MASTERS BY RESEARCH

Wastewater-based Epidemiology: Estimating the number of viral cases in a hospital setting using wastewater analysis

Pantea, Igor

Award date:
2024

Awarding institution:
Bangor University

Wastewater-based Epidemiology: Estimating the number of viral cases in a hospital setting using wastewater analysis

Master of Science by Research Candidate:

Igor Pântea

500510475

Research Supervisors:

Dr. Kata Farkas

Prof. David L. Jones

Declaration

I hereby declare that this thesis is the results of my own investigations, except where otherwise stated. All other sources are acknowledged by bibliographic references. This work has not previously been accepted in substance for any degree and is not being concurrently submitted in candidature for any degree unless, as agreed by the University, for approved dual awards.

Yr wyf drwy hyn yn datgan mai canlyniad fy ymchwil fy hun yw'r thesis hwn, ac eithrio lle nodir yn wahanol. Caiff ffynonellau eraill eu cydnabod gan droednodiadau yn rhoi cyfeiriadau eglur. Nid yw sylwedd y gwaith hwn wedi cael ei dderbyn o'r blaen ar gyfer unrhyw radd, ac nid yw'n cael ei gyflwyno ar yr un pryd mewn ymgeisiaeth am unrhyw radd oni bai ei fod, fel y cytunwyd gan y Brifysgol, am gymwysterau deuol cymeradwy.

Signature

Date 01/12/2022

ABSTRACT

Wastewater-based epidemiology (WBE) represents an effective complementary tool for the surveillance of infectious diseases when the capacity of conventional clinical testing is limited or when there is a significant proportion of infectious asymptomatic cases in the community being monitored. Two of the current challenges of WBE is understanding how the wastewater matrix affects virus recovery and how to translate the wastewater results into a number of infected individuals.

To address these challenges, wastewater surveillance was undertaken at a large regional hospital (Ysbyty Gwynedd) in Wales, UK. The samples were analysed using qPCR for three viruses of public health concern: influenza A virus, norovirus GII (NoVGII) and SARS-CoV-2. Additionally, the concentrations of ammonium-nitrogen and orthophosphate and the faecal indicator, namely crAssphage, were trialled as estimates for human urine or faecal load (i.e. population captured). The mathematical models to estimate the population numbers included empirical equations following the Central Limit Theorem premise (EMCLT) and three separate Monte-Carlo-Bayesian approaches (MCBA). Additionally, pH, electrical conductivity (EC) and turbidity were also measured to characterise the wastewater matrix.

Modelling based on an adaptive neuro-fuzzy interference system was used to assess the impact of wastewater parameters on virus recovery, and literature research was used to interpret the relationship. Ammonium and phosphate had a positive modelled association on virus recovery, likely due to their concentrations being correlated with the number of individuals shedding into the wastewater system. SARS-CoV-2 and crAssphage detection was greatest in the pH range 7.6 – 8.5 and from 6.8 – 8.5, respectively. Turbidity was not associated with SARS-CoV-2 but was with crAssphage, which was attributed to the association of the latter with faecal matter. EC was positively associated with the recovery of both viruses, starting at >0.75 mS/cm due to soluble organic matter and ions inhibiting virus sorption. However, it also reflects the capture of human effluent, which has a high EC value (ca. 20 mS/cm).

CrAssphage-based total population estimates were within the expected boundaries in 35.7% of cases, whereas ammonium and phosphate gave significant overestimates with considerable fluctuations likely due to unknown sources of discharge. One MCBA provided realistic results for NoVGII and influenza A virus, whereas two other MCBA and the EMCLT provided promising results for SARS-CoV-2 and crAssphage. For validation, the SARS-CoV-2 estimates were compared with COVID-19 clinical cases in the region of Gwynedd, whereas the influenza A virus estimates were compared with national cases. The infection numbers were converted into proportion of the total population by dividing the estimates with the crAssphage-based estimates or the clinical numbers by the census data of Gwynedd or Wales. For influenza A, RNA concentrations in wastewater did not coincide with the increase of cases in Wales, probably due to low virus detection rates. Hence, the clinical validation was not successful. When the SARS-CoV-2 estimates were overestimated, it was likely due to underestimated crAssphage estimates of the total population. When converting to percent

changes, there was a significant correlation between the estimated and reported trends, pointing out to the potential suitability of the models. However, validation with clinical data may be difficult in a hospital setting due to a dynamic population, a residential area being more suitable.

The study has expanded on the usefulness of MCBA and EMCLT for WBE, and providing their adjustments, particularly normalising virus concentration, they could be implemented for routine surveillance. They are useful in settings with dynamic populations, such as a hospital, where predictions cannot be made about the population infection susceptibility. Despite the assessed impact of physico-chemical parameters on virus recovery, the wastewater matrix is complex and requires a more comprehensive study that explores other parameters (e.g., bio-chemical oxygen demand, total organic carbon) and investigates mathematical relationships.

Acknowledgments

The WEWASH project is funded by the Welsh Government surveillance programme. The Centre for Environmental Biotechnology Project was funded through the European Regional Development Fund (ERDF) by Welsh Government.

I would like to offer special thanks to my supervisors Kata Farkas and Davey Jones not only for the provided career opportunities and the integrated master's degree, but for the extensive freedom of thought and choice. Coupled with mentoring and research support, their contribution has been invaluable, for which I am extremely thankful. The possibility to combine work responsibilities with pursuing research ideas has made my excellent experience for what it was, an accelerated course of character development.

I would like to acknowledge the contribution of my team at the WEWASH project for helping to run the laboratory analyses, particularly Nick Woodhall, whose friendship and work ethic has offered me a very valuable perspective. Their contribution was extended beyond work relationship etiquettes, making almost every day involving, engaging and rewarding.

My Family was always there. Probably, this thesis will be read with a pinch of juridical scepticism. The significant time and effort that was put in completing it, was derived from their sacrifice. I will eternally be grateful for the foundation they built for me and the continuous support throughout all the time.

I would also like to acknowledge the natural and cultural beautifulness of extended Bangor and North Wales. It is a source of inspiration, motivation, harmony and mind peace.

Table of Contents

ABBREVIATIONS

(RT-)qPCR – (Reverse Transcriptase-) quantitative Polymerase Chain Reaction

ACE2 – Angiotensin-converting enzyme 2

ANFIS – Adaptive neuro fuzzy inference system

CDC – Centres for Disease Control and Prevention

CLT – Central Limit Theorem

EMCLT – Equation-based model following the Central Limit Theorem premise

COVID-19 – Coronavirus disease 2019

CrAssphage – Cross-Assembly phage

DNA, RNA – Deoxyribonucleic acid, ribonucleic acid

dsDNA, ssDNA, dsRNA, ssRNA – Double-strand DNA, single-strand DNA, double-strand RNA, single-strand RNA

DTW – Dynamic time warping

EDTA – Ethylenediaminetetraacetic acid

gc, gc per g stool – Gene copies, gene copies per gram stool

MCBA – Monte-Carlo-Bayesian approach

MF – Membership function

NADH – Nicotinamide adenine dinucleotide hydrogen

NoVGII – Norovirus GII

NTU – Nephelometric turbidity unit

P/P mix – Primer and probe mix

PMMoV – Pepper mild mottle virus

RMSE – Root-mean-square error

SARS-CoV-2 – Severe acute respiratory syndrome coronavirus 2

SEIR model – Susceptible-exposed-infectious-removed model

TMPRSS2 – Transmembrane protease, serine 2

WBE – Wastewater-based epidemiology

WEWASH – Wales Environmental Wastewater Analysis & Surveillance for Health

WWTP – Wastewater treatment plant

EQUATION LIST

$$\frac{z_{\alpha/2}\times\sqrt{\frac{\hat{p}\hat{q}}{n}+\frac{z^2_{\alpha/2}}{4n^2}}}{1+(z^2_{\alpha/2})/n} < \Delta \quad (1)$$

$$C_1V_1 = C_2V_2 \quad (2)$$

$$P = \frac{C\times F}{M} \quad (3)$$

$$N = \frac{C_{RNA}\bar{F}}{\bar{\alpha}\bar{\beta}(1-\bar{\gamma})} \quad (4)$$

$$\bar{M}_{RNA} = N\bar{\alpha}\bar{\beta}(1-\bar{\gamma}) \quad (5)$$

$$\delta M_{RNA}{}^2 = N\left((\bar{\alpha}^2+\delta\alpha^2)(\bar{\beta}^2+\delta\beta^2)((1-\bar{\gamma})^2+\delta\gamma^2) - \left(\bar{\alpha}\bar{\beta}(1-\bar{\gamma})\right)^2\right) \quad (6)$$

$$\delta C^2_{RNA} = (\bar{M}^2_{RNA}+\delta M^2_{RNA})\left(\frac{1}{\bar{F}^2}+\frac{\delta F^2}{\bar{F}^4}\right) - \frac{\bar{M}^2_{RNA}}{\bar{F}^2} \quad (7)$$

$$(\delta C^2_{RNA})_{Total} = (\delta C^2_{RNA})_{Intrinsic} + (\delta C^2_{RNA})_{Measurement} \quad (8)$$

$$\delta N = \frac{\delta C_{RNA}\bar{F}}{\bar{\alpha}\bar{\beta}(1-\bar{\gamma})} \quad (9)$$

$$\gamma_{min} = 1 - (C_{measured}/C_{estimated\_min}) \quad (10)$$

$$C_{RNA} = \frac{\sum_{i=1}^{N}\alpha_i\beta_i(1-\gamma_i)}{F_i} \quad (11)$$

$$P(N|C_{RNA}) = \frac{P(C_{RNA}|N)P(N)}{P(C_{RNA})} \quad (12)$$

$$\rho(A,B) = \frac{1}{N-1}\times\sum_{i=1}^{N}\left(\frac{A_i-\mu_A}{\sigma_A}\right)\times\left(\frac{B_i-\mu_B}{\sigma_B}\right) \quad (13)$$

$$RMSE = \sqrt{\sum_{i=1}^{N}\frac{(A_i-B_i)^2}{N}} \quad (14)$$

1. Introduction

Epidemiology is a field concerned with the prevalence of various diseases in populations and allows healthcare practitioners to take informed decisions, plan strategies, and adopt policies in order to prevent the spread of diseases and to improve public health.

Within the context of epidemiology, wastewater-based epidemiology (WBE) has emerged, and it is a rapidly developing field. One of the earliest uses of WBE can be dated back to 1939 for poliovirus surveillance, and it has been the major monitoring tool for poliovirus since (Singer et al., 2023; Klapsa et al., 2022). Later, in the 1990s, it evolved to allow the monitoring of household liquid waste and illicit drug use (EMCDDA, 2022). Currently, one of its largest applications is the surveillance of infectious diseases. During the COVID-19 pandemic, the monitoring of SARS-CoV-2 through wastewater proved to be essential in many countries for informing public health responses (Figure 1) (Singer et al., 2023; O'Keeffe, 2021). Other applications include for surveillance of antimicrobial resistance, chemicals exposure and for estimating drug and pharmaceutical consumption (O'Keeffe, 2021).

Generally, WBE consists of analysing wastewater to gather information on community health. With the COVID-19 pandemic, where the proportion of unreported cases was estimated at 92.9% (Maugeri et al., 2020), and with the need to prepare for future endemics or pandemics, the importance of alternative epidemiological surveillance tools became more evident. Thereby, WBE may represent an effective tool for epidemiological surveillance especially when used in conjunction with the conventional methods.

Currently, WBE provides data about geographical and temporal trends, however, this paper aims to go beyond trends, and to rationalise the data collected from the wastewater analysis, the final aim being whether the actual number of infection cases can be inferred.

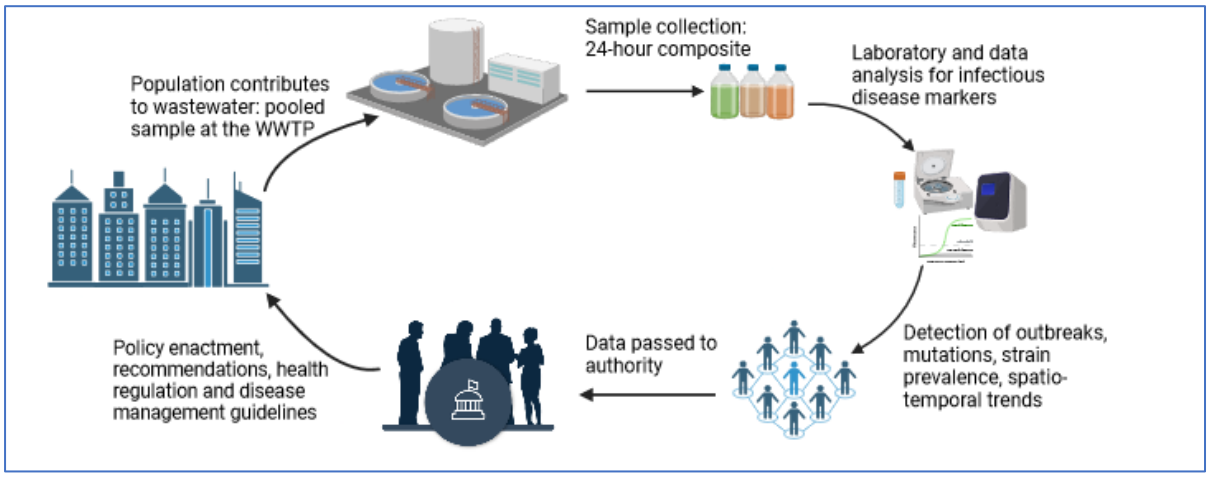

Figure 1. Infectious disease monitoring with wastewater-based epidemiology. *Upon exposure of the population to infectious diseases (or chemicals) and subsequent infection spread, the population contributes to wastewater with specific markers that can be identified. The wastewater is either sampled at a wastewater treatment plant (WWTP) or at any point before that. After sampling, it is transferred to a laboratory for analysis which allows for an estimation of the disease spread in the*

community, data which can be analysed complementary to clinical-based epidemiological surveillance methods. Image created with BioRender (2024).

1.1 Comparative evaluation of wastewater-based epidemiology

Wastewater samples can provide time-averaged estimates of disease load in the whole population within an individual sewage catchment (i.e., sewershed), especially when composite samples (i.e. small volumes collected and mixed at regular time intervals) are collected (EMCDDA, 2022). This provides a cost-efficient and quick alternative to mass clinical testing when that is not viable, such as due to economic or logistics reasons (Shrestha et al., 2021). Additionally, since wastewater testing is anonymous and less invasive, WBE may overcome ethical barriers (Sims & Kasprzyk-Hordern, 2020). For example, the advantage over epidemiological surveillance based on questionnaires or surveys is that it eliminates the risk of incomplete disclosure or data fabrication by respondents due to fear of stigmatization (O'Keeffe, 2021). When comparing to sentinel surveillance or clinical-based surveillance, WBE has an advantage when the clinical testing capacity is limited, or the asymptomatic cases represent a significant portion of infections. Some studies have also used WBE to initiate mass clinical testing (Diemert & Yan, 2020; Schmitz et al., 2021; Shrestha et al., 2021). WBE can also offer more timely epidemiological data than sentinel or clinical-based surveillance, which allows to detect an outbreak 4-5 days earlier (Morvan et al., 2022). It is generally explained by a bigger delay from specimen collection to reporting date in clinical-based surveillance and due to the transmission and viral shedding occurring before the onset of symptoms (He et al., 2020; Peccia et al., 2020). Therefore, public health responses, such as increased testing or lockdown measures, can be implemented in a timely manner.

