

# Final Summary

## COVID-19 Confounding Effects and Normalisation Research

Toby Hayward      Thomas Lumley

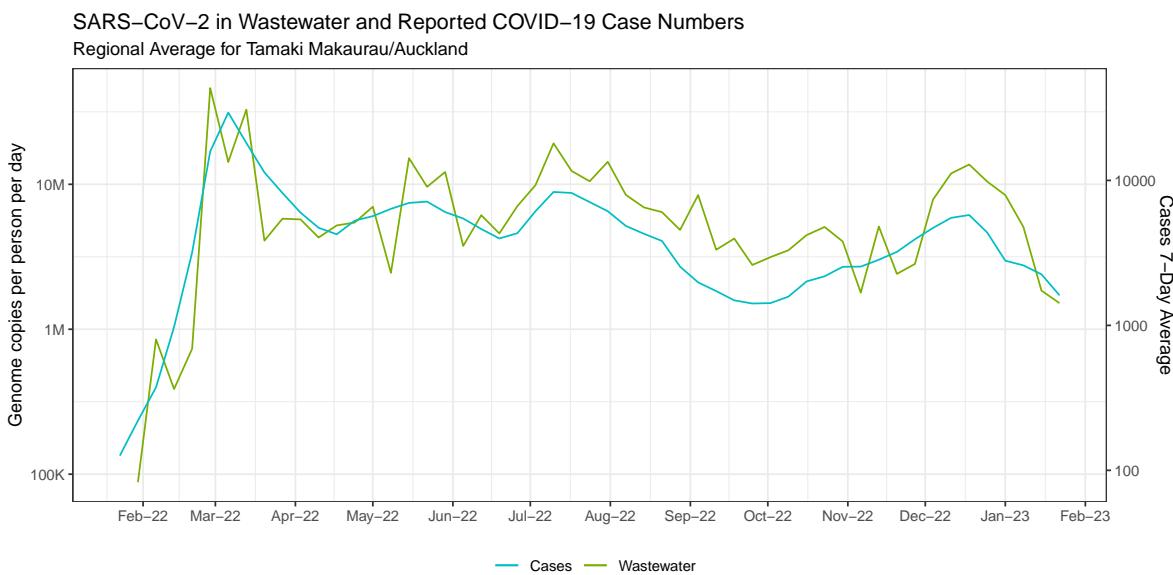
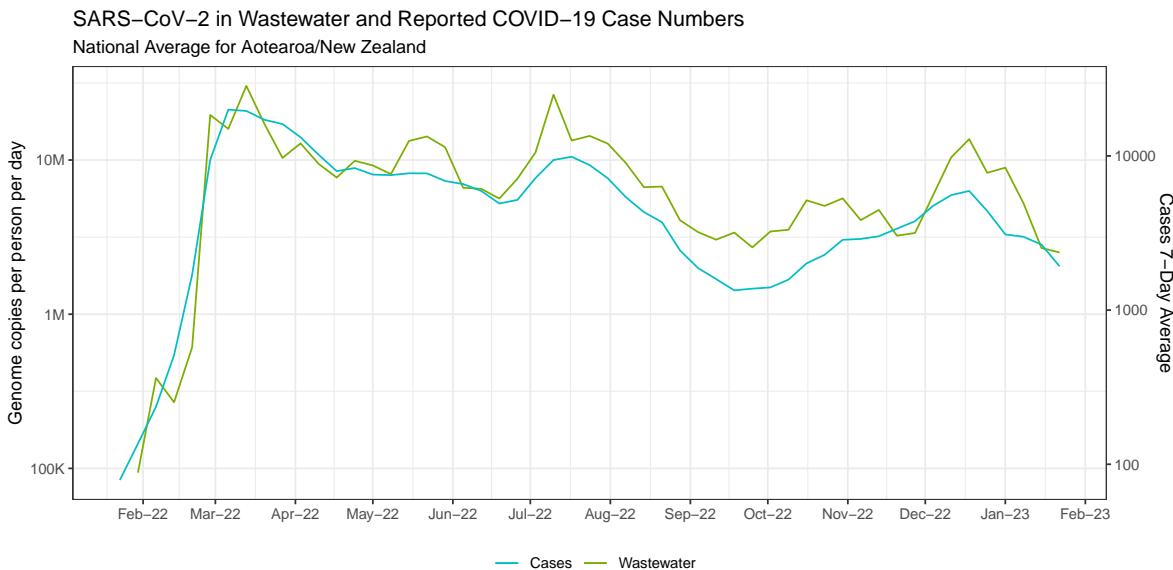
1/29/23

### Background

As self-reported COVID-19 cases are becoming increasingly less useful to measure the prevalence of **COVID-19** around Aotearoa/New Zealand, scientists and the government has an increased interest in using *Wastewater* data in place of case count. The benefits to this are a simple steady supply of prevalence data that isn't dependent on the population's ability to diagnose themselves, and the ability to detect an emergence of an outbreak the moment people start releasing SARS-CoV2 genome copies into their waste; which could be days before they even suppose COVID-19 symptoms.

And whilst Wastewater data appears somewhat accurate at estimating, there is often times a large amount of variance between what the Wastewater is estimating and the actual case average; especially as we narrow down by the location.

Hence arises the problem this research aimed to tackle: **What confounding effects for Wastewater sampling exist and how can we account for them?**



## Areas of Investigation

In the research that has been completed so far, we investigated three potential correlations:

1. Rain
2. Flow
3. Population Variance

Using linear modelling techniques, we aim to identify significance between these variables and the concentration of SARS-CoV2 genome copies.

We're interested in measuring this potential correlation in all areas that we have both Wastewater data, and Rain/Flow/Population data for.

## Rain

Although an ideal assumption is that Rainwater is processed separately to Wastewater, **it is often not reliably the case**.

It is also a simple first choice to investigate, as rain data can be easily pulled from *Cliflo's (NIWA) own database* and requires no permission to access other than a free Cliflo account (which entitles you to pull up to 2 million rows of data).

