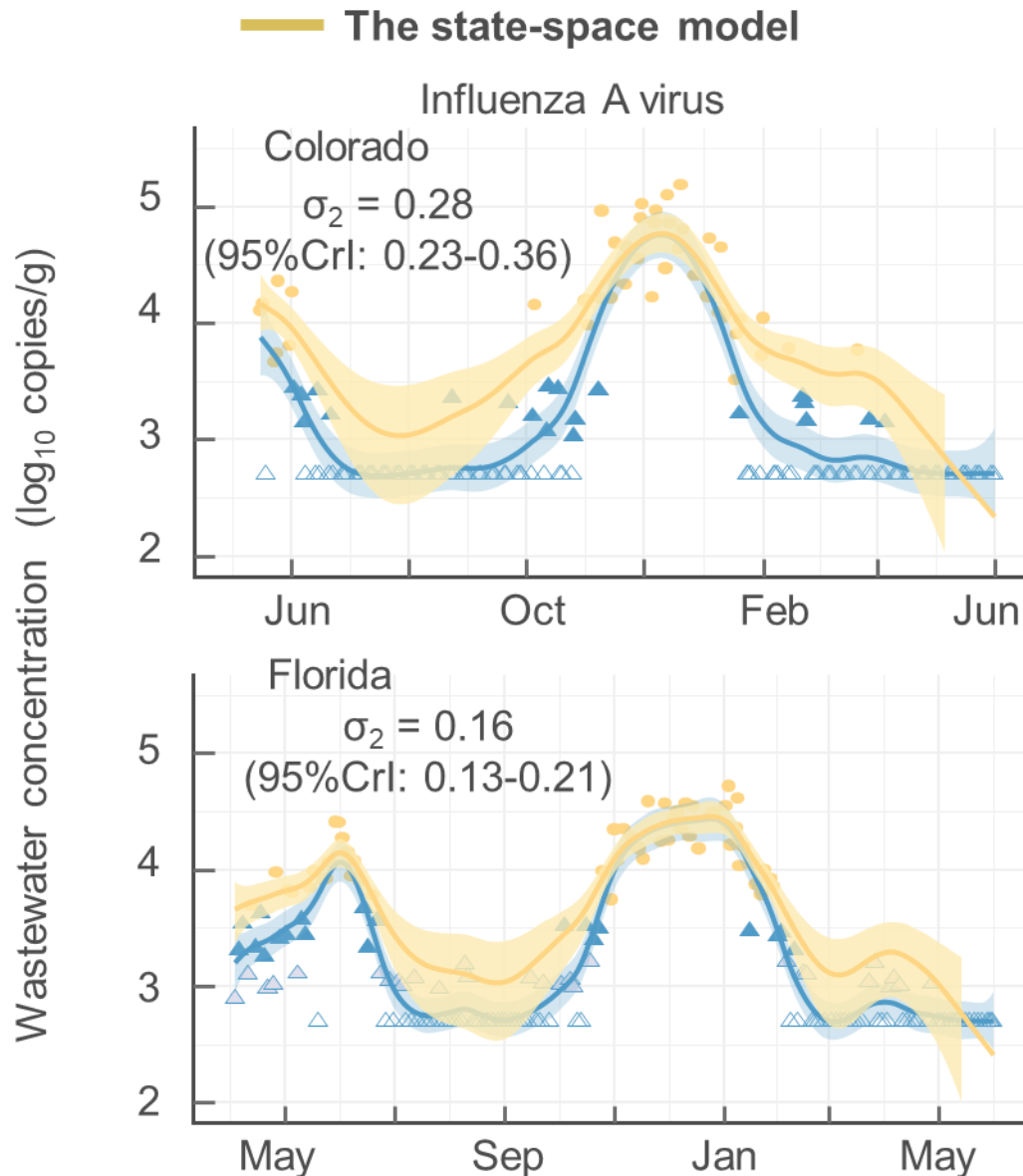


# Brief guide for using the state-space model



## 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.

The screenshot shows a GitHub repository page for '202410\_WBE\_censored\_data' by user 'Hiroki-Ando1998'. The repository is private and has 2 branches and 0 tags. The file list includes:

- Analysis of real-world data
- Figure
- Raw data on wastewater concentration of IAV ...
- Simulation
- Protocol for the state\_space\_model.pdf
- state\_space\_model\_with\_logistic.R
- state\_space\_model\_with\_logistic.stan
- state\_space\_model\_with\_logistic\_highspeed.R
- state\_space\_model\_with\_logistic\_highspeed.stan
- template\_example\_file.xlsx
- template\_file.csv

The 'Simulation' folder is highlighted with a red box. Below it, the file 'state\_space\_model\_with\_logistic.R' is shown in the viewer, also highlighted with a red box. The file content includes R code for installing packages and loading libraries. The 'Download' button is highlighted with a red box.