On the other hand, disadvantages are that there is variability and uncertainty in the pathogen degradation rate, pathogen residency in the sewage system, infection incubation period, infection stage and pathogen excretion, environmental factors, and variation in the concerned population (Li et al., 2021; Wade et al., 2022). While uncertainty was previously quantified (20-40%) and disentangled in numerous modelling papers, it remains a topic of further research to allow for an accurate translation of wastewater concentration into prevalence in the population (Joung et al., 2023; Li et al., 2021; Nourbakhsh et al., 2022). Depending on the target, it can also be challenging to select the biomarker since it should be relatively stable and sufficiently abundant in wastewater (Sims & Kasprzyk-Hordern, 2020). A disadvantage identified in some studies is the significant time-lag between data collection and data analysis (Sims & Kasprzyk-Hordern, 2020). However, with proper logistics as demonstrated at the "Wales Environmental Wastewater Analysis for Surveillance and Health" (WEWASH) (Welsh Government, 2023), wastewater samples can be processed and analysed within 24 – 48 hours from collection, allowing for timely data reporting.

1.2 Targeted Viruses

The quantitative detection of viruses in wastewater provides an assessment of viral prevalence and distribution in a population (Morvan et al., 2022; Singer et al., 2023).

However, it may also provide other information, such as the efficiency of disinfection procedures (e.g., WWTP processes) and an assessment of the down-stream risk of waterborne viral contamination of recreational and bathing waters (Corpuz et al., 2020).

The focus of this study will be influenza A virus and SARS-CoV-2 as they represent major respiratory viruses of increasing public concern due to their endemic and pandemic potential, and the damage it can cause to society in terms of economic losses, public health and life satisfaction. Norovirus GII (NoVGII) is another target of this study as this is the predominant norovirus genotype responsible for gastroenteritis (Huhti et al., 2011). Noroviruses have several transmission routes (e.g., food-borne, person-to-person), a low infectious dose, high shedding rates, and relatively high environmental stability, hence are difficult to prevent and contain (Barclay et al., 2014). This renders norovirus the most common cause of sporadic gastroenteritis and outbreaks of acute gastroenteritis in the UK (Barclay et al., 2014). Hence the rapid identification of norovirus outbreaks is essential in containing its further spread.

Viral abundance and diversity in wastewater are dependent on the virus pathophysiology such as the organ system infected (e.g., noroviruses and the gastrointestinal system), the route of infection, and the extent of the spread within the population (Corpuz et al., 2020). It also should be noted that virus structure (e.g. enveloped vs non-enveloped) also determines its prevalence in wastewater by defining its environmental stability. Therefore, to understand virus behaviour in the sewage system, it is necessary to understand infection at both an individual level and at a population level.

1.2.1 Influenza A Virus

Influenza A virus is an enveloped ssRNA virus and was responsible for the 1918 Spanish Flu pandemic which resulted in an estimated >50 million deaths and 500 million infection cases globally (CDC, 2018; Blümel et al., 2009). Influenza is a respiratory infection, individuals being infectious commonly for up to four days after the onset of symptoms (Mayo Clinic, 2022a). Common symptoms include headaches, fever, dry persistent cough, rhinitis, myalgia, pharyngitis and fatigue (Mayo Clinic, 2022a). Influenza viruses are capable of antigenic evolution, or mutations, through drift and shift events, which are characterised by small changes and, respectively, considerable genome changes (Blümel et al., 2009), hence being of significant public health concern. Unlike antigenic drifts to which the population is likely to have already formed partial immunity, antigenic shifts strains are more dangerous due to most of the population

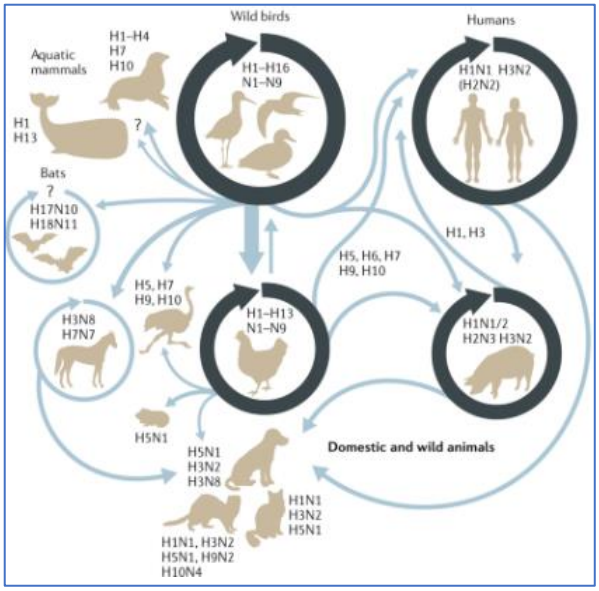

Figure 2. Influenza A virus cross-species transmission. *The cross-species transmission of influenza A virus depends on the strain and is usually coupled with antigenic drift or shift events, thereby eliciting fewer immune responses. Image and Data Source: (Long et al., 2019).*

having little to no immunity against them. Additionally, influenza A virus can infect other species than humans (i.e., birds, horses, pigs and humans), a cross infection likely contributing to an antigenic shift (Figure 2) (Blümel et al., 2009). For example, the 1918 H1N1 "Spanish Flu" and the 1968 H3N2 "Hong Kong Flu" were of avian origin (Short et al., 2015), while the 2009 N1H1 "Swine Flu" pandemic was of combined avian, swine and human origin (Mayo Clinic, 2021). And it is even believed that the cross-species transmission such as of the H5N1 avian influenza not only has the future capacity to become a human infectious disease but also that of global pandemic level (Riedel, 2006).

Two potential ways in which influenza A virus localises in the gastrointestinal system and is being shed in faeces, and later in wastewater, are through haematogenous dissemination to organs and through replication in intestinal cells (Minodier et al., 2015). In the latter case, as demonstrated *in vitro*, specific receptors, i.e., sialic acid-α2,6-galactose-terminated saccharides, localised on the epithelial cells of the human gastrointestinal system allow for virus binding and subsequent cellular infection and viral replication (Minodier et al., 2015).

A study detected influenza A and B viruses in 36.4% of faecal samples from 22 individuals with the mean virus copy number being $1.0x10^6$ gene copies (gc) per g stool (Hirose et al., 2016). Four studies recorded lower mean values, these being 4 ± 0.8 $\log_{10}$ copies/g stool, $6.5 \pm 2.7 \times 10^4$ copies/ml of faecal suspension, 3.2×10^4 copies per gram stool, and 1.44×10^4 gc/ml (Chan et al., 2011; al Khatib et al., 2021; Chan et al., 2009; To et al., 2010). This is consistent with another study recording varying values of $2.5x10^4$-$4.2x10^6$ gc/g of stool (Arena et al., 2012). Therefore, the high titres and abundance of virus in stool makes influenza viruses a good WBE target.

1.2.2 SARS-CoV-2

SARS-CoV-2 is an enveloped ssRNA virus and is responsible for the ongoing COVID-19 pandemic with more than 627 million confirmed cases and 6.5 million deaths as of 26th Oct 2022 (WHO, 2022). Moreover, as new variants continue to appear, it is likely to that the infection cases will surge in a seasonal pattern (Murray & Piot, 2021). Symptoms are flu-like or cold-like, commonly lasting up to two weeks, and include fever, continuous tussis, ageusia, hyposmia, fatigue, myalgia, headache, pharyngitis, rhinitis (NHS, 2022).

SARS-CoV-2 is occasionally shed in urine (<5% of COVID-19 cases) but more often it can lead to a gastrointestinal infection and is shed in faeces (30-60% of cases) (Gupta et al., 2020; Zhang et al., 2021; Guo et al., 2021; Jones et al., 2020; Anjos et al., 2022). SARS-CoV-2 infects cells through binding to the angiotensin-converting enzyme 2 (ACE2) receptor, followed by activation of the cell entry process mediated by the host-cell protease enzymes, i.e., TMPRSS2 and lysosomal proteases cathepsins (Scudellari, 2021). These are expressed on the surface of respiratory cells, particularly ACE2 and TMPRSS2 on lung alveolar type 2 cells, oesophageal upper epithelial and gland cells (Scudellari, 2021; Zhang et al., 2020). An analysis of single cell co-expression of ACE2 and TMPRSS2 found that these proteins are abundantly expressed in

the enterocytes of the ileum and colon regions of the gastrointestinal systems (Zhang et al., 2020). Moreover, ACE2 is expressed on all regions of the small intestine, and in an experiment involving human small intestinal organoids, SARS-CoV-2 demonstrated its capacity to effectively infect all enterocytes (Hamming et al., 2004; Lamers et al., 2020). Together, this evidence demonstrates SARS-CoV-2 infection and presence in the gastrointestinal system, and subsequent SARS-CoV-2 faecal shedding.

Different studies report different prevalence of SARS-CoV-2 in faeces. For example, a study in China from data gathered on a 93 patients cohort reported a detection rate of 59% while another study in Liverpool, UK reported detecting SARS-CoV-2 in 18% of COVID-19 infections (Guo et al., 2021; Vaselli et al., 2021). The shedding rate in faeces varies among different studies from 10^2 to 10^7 gc/ml, the variation depending on the severity of the infection (Jones et al., 2020). A study found that 3 patients had a SARS-CoV-2 concentration of $6x10^5$ to $7x10^6$ gc/g stool (Zang et al., 2020). While another two studies reported concentrations of up to 7.5 $\log_{10}$ gc/g stool (Wölfel et al., 2020; Han et al., 2020).

A study found that the sputum is a significant source of SARS-CoV-2 in wastewater (Li et al., 2022a). Simulations of maximum theoretical virus concentration in wastewater that account for sputum and faecal contribution instead of only faecal contribution demonstrated that 91% of WBE literature data were within the maximum theoretical concentration as opposed to only 63% (Li et al., 2022a). However, sputum contribution was not considered in the present study due to unknown total sputum shedding per capita.

Additionally, SARS-CoV-2 with the concentration of $3.2x10^2$ gc/ml was detected in urine for 1/9 patients (Peng et al., 2020), while another study reported SARS-CoV-2 with a concentration of 5.48-5.79 $\log_{10}$ gc/ml for 2 patients (Yoon et al., 2020). However, urinary shedding is very rare, and it is most likely to occur in hospitalised patients with mild to severe COVID-19 infections (Jones et al., 2020). At the same time, another study reported no SARS-CoV-2 in 27 urine samples derived from COVID-19 patients (Wölfel et al., 2020).

1.2.3 Norovirus GII

Norovirus, particularly the GII strains (NoVGII), also known as the “winter-vomiting bug” is responsible for an estimated of 685 million cases worldwide of acute gastroenteritis annually (CDC, 2021). Gastroenteritis is characterised by diarrhoea, emesis and dyspepsia, symptoms usually lasting up to three days (Mayo Clinic, 2022b). Genogroups I and IV are also associated with gastroenteritis in humans (Hassan & Baldridge, 2019). Recently, in England, the UK Health Security Agency (UKHSA) reported an increase in norovirus outbreaks from 24 to 40 in a short time period, specifically the week commencing 6th Feb to the week commencing 14th Feb 2022, hence emphasising the importance of its timely epidemiological control (UKHSA, 2022b).

Noroviruses are non-enveloped and single-stranded RNA (ssRNA) viruses, being transmitted primarily through the faecal-oral route (Seitz et al., 2011). VP1, a norovirus capsid protein produced upon cell infection, was detected in enterocytes, and in macrophages, dendritic cells and B/T lymphocytes within gut-associated lymphoid tissue, pointing to a possible viral capacity to infect these cell types (Hassan & Baldridge, 2019). The gut epithelial barrier is overcome potentially though microfold cells which allow human norovirus transcytosis (Hassan & Baldridge, 2019). For infection initiation, human noroviruses bind to histo-blood group antigens, localised on the surface of enterocytes, as well as in saliva and other secretions (Hassan & Baldridge, 2019). However, another working model of norovirus intestine infection proposes that noroviruses are only transcytosised across the epithelium barrier (enterocytes) in order to gain access and to infect immune cells within the gut lamina propria (Karst & Wobus, 2015).

Based on data collected from 40 patients, a study found that the median norovirus concentration was 8.48 $\log_{10}$ to 8.97 $\log_{10}$ copies per gram of faeces (Lee et al., 2007). These findings are consistent with another study which demonstrated that 7 employees with acute gastroenteritis had a mean norovirus concentration of 8.33 $\log_{10}$ gc/g stool while elderly residents with the mean age of 78.7 years had a mean norovirus concentration of 9.63 $\log_{10}$ gc/g stool (Lai et al., 2013). Also, the concentration reaches as high as 10.5 $\log_{10}$ copies/g of faeces in patients with prolonged diarrhoea (≥4 days) (Lee et al., 2007).

NoVGII is relatively stable in the environment enabling it to maintain its infectivity in groundwater for more than 2 months but being detectable for over 3 years at 25^{o}C (Seitz et al., 2011). It is also expected to be periodically prevalent in hospital wastewater in response to localised outbreaks (UKHSA, 2022b). This is supported by a systematic review reporting a 82.1% detection rate of noroviruses in wastewater, consistent with an experimental study detecting NoVGII in 96.1% of raw wastewater samples at a WWTP (Huang et al., 2022; Fumian et al., 2019).

1.3 Population Estimation

For the estimation of the population served by the sewage draining point, ammonium, phosphate and crAssphage concentrations are investigated.

1.3.1 Ammonium and Phosphate

Ammonium and phosphate population markers are often used due to the low-cost and speediness of their measurement in wastewater requiring only a centrifugation step and a simple colorimetric-based quantification method. Despite of its input from agricultural and industrial sources, ammonium has been validated previously for normalising virus concentrations when estimating viral carriage rate in the population (Been et al., 2014; Hutchison et al., 2022). Although research estimating contribution per capita is available (Alexander & Stevens, 1976), a similar exercise has not been undertaken for the use of phosphate as a human faecal or urine marker. Due to their widespread use of lifecare products, such as synthetic detergents, phosphate may be a poor indicator of human faecal

or urine load and population number estimation (LennTech, 2022). Nonetheless, if the human exogenous phosphate discharge is relatively constant, then the phosphate fluctuations could originate from a change in the population number.

1.3.1 CrAssphage

CrAssphage is a bacteriophage associated with the human gut and is used as a tracking marker for human faecal contamination, reaching concentrations of $7.3x10^4$-$4.9x10^7$ gc/100ml in wastewater samples (Sala-Comorera et al., 2021; Park et al., 2020). This is consistent with another study recording a crAssphage concentration of 5.28-7.38 log_{10} gc/100ml in wastewater samples from a hospital and residential buildings (Kongprajug, Mongkolsuk & Sirikanchana, 2019). CrAssphage can be detected in 48%-68.5% of healthy human stool samples, and the prevalence can increase up to 71.4% during norovirus outbreaks (Park et al., 2020). It is estimated to be present in ca. 75% of faeces from healthy individuals (Mertz & Speicher, 2020). However, other reports suggest a CrAssphage prevalence of only 35.7% among U.S. residents (Honap et al., 2020). The study recorded a mean concentration of 8.1 log_{10} gc/g stool for individuals without acute gastrointestinal symptoms and 8.4 log_{10} gc/g stool for healthy children with the age ≤ years old, while the mean concentration decreased to 5.9 log_{10} gc/g in faeces sampled during a norovirus outbreak (Park et al., 2020).

