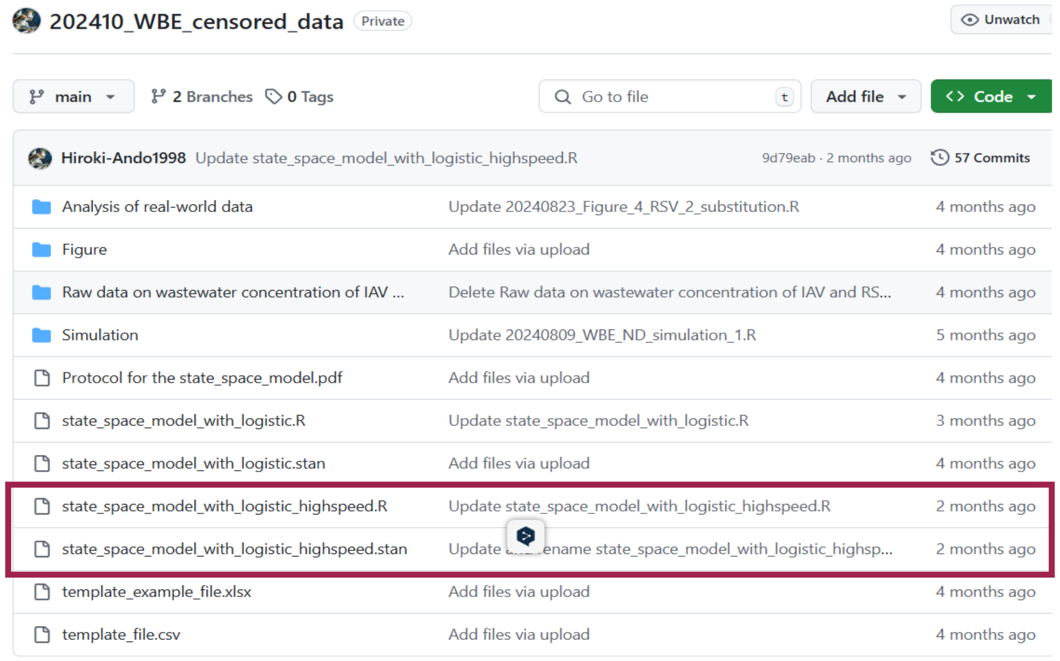
```r
  axis.line = element_line(linewidth = 1.0, lineend = "square"),
  text = element_text(colour ="black", size = 14),
  legend.position = "none",
  axis.ticks = element_line(linewidth = 1.0),
  axis.ticks.length = unit(-2, "mm"))
plot


#Export estimation result
#confirm the place where file should be loaded by using "getwd()": "C:/XXXX/"
#Write file name: "2024XXXX_XX_XX,csv"
write.csv(x = data_estimated_concentration, file = "C:/XXXX/2024XXXX_XX_XX.csv")
```