- 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	

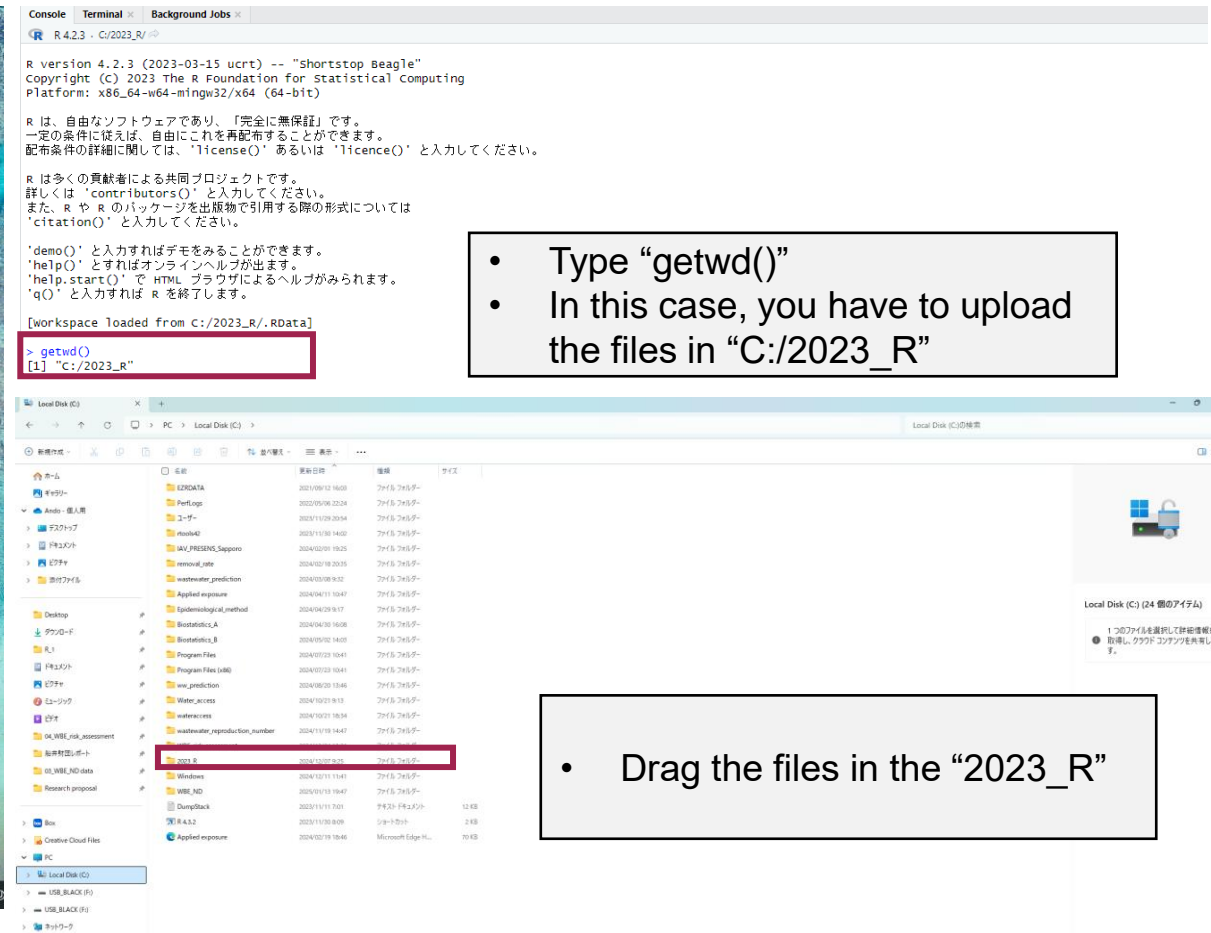
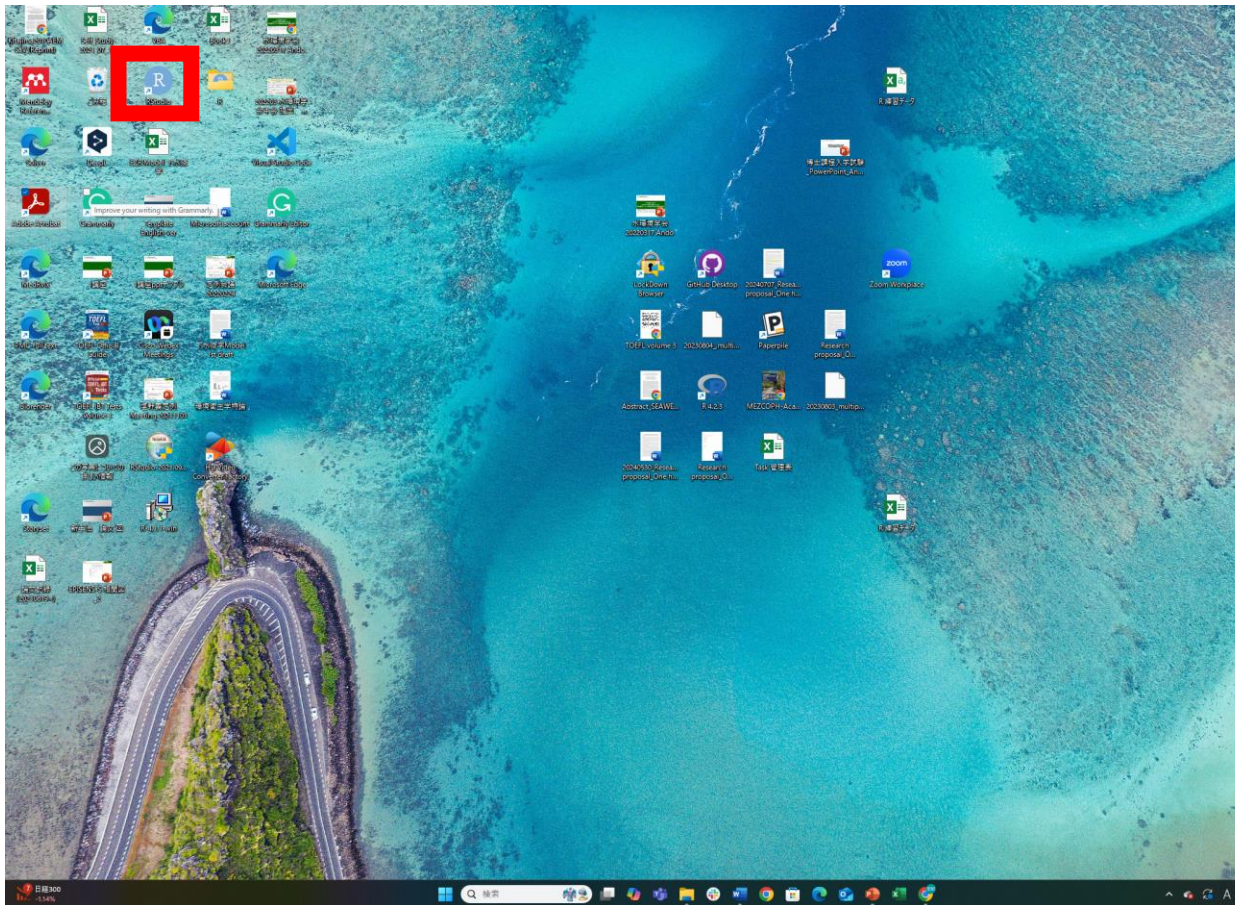
Data in non sampling day should be blank or "NA".  
(採水していない日は、空白かNAに)

↑  
Integer  
(整数)

↑  
Integer  
(整数)