The advantage of using crAssphage as an indicator of human faecal contamination is that it is specific to humans. While some cross-reactivity has been shown with dog, gull, poultry, pig and cattle faeces, the levels of crAssphage found in these non-human sources were several orders of magnitude lower than that of human sources (Park et al., 2020; Ahmed et al., 2018; García-Aljaro et al., 2017; Stachler et al., 2017). Also, crAssphage is present exclusively in faeces, and not in other excretion or secretion sources, such as saliva, emesis matter or nasal secretions (Park et al., 2020).

The estimation of the population based on the crAssphage concentration will be done according to the modelling method outlined in the Materials and Methods section.

1.4 Mathematical Modelling

Mathematical models are useful to understand how viral particles behave in sewerage systems (i.e. physical hydrological models) but also to estimate the probability of infection cases in relation to measured RNA/DNA concentrations (i.e. epidemiological models) (Joung et al., 2023; Sonnenwald et al., 2023). Two mathematical approaches are investigated for the potential to give an accurate estimation of the number of infection cases based on the detected viral concentration, an equation-based model following the Central Limit Theorem premise (EMCLT) and the Monte-Carlo-Bayesian approach (MCBA). Briefly, they rely on the volumetric flow rate and the viral contribution into wastewater of the average individual.

Since the researched wastewater site encompasses a hospital setting, the epidemiological dynamics would differ from a residential or urban area. Significant population movement, higher immunisation profile of frontline workers, more preventive measures (frequent disinfections and hand washing, face covering) and, simultaneously, higher rate of infectious patients than in the wider community would yield epidemiological models accounting for

person-to-person transmission uncertain within a hospital setting. Although person-to-person transmission models would facilitate the simulation of infection transmission among in-patients, it would be difficult to account for virus transmission or shedding from individuals visiting the hospital (Cooper et al., 2023). EMCLT and MCBA don't require prior or current knowledge of the transmission profile or population movement, and allow for a simple to implement, real-time representation of the infection burden within the setting, which made them, the models of choice for this study.

1.4.1 The Central Limit Theorem

The Central Limit Theorem (CLT) is a statistical method to quantify uncertainty by inferring the population mean based on the sample mean (LaMorte, 2016). If multiple samples are taken from the same population, then it is likely that it would result in different estimates of means, creating a distribution of the sample means (LaMorte, 2016). In CLT, the distribution of sample means is approximately normal providing a sufficiently large random sample size (n≥30), regardless of the population distribution except special cases, such as for a Cauchy distribution due to its infinite variance (Figure 3) (Chang et al., 2006). Furthermore, the sample size requirement of n≥30 can be ignored in the case of a normal population or of a binomial population (e.g., dichotomous outcome) (LaMorte, 2016).

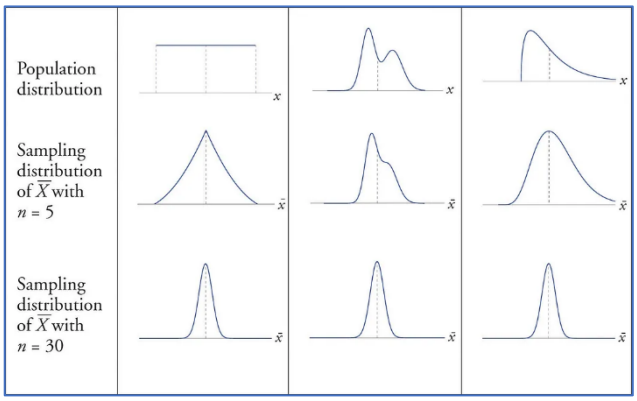

Figure 3. Distribution under Central Limit Theorem. *Providing a finite variance of the population and independent and random sampling of the means ($\bar{X}$), the sampled means will have a normal distribution regardless of the population distribution. This requires a sufficiently large sample size of means, n≥30. Image Source: (Singh, 2018). Data Source: (LaMorte, 2016; Chang et al., 2006).*

In the context of EMCLT, an important assumption is that, similarly to multiple sample means giving an estimate of the population mean, the CLT permits the estimation of the mean for the probability distribution of the virus concentration in wastewater for any given population size (N) (Saththasivam et al., 2021). There, the probability distribution of the virus concentration in wastewater only for one sample or one individual is not necessarily normal, and its mean is not necessarily an accurate estimate of the population mean. However, with a sufficiently large N, the probability distribution becomes normal, and its approximation of the mean becomes more accurate. Similarly, the probability of all the other variables related to this large population, such as viral shedding, approximate a normal distribution. This assumption of a normal distribution for all the parameters of N, providing a sufficiently large size of N, will allow to perform an inversion procedure to obtain a normal probability distribution for N, the standard deviation and mean of which can be estimated with equations as outlined in "2.2.2 Modelling Method 1 (Equation-based Model Following the Central Limit

Theorem Premise)". In this study, the EMCLT is an alternative to the MCBA since the latter involves lengthy calculations requiring relatively significant computational power.

1.4.2 The Monte-Carlo-Bayesian Approach

In the presence of uncertainty associated with variables, which are hard to predict (e.g. wastewater flow rate, faecal load or virus faecal shedding), a method to create a probability distribution of possible outcomes is through Monte-Carlo simulations (Kenton, 2022). It is not a substitute for empirical evidence, but an approach for modelling probability. Generally, Monte-Carlo simulations are repeated calculations using each time a different set of randomly generated values, which themselves can also be derived from a probability distribution (e.g., uniform or normal) (IBM, 2020). As the number of recalculations increases, the accuracy of the forecast model also increases. The randomly generated sets as a representation of the variable uncertainty are between minimum and maximum values as determined through historical experimental data. It should also be noted that the random values are in reality pseudo-random numbers due to most algorithms using deterministic mathematical functions. For it to be a truly random variable, the algorithm must use an external variable that is unpredictable and relates to a random behaviour (e.g., radioactive decay, or a quantum level process). In this mathematical model, full Monte-Carlo simulations were used for multiple population estimates to generate the corresponding probability distribution of virus concentrations in wastewater.

Bayes' rule offers the framework to solve conditional probability problems through a mathematical equation (Westbury, 2010). The equation describes how to calculate probability of event A given condition B, which is referred to as the posterior probability. This requires knowledge about prior probabilities of each event independently, and the probability of event B given A (Westbury, 2010). Within the context of MCBA, the Bayes' rule results in equation (12). Since the virus concentration in wastewater is a variable measured experimentally, the Bayesian aspect of the model uses the probability of the simulated concentrations for each population number to derive the probability distribution of the total population for the corresponding measured virus concentration (Saththasivam et al., 2021). This can be thought of as an inversion procedure. A visual representation of this Bayesian transformation can be observed in the example provided in the "9. Appendix: Supplementary Materials".

Although there are several proposed approaches to estimate the number of required Monte-Carlo simulations, and even though some suggest using an absolute number of approximately 20000 simulations (Quinlan, 2015), Oberle (2015) developed a procedure based on Dunnigan and Agresti critique of the Wald method, offering an iterations number estimation method individual to the problem (Oberle, 2015).

It can be summarized as follows: while the inequality in the equation

$$\frac{z_{\alpha/2}\times\sqrt{\frac{\hat{p}\hat{q}}{n}+\frac{z_{\alpha/2}^2}{4n^2}}}{1+(z_{\alpha/2}^2)/n} < \Delta \quad (1)$$

is false, the Monte-Carlo simulation continues, where n – the number of iterations, $z_{\alpha/2}^2$ – the z-multiplier (has a value of 1.96 for a CI of 95%), p – the population proportion where $\hat{p}$ is its natural estimator, $\hat{q} = 1 - \hat{p}$, and Δ – half the length of the confidence interval. However, this approach may result in an iteration number of the order of hundreds of thousands of iterations which is not feasible to be implemented without significant procession power.

1.4.3 Normalisation of Pathogen Concentrations

An unknown factor is the rate at which pathogens are uncaptured during sample pre-collection (e.g., degradation and adherence to solid particles) and post-collection, including laboratory analysis (e.g., poor virus recoveries for wastewater concentration and RNA extraction, or inhibited virus detection). If known, the value of losses can be used for the normalisation of the detected concentration and a more accurate estimate of the population.

While using the EMCLT, viral losses are assumed to be 0 due to uncertainty. However, in the MCBA, viral losses are accounted for by (pseudo-) random number generation between the values of 0 (no degradation) to 1 (all viruses were lost) inclusive. This would vary considerably from sample to sample due to varying flow rate, wastewater matrices, distance from the source to the sampling point, temperature, presence of disinfectants, time between sampling and analysis, varying pathogen residency time in the sewer etc. The whole breadth of factors is impossible to estimate at the moment, but it is possible to measure and account for factors such as turbidity and electrical conductivity as indicators of increased residency in the sewer and adherence to solid particles. While the maximum viral loss is to remain 1 (all the particles were lost), the minimum viral loss can be established.

The concentration normalisation approach should also account for viral losses related to the laboratory methodology. For example, while concentrating the wastewater sample (as outlined in the methodology section), the pellet which contains a small fraction of the viral quantity, is typically discarded. Different experiments report different viral recovery in the pellet. A recent article reports that at turbidity of approximately 100 NTU, the virus recovery in the pellet represents approximately 1% comparative to the viral recovery from the wastewater concentrate (Farkas et al., 2021), hence if the turbidity of the sample is >100 NTU, then as a first step of the concentration normalisation, the detected concentration can be multiplied by 1.01. However, some research reports a SARS-CoV-2 recovery in the pellet of up to ~9% (Ahmed et al., 2021) and even 23% (Forés et al., 2021).

Several chemical parameters, such as ammonium and phosphate concentrations, turbidity, electrical conductivity (EC), and pH, will be studied to understand the spatial and temporal

heterogeneity of the wastewater matrix, potentially allowing for normalisation of the detected pathogen concentration for anomalies in the flow.

Different liquid concentration methods are available, and since they rely on different mechanisms for concentration, a variability in the recovery can be observed. For this study, an ammonium sulphate-based precipitation method was used. The procedure is identical to the method used by Farkas et al. (2022), who reported a viral recovery of 6.2%. Therefore, it may be possible to normalise the concentration by the corresponding factor, hence accounting for viral losses related to the laboratory method.

1.5 Objectives of the project

The main hypothesis is that the number of infectious disease cases can be calculated using wastewater data analysis. The chosen targets include SARS-CoV-2, influenza A virus and norovirus GII. The concentrations of ammonium and phosphate, as well as that of crAssphage, were tested as indicators of the population total number contributing to the collected wastewater sample.

A rational correlation between virus concentration and electrical conductivity as an indicator of salinity and dissolved organic mass, and turbidity was sought as an indicator of the total amount of solid particles. If found, the correlation can be used to normalise wastewater viral concentration data.

If the number of infection cases can be calculated with sufficient confidence, then it will contribute positively to the development of WBE, and it can prove to be an effective tool for epidemiological insight.

2. Materials and Methods

2.1 Laboratory Analysis

2.1.1 Sample Collection

The sampling was performed at a wastewater drainage point located near the regional hospital Ysbyty Gwynedd, Wales, under the aegis of the Wales Environmental Wastewater Analysis for Surveillance and Health programme (Welsh Government, 2020 – 2023). The catchment area included the hospital and a small amount of the nearby residential area dedicated to nurses working in the hospital. The wastewater system was not completely isolated from the surface, hence there was an additional surface water contribution during rainfall events.

Sampling was performed in two stages: 28th Oct 2021 till 16th Dec 2021, then it resumed on 11th Jan 2022 till 20th May 2022. Between 28th Oct and 16th Dec 2021, 12 individual 2-hour composites (frequency of 15 ml every 10 minutes) were collected over 24 hours taken Monday to Friday, except for 14th Dec between 8:00 and 16:00 when the samples were a one-go grab every 10 minutes. After 11th Jan 2022, the sampling was consistent with composites being collected regularly over a 24-hour period (50 ml every 10 minutes) by the installed refrigerated autosampler (Teledyne ISCO Avalanche), providing 1 sample per 24 hours, after which they were transported to the laboratory at 4^{o}C. As the weekly composite sample collection started on Monday, they were transported to the laboratory on Tuesday to Friday (4 samples per week). There, the samples were processed in a biosafety level 2 laboratory.

Due to logistics, it was not always possible to collect samples every Monday to Friday between 28th Oct 2021 and 16th Dec 2021, or every Tuesday to Friday between 11th Jan 2022 and 20th May 2022. Moreover, occasionally, the collection for 2-hours composites was unsuccessful for specific different hours during the day.

The total number of collected samples was 377. During the first stage of the collection, the sampling was at times inconsistent in terms of the total collected volume due to autosampler issues, and hence it was not always possible to perform all the laboratory analyses on the samples. During the second stage, the sampling was consistent with the same volume being collected throughout the whole period, and all the laboratory analyses were performed. Among the collected 377 samples, only 344 samples were processed completely in terms of virological and chemical parameters analyses.

During sample collection, the volumetric flow rate was measured. The collected data was then compared and adjusted with the total water use data provided by the Ysbyty Gwynedd estate manager.

2.1.2 Chemical Parameters

Wastewater pH and electrical conductivity were measured with a SevenCompact pH/Cond S213 meter (Mettler Toledo). Turbidity and free Cl^- was measure with a HI-83414-02 Turbidity and Free and Total Chlorine Meter (Hanna Instruments), respectively.

For ammonium (NH_4^+) and orthophosphate (PO_4^{3-}) measurements, spectrometry-based methods were used, and the optical density (absorbance) readings were performed on a SPECTROstar® Nano spectrometer (BMG Labtech). The phosphate determination method was based on the reaction of phosphate within the sample (80 µL) with ammonium molybdate (180 µL, 14 ml of >95% H_2SO_4 and 2.1 mg of ammonium molybdate tetrahydrate diluted with H_2O to a total volume of 500 ml) and ascorbic acid (30 µL, 1 g $C_6H_7O_6$ diluted in 10 ml of H_2O), followed by 30 minutes incubation and spectrometry reading at 820 nm.

For ammonium determination, the sample (150 µL) was reacted with 15 µL of EDTA (C(EDTA) = 0.06 g/ml) to prevent metal ions interference (e.g., Mg^{2+}), 60 µL of Sal-Na-nitroprusside (6.252 g sodium salicylate and 0.1 g of sodium nitroprusside dihydrate in 80 ml of H_2O), and 30 µL of Na-K_2PO_4 (1.18 g NaOH and 3.98 g K_2PO_4 in 36 ml of H_2O and 4 ml of NaClO). The incubation time was 10 minutes followed by absorbance reading at 667 nm.

For more details on the approach, refer to the methods described elsewhere (Murphy & Riley, 1962; Mulvaney, 2018).