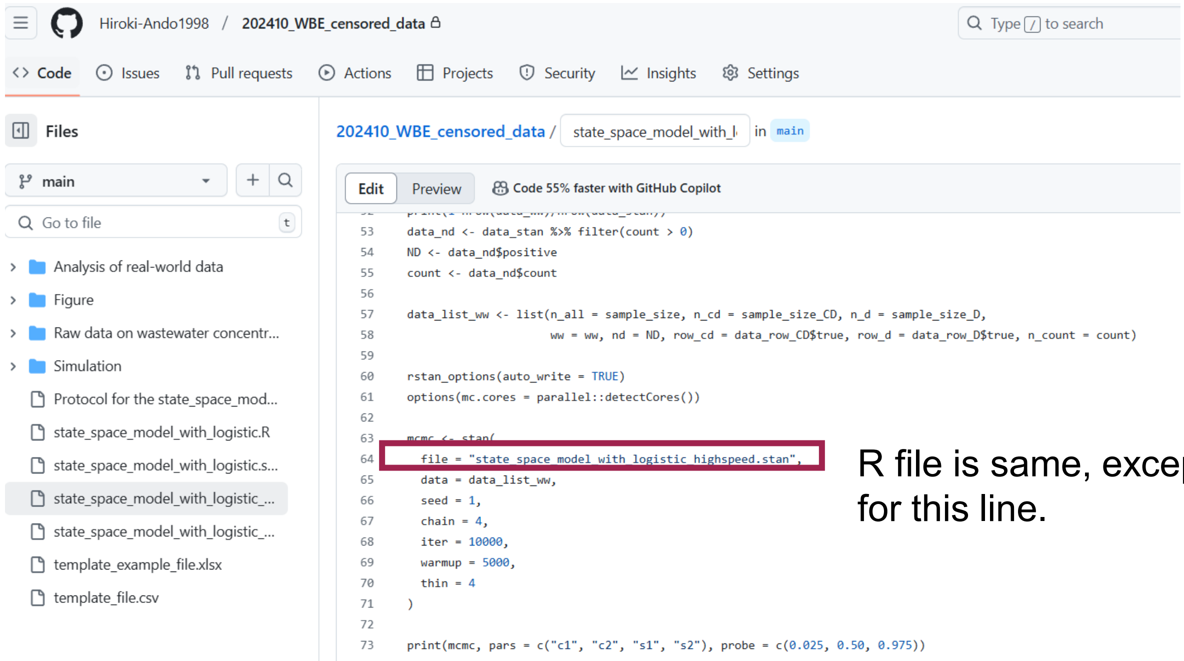
Decide the name of the CSV file to save (ex, "C:/2023_R/result_1.csv")

High speed analysis

- The above approach (file name: state_space_model_with_logistic.R & state_space_model_with_logistic.stan) requires a lot of analytical time.

- This problem is solved by using re-parameterization approach.
  - https://mc-stan.org/docs/2_18/stan-users-guide/reparameterization-section.html
  - https://mc-stan.org/docs/stan-users-guide/reparameterization.html





R file is same, except for this line.

High speed analysis

The purpose of reparameterization is to standardize the scale of the prior distributions so that the sampling space becomes narrower and more isotropic, allowing the sampler to explore the parameter space more efficiently.

state_space_model_with_logistic_highspeed.stan (Stan file)

```
parameters {
  vector<lower=-4, upper=4>[n_all] mu_raw;
  real<lower=-4, upper=4> s1_raw;
  real<lower=-4, upper=4> s2_raw;
  real<lower=-4, upper=4> c1_raw;
  real<lower=-4, upper=4> c2_raw;
}
```

Range of parameters

```
transformed parameters{
      vector[n_all] mu;
  for (i in 1:n_all) {
    mu[i] = 3.5 + 1.5 * mu_raw[i];
  }
    real s1;
    s1 = exp(s1_raw - 2);

    real s2;
    s2 = exp(s2_raw - 2);

    real c1;
     c1 = -22 + 5*c1_raw;

    real c2;
    c2 = 6 + 3*c2_raw;

}
```

Parameters are converted into reasonable scale values, which is expected from previous studies.

- In this case, typical range of wastewater concentration is expected to be 0.5 to 6.5 $\log_{10}$ copies/L.
- Typically, S_1 and S_2 are positive values, ranging from 0 to 1.
- C_1 and C_2 are probably in this range, according to a previous study.

These values can be changed, according to your dataset.
(Do not forget to save file after changing values.)

```
model {
  // Priors
  mu_raw ~ normal(0, 1);
  s1_raw ~ normal(0, 1);
  s2_raw ~ normal(0, 1);
  c1_raw ~ normal(0, 1);
  c2_raw ~ normal(0, 1);

  // Autoregressive prior for mu
    mu[3:n_all] ~ normal(2 * mu[2:n_all-1] - mu[1:n_all-2], s1);
```

The same prior distribution

- Run MCMC sampling from R file
- Check R-hat and n_eff
- Confirm estimated wastewater concentrations.