Assuming that some amount of rain is leaked into the wastewater system every day, we can estimate the following:

If we let:

- $N(t)$  = # SARS genomes in Wastewater at time  $t$  (count).
- $V(t)$  = Wastewater volume from waste sources, excluding rain (litres).
- $\hat{V}$  = Estimated Total Wastewater flow. (Not dependant on time)
- $R(t)$  = Additional volume of water from Rain runoff (litres).
- $C(t)$  = Concentration of SARS genomes per litre Wastewater.

Then:

$$\begin{aligned} C(t) &= \frac{N(t)}{V(t) + R(t)} \\ \implies \hat{N}(t) &= C(t) \times \hat{V} = \frac{N(t) \times \hat{V}}{V(t) + R(t)} \\ \implies \log(\hat{N}(t)) &= \log(N(t)) + \log\left(\frac{\hat{V}}{V(t) + R(t)}\right) \approx \log(N(t)) + \log\left(\frac{1}{1 + \frac{R(t)}{V(t)}}\right) \\ \text{for } \hat{V} &= \overline{V(t) + R(t)} \approx V(t) \end{aligned}$$

$$\begin{aligned} \text{where } \log\left(\frac{1}{1 + \frac{R(t)}{V(t)}}\right) &= -\log\left(1 + \frac{R(t)}{V(t)}\right) \approx -\frac{R(t)}{V(t)} \\ \implies \log(\hat{N}(t)) &\approx \log(N(t)) - \frac{R(t)}{V(t)} \end{aligned}$$

So assuming that rain confounds over the course of up to a week, we can manipulate the rain data into a series of *lagged days* as separate variables, fit a *smoother* to account for temporal variation (outbreaks and such) and estimate the effect that rain has on the observed concentration of SARS gcs.

$$C_t \sim \text{Quasipoisson}(\lambda_t, k)$$

$$\log(\lambda_t) = s(\text{Date}_t) + \beta_1 R_t + \beta_0$$

The number of *knots* in the smoother we choose to be equal to the number of months that we have data for (which was usually about 12). The smoother wrapped around the date should account plenty enough for the variation of outbreaks over time so that we can get a rough estimate of how much rainfall impacts the concentration.

### Locations of Interest

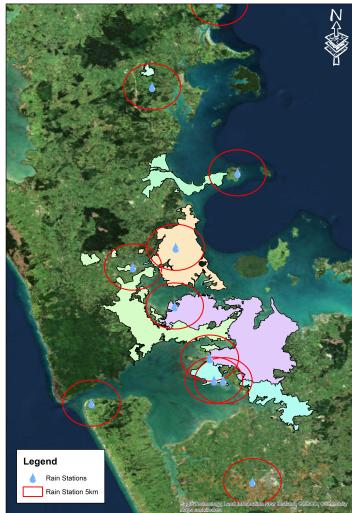


Figure 1: Auckland

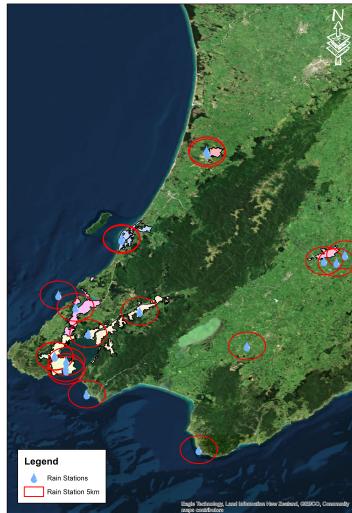


Figure 2: Wellington

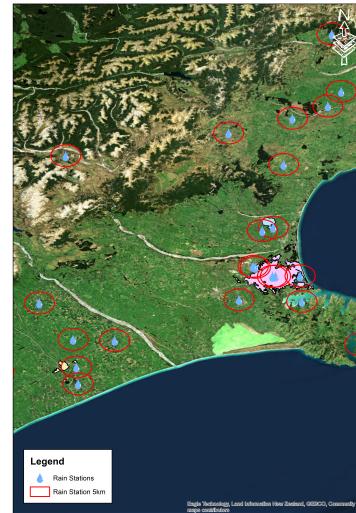


Figure 3: Christchurch

Figure 4: Maps of major Aotearoa cities with waste catchments and mapped rain catchments.

We supposed that a rain catchment that was within 5Km of a waste catchment was considered relevant, and the daily rain observed was averaged over the catchments relevant.

The catchments we were interested in are listed below:

Sample Location
WG_Seaview
AU_Warkworth
AU_Western
AU_Rosedale
WG_Porirua
CA_Christchurch
WG_Paraparaumu
CA_Ashburton
WG_Masterton
WG_MoaPoint
MW_Levin
WK_Hamilton
CA_Rangiora
WG_Karori
MA_Blenheim
AU_Eastern
AU_SouthWestern

WK\_Thames was considered but had insufficient data to support the model.

To see the full analysis, see `wastewater_rain_analysis/WastewaterRain.html` in the repo.