The data collected from ammonium and phosphate measurements were considered sufficiently accurate, and the measurement did not require repetition when the R^2 of the standards linear regression was >0.95.

2.1.3 Wastewater Concentration

In order to reduce the sample volume from 150 ml to 1 ml, the samples were concentrated using an ammonium-sulphate-based concentration method (Kevill et al., 2022). This method previously demonstrated the most optimal combination of parameters for viral recovery, concentration factor, procession time and cost per sample when compared with polyethylene glycol-based precipitation methods and InnovaPrep or Amicon ultrafiltration methods (Farkas et al., 2022).

The first laboratory analysis step was to separate the solid matter, which was achieved by centrifuging 200 ml of wastewater at 10000 RCF for 10 minutes at 4°C. From this, 150 ml of supernatant was kept, while the rest, including the solid matter, was discarded. The 150 ml sample was transferred to another centrifuge bottle containing 60 g of ammonium sulphate salt, $(NH_4)_2SO_4$. After inverting the sample sufficiently to allow the salt to dissolve completely, achieving an ammonium sulphate concentration of 40% w/v, the bottles were incubated for 1 hour at 4°C. The next step was sample centrifugation at 10000 RCF for 30 minutes at 4°C in order to achieve organic molecules precipitation. The supernatant was discarded, while the precipitate was resuspended in 850 µl of NucliSens® lysis buffer (BioMerieux), achieving an approximate 1 ml of concentrate. The 1 ml of wastewater concentrate was transferred into a 2.2 ml 96-well Kingfisher deep well plate (Thermo Scientific) for nucleic acids extraction.

2.1.4 Nucleic Acid Extraction

The nucleic acids extraction, and separation from other organic molecules, was based on a magnetic silica beads method using NucliSens® easyMAG® reagents (bioMérieux, 2022). The extraction process was performed automatically on a Kingfisher 96 Flex System (Thermo Scientific) as described by Kevill et al. (2022).

The extraction process began with adding 50 µl of beads into the 1 ml sample and incubating for 10 minutes, hence binding the beads to polar molecules including organic molecules. It followed by "washing" the beads twice with 385 µl of wash buffer 1, then twice with 485 µl of wash buffer 2, and once with 500 µl of wash buffer 3. Once the separation from impurities was completed, the beads were introduced into an elution solution and incubated at 60^{o}C for 5 minutes where the nucleic acids were separated from the beads. Upon the removal of the magnet from the elution plate, the nucleic acids remained in the elution plate while the beads were removed and discarded. The elution plate with 100 µl of eluate containing nucleic acids was then ready for analysis with quantitative polymerase chain reaction (qPCR) for DNA virus (crAssphage) or reverse-transcription qPCR (RT-qPCR) for RNA viruses.

2.1.5 Nucleic Acids Quantification

For the identification of SARS-CoV-2, influenza A virus, NoVGII RNA genetic material, RT-qPCR was used, and for the identification of crAssphage bacteriophage DNA genetic material, qPCR was used. The qPCR reactions were performed on a QuantStudio Flex 6 real-time PCR machine. The protocols were devised by Farkas et al. (2022a).

The primers and probes (Table 1) were mixed into 1 ml solutions, called P/P mixes, with the following final concentrations**: SARS-CoV-2 P/P** – 10 µM forward primer, 20 µM reverse primer, 2.5 µM probe; **NoVGII P/P** – 10 µM forward primer, 20 µM reverse primer, 5 µM probe; **influenza A virus P/P** – 10 µM forward primer, 20 µM reverse primer, 5 µM probe; **crAssphage bacteriophage P/P** - 5 µM forward primer, 10 µM reverse primer, 5 µM probe.

Table 1. Virus target sequences primers and probes.

Virus Type	Forward primer sequence	Reverse primer sequence	Probe sequence	Probe dyes	Ta	Target DNA sequence	Standard Type
SARS-CoV-2	GACCCCAAAATCAGCGAAAT	TCTGGTTACTGCCAGTTGAATCTG	ACCCCGCATTACGTTTGGTGGACC	FAM-MGB	60	GACCCCAAAATCAGCGAAATGCACCCCGCATTACGTTTGGTGGACCCTCAGATTCAACTGGCAGTAACCAGAATGGTGGGGCGCGATCAAAACAACGTCGGCCCCA	Synthetic RNA (Kevill et al., 2022)
Norovirus GII	ATGTTCAGRTGGATGAGRTTCTCWGA	TCGACGCCATCTTCATTCACA	AGCACGTGGGAGGGCGATCG	FAM-QSY	60	ATGTTCAGATGGATGAGATTCTCAGATCTGAGCACGTGGGAGGGCGATCGCAATCTGGCTCCCAGTTTTGTGAATGAAGATGGCGTCGA	plasmid DNA (Farkas et al., 2017)
Influenza A Virus	CAAGACCAATCYTGTCACCTCTGAC	GCATTYTGGACAAAVCGTCTACG	TGCAGTCCT	FAM-ZEN-IABkFQ	60	AAAGACAAGACCAATCCTGTCACCTCTGACTAAGGGGATTTTAGGATTTGTGTTCACGCTCACCGTGCCCAGTGAGCGAGGACTGCAGCGTAGACGCTTTGTCCAAAATGCCCTAAATGGG	Synthetic Influenza H1N1 (2002) RNA control (Twist Bioscience, USA)
CrAssphage	CAGAAGTACAAACTCCTAAAAAACGTAGAG	GATGACCAATAAACAAGCCATTAGC	AATAACGATTTACGTGATGTAAC	FAM-TAMRA	60	CAGAAGTACAAACTCCTAAAAAACGTAGAGGTAGAGGTATTAATAACGATTTACGTGATGTAACTCGTAAAAAGTTTGATGAACGTACTGATTGTAATAAAGCTAATGGCTTGTTTATTGGTC	Plasmid DNA (Farkas et al., 2019)

Through a 1:10 serial dilution, DNA or RNA standards (Table 1) were obtained for the corresponding virus with concentrations of 10^5 copies/µl to 10^0 copies/µl, and an additional two negative control standards were used.

The reaction components for a total final reaction volume of 20 µl per well for crAssphage qPCR were 6.3 µl of H_2O, 8 µl of Quantifast no ROX qPCR mix, 2 µl of Quantifast ROX qPCR mix, 1 µl of BSA (1 mg/l), 0.7 µl of P/P mix, 2 µl of sample/standard.

The reaction components for a total final reaction volume of 20 µl per well for SARS-CoV-2, NoVGII and influenza A virus RT-qPCRs were 8.68 µl of H_2O, 5 µl of TaqMan Virus Fast one-step RT-qPCR mix, 1 µl of BSA (1 mg/l), 0.32 µl of $MgSO_4$ (50 mM), 1 µl of P/P Mix and 4 µl of sample/standard.

The reaction conditions for QuantiFast Probe DNA qPCR were 1 cycle of 5 min at 95°C for denaturation, 40 cycles of 15 sec at 95°C for denaturation followed by 1 min at 60°C for annealing-extension. The fluorescence was captured during annealing-extension.

The reaction conditions for TaqMan RNA RT-qPCR were 1 cycle of 30 min at 50°C for reverse transcription, 1 cycle of 20 sec at 95°C for denaturation, 45 cycles of 3 sec at 95°C followed by 45 sec 60°C for annealing-extension. The fluorescence was captured during annealing-extension.

The file output from the QuantStudio Flex 6 real-time PCR machine was then analysed using the QuantStudio Real-time PCR software v1.7 (Applied Biosystems).

Each sample, standard or negative control, had two replicates, which were used to calculate the mean value. To accept the standard curve, the following quality criteria were set: slope between -3.1 and -3.6, efficiency between 90% and 110% and R^2 >0.95. Whenever the negative control of the concentration, extraction or qPCR phase were contaminated, single positive samples with a concentration lower than two times the contamination concentration were removed.

2.2 Data Analysis

The RT-qPCR or qPCR results, representing the viral concentration in nucleic acid eluent (gc/µl), were used to obtain the initial sample concentration with the formula:

$$C_1V_1 = C_2V_2 \quad (2),$$

where C_1 and V_1 – concentration and volume of the initial sample, C_2 and V_2 – concentration and volume of the concentrate.

After calculating the concentrations of the non-concentrated initial wastewater samples, the values were ready to be integrated in the modelling methodology.

The two tested modelling methods were the EMCLT and the MCBA, which were compared against each other and against the available clinical data. It should be noted that the EMCLT was an intermediate step in the MCBA as well as a separate model.

The modelling methodologies required the following variables: volume flow rate, virus faecal load, faecal shedding rate, viral losses, ammonium and phosphate production per day per capita. For these variables, the literature was researched to find mean values and the range within which each of the variables do fit. For the flow rate values, experimental measurements of the sewage sampling site were conducted which were used in conjunction with data for the total used water kindly provided by the hospital management.

2.2.1 Population Estimation with Ammonium and Phosphate

The population was estimated based on wastewater ammonium and phosphate concentrations according to the equation:

$$P = \frac{C \times F}{M} \quad (3),$$

where P – the population estimate, F – volumetric flow rate (l/day), C – measured ammonium or phosphate concentration (mg/l), M – estimated average amount of daily NH_4^+ or PO_4^{3-} production per person (mg/day/person).

However, there is a very high variation and sometimes conflicting evidence in the amount of ammonium and phosphate discharged by the average individual in a day. For example, the total average phosphorus in urine, including phosphate, was determined to be between 800 and 2000 mg/L, and it reflects the phosphorus intake and metabolism rather than having a relatively static value (Münch & Winker, 2011). However, the high variability in phosphate concentrations could be due to Münch & Winker (2021) using a measurement of the concentration in a single sample per person.

In contrast, other studies advise that for a measurement with a reliability index of >80%, it is required to collect two to three 24-hour urine samples, and to calculate the average concentration (Cupisti & Gallieni, 2018). According to three different studies measuring urinary biomarkers in multiple 24h urine samples, the authors reported mean phosphate concentrations (±SD) of 1032 ± 300 mg/d, 738 ± 221 mg/d and 850 ± 256 mg/d (Sun et al., 2017). For population estimation, the phosphate results collected by the NHSII study was selected (mean of 850 ± 256 mg/d).

An overview of renal ammonia metabolism has estimated that the average urinary ammonia excretion is 30-40 mmol/d, which is equivalent to 511-681 mg/d (Weiner & Verlander, 2013). This is consistent with a report of urine composition by the German Society for International Cooperation that reports an ammonium/ammonia-N average composition of 460 mg/L (Münch & Winker, 2011). Considering a normal 24-hour urine volume being 800-2000 mL (MedlinePlus, 2022), 460 mg/L would be equal to 368-920 mg/d. Due to urea hydrolysis, the ammonium concentration in stored urine can increase up to 8.1 g/L within 2 – 4 weeks, however, the sampling and the laboratory analysis was performed within 1 – 3 days and hence the measurement was not affected by hydrolysis.

An ammonia excretion value of 511-681 mg/d and 368-920 mg/d as outlined above conflicts with other studies reporting an ammonia discharge of approximately 6 g/d/capita (Zheng et al., 2017). A survey in Lausanne, Switzerland with a low ammonia industrial input, reported an ammonia-N discharge of 6.9 ± 0.4 g/day/capita (Been et al., 2014). This is consistent with two other studies performed in China reporting a discharge of 5.78-7.57 g/d/capita and 6.4 g/d/capita (Zheng et al., 2017; Xie et al., 2008; Zhu et al., 2010). The reason for such a big difference in ammonia-nitrogen estimates (~0.6 g/d and 6 g/d) per capita is very likely to due to the fact that the studies reporting ~600 mg/d/capita were referring specifically to the concentration in urine, while the ~6 g/d/capita estimates were referring to the total household ammonia-nitrogen discharge in wastewater, in the latter case it could have been derived from household products and also include faecal inputs.

Since the studies reporting approx. 6 g/d/capita were performed for the purpose of wastewater-based epidemiology, which accounts for total ammonia-nitrogen discharge per capita rather than only ammonium urine concentration, and since the wastewater sampling point serves not only the hospital (approximately 500 patients' beds) but also the nurse

residential area in the hospital proximity, the ~ 6 g/d/capita studies were used for further population estimation.

2.2.2 Modelling Method 1 (Equation-based Model Following the Central Limit Theorem Premise)

As explained in "1.4.1 The Central Limit Theorem", let us assume that the probability distribution of the virus concentration in wastewater becomes normal for a sufficiently large population. According to the same CLT-based assumption, all the other variables related to this population will also approximate a normal distribution.

The estimate of the population will be centred at:

$$N = \frac{C_{RNA}\bar{F}}{\bar{\alpha}\bar{\beta}(1-\bar{\gamma})} \text{ (4)},$$

where N – estimate of population, C_{RNA} – measured virus concentration in wastewater (gc/L), F – mean volumetric flow rate (L/day), α – mean faecal load (g/day/person), β – mean faecal shedding (viral copies/g), γ – mean viral losses (Saththasivam et al., 2021).

Following the estimation of N, the next steps require the estimation of its standard deviation. As described by Saththasivam et al. (2021), initially, the mean total number of viral genome copies is calculated with the equation:

$$\bar{M}_{RNA} = N\bar{\alpha}\bar{\beta}(1-\bar{\gamma}) \text{ (5)},$$

where the variance was given by:

$$\delta M_{RNA}{}^2 = N\left((\bar{\alpha}^2+\delta\alpha^2)(\bar{\beta}^2+\delta\beta^2)((1-\bar{\gamma})^2+\delta\gamma^2) - \left(\bar{\alpha}\bar{\beta}(1-\bar{y})\right)^2\right) \text{(6)}.$$

The C_{RNA} described in (4) also follows a normal distribution, which variance of C_{RNA} is given by:

$$\delta C_{RNA}^2 = (\bar{M}_{RNA}^2 + \delta M_{RNA}^2)\left(\frac{1}{\bar{F}^2}+\frac{\delta F^2}{\bar{F}^4}\right) - \frac{\bar{M}_{RNA}^2}{\bar{F}^2} \text{ (7)}.$$

The above variance of C_{RNA} represents uncertainty of population origin, and it can be referred to as intrinsic $\delta C_{RNA}{}^2$. Uncertainty can also originate from laboratory measurements, and that can be referred to as measurement $\delta C_{RNA}{}^2$. When the latter is available, it can be included in the calculation of the total C_{RNA} variance with the formula:

$$(\delta C_{RNA}^2)_{Total} = (\delta C_{RNA}^2)_{Intrinsic} + (\delta C_{RNA}^2)_{Measurement} \text{ (8)}.$$

In this study, the total C_{RNA} variance is given only by the intrinsic $\delta C_{RNA}{}^2$ (7), because the laboratory analysis produced only two final replicates, which were used to calculate the mean C_{RNA} but, which, is not enough to calculate a meaningful measurement variance.

Using a fixed estimate of N in (4), we calculated the total variance of C_{RNA}. Under the assumption of a normal distribution for N, we can finalise by estimating its standard deviation with:

$$\delta N = \frac{\delta C_{RNA} \bar{F}}{\bar{\alpha}\bar{\beta}(1-\bar{\gamma})} \quad (9).$$

Finally, the two main outputs of the EMCLT are the population estimate, N, and its standard deviation, δN. The estimates represent the total population when the virus concentration is that of crAssphage, and they represent the infected population when the virus concentration is that of SARS-CoV-2, influenza A virus and NoVGII. The assumption of a normal distribution under CLT is essential to argument that the standard deviation is a valuable indicator of the probability dispersion of the population estimate (Madadizadeh et al., 2015).