↑  
Copies/L

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

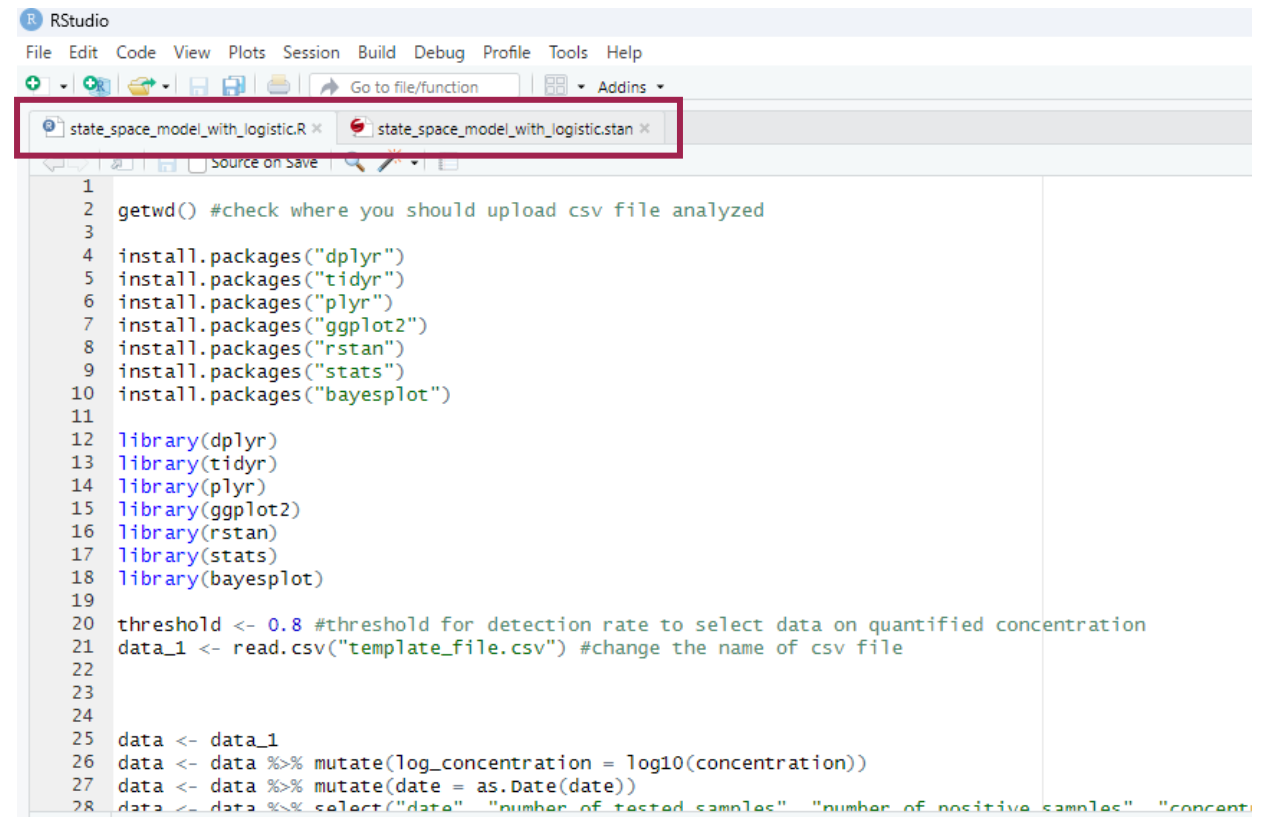
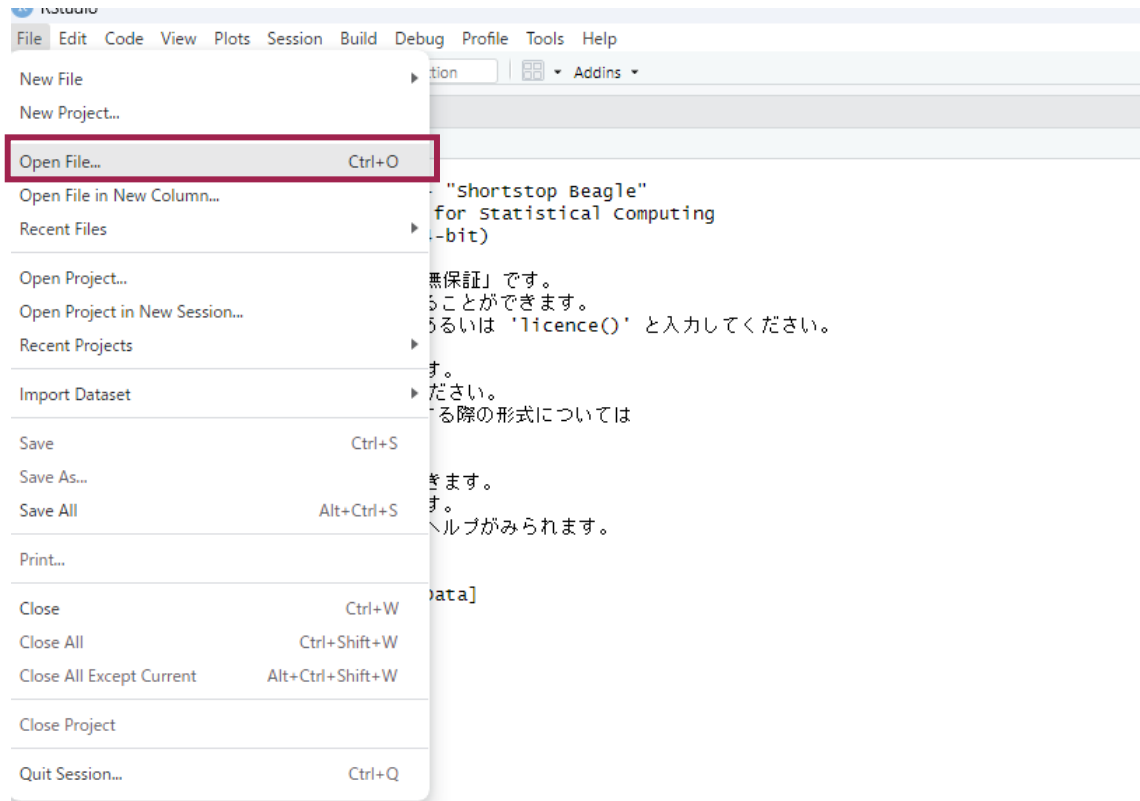


- Type “getwd()”
- In this case, you have to upload the files in “C:/2023 R”