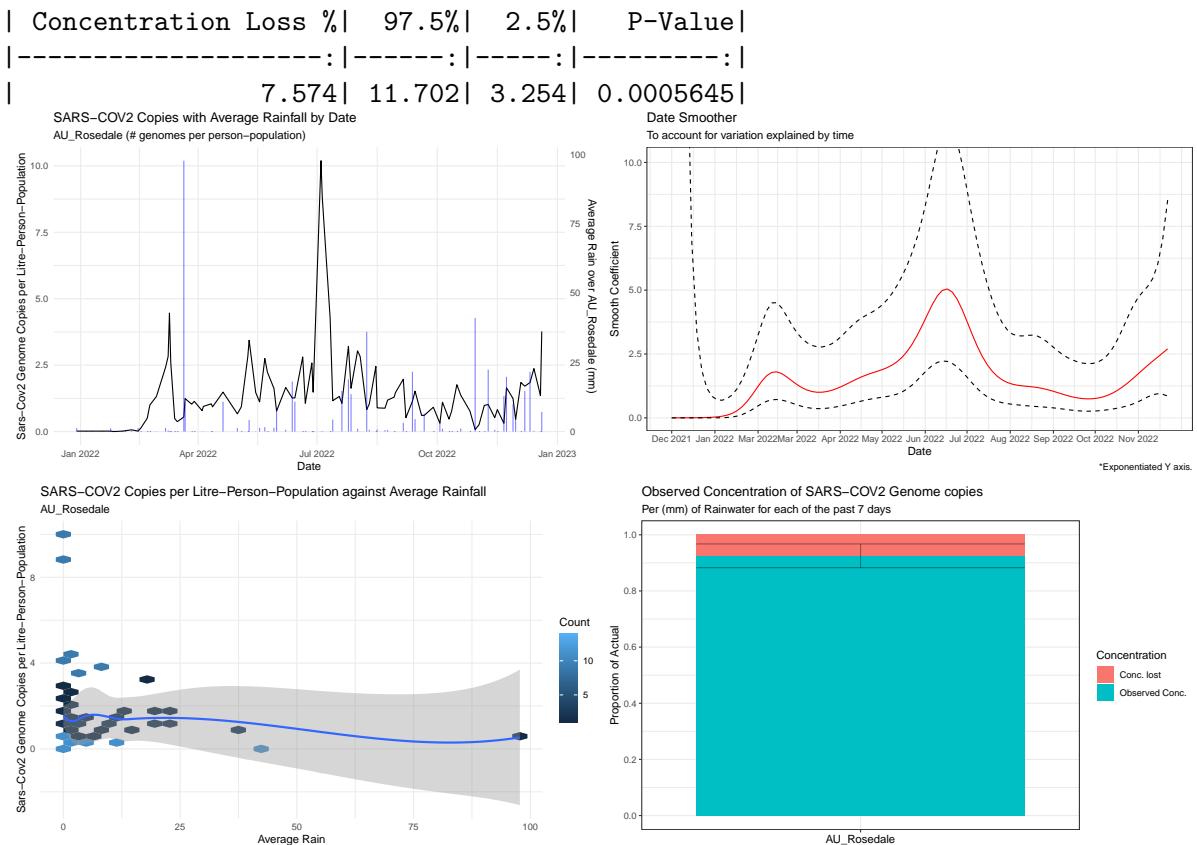
## Results

Each catchment was fitted to the same model as listed above with 12 knots. The script `produce_modelling_summaries.R` contains the functions that I used to fit these models and obtain summary statistics.

An example fit would look like this:

```
data.model = read_csv('flowraindata.csv')
data.model = data.model %>% split(data.model$SampleLocation)
source('../produce_modelling_summaries.R')
model.AU_Rosedale = produce_rainsummary.gam(data.model$AU_Rosedale)
print_rainsummary.gam(model.AU_Rosedale)
```

Table: AU\_Rosedale



From above, the catchment observed was AU\_Rosedale. The model shows that there is high significance in the effect of rainfall on the observed number of SARS-CoV2 Genome Copies (p-value = 0.0006), and estimates a combined effect of rainfall over the past 7 days to correspond with a decrease in concentration between 3% and 12%. The first two figures show that by eyeball we can see that the smoother that is approximating the variance over time seems appropriate. The third figure represents the trend of observed concentration of genome copies per person to the amount of rainfall, and the final figure is a visual representation of how much the concentration has decreased.

Below is a summary table of all the fitted models to locations, ordered by level of significance.

Sample Location	Estimated Loss in Concentration*	p-value
WG_Seaview	0.090	0.000
AU_Warkworth	0.075	0.000
AU_Western	0.028	0.000
AU_Rosedale	0.076	0.001
WG_Porirua	0.077	0.001
CA_Christchurch	0.091	0.002
WG_Paraparaumu	0.109	0.002
CA_Ashburton	0.142	0.006
WG_Masterton	0.120	0.006
WG_MoaPoint	0.077	0.011
MW_Levin	0.057	0.023
WK_Hamilton	0.057	0.023
CA_Rangiora	0.125	0.110
WG_Karori	0.046	0.156
MA_Blenheim	0.039	0.280
AU_Eastern	0.008	0.672
AU_SouthWestern	0.010	0.728

\*The combined effect of each lagged day's effect on the observed number of genome copies per litre, per person population.

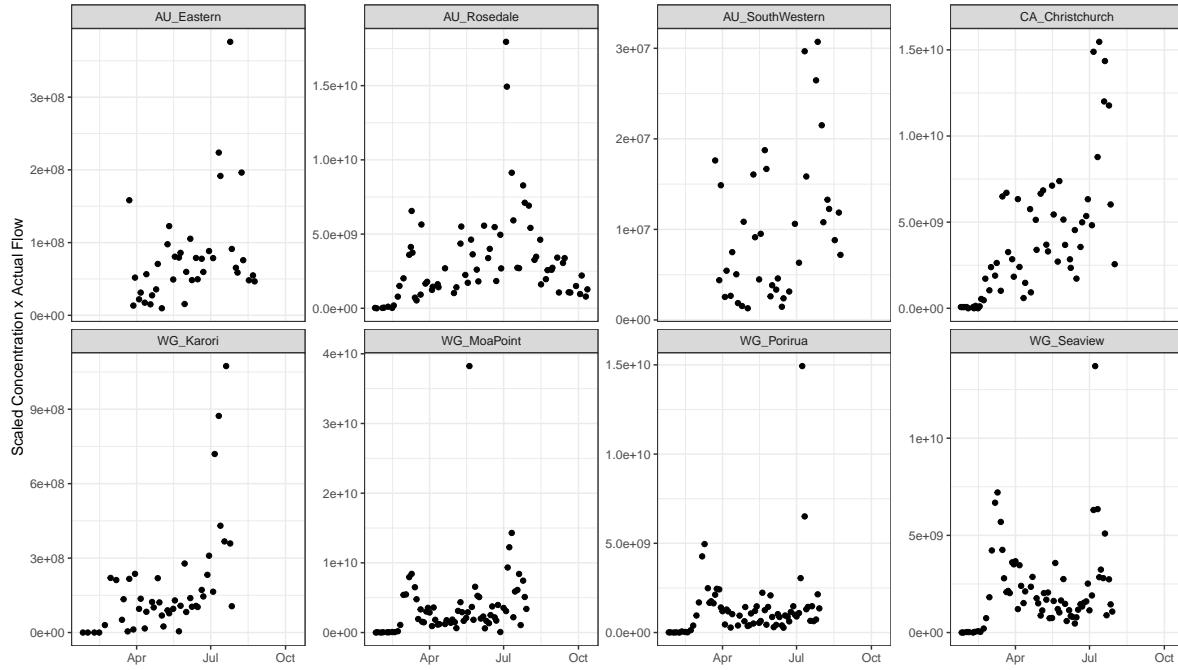
From above, it is clear that rainfall is significant in it's effect on the observed concentration of SARS-CoV2 genome copies per litre. This means that it is likely that rain overflow is making it's way into our sewage system.