2.2.3 Modelling Method 2 (Monte-Carlo-Bayesian Approach)

The MCBA can be summarised in the following way: Monte-Carlo calculations were applied to create a set of probability distributions of simulated virus concentrations for various numbers of the population, N. After this, the Bayes' rule was applied to obtain the posterior probability of N given the experimentally measured viral concentration.

Three MCBAs were employed, with the differences relating to (pseudo-) random number generation for viral losses. MCBA 1 assumes viral losses between [0; 1] corresponding to 0% to 100%. MCBA 2 assumes viral losses between a minimum X viral loss and maximum 100%. The minimum X viral loss was determined by estimating N with the measured concentration, followed by using a full Monte-Carlo simulation intermediate step where a viral distribution with viral losses [0; 1] was determined after which the minimum X viral losses were calculated with the formula:

$$\gamma_{min} = 1 - (C_{measured}/C_{estimated\_min}) \quad (10),$$

where γ_{min} – estimated minimum viral losses, $C_{measured}$ – measured viral concentration, $C_{estimated\_min}$ – minimum concentration as simulated with the full Monte-Carlo approach assuming viral losses between [0; 1]. The MCBA 2 continued by applying the estimated γ_{min} as the lower boundary for all the subsequential estimation of N values within the second MCBA simulation. The MCBA 3 was based on calculating a minimum X viral loss for each measured concentration individually through an intermediate full Monte-Carlo simulation where the losses were initially assumed to be between 0% and 100%. After that, the approach continued with using the estimated losses [γ_{min}; 1] for a second Monte-Carlo simulation where all the subsequential N values were calculated.

The steps can be summarised as follows for all the 3 variations of the MCBA:

1. An initial estimate for N, which will be referred to as $N_{estimate}$, was obtained similarly to the EMCLT (4):

$$N = \frac{C_{RNA}\bar{F}}{\bar{\alpha}\bar{\beta}(1-\bar{\gamma})} \ .$$

2. A dataset of N values was calculated as follows:

$$N_i = f \times m \times N_{estimate}, m = 1, 2, \ldots, j,$$

where f – the frequency of the dataset, which was chosen to be 0.1 due to the limited computational power, m – the position of the N value in the dataset being generated, $N_{estimate}$ – calculated in the previous step using (4), N_m – the simulated population value, j – the size of the dataset of N values. It is recommended to choose the value for the dataset size, j, so that during the last iteration when

$$m = j$$

the value of

$$f \times m$$

will be equal to two. This will ensure that the dataset of N values will include the maximum value of 2 x $N_{estimate}$. For example, assume a frequency of 0.1 and a dataset size of 20. During the last iteration, it will result in an N_m of 2 x $N_{estimate}$. This recommended setting ensures that, at a later stage, the dataset of N values will be wide enough to allow for an accurate inference of the probability distribution of the population numbers based on a measured virus concentration.

The frequency was set at 0.1 but was also increased to 0.01 or 0.001 when the computational power permitted, specifically for small $N_{estimate}$ <500 individuals. Also, in the case of $N_{estimate}$ being a relatively low value, such as 25 individuals, the algorithm was adjusted to establish the frequency of 1 individual, e.g., the N dataset being [0, 1, 2, 3, 4 …, 48, 49, 50]. The maximum value of the N dataset can also be increased from 2 x $N_{estimate}$ to, e.g., 5 x $N_{estimate}$, depending on the maximum experimentally measured concentrations since at times the measured concentrations can have outliers outside the normal range.

3. For the obtained values N_1 to N_m in step 2, viral concentrations were calculated for each individual N separately as follows:

$$C_{RNA} = \frac{\sum_{i=1}^{N} \alpha_i \beta_i (1-\gamma_i)}{F_i} \ (11),$$

where "*i*" is a counter for the number of infected individuals up to the number of individuals as indicated by the value of N.

Sets for α_i, β_i, γ_i, F_i variables were generated (pseudo-) randomly within a predefined range, following a uniform distribution. For α_i, the range depended on the virus, and was within the minimum and maximum viral concentration shed in faeces as established by several articles. For β_i, the range was 51-796 g/cap/day (Rose et al., 2015). For γ_i, the range was 0 to 1. For the generation of one Monte-Carlo datapoint (one simulated concentration), α_i, β_i, γ_i, F_i datasets were generated, containing random values the count of which corresponded to the total number of individuals, N. The number of Monte-Carlo simulations or datapoints (M) was chosen manually, with at least 250 simulations. For example, for the MCBA 1, 2000 simulations were chosen as the most feasible number of iterations in terms of computational power,

The generated Monte-Carlo datapoints will represent the probability distribution of C_{RNA} for each individual value of N from the dataset obtained in step 3.

4. In the case of MCBA 2 and 3, the probability distribution obtained in step 3 was used to estimate a minimum value of viral losses, using (10). In the case of MCBA 1, this step is omitted.

   For MCBA 2, one unique γ_{min} was obtained, using (10), where $C_{measured}$ is the average of all measured viral concentrations in wastewater (Table 5). In the case of MCBA 3, γ_{min} was calculated individually for each measured viral concentration. For both, MCBA 2 and 3, step 3 was then rerun using the new lower boundary for viral losses for random number generation, while the upper boundary remained 1, corresponding to 100%.

   The advantage of MCBA 2 is that it is almost twice as efficient in terms of the algorithm computational time when comparing to MCBA 3, due to estimating γ_{min} only once. Whereas, MCBA 3 may prove more flexible in the estimation of γ_{min}, but will require more computational power.

5. In order to obtain a probability distribution instead of a datapoints distribution, the total number of Monte-Carlo data points within an interval was divided by the total number of Monte-Carlo data points.

   Therefore, a histogram can be obtained giving the probability distribution (or density) of C_{RNA} for a given value of N.

6. Since C_{RNA} was a measured value and not a random one, the Bayes' rule was used to obtain a probability distribution of N given the known measured value of C_{RNA}.

The probability that the number of infected individuals is N based on the measured concentration (C_{RNA}) was calculated with the formula:

$$P(N|C_{RNA}) = \frac{P(C_{RNA}|N)P(N)}{P(C_{RNA})} \quad (12).$$

The probability of obtaining the measured concentration given a value N, expressed as $P(C_{RNA}/N)$, was calculated by taking the number of Monte-Carlo points in an interval where the measured C_{RNA} fits and dividing it by the total number of Monte-Carlo points. The interval width, in this case, depended on a value, which was chosen manually, and we will refer to as acceptable deviation, the interval width corresponding to C_{RNA} ± acceptable deviation. The chosen acceptable deviation depends on the frequency of the simulated data points. For example, viruses with higher wastewater concentration and faecal virus shedding had a more disperse distribution for the same number of MCBA simulations than a virus with lower concentration and faecal shedding. Therefore, such viruses require a higher value for acceptable deviation. As a result, the earlier mentioned calculation of $P(C_{RNA}/N)$ will be up to the uncertainty which equals this interval width.

For influenza A virus, the acceptable deviation was 10^2 (all approaches). For norovirus GII, the acceptable deviation was 10^4 (all approaches). For SARS-CoV-2, the acceptable deviation was 10^4 (MCBA 1 and 3) and $2x10^3$ (MCBA 2). For crAssphage, the acceptable deviation was 10^5 (MCBA 1), $5x10^4$ (MCBA 2) and 10^4 (MCBA 3). These are arbitrarily chosen values which were found to be the most suitable for this experiment as to be able to include and count datapoints values within a highly dispersed range. However, the higher the number of simulations, the smaller the acceptable deviation, and the MCBA 3 allows for a smaller acceptable deviation as in the case of crAssphage since the simulations are generated for the N values that are in a proximity of the N estimate as determined by the measured concentration.

Since C_{RNA} was measured, there was no uncertainty regarding its value, the probability being 100% and the denominator being 1. The formula is reduced to:

$$P(N/C_{RNA}) = P(C_{RNA}/N) \times P(N),$$

The right hand side of the equation was calculated as follows. For every value of N, the number of Monte-Carlo points that were within the range of the measured C_{RNA} and its chosen acceptable deviation were counted. The obtained counts for each individual value of N were then divided by the total number of counts among all values of N, which can be plotted as the probability distribution for N given the measured C_{RNA}.

The modelling was performed in the Python programming language (version 3.10.0), and the code for MCBA 1, 2 and 3 can be found in the Appendix, Supplementary Materials (Figure S1-S3). The mean, variance, standard deviation, and a 99% confidence interval of the N probability datasets by measured concentration was also obtained as detailed in the MCBA Python codes.

2.2.4 Gaussian Filter and Polynomial Trendline

Two main approaches for data transformation to visualise the underlying trends were a Gaussian filter and a polynomial trendline.

A Gaussian filter is a tool used to compute derivatives of an image, or of y-axis values in this experiment, thereby smoothing and reducing the noise. Since all the coding is performed in Python, the library *SciPy* was used, specifically *scipy.ndimage*. To import the package, the following command is applied "*from scipy.ndimage.filters import gaussian_filter1d*", where *scipy.ndimage.gaussian_filter1d* takes the following arguments: *input* and *sigma*, for example "*output_dataset = gaussian_filter1d(input_dataset, sigma)*". The argument *sigma* is the standard deviation for the Gaussian kernel, which can be perceived as the "smoothing" factor (SciPy v1.9.0 Manual, 2022). In this case the function implements a 1-dimension filter.

For generating a polynomial trendline in Python, the library "NumPy" was used, specifically "*numpy.poly1d*" and "*numpy.polyfit*" packages. The polynomial order in all cases was chosen to be 10. Additionally, from the "*sklearn.metrics*" package it was used the function "*r2_score*" so that to estimate R^2. The used function can be observed in Figure S5.

2.2.5 Comparison of Estimates with Clinical Data

For ethical and patient confidentiality reasons, it was not possible to obtain the data of clinically diagnosed in-care cases of SARS-CoV-2, influenza A virus and NoVGII at the Ysbyty Gwynedd hospital, hence information about the diagnosed infectious cases for Wales and Gwynedd was researched in publicly available sources, such as public official reports issued by Public Health Wales, UK Health Security Agency, or Office for National Statistics. We assume that cases in the community are closely correlated with hospitalisations.

The number of reported infectious cases were divided by the total number of the population as to estimate the prevalence in the community, whereas the viral cases estimated with the EMCLT or MCBA were divided by the population number estimated with the crAssphage concentration, and the value was multiplied by 100 for both to obtain the percentage. The datasets were integrated into data frames with the library "*pandas*". The proportion obtained previously was converted to percent changes with the function "*pct_change()*" as to obtain the relative change in time of the reported cases proportion and estimated cases proportion. The results were processed with the Gaussian filter of varying *sigma* argument (see "Gaussian Filter and Polynomial Trendline") section.

The Pearson pairwise correlation between the percent changes of the reported and estimated cases proportion was computed with the function "*DataFrame.corr()*".

A second method to capture and quantify synchronicity was dynamic time warping (DTW), usually used in economic time-series (Franses & Wiemann, 2020). The employed code is provided in Figure S4. The used library is "*dtw*" and specifically the function "*accelerated_dtw*". Due to the collected data being discontinues, DTW was applied only for data collected after 10th Jan 2021. For the visualisations of the DTW results, a map is provided which describes the warping path and the distance matrix.

2.2.6 Association between Wastewater Physico-chemical Characteristics and Detected Virus Concentration

To model the impact of wastewater characteristics on virus recovery, a Sugeno adaptive neuro-fuzzy interference system (ANFIS) was used, specifically fuzzy subtractive clustering. Generally, subtractive clustering organises data into clusters based on the similarity of the data to the cluster through specific membership functions (MFs), such as Gaussian MFs (Keshavarzi et al., 2017). Since MFs also include the association strength of datapoints to various clusters, one datapoint can belong to multiple clusters but of varying levels of strengths. For example, assume the system has 10 clusters expressed as Gaussian MFs. The datapoint x from input dataset 1 can belong to input cluster number 6 with a strength of 80% and, at the same time, it can belong to cluster 7 with a strength of 85%, while not belonging at all to other input clusters. If input 1 was the only input we had, then datapoint x will correspond to output cluster 6 and 7 with the same level of membership strength, 80% and 85%. If there was input dataset 2, and we take a datapoint y, let's assume it belonged to input cluster 7 with a strength of 60% and to cluster 8 with a strength of 65%. If the relationship between inputs 1 and 2 is characterised by the expression "and" (Figure 4), datapoints x and y will result in association with only the output cluster 7 with a strength of 60%. During the final layers of defuzzification and aggregated output, the algorithm sums up the MF strengths of those two datapoints for all clusters, and then derives a corresponding output value. The full structure of the employed ANFIS can be observed in Figure 4.

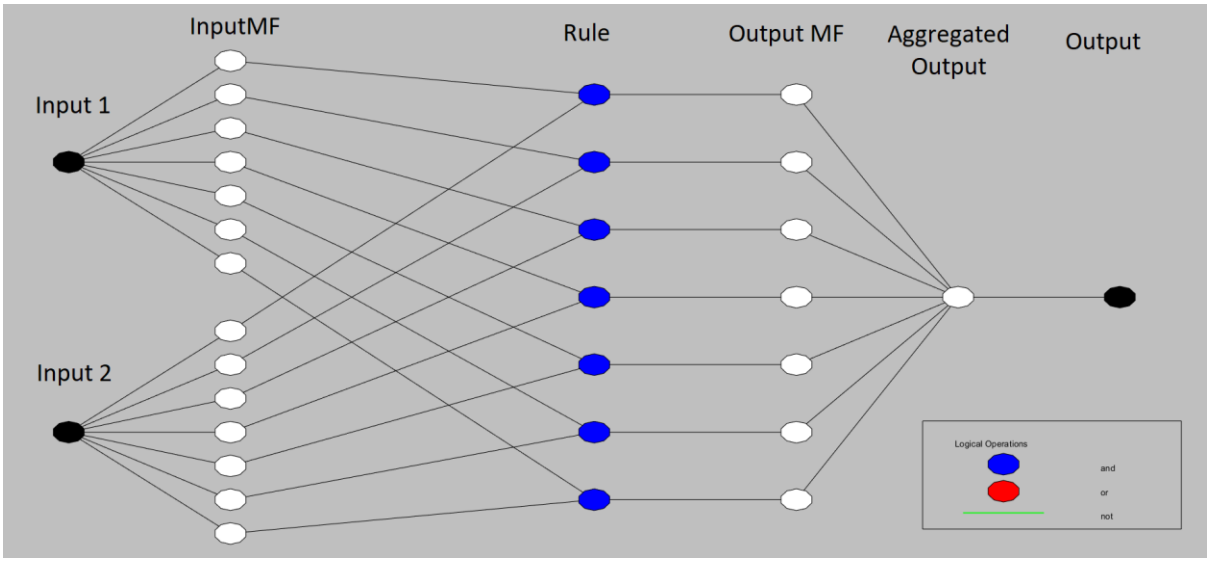

Figure 4. Structure of the Adaptive Neuro-Fuzzy Interference System (ANFIS) based on a Subtractive Clustering with Two Input Variables. *The ANFIS structure incorporates 5 layers:* ***fuzzification layer*** *(receives the input data and applies the membership functions (MF) as node functions),* ***rule layer, also called interference layer*** *(the node output indicates the firing strength of the rule),* ***normalisation layer*** *(normalises the firing strength of each rule),* ***defuzzification layer*** *(receives the normalised values and the consequence parameters),* ***total output layer*** *(calculates the sum of the rules' outputs and returns it as a final output). The rule layer and the normalisation layer are represented as the layer "Rule". The defuzzification layer "Output MF" is an adaptative node containing functions that indicate the contribution or strength of the membership rules' towards the final aggregated output.*

Coding in MATLAB R2022a (The Mathworks, Natick, USA) and the embedded Neuro-Fuzzy Designer were used to develop the model. Two variables were used as input data at the same time and related to the virus concentration (Table 4). In the employed system, the two inputs were related to each other with an "*and"* logical operator. The model training was comprised of 2 stages: initial FIS training and FIS tuning. The initial FIS training was performed with 2 statements:

"*fisOpt = genfisOptions("SubtractiveClustering", "ClusterInfluenceRange", 0.5, "SquashFactor", 0.5, "AcceptRatio", 0.5, "RejectRatio", 0.01);*

fis = genfis(datain_train, dataout_train, fisOpt);" .