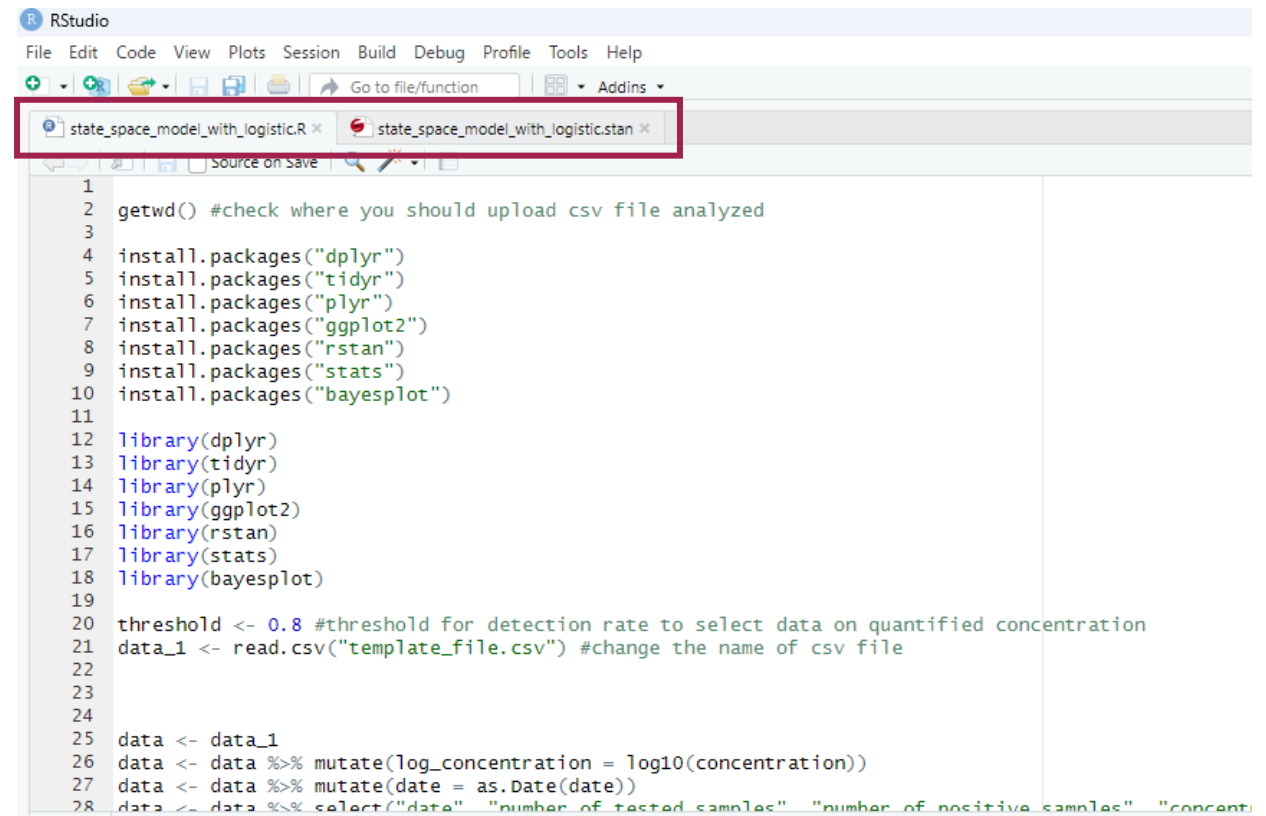
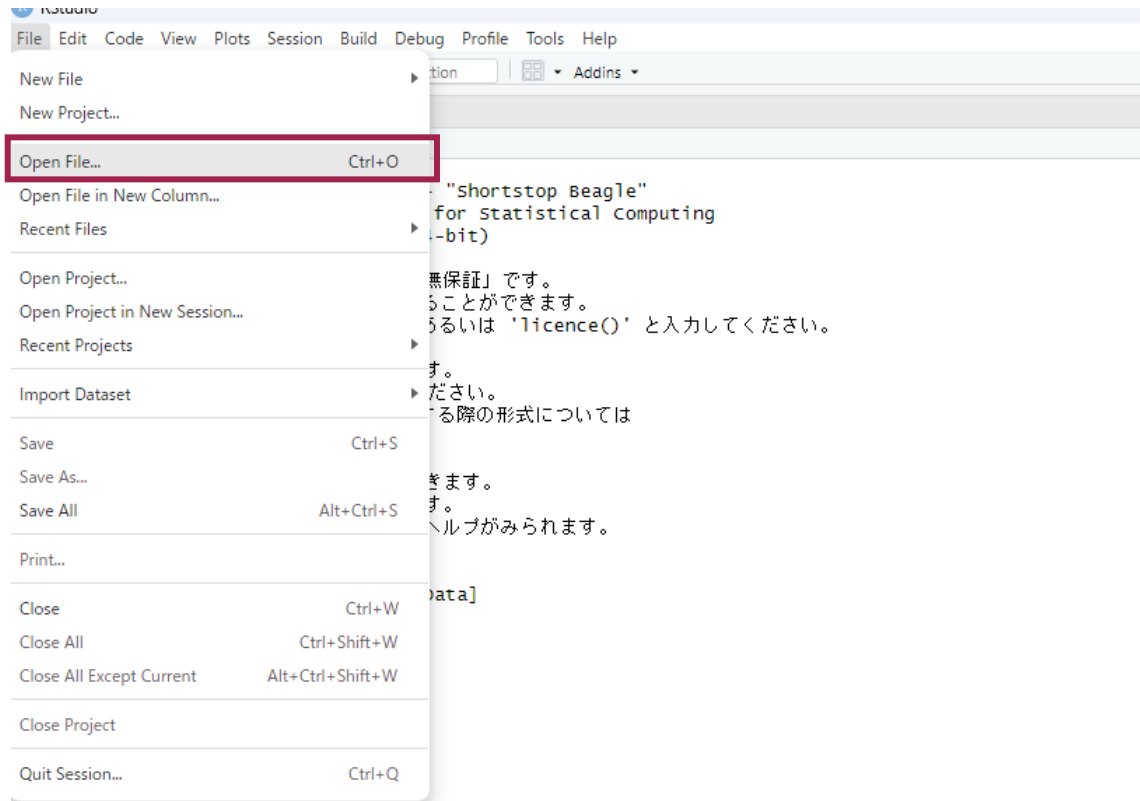
- Drag the files in the “2023\_R”