We'd like to assume that rain accounts for all of the variation in *wastewater flow*, however we

know this is not the case as there are many more impactful factors which affect overall flow in wastewater. Hence we suspect that a more reliable means for normalising the concentration of genome copies will be to account for **Flow**; which some sampling sites have the measurements for.

## Flow

In the data, the scaled number of SARS-CoV2 genome copies per litre wastewater is estimated by dividing the observed concentration of genome copies per litre by the *yearly average amount of flow*. If this is a fair assumption, then we should see little variance year round for all locations when dividing the genome copies per litre by the actual amount of flow.



So it is obvious that the flow varies largely over time and daily flow must be accounted for when scaling the concentration of genome copies per litre.

If we are interested in the effect of flow on the observed number of genome copies, we can perform a similar analysis as what we did with *rainfall* and fit a linear model with the same formula.

$$C_t \sim \text{Quasipoisson}(\lambda_t, k)$$

$$\log(\lambda_t) = s(\text{Date}_t) + \beta_1 F_t + \beta_0$$

Then determine the significance of flow for each location.

## Locations of Interest

Because of limited flow data, we are limited to analysing just the following locations:

Sample Location
CA_Christchurch
AU_Rosedale
WG_MoaPoint
WG_Porirua
AU_Eastern
WG_Seaview
AU_SouthWestern
WG_Karori

**AU\_Western** was considered but had insufficient data to support the model (no flow data within dates of interest).

To see the full analysis, see `wastewater_flow_analysis/WastewaterFlow.html` in the repo.

## Results

Each catchment was fitted to the same model as listed above with 12 knots.

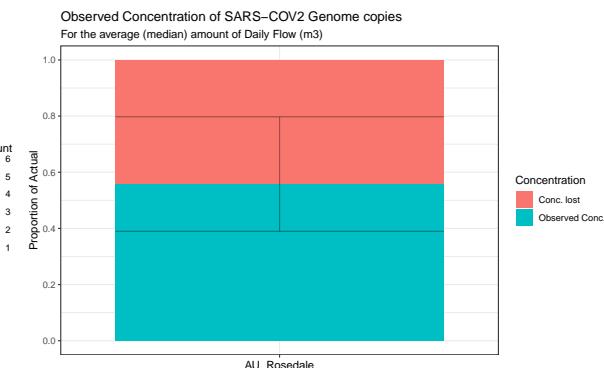
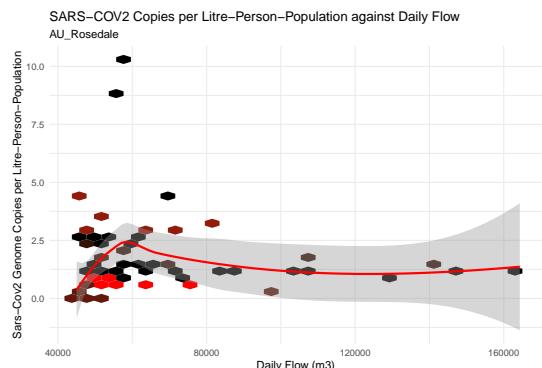
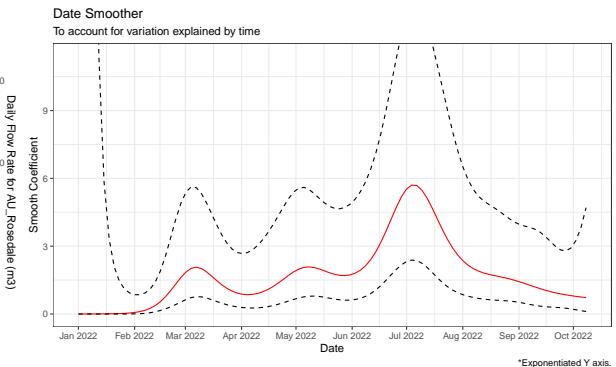
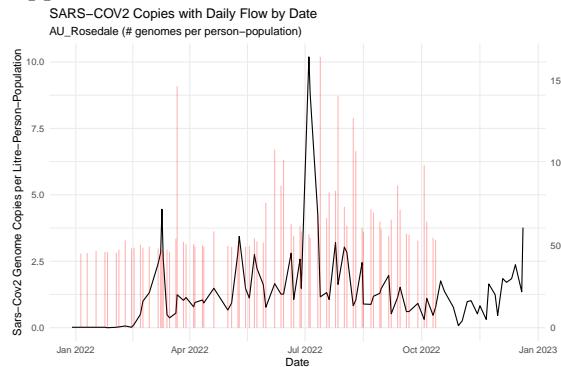
An example fit would look like this:

```
model.AU_Rosedale.flow = produce_flowsummary.gam(data.model$AU_Rosedale)
print_flowsummary.gam(model.AU_Rosedale.flow)
```

From above, the catchment observed was **AU\_Rosedale**. The model shows that there is high significance in the effect of rainfall on the observed number of SARS-CoV2 Genome Copies (p-value = 0.0006), and estimates a combined effect of rainfall over the past 7 days to correspond with a decrease in concentration between 3% and 12%. The first two figures show that by eyeball we can see that the smoother that is approximating the variance over time seems appropriate. The third figure represents the trend of observed concentration of genome copies per person to the amount of rainfall, and the final figure is a visual representation of how much the concentration has decreased.