The obtained FIS was tuned with the statements:

"*opt = anfisOptions('InitialFIS', fis, 'EpochNumber', 40, 'InitialStepSize',0.1);*

fis2 = anfis([datain_train dataout_train], opt);".

For each individual combination of input and output parameters (Table 6), various combinations of the squash factor, accept ratio, reject ratio, epoch number and initial step size were trialled to achieve the lowest error. An important aspect of the tuning stage was to not overfit the model to the training data, hence the epoch number was kept relatively low at approximately 30-40 epochs.

To obtain an initial evaluation of the model, the statement "*fuzout = evalfis(fis, datain_train);*" was used where *"fuzout"* are the modelled virus concentrations and "*datain-train"* are the measured virus concentrations, both being 1-D arrays. Therefore, the variables *"fuzout"* and "*datain-train"* can be plotted against each other.

To exclude the outliers, only the data within 2 standard deviations from the mean value was used, corresponding to approximately 95% of the distribution.

The Pearson correlation coefficient (R value) was calculated with the MATLAB function "*corrcoef(A,B)*" where the variables A and B are the modelled and measured viral concentrations, corresponding to the formula:

$$\rho(A,B) = \frac{1}{N-1} \times \sum_{i=1}^{N} \left(\frac{A_i - \mu_A}{\sigma_A}\right) \times \left(\frac{B_i - \mu_B}{\sigma_B}\right) \quad (13),$$

where μ_A and σ_A are the mean and standard deviation of A, and μ_B and σ_B are the mean and standard deviation of B. The A values corresponds to the measured values whereas the B values correspond to the predicted values.

The root mean square error (RMSE) was calculated with the expression:

$$RMSE = \sqrt{\sum_{i=1}^{N} \frac{(A_i - B_i)^2}{N}} \quad (14).$$

3. Results

3.1 Characteristics of Wastewater Physico-chemical Parameters

The results for the wastewater physico-chemical parameters pH, electrical conductivity, turbidity, ammonium and phosphate are recorded in Table 4. Overall, the data demonstrated a significant variation rather than data being centred around a mean value. The change of the sampling approach from 2-hour 12 composite samples a day to a one 24-hour composite sample a day did not influence the chemical parameters results significantly. The mean and standard deviation values between the two sampling regimes were relatively close to each other, and the differences were likely to result from day-to-day variability.

Table 4. Physico-chemical Parameters Results of Hospital-Derived Wastewater

Physico-chemical Parameter	Mean	SD	Min	Max
pH	7.69	0.46	6.21	8.75
Electrical Conductivity (µS/cm)	917	529	172	7179
Turbidity (NTU)	119	217	2.5	2985
Ammonium Concentration (mg/l)	32.3	17.8	0.1	84.8
Phosphate Concentration (mg/l)	3.45	1.91	0.03	12.56

NTU – Nephelometric turbidity units, SD – standard deviation.

3.2 Virus Detection

The qPCR or RT-qPCR results are recorded in Table 5. CrAssphage had a detection rate of 100% and a mean concentration in wastewater of 7.36 $\log_{10}$ gc/l, indicating that all samples contained faecal material, and the suitability of using crAssphage bacteriophage as a human marker. SARS-CoV-2 had the second highest detection rate and norovirus GII had the third (Table 5). This indicates that SARS-CoV-2 and NoVGII were extensively present within the hospital catchment area (sewershed) during the study period (Table S3; Table S4). The SARS-CoV-2 detection rate increased by 8.5% and the NoVGII detection rate decreased by 29% in the second sampling period. The mean concentrations were also not consistent between the two sampling regimes. It decreased from 7.26 $\log_{10}$ gc/l in the first sampling regime to 6.84 $\log_{10}$ gc/l in the second regime for SARS-CoV-2, and from 7.43 $\log_{10}$ gc/l to 7.04 $\log_{10}$ gc/l for crAssphage. For NoVGII, the mean concentration increased from 5.58 $\log_{10}$ gc/l to 6.33 $\log_{10}$ gc/l. In contrast, Influenza A virus had the lowest detection rate of 8.8% at relatively low concentrations, all positive detections taking place in 2021 during the first sampling regime (Table S1). The positive samples, however, did not correlate well with the seasonality of influenza A as there were similar or higher number of acute respiratory incidents and higher hospital admission rates for Flu A in UK in the second sampling regime, with the positivity rate peaking in the week of 07/03/2022 (UKHSA, 2022c). Other possible reasons include the hospital having a relatively much lower admission rate, low viral shedding rates (Table 7) and a relatively lower viral resistance to environmental stress (i.e., loss of signal due to RNA degradation in the wastewater).

Table 5. Wastewater Virus Detection Parameters

Virus Type	Number of Samples	Detection Rate (%)	Mean	Median	SD	Q1	Q2	Q3	IQR
Influenza A Virus	352	8.8	4.21	2.82	4.69	2.59	2.82	3.39	3.32
Norovirus GII	113	74.3	6.09	4.87	6.61	4.33	4.87	5.56	5.54
SARS-CoV-2	364	78.6	4.49	3.38	5.23	2.80	3.38	4.02	3.99
CrAssphage	376	100	7.36	6.31	7.82	5.88	6.31	6.92	6.88

The mean, median, standard deviation (SD), 25th percentile (Q1), 50th percentile (Q2), 75th percentile (Q3), and the interquartile range (IQR) were calculated based only on the positive results and are expressed in log_{10} gene copies per litre.

3.3 Association between Wastewater Physico-chemical Characteristics and Detected Concentrations of SARS-CoV-2 and CrAssphage

Various ANFIS structures, such as Gaussian functions or subtractive clustering, were trialled for different input combinations (Table 6) to identify the most suitable structure.

It was decided to implement subtractive clustering with subsequent conversion to a 3-D representation for the evaluation of the wastewater characteristics for the following reasons. The logic of subtractive clustering (see Materials and Methods section 2.2.6) is the most suitable because data is organised in clusters based on similarity and, hence points to trends, and due to the fact that this method could be applied consistently for all paired input combinations as opposed to different structures fitting different input combinations.

The performance results of the initial FIS and the tuned FIS are presented in Table 6. When the tuning stage had resulted into an unacceptable error rate and a high RMSE, the 3-D model of the initial untuned FIS was presented for visual evaluation.

Interestingly, the performance evaluation and the correlation between all the modelled and measured values (Figure S17 – Figure S26) indicates that the model consistently underestimated at higher concentration and overestimated at lower concentrations the virus output concentrations for both SARS-CoV-2 and crAssphage, potentially due to a non-uniform data distribution and an insufficiently large training dataset.

Table 6. ANFIS Model Input-Output Combinations and Their Performance.

Virus Output	Input Parameters	Initial FIS Training Data		Tuned FIS Training Data	
		R	RMSE	R	RMSE
SARS-CoV-2	pH and ammonium	0.352	32476.7	0.349	32512.0
	pH and phosphate	0.448	29107.5	0.596	26135.2
	ammonium and phosphate	0.480	28984.3	0.418	30010.0
	electrical conductivity and turbidity	0.166	34342.4	0.168	34328.6
	electrical conductivity and pH	0.286	32946.9	0.392	31633.1
	crAssphage and ammonium	0.174	34716.7	-0.006	4.3E+11
	crAssphage and phosphate	0.193	32438.3	0.182	32664.7
	crAssphage and turbidity	0.111	34914.8	0.017	2201382.7
	crAssphage and pH	0.190	34144.9	0.195	34115.6
	crAssphage and electrical conductivity	0.135	34657.6	0.135	34657.7
CrAssphage	pH and ammonium	0.510	26755635.2	0.553	25913118.1
	pH and phosphate	0.485	26168140.2	0.515	25651357.2
	ammonium and phosphate	0.535	26068188.0	0.557	25632331.8
	electrical conductivity and turbidity	0.445	27137828.9	0.454	26994247.3
	electrical conductivity and pH	0.355	27471025.5	0.507	25334406.7
	SARS-CoV-2 and ammonium	0.271	24727624.6	0.335	24199922.1
	SARS-CoV-2 and phosphate	0.300	23123446.9	0.300	23123522.8
	SARS-CoV-2 and turbidity	0.360	23904951.3	0.373	23771063.2
	SARS-CoV-2 and pH	0.424	22235103.4	0.372	22794337.3
	SARS-CoV-2 and electrical conductivity	0.269	23071293.7	0.284	22966868.2

The input parameters and virus output refer to the input pair and single output combinations used for the training of the adaptive neuro-fuzzy interference system (ANFIS) which occurred in 2 stages: initial FIS training and FIS tuning. The performance results for pre-tuning and post-tuning model are presented, indicating how well the measured output correlated with the modelled output.

The obtained 3-D models can be observed in Figure 5 – Figure 12. When visually comparing the models where SARS-CoV-2 was as an output against crAssphage as an output, it is essential to consider that the crAssphage axis magnitude reached 8 $\log_{10}$ gc/l whereas for SARS-CoV-2, it typically reached 4-5 $\log_{10}$ gc/l.

The highest SARS-CoV-2 modelled recoveries were observed at pH 7.6 – 8.3 and ammonium >25 mg/l (Figure 5). High crAssphage modelled concentrations were observed at pH 6.8 – 7.2 and ammonium 0 – 50 mg/l, the virus concentration increasing drastically at an ammonium concentration >60 mg/l (Figure 5).

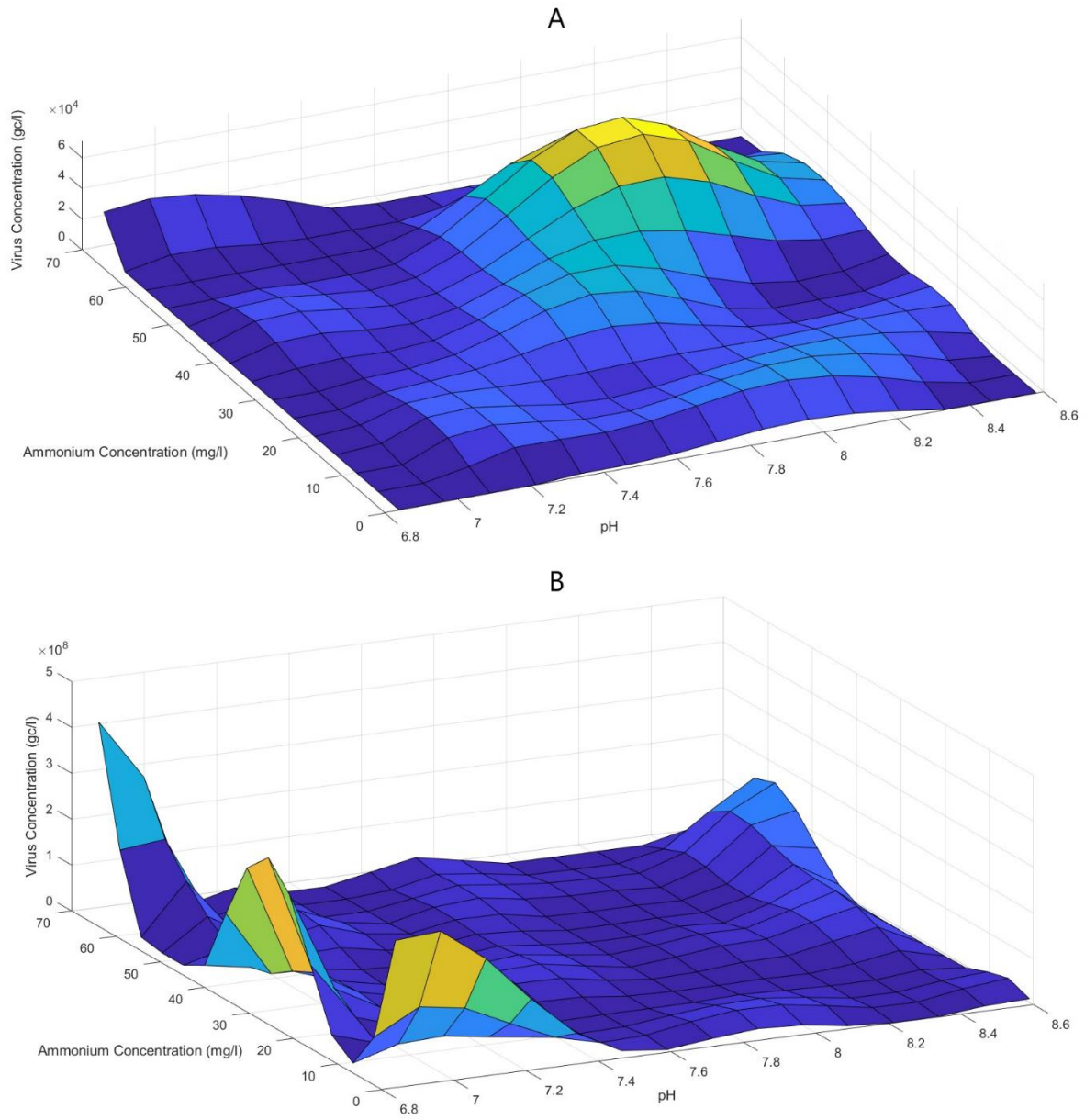

Figure 5. The Modelled Impact of Ammonium and pH on Virus Concentration. *The modelling is based on a Sugeno adaptive neuro-fuzzy interference system (ANFIS), the sub-method being subtractive clustering followed by tuning with the training data, where the inputs are ammonium and pH, and the outputs are the SARS-CoV-2 concentration (A) or crAssphage concentration (B).*

With an increasing phosphate concentration, specifically >3.5 mg/l, and pH of 7.6 – 8.4, also increased the SARS-CoV-2 concentration (Figure 6), whereas for crAssphage, the concentrations were the highest at pH 8.15 – 8.5, the concentration increase having occurred uniformly to phosphate starting at 1.5 mg/l. CrAssphage also solely demonstrated high modelled recoveries at pH 6.9 – 7.5 and phosphate concentrations of 2.50 – 6.75 mg/l.