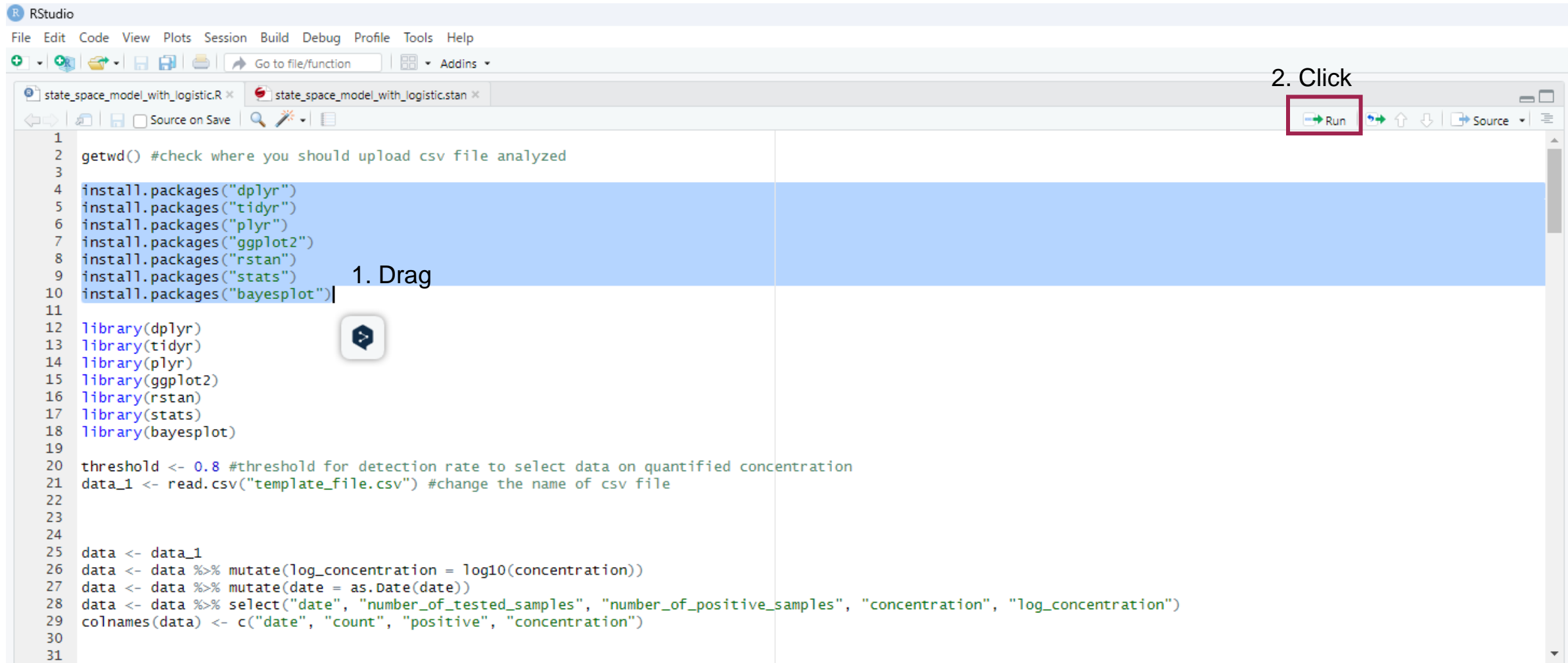
- Open R file and Stan file in Rstudio



- Open R file and Stan file in Rstudio



- Install R packages used in the analysis



RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins

state\_space\_model\_with\_logistic.R x state\_space\_model\_with\_logistic.stan x

Source on Save

1. Drag

```
1 getwd() #check where you should upload csv file analyzed
2
3
4 install.packages("dplyr")
5 install.packages("tidyr")
6 install.packages("plyr")
7 install.packages("ggplot2")
8 install.packages("rstan")
9 install.packages("stats")
10 install.packages("bayesplot")
11
12 library(dplyr)
13 library(tidyr)
14 library(plyr)
15 library(ggplot2)
16 library(rstan)
17 library(stats)
18 library(bayesplot)
19
20 threshold <- 0.8 #threshold for detection rate to select data on quantified concentration
21 data_1 <- read.csv("template_file.csv") #change the name of csv file
22
23
24
25 data <- data_1
26 data <- data %>% mutate(log_concentration = log10(concentration))
27 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
```

2. Click

Run

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

- Upload library used in the analysis

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins

state\_space\_model\_with\_logistic\_highs... 20240917\_simulation\_2\_without\_round.R Figure\_5.R Figure\_4.R

Source on Save

1. Drag

```
1  
2 getwd() #check where you should upload csv file analyzed  
3  
4 install.packages("dplyr")  
5 install.packages("tidyr")  
6 install.packages("plyr")  
7 install.packages("ggplot2")  
8 install.packages("rstan")  
9 install.packages("stats")  
10 install.packages("bayesplot")  
11  
12  
13  
14 library(dplyr)  
15 library(tidyr)  
16 library(plyr)  
17 library(ggplot2)  
18 library(rstan)  
19 library(stats)  
20 library(bayesplot)  
21  
22 threshold <- 0.8 #threshold for detection rate to select data on quantified concentration  
23 d <- read.csv("template_file.csv") #change the name of csv file  
24  
25  
26  
27 data <- data_1  
28 data <- data %>% mutate(log_concentration = log10(concentration))  
29 data <- data %>% mutate(date = as.Date(date))  
30 data <- data %>% select("date", "number_of_tested_samples", "number_of_positive_samples", "log_concentration")  
31 colnames(data) <- c("date", "count", "positive", "concentration")  
32  
33  
34  
35 #state-space model with logistic  
36 data_stan <- data  
37 data_stan <- data_stan %>% mutate(positive_rate = positive/count)  
38 sample_size <- nrow(data_stan)  
39
```

2. Click

Run

21:1 (Top Level) R Script



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

```

1 getwd() #check where you should upload csv file analyzed
2
3
4 install.packages("dplyr")
5 install.packages("tidyr")
6 install.packages("plyr")
7 install.packages("ggplot2")
8 install.packages("rstan")
9 install.packages("stats")
10 install.packages("bayesplot")
11
12 library(dplyr)
13 library(tidyr)
14 library(plyr)
15 library(ggplot2)
16 library(rstan)
17 library(stats)
18 library(bayesplot)
19
20 threshold <- 0.8 #threshold for detection rate to select data on quantified concentration
21 data_1 <- read.csv("template_file.csv") #change the name of csv file

```



Check the name of CSV file you uploaded

```

1 getwd() #check where you should upload csv file analyzed
2
3
4 install.packages("dplyr")
5 install.packages("tidyr")
6 install.packages("plyr")
7 install.packages("ggplot2")
8 install.packages("rstan")
9 install.packages("stats")
10 install.packages("bayesplot")
11
12 library(dplyr)
13 library(tidyr)
14 library(plyr)
15 library(ggplot2)
16 library(rstan)
17 library(stats)
18 library(bayesplot)
19
20 threshold <- 0.8 #threshold for detection rate to select data on quantified concentration
21 data_1 <- read.csv("template_file.csv") #change the name of csv file