Table: AU\_Rosedale

Quantile	Flow Rate	% Concentration Loss Explained	97.5%	2.5%	P-Value
Lower 25%	49187.00		41.05229	57.36412	18.49982   0.0017842
Median	54386.50		44.25534	61.03818	20.24329   0.0017842
Upper 75%	66220.25		50.91141	68.26284	24.07354   0.0017842



Below is a summary table of all the fitted models to locations, ordered by level of significance.

Sample Location	Median Flow (m3)	Estimated Loss in Concentration*	p-value
CA_Christchurch	145670.919	0.759	0.001
AU_Rosedale	54386.500	0.794	0.002
WG_MoaPoint	66666.725	-0.407	0.048
WG_Porirua	23933.778	-0.159	0.165
AU_Eastern	2220.906	0.336	0.210
WG_Seaview	55110.337	0.184	0.265
AU_SouthWestern	362.000	0.309	0.308
WG_Karori	3674.000	0.166	0.432

\*Estimated Loss in Concentration given the median amount of flow for that location.

Flow does not appear to be significant in most areas, however in those where it is a significantly large quoted effect for median flow is mentioned. **WG\_MoaPoint** shows some significance (p-value = 0.048) with a *positive* effect of flow on the amount of observed concentration of genome copies.

Admittedly, not a lot of care was taken to the data when conducting this part of the analysis and hence these results should be taken with a large grain of salt. **Definitely worth someone redoing!**

## Rain Versus Flow

Flow in is always going to prove more useful than rainfall, since the only way that rain can have an impact on the concentration of genome copies is by being introduced through *flow*. But from the analysis above, it does appear to have some significance in some places (and more so than flow for the time being, again the flow investigation deserves another shot), and could potentially be used in places where flow data is not readily available.

So begs the question of in which places can we get away with considering Rain in place of Flow.

To investigate this, we can fit a similar model as the ones above, only consider both Rain and Flow in the model. We can then extract the effect of Flow when Rain is controlled, and if it proves significant then we get an understanding of just how much more variation in the concentration of SARS-CoV2 gcs is explained by Flow.

$$C_t \sim \text{Quasipoisson}(\lambda_t, k)$$
$$\log(\lambda_t) = s(\text{Date}_t) + \beta_1 F_t + \beta_2 R_t + \beta_0$$

## Locations of Interest

Because of limited flow data, we are limited to analysing the same locations as flow such that we also have rain data for those locations:

Sample Location
WG_MoaPoint
CA_Christchurch
AU_Rosedale
WG_Karori
WG_Porirua
AU_SouthWestern
WG_Seaview
AU_Eastern

To see the full analysis, see `wastewater_flow_analysis/WastewaterFlow.html` in the repo.

## Results

Each catchment was fitted to the same model as listed above with 12 knots.  
An example fit would look like this:

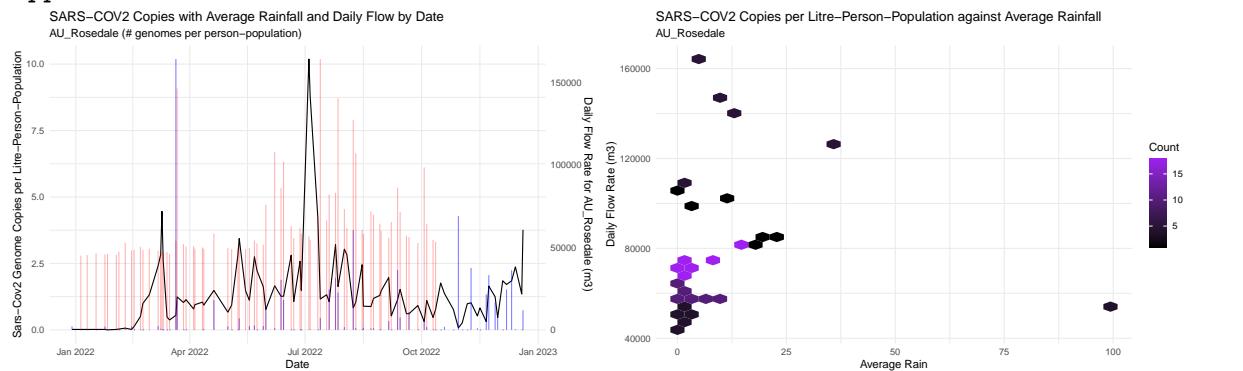
```

model.AU_Rosedale.rainflow = produce_flowrainsummary.gam(data.model$AU_Rosedale)
print_flowrainsummary.gam(model.AU_Rosedale.rainflow)

```

Table: AU\_Rosedale

Quantile	Flow Rate	% Concentration Loss Explained	97.5%	2.5%	P-Value
Lower 25%	49187.00		46.85721	73.49965	-6.570539   0.074627
Median	54386.50		50.29256	76.97052	-7.289854   0.074627
Upper 75%	66220.25		57.30599	83.26884	-8.945124   0.074627



Important to note there is that the % Concentration Loss Explained column explains the percentage loss in concentration *solely due to flow*. The estimates may seem wild for some of the locations, but what is more important here is determining the significance of the given value (p-value).

Below is a summary table of all the fitted models to locations, ordered by level of significance.