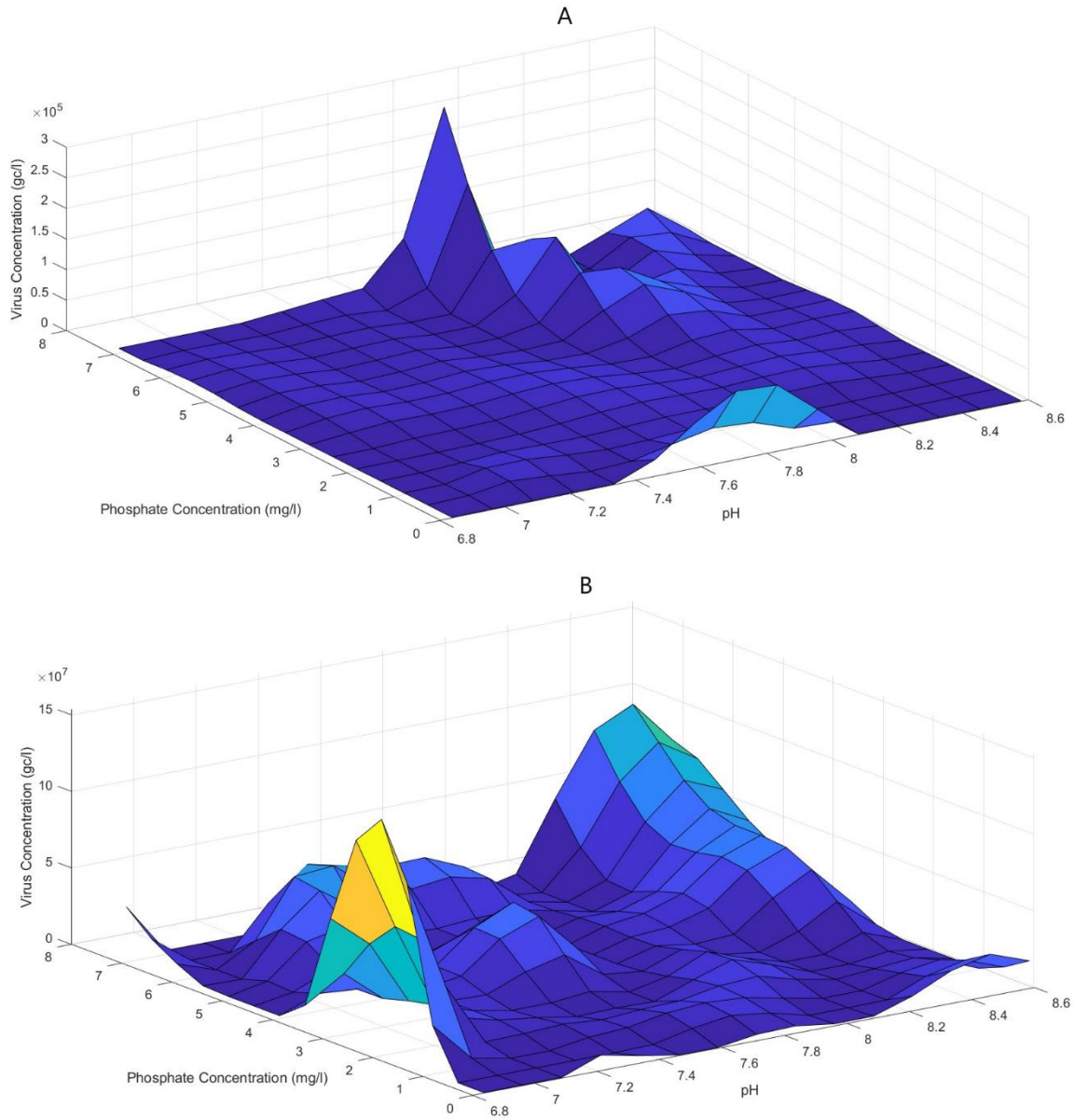

Figure 6. The Modelled Impact of Phosphate and pH on Virus Concentration. *The modelling is based on a Sugeno adaptive neuro-fuzzy interference system (ANFIS), the sub-method being subtractive clustering followed by tuning with the training data, where the inputs are phosphate and pH, and the outputs are the SARS-CoV-2 concentration (A) or crAssphage concentration (B).*

Interestingly, the impact trends of ammonium and phosphate (Figure 7) differed for the two viruses. The SARS-CoV-2 recovery trend demonstrated a continuous increase at a phosphate concentration of >6.5 mg/l and ammonium concentration of 10 – 50 mg/l, and a much significant virus concentration increase at ammonium concentration of >10 mg/l and low phosphate concentration, 0 – 1 mg/l. A relatively lower recovery (up to 4.5 log10 gc/l) with a dynamic increase occurred simultaneously with the increase in ammonium and phosphate concentration. The impact trend on crAssphage recovery (Figure 7) demonstrated that with an increasing phosphate concentration increased the virus concentration regardless of the ammonium concentration, suggesting no direct correlation between ammonium and crAssphage. However, a higher peak can be observed at ammonium concentrations of approximately 56 mg/l.

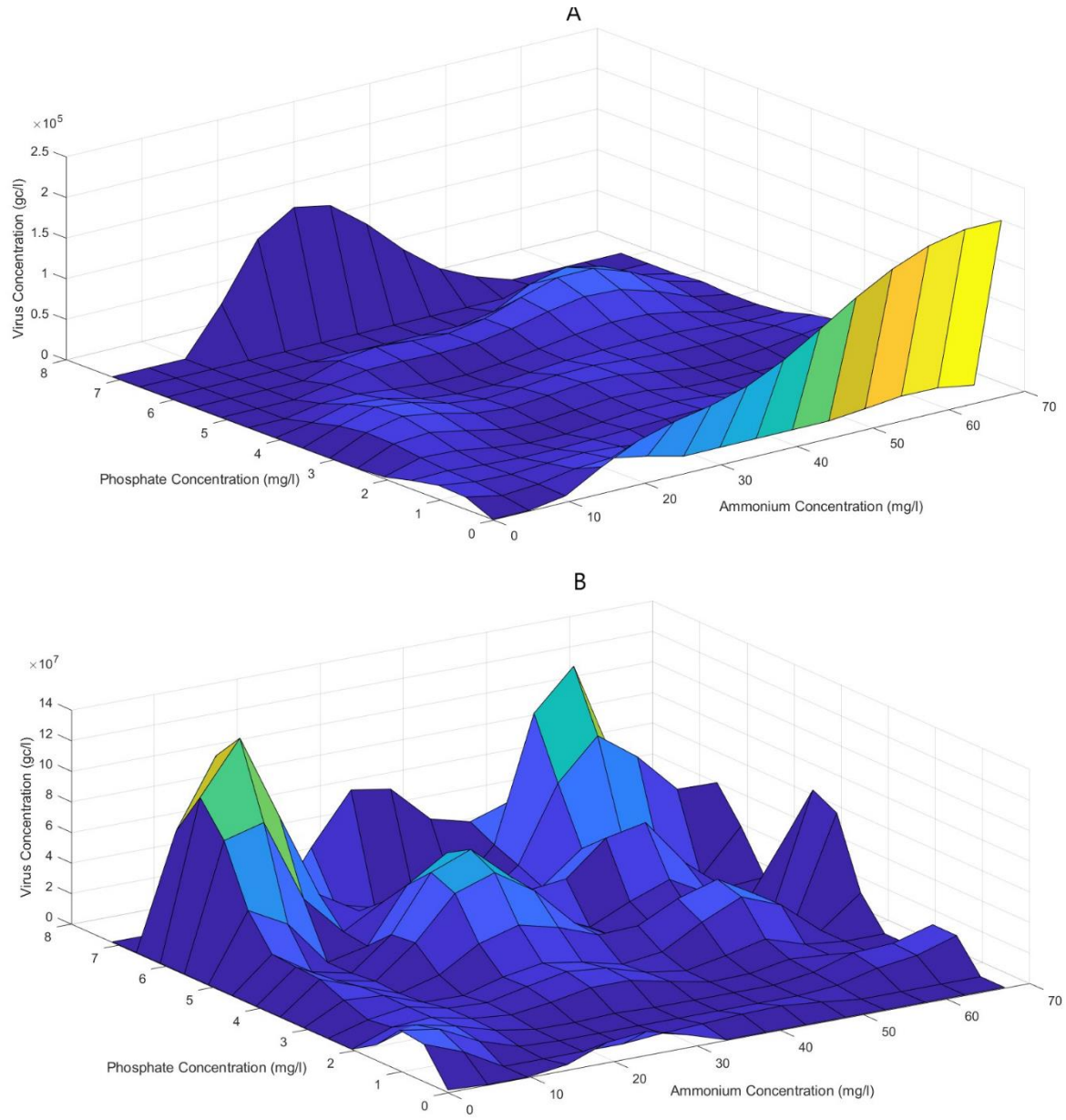

Figure 7. The Modelled Impact of Ammonium and Phosphate on Virus Concentration. *The modelling is based on a Sugeno adaptive neuro-fuzzy interference system (ANFIS), the sub-method being subtractive clustering followed by tuning with the training data, where the inputs are phosphate and ammonium, and the outputs are the SARS-CoV-2 concentration (A) or crAssphage concentration (B).*

The turbidity and electrical conductivity modelled impact (Figure 8) demonstrated that an increase in turbidity did not result in an increase in virus recovery whereas an increase in electrical conductivity resulted in a positive trend of virus concentration. The optimal conditions for a higher SARS-CoV-2 recovery as modelled below are approximately <200 NTU and an increasing electrical conductivity with the peak having been at approximately 1550 – 1825 µS/cm, whereas for crAssphage, a more significant concentration increase happened at 100 – 400 NTU and at higher electrical conductivities.

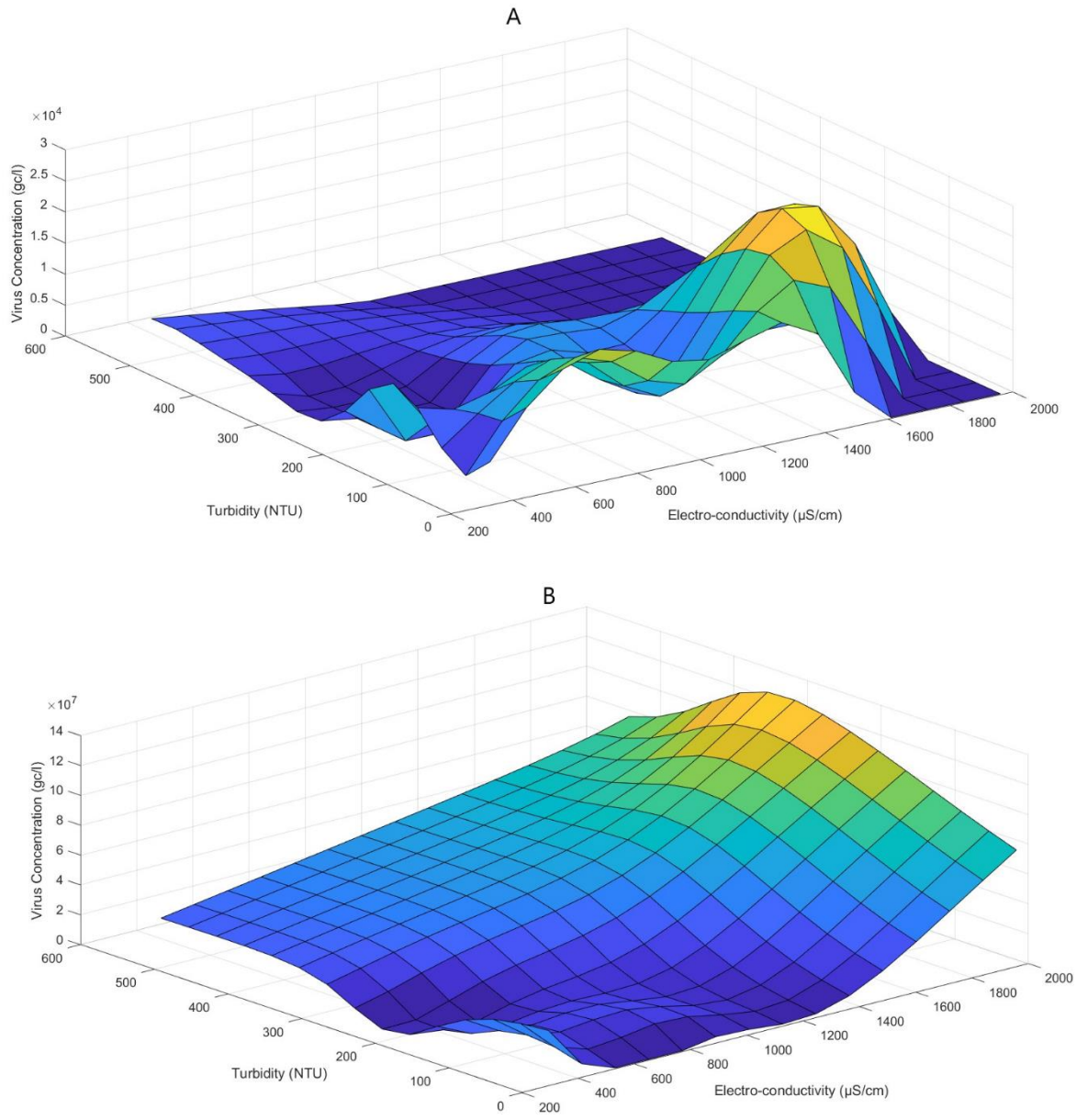

Figure 8. The Modelled Impact of Turbidity and Electrical conductivity on Virus Concentration. *The modelling is based on a Sugeno adaptive neuro-fuzzy interference system (ANFIS), the sub-method being subtractive clustering followed by tuning with the training data, where the inputs are turbidity and electrical conductivity, and the outputs are the SARS-CoV-2 concentration (A) or crAssphage concentration (B).*

Interestingly, in the case of both viruses, the modelled virus concentration was static for specific electrical conductivities when the turbidity increased. SARS-CoV-2 concentration was relatively static at >350 NTU and 0 – 600 µS/cm, whereas crAssphage concentrations at lower electrical conductivities started increasing at approximately 180 NTU, remaining static at >280 NTU, whereas at higher electrical conductivities, it remained static at >450 NTU. In case of crAssphage, the specific values at which the virus concentration was static depended on the electrical conductivity (i.e., an increasing electrical conductivity also increased the value at which the virus concentration is static).

An increase in electrical conductivity at a specific range of pH correlated well with the modelled virus concentrations (Figure 9).