```

Console

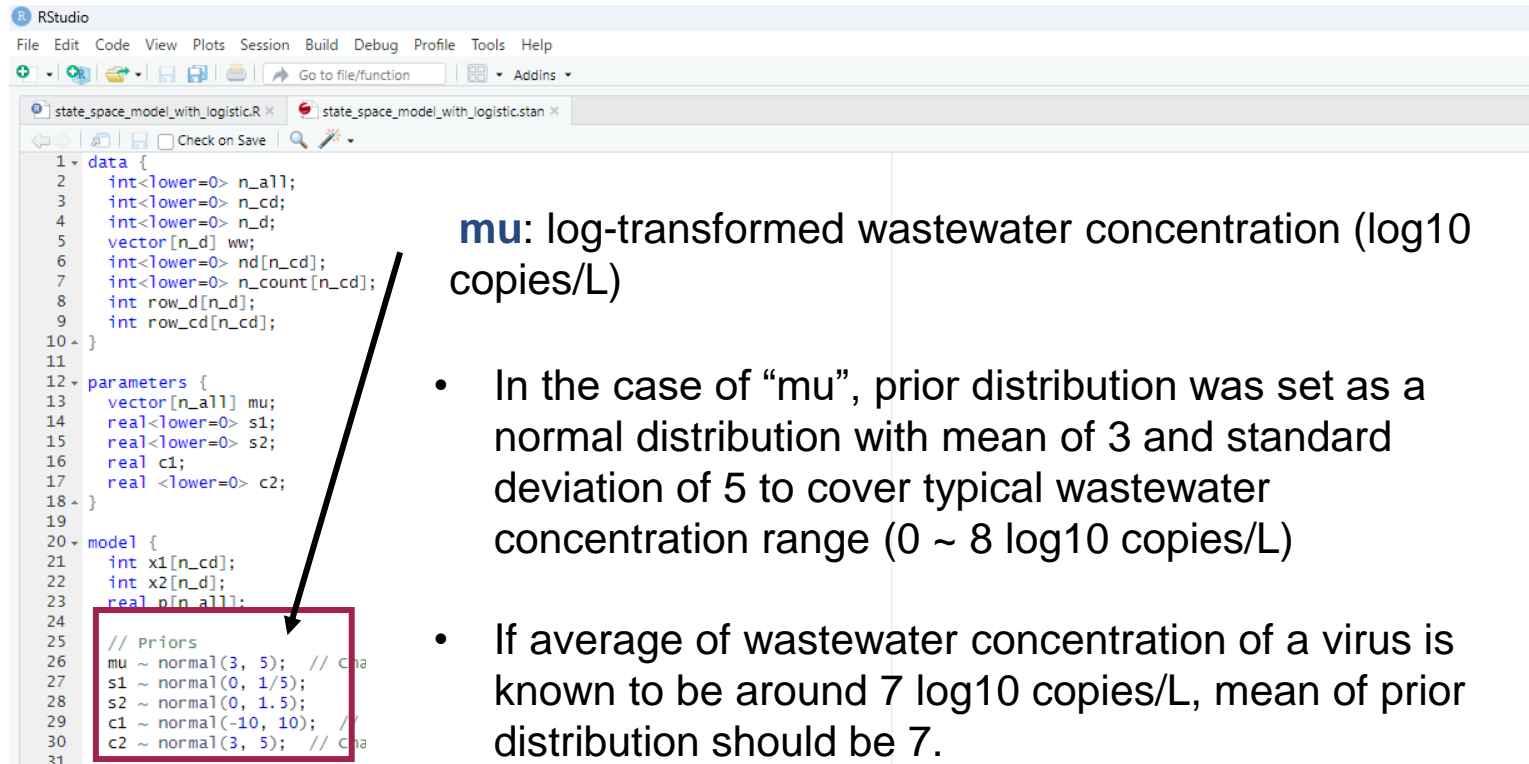
```

R 4.2.3 - C:/WBE/ND/
> See ?bayesplot_theme_set for details on theme setting
警告メッセージ:
パッケージ 'bayesplot' はバージョン 4.3.2 の R の下で壊れました
> 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
> data_1

```

	date	number_of_tested_samples	number_of_positive_samples	concentration
1	2022/1/1	0	0	NA
2	2022/1/2	0	0	NA
3	2022/1/3	0	0	NA
4	2022/1/4	0	0	NA
5	2022/1/5	0	0	NA
6	2022/1/6	0	0	NA
7	2022/1/7	0	0	NA
8	2022/1/8	2	1	1334.9637
9	2022/1/9	3	3	1842.5722
10	2022/1/10	5	3	2632.2921
11	2022/1/11	5	3	3171.8280
12	2022/1/12	4	3	2149.0025
13	2022/1/13	5	2	4058.2536
14	2022/1/14	4	1	13460.9749
15	2022/1/15	5	1	8864.0000
16	2022/1/16	5	2	3543.9927
17	2022/1/17	5	1	10151.7241
18	2022/1/18	5	1	7078.2516