Sample Location	Median Flow (m³)	Estimated Loss in Concentration*	p-value
WG_MoaPoint	66666.725	-1.656	0.011
CA_Christchurch	145670.919	0.768	0.052
AU_Rosedale	54386.500	0.503	0.075
WG_Karori	3674.000	0.327	0.165
WG_Porirua	23933.778	-0.215	0.253
AU_SouthWestern	362.000	-0.374	0.510
WG_Seaview	55110.337	-0.089	0.736

Sample Location	Median Flow (m3)	Estimated Loss in Concentration*	p-value
AU_Eastern	2220.906	0.126	0.868

\*Estimated Loss in Concentration due solely to the median amount of flow for that location.

This tells us that at least for the case of **WG\_MoaPoint**, it is likely that we will need to use the Flow data in place of rain. From the rest of the analysis above, it is clear that at least from this location Flow is a more useful indicator of difference in concentration. It is also important to note that the corresponding estimate is actually an *increase* in concentration which doesn't seem right.

However for the majority of the other locations that we have flow and rain data access to, flow doesn't seem to explain much more variance over what rain already does. This is likely because flow is somewhat accounted for in the observed concentration of gcs when the number is divided by the yearly average flow; and hence any increase in variation is likely due to large rain spells.

## Population Variance

Although we expect flow to vary proportional to the population of that caught by the waste catchment, it is worth investigating the case of where a large increase/decrease in population causes any change to the observed number of gcs caught by the catchment.

One particular location that we know has a regularly changing population size is **Dunedin**; where we know approximately **15% of the population are students attending Otago University**. If we assume some proportion of these students leave over the holiday period, then we can estimate the effect that it has on the observed number of gcs.

```
# Define Term dates
# https://www.otago.ac.nz/news/events/keydates/archive/index.html
# Deciding that 0-Week is the official start since it's likely most students would be in Dunedin
# Had to use Wayback Machine to get SS dates for 2021-22
# https://web.archive.org/web/20220319234748/https://www.otago.ac.nz/summerschool/study/ke
data.dates = tibble(
  date = dmy(c('10/01/22', '24/02/22', '21/02/22', '15/04/22', '26/04/22', '22/06/22', '04/07/22')),
  event = c('Summer School Start', 'Summer School End', 'Sem 1 Start', 'Sem 1 Mid Sem-Break', 'Sem 1 End', 'Sem 2 Start', 'Sem 2 Mid Sem-Break', 'Sem 2 End')
)
# Vis
```

```
data.dates %>% knitr::kable(caption = 'OTAGO UNI KEY DATES')
```

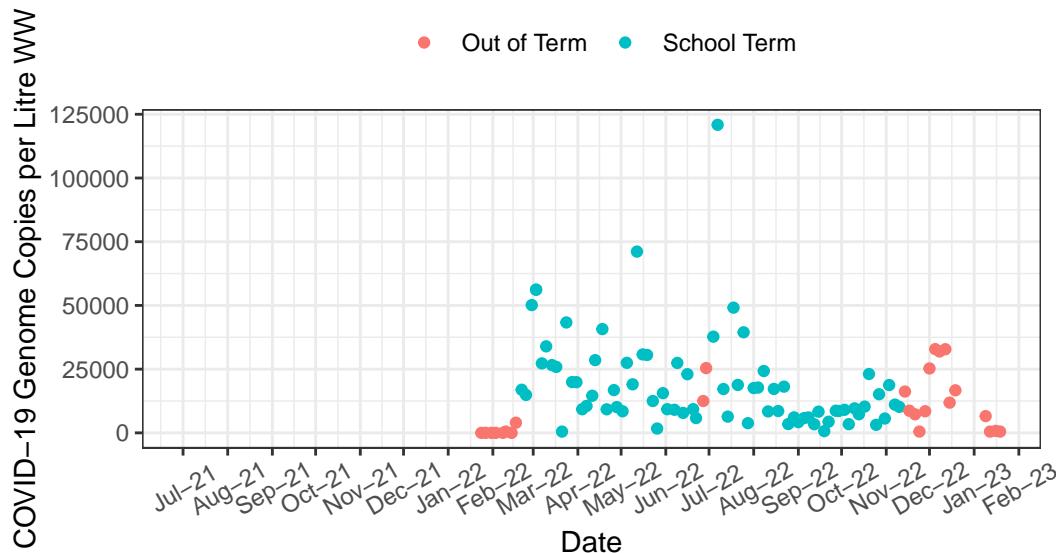
Table 7: OTAGO UNI KEY DATES

date	event
2022-01-10	Summer School Start
2022-02-24	Summer School End
2022-02-21	Sem 1 Start
2022-04-15	Sem 1 Mid Sem-Break Start
2022-04-26	Sem 1 Resume
2022-06-22	Sem 1 End
2022-07-04	Sem 2 Start
2022-08-27	Sem 2 Mid Sem-Break Start
2022-09-05	Sem 2 Resume
2022-11-12	Sem 2 End
2023-01-09	Summer School Start