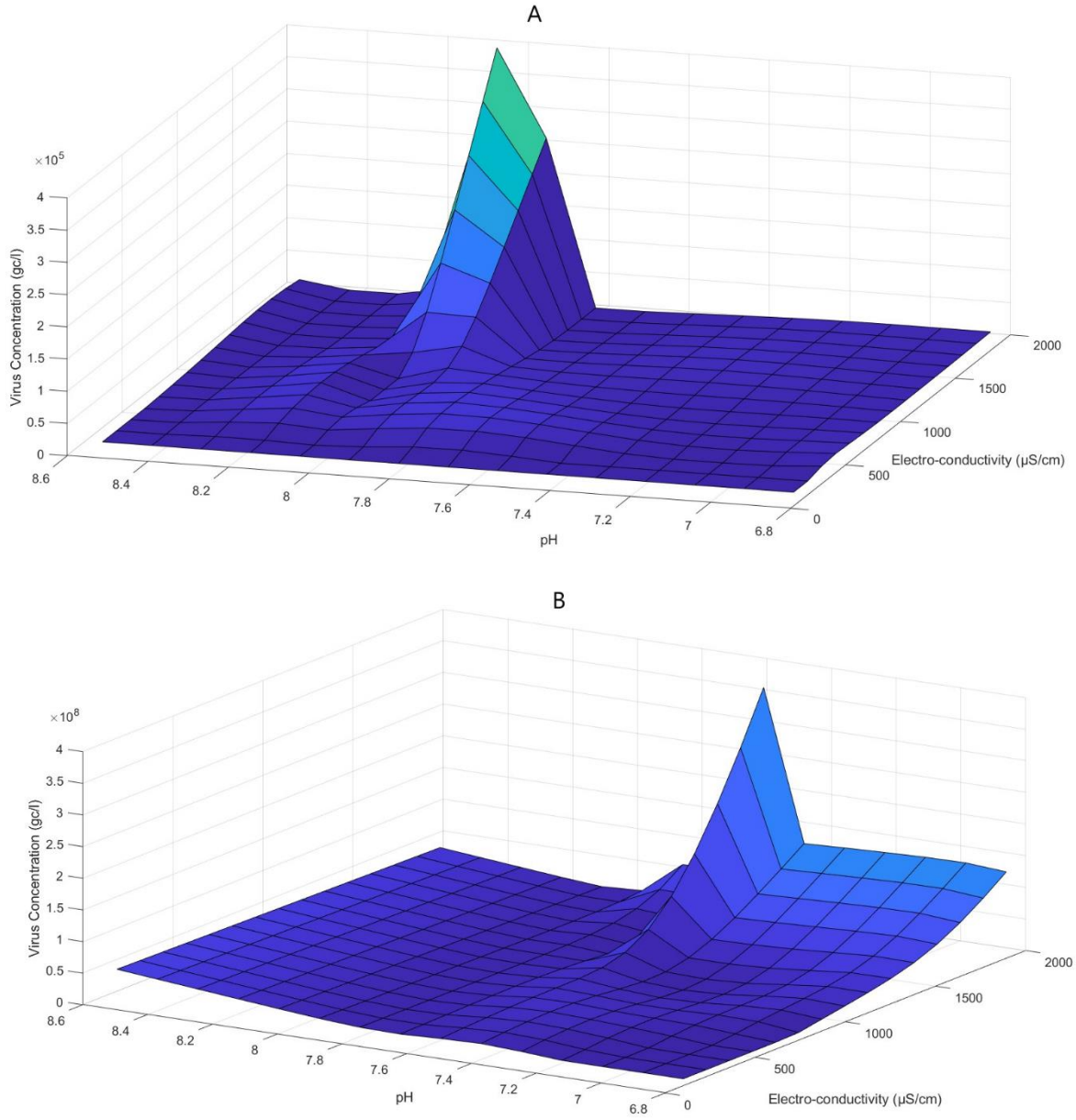

Figure 9. The Modelled Impact of Electrical conductivity and pH on Virus Concentration. *The modelling is based on a Sugeno adaptive neuro-fuzzy interference system (ANFIS), the sub-method being subtractive clustering followed by tuning with the training data, where the inputs are pH and electrical conductivity, and the outputs are the SARS-CoV-2 concentration (A) or crAssphage concentration (B).*

Both SARS-CoV-2 and crAssphage modelled concentrations demonstrated a trend of increasing concentrations simultaneously with electrical conductivity starting at approximately 750 µS/cm, the optimal pH being 7.8 – 8.2 for the former and 6.8 – 7.8 for the latter. In the case of crAssphage it should also be noted that the concentration peak was specifically between pH 7.4 – 7.6.

Ammonium and one of the two virus concentrations impact on the second virus concentrations differed for both modelled viruses' concentrations (Figure 10).

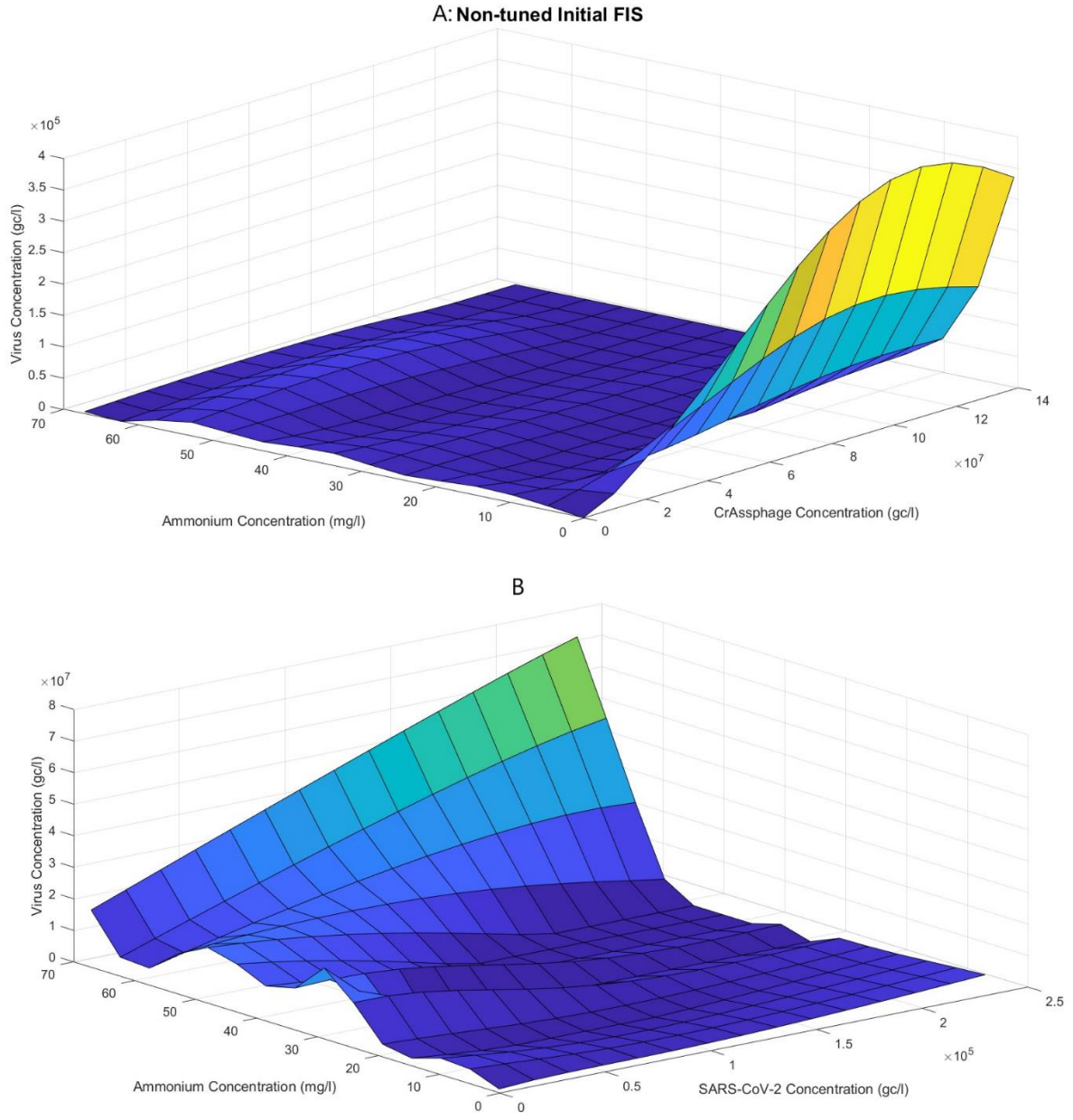

Figure 10. The Modelled Impact of Ammonium and CrAssphage (or SARS-CoV-2) on SARS-CoV-2 (or CrAssphage) Concentration. *The modelling is based on a Sugeno adaptive neuro-fuzzy interference system (ANFIS), the sub-method being subtractive clustering followed by tuning with the training data, where the inputs are ammonium and crAssphage concentration (panel A) or ammonium and SARS-CoV-2 concentration (panel B), and the outputs are the SARS-CoV-2 concentration (A) or crAssphage concentration (B). Due to an unacceptable error of the tuned model B, the non-tuned initial FIS is presented for panel B.*

Due to an increased error associated with the tuning process for the model, it was presented the untuned FIS which also should illustrate the model's trends (Figure 10, Panel A), however, it is expected to be less accurate in relation to fitting the training data. The modelled SARS-CoV-2 recovery demonstrated increased values of up to 4.7 $\log_{10}$ gc/l at ammonium concentrations of 45 – 58 mg/l and crAssphage concentrations up to 7.95 $\log_{10}$ gc/l. Interestingly, the modelled SARS-CoV-2 concentration increased simultaneously with the crAssphage concentration within the ammonium concentration range of 0 – 10 mg/l, while at higher ammonium concentrations, the trend was not maintained. The increased SARS-CoV-2 concentration at ammonium of 0 – 10 mg/l displayed similarities to the previously observed models' trends (Figure 5; Figure 7), although a relatively lower modelled SARS-CoV-2 concentration, 4.3-4.6 $\log_{10}$ gc/l (Figure 5; Figure 7), as opposed to 5.6 $\log_{10}$ gc/l (Figure 10). These observations differ from the modelled crAssphage concentration (Figure 10) which showed that with an increase in the detected SARS-CoV-2 concentration and an increase in the ammonium concentration also increased the crAssphage concentration. However, the modelled crAssphage concentration still increased gradually and simultaneously with the increase in ammonium concentration, albeit slower, at low SARS-CoV-2 concentrations (0 – 4.7 $\log_{10}$ gc/l).

Some of the trends observed in the modelled impact of phosphate and one of the viruses on the second virus did agree (Figure 11). At a phosphate concentration of 0 – 2 mg/l, and an increasing crAssphage (or SARS-CoV-2) concentration also increased the SARS-CoV-2 concentration (or crAssphage respectively). There was also a common trend of increased modelled virus recovery at phosphate concentrations of 4 – 7 mg/l and lower concentrations for the input virus.

An important observation is that when SARS-CoV-2 and crAssphage were switched among each other in terms of being the output in the ANFIS model, only a few of the trends were preserved, pointing out to the differences in the optimal conditions required for relatively higher SARS-CoV-2, or respectively, higher crAssphage recoveries (Figure 5 – Figure 12).

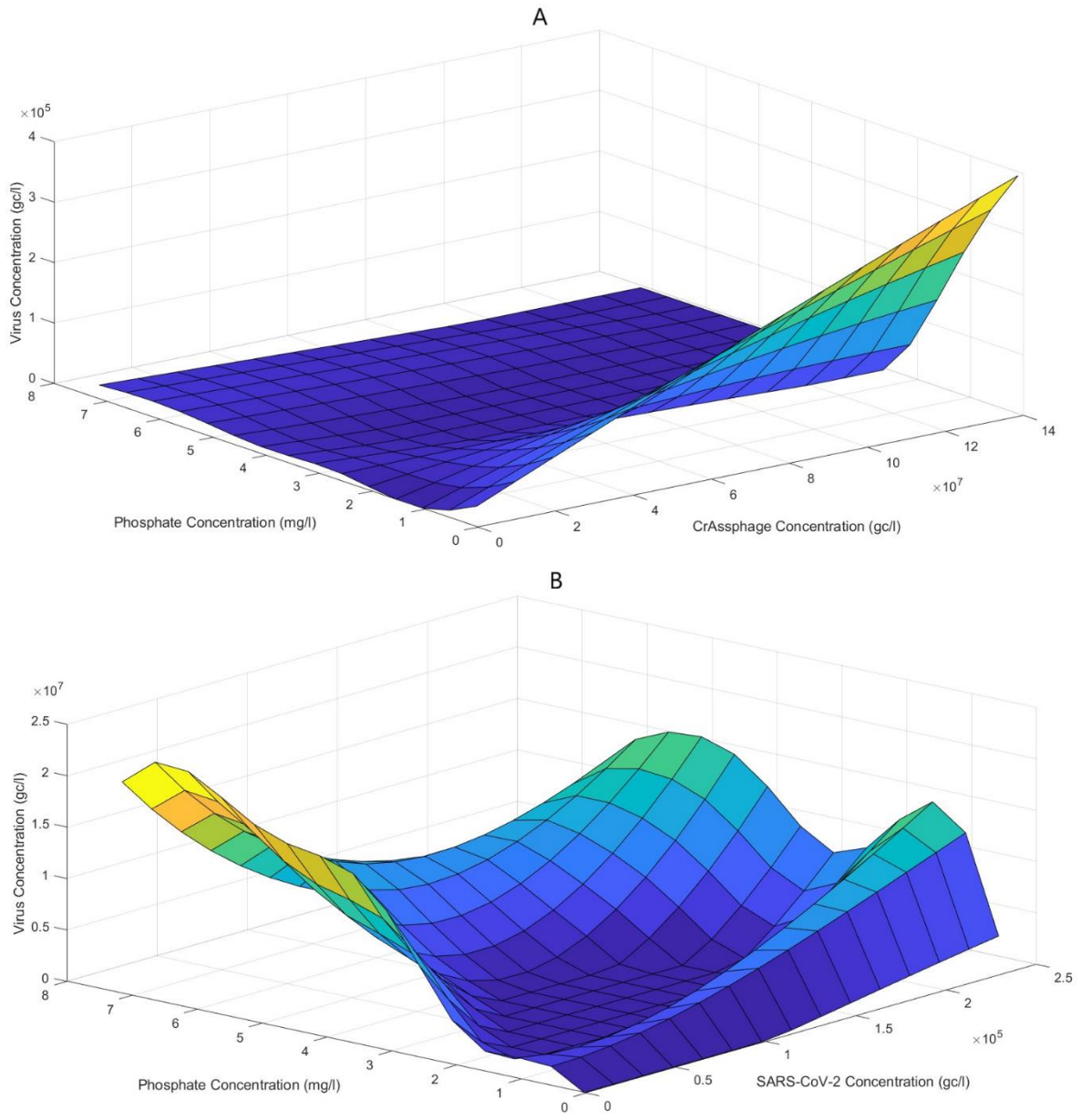

Figure 11. The Modelled Impact of Phosphate and CrAssphage (or SARS-CoV-2) on SARS-CoV-2 (or CrAssphage) Concentration. *The modelling is based on a Sugeno adaptive neuro-fuzzy interference system (ANFIS), the sub-method being subtractive clustering followed by tuning with the training data, where the inputs are phosphate and crAssphage concentration (panel A) or phosphate and SARS-CoV-2 concentration (panel B), and the outputs are the SARS-CoV-2 concentration (A) or crAssphage concentration (B).*

SARS-CoV-2 modelled concentration (Figure 12) demonstrated that the optimal pH conditions were 7.6 – 8.4 (up to 4.34 $\log_{10}$ gc/