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



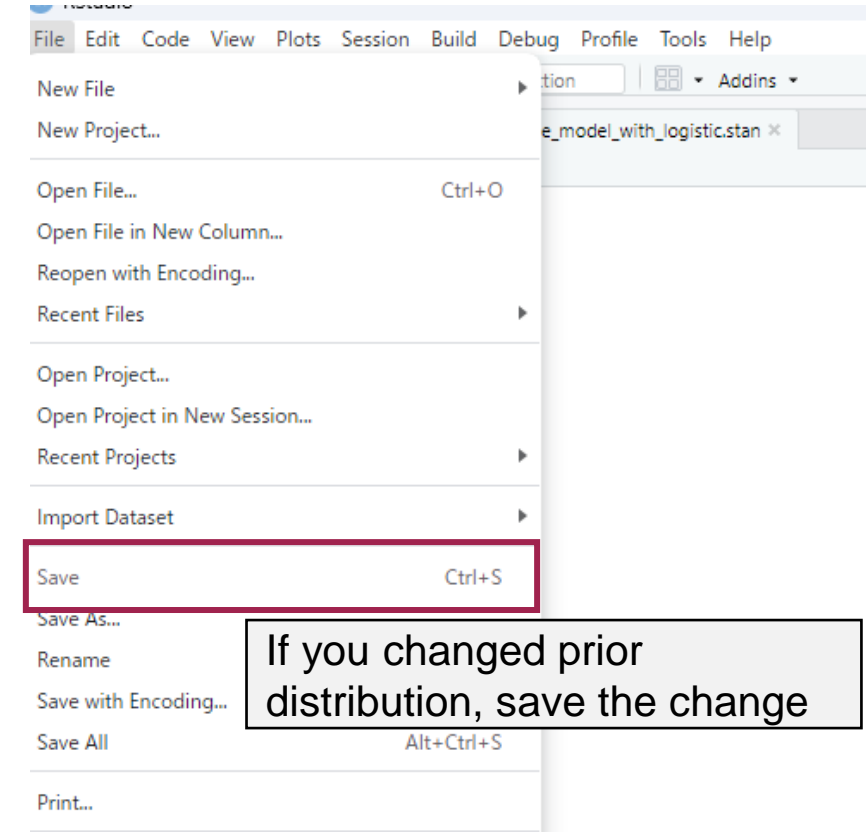
```

1 data {
2   int<lower=0> n_all;
3   int<lower=0> n_cd;
4   int<lower=0> n_d;
5   vector[n_d] ww;
6   int<lower=0> nd[n_cd];
7   int<lower=0> n_count[n_cd];
8   int row_d[n_d];
9   int row_cd[n_cd];
10 }
11
12 parameters {
13   vector[n_all] mu;
14   real<lower=0> s1;
15   real<lower=0> s2;
16   real c1;
17   real<lower=0> c2;
18 }
19
20 model {
21   int x1[n_cd];
22   int x2[n_d];
23   real p[n_all];
24
25   // Priors
26   mu ~ normal(3, 5); // C1
27   s1 ~ normal(0, 1/5);
28   s2 ~ normal(0, 1.5);
29   c1 ~ normal(-10, 10); // C2
30   c2 ~ normal(3, 5); // C1
31 }

```

**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.



- $S_1$  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 x state_space_model_with_logistic.stan x
Source on Save
25 data <- data_1
26 data <- data %>% mutate(log_concentration = log10(concentration))
27 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))
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
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 )
69
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)

```

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

- 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.
```

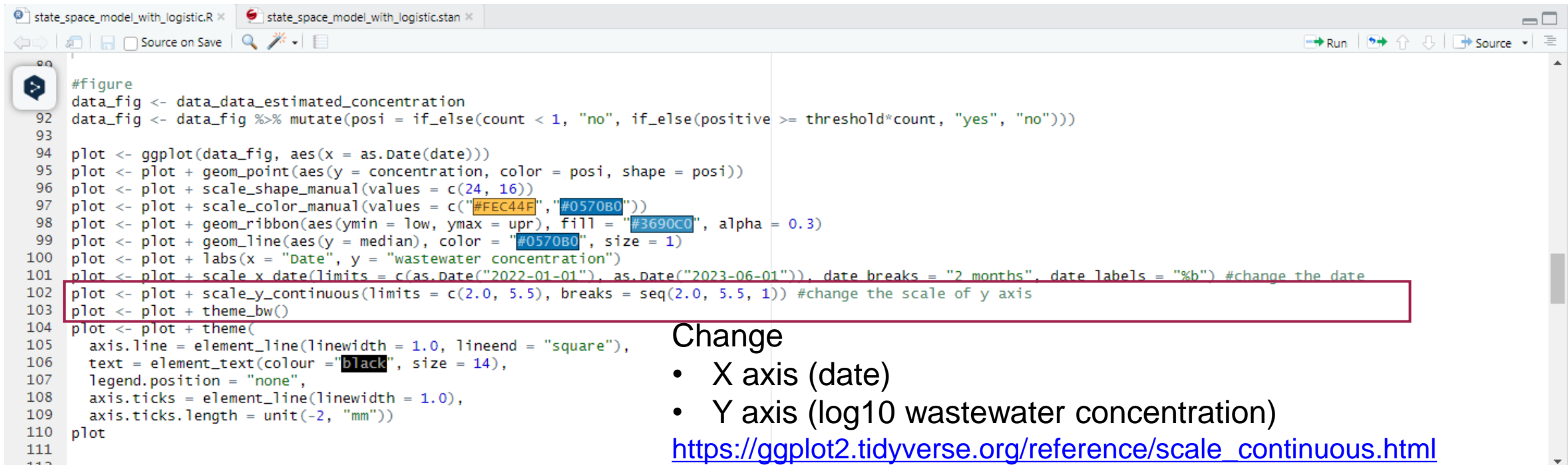
	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
c1	-24.55	0.11	3.76	-32.08	-27.04	-24.37	-21.96	-17.60	1208	1.00
c2	8.04	0.03	1.18	5.82	7.21	8.01	8.82	10.39	1565	1.00
s1	0.04	0.00	0.01	0.02	0.03	0.04	0.05	0.07	272	1.02
s2	0.96	0.00	0.15	0.71	0.85	0.94	1.04	1.31	1805	1.00

- 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**)

- Check the estimation of wastewater concentration



```

#figure
data_fig <- data_data_estimated_concentration
92 data_fig <- data_fig %>% mutate(posi = if_else(count < 1, "no", if_else(positive >= threshold*count, "yes", "no")))
93
94 plot <- ggplot(data_fig, aes(x = as.Date(date)))
95 plot <- plot + geom_point(aes(y = concentration, color = posi, shape = posi))
96 plot <- plot + scale_shape_manual(values = c(24, 16))
97 plot <- plot + scale_color_manual(values = c("#FEC44F", "#0570B0"))
98 plot <- plot + geom_ribbon(aes(ymin = low, ymax = upr), fill = "#3690C0", alpha = 0.3)
99 plot <- plot + geom_line(aes(y = median), color = "#0570B0", size = 1)
100 plot <- plot + labs(x = "Date", y = "wastewater concentration")
101 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
102 plot <- plot + scale_y_continuous(limits = c(2.0, 5.5), breaks = seq(2.0, 5.5, 1)) #change the scale of y axis
103 plot <- plot + theme_bw()
104 plot <- plot + theme(
105   axis.line = element_line(linewidth = 1.0, lineend = "square"),
106   text = element_text(colour = "black", size = 14),
107   legend.position = "none",
108   axis.ticks = element_line(linewidth = 1.0),
109   axis.ticks.length = unit(-2, "mm"))
110 plot
111
112

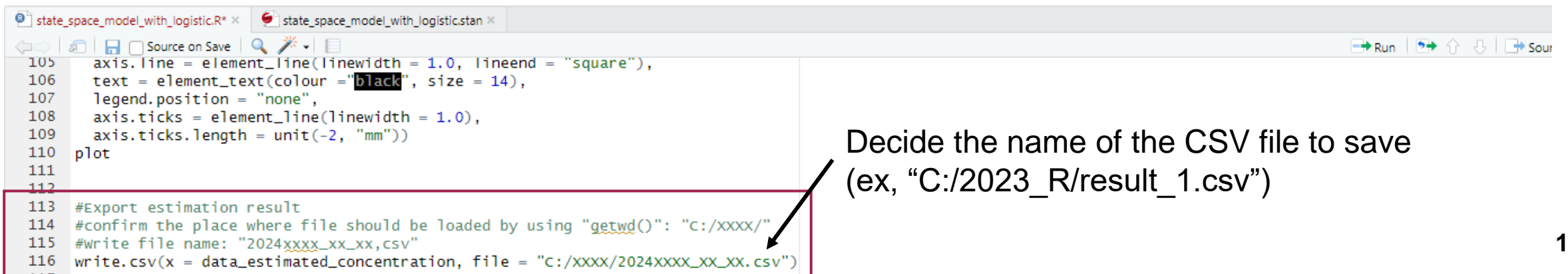
```