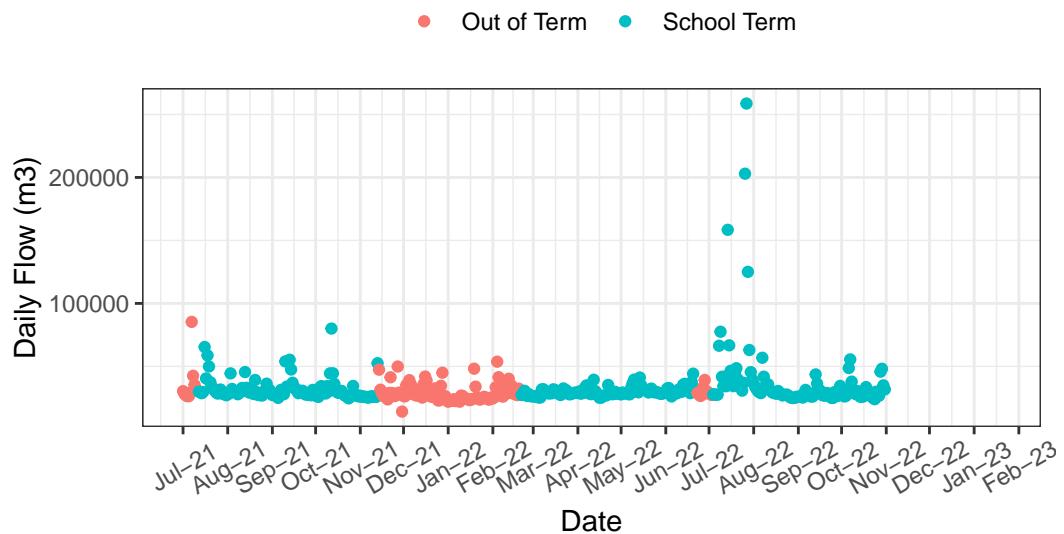


Currently there is insufficient flow data for any decent analysis to be conducted. Once we receive flow and concentration data for the 2022-2023 holiday period we might be able to do a proper analysis.

## Observed SARS-CoV2 Concentration per L OT\_DunedinTahuna

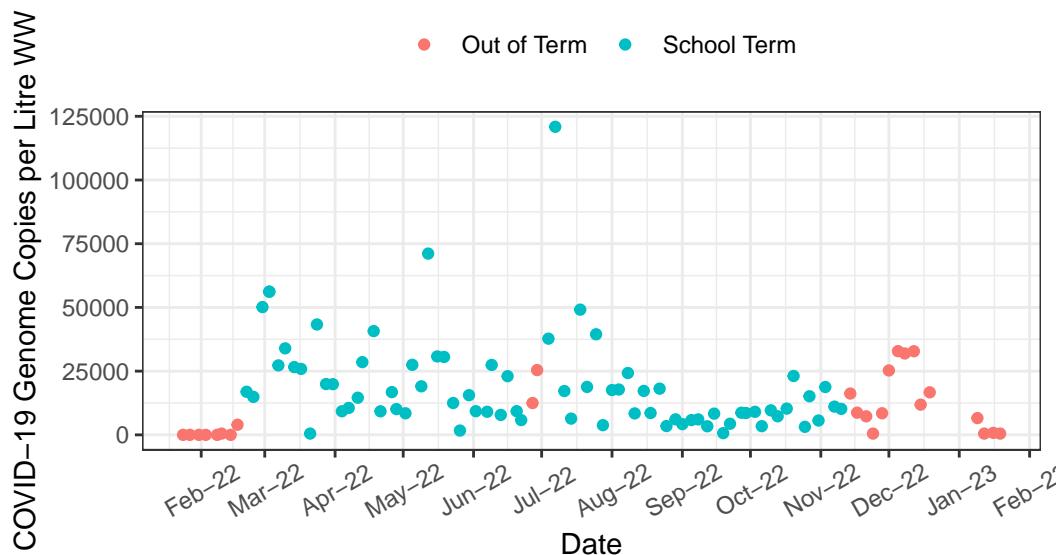


## Daily Flow OT\_DunedinTahuna

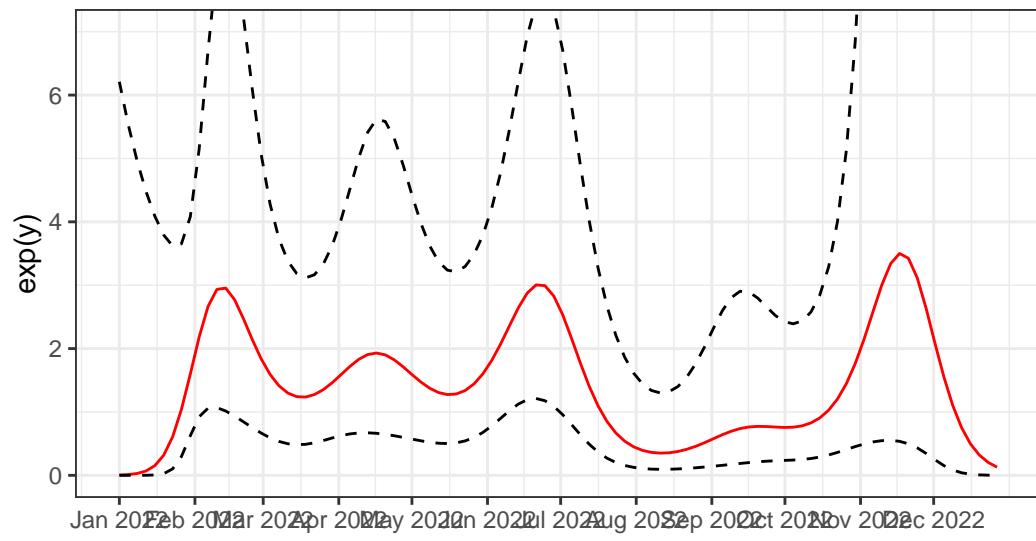


To estimate the effect that students being in school has on the observed concentration of genome copies, we will fit a similar model as above; accounting for all the seasonal variation of the virus outbreak with a smoother and letting the rest of the information be explained by the simple categorical `students_in_session`.

## Observed SARS-CoV2 Concentration per L OT\_DunedinTahuna



Smooth Mean to explain Monthly variation



\*Exponentiated Y axis.

Family: quasipoisson  
Link function: log

```

Formula:
sars_gcl ~ s(date.n, k = 11) + students_in_session

Parametric coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 8.8536    0.3768  23.499 <2e-16 ***
students_in_sessionTRUE 0.6327    0.4645   1.362    0.177
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
edf Ref.df F p-value
s(date.n) 9.777 9.985 5.412 3.76e-06 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) = 0.319 Deviance explained = 53.2%
GCV = 8952.9 Scale est. = 8868.1 n = 100

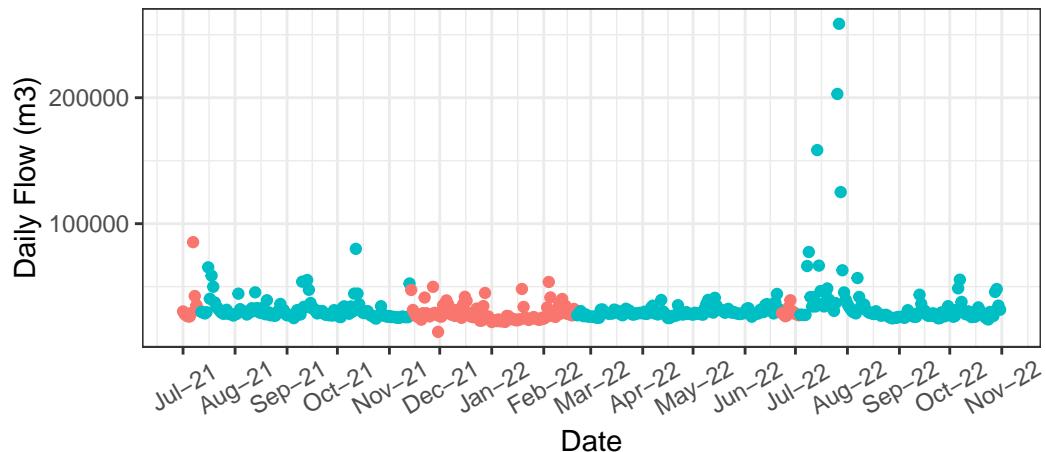
```

From this output, the model estimates with little to no significance ( $p = 0.177$ ) that there is an effect of students on the level of concentration of SARS-CoV2 gcs. The smoother applied is however likely explaining a lot of the effect from students; since it accounts for *monthly variation which may or may not be a result of lots of sick students in the area*, and hence it is unwise to trust this estimate yet. Applying a lower-degree smoother would allow more information to be explained, but it will fail to do its purpose of modelling the outbreak thus far and may be unfair toward the student effect.

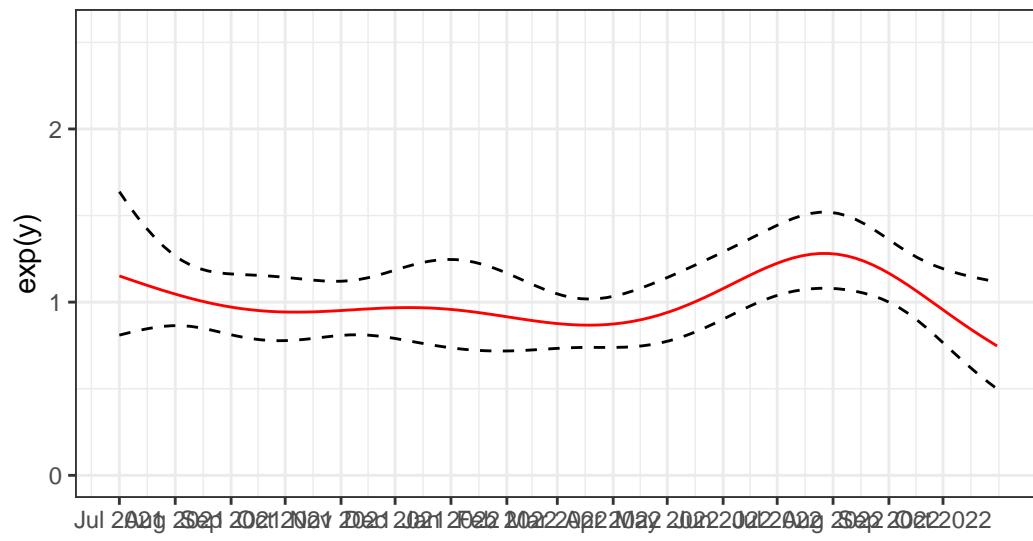
Before summarising, we can observe the effect that students have on Flow. Here however, since we expect flow to naturally fluctuate on a *seasonal* basis, we will instead just fit a smoother accounting for the number of seasons that the data covers.

## Daily Flow OT\_DunedinTahuna

● Out of Term ● School Term



## Smooth Mean to explain Seasonal variation



\*Exponentiated Y axis.

Family: quasipoisson  
Link function: log

```

Formula:
daily_flow_m3 ~ s(date.n, k = 6) + students_in_session

Parametric coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.30692 0.05971 172.615 <2e-16 ***
students_in_sessionTRUE 0.08439 0.07397 1.141 0.254
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
edf Ref.df F p-value
s(date.n) 4.841 4.984 7.348 1.43e-06 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) = 0.0661 Deviance explained = 12.8%
GCV = 4141.1 Scale est. = 6597.2 n = 488

```

Between July 2021 and November 2022; the data stretches across 6 seasons and hence an appropriate number of knots was applied.

The results show no significance ( $p = 0.254$ ) in the idea that there is an effect of students on the amount of Daily Flow.

This is interesting since we would usually expect the influx of students in the area should impact the amount of wastewater entering the plants in a positive way.

## Results

As much as the investigation into population effects is concerned, to continue we would require a couple data sets:

1. Flow data for the 2022-2023 holiday period (see the graphs above; currently there is insufficient data to support an analysis on scaled flow from population).
2. Population data for the area.

## **Summary**

This report contains the results of Toby and Thomas's investigation into confounding effects on the Concentration of SARS-CoV2 Genome Copies in Wastewater. In the process we investigated **three** different potential effects:

1. Rain
2. Flow
3. Population Variance

We found that **Rain** in most areas has a negative relationship with the observed number of SARS-CoV2 Gcs.

Whilst we have faith in the fact that controlling the effect of Rain in models using the Concentration of Gcs to estimate COVID-19 prevalence, we think it is more wise where possible to instead use **Daily Flow** measurements instead of a *yearly average*; since this factor will contain the information presented by rainfall.

Using *Dunedin* as a case study, we found that there is potential for the population to have an effect on both Concentration and Flow, however there is insufficient data at this point to complete the analysis.