Change

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

[https://ggplot2.tidyverse.org/reference/scale\\_continuous.html](https://ggplot2.tidyverse.org/reference/scale_continuous.html)

- Export the result in CSV file



```

105 axis.line = element_line(linewidth = 1.0, lineend = "square"),
106 text = element_text(colour = "black", size = 14),
107 legend.position = "none",
108 axis.ticks = element_line(linewidth = 1.0),
109 axis.ticks.length = unit(-2, "mm"))
110 plot
111
112
113 #Export estimation result
114 #confirm the place where file should be loaded by using "getwd()": "c:/xxxx/"
115 #write file name: "2024xxxx_xx_xx.csv"
116 write.csv(x = data_estimated_concentration, file = "c:/xxxx/2024xxxx_xx_xx.csv")
117

```

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) require a lot of analytical time.  
(上の方法では、解析時間が長いのが難点)
- This problem is solved by using re-parameterization approach.  
(再パラメータ化というモデリング技術を使って、解析時間を短縮する)
  - [https://mc-stan.org/docs/2\\_18/stan-users-guide/reparameterization-section.html](https://mc-stan.org/docs/2_18/stan-users-guide/reparameterization-section.html)
  - <https://mc-stan.org/docs/stan-users-guide/reparameterization.html>

202410\_WBE\_censored\_data Private Unwatch

main 2 Branches 0 Tags

Go to file t Add file <> Code

Hiroki-Ando1998 Update state\_space\_model\_with\_logistic\_highspeed.R 9d79eab · 2 months ago 57 Commits

Analysis of real-world data	Update 20240823_Figure_4_RSV_2_substitution.R	4 months ago
Figure	Add files via upload	4 months ago
Raw data on wastewater concentration of IAV ...	Delete Raw data on wastewater concentration of IAV and RS...	4 months ago
Simulation	Update 20240809_WBE_ND_simulation_1.R	5 months ago
Protocol for the state_space_model.pdf	Add files via upload	4 months ago
state_space_model_with_logistic.R	Update state_space_model_with_logistic.R	3 months ago
state_space_model_with_logistic.stan	Add files via upload	4 months ago
state_space_model_with_logistic_highspeed.R	Update state_space_model_with_logistic_highspeed.R	2 months ago
state_space_model_with_logistic_highspeed.stan	Update state_space_model_with_logistic_highsp...	2 months ago
template_example_file.xlsx	Add files via upload	4 months ago
template_file.csv	Add files via upload	4 months ago

Hiroki-Ando1998 / 202410\_WBE\_censored\_data

Code Issues Pull requests Actions Projects Security Insights Settings

Files

main + Q

Go to file t

202410\_WBE\_censored\_data / state\_space\_model\_with\_l in main

Edit Preview Code 55% faster with GitHub Copilot

```

53 data_nd <- data_stan %>% filter(count > 0)
54 ND <- data_nd$positive
55 count <- data_nd$count
56
57 data_list_vw <- list(n_all = sample_size, n_cd = sample_size_CD, n_d = sample_size_D,
58                    vw = vw, nd = ND, row_cd = data_row_CD$true, row_d = data_row_D$true, n_count = count)
59
60 rstan_options(auto_write = TRUE)
61 options(mc.cores = parallel::detectCores())
62
63 mcmc <- stan(
64   file = "state_space_model_with_logistic_highspeed.stan",
65   data = data_list_vw,
66   seed = 1,
67   chain = 4,
68   iter = 10000,
69   warmup = 5000,
70   thin = 4
71 )
72
73 print(mcmc, pars = c("c1", "c2", "s1", "s2"), probe = c(0.025, 0.50, 0.975))

```

R file is almost same, except for this line.

## High speed analysis

- Re-parameterization enable to standardize prior distributions of parameters  
(再パラメータ化では、全てのパラメータの事前分布のスケールを同一にする)

## state\_space\_model\_with\_logistic\_highspeed.stan (Stan file)

```

12 parameters {
13   vector<lower=-4, upper=4>[n_all] mu_raw;
14   real<lower=-4, upper=4> s1_raw;
15   real<lower=-4, upper=4> s2_raw;
16   real<lower=-4, upper=4> c1_raw;
17   real<lower=-4, upper=4> c2_raw;
18 }
19

```

Range of parameters

```

20 transformed parameters {
21   vector[n_all] mu;
22   for (i in 1:n_all) {
23     mu[i] = 3.5 + 1.5 * mu_raw[i];
24   }
25   real s1;
26   s1 = exp(s1_raw - 2);
27   real s2;
28   s2 = exp(s2_raw - 2);
29   real c1;
30   c1 = -22 + 5*c1_raw;
31   real c2;
32   c2 = 6 + 3*c2_raw;
33 }
34

```

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<sub>10</sub> 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.)

```

40 model {
41   // Priors
42   mu_raw ~ normal(0, 1);
43   s1_raw ~ normal(0, 1);
44   s2_raw ~ normal(0, 1);
45   c1_raw ~ normal(0, 1);
46   c2_raw ~ normal(0, 1);
47
48   // Autoregressive prior for mu
49   mu[3:n_all] ~ normal(2 * mu[2:n_all-1] - mu[1:n_all-2], s1);
50

```

The same prior distribution



- Run MCMC sampling from R file
- Check R-hat and n\_eff
- Confirm wastewater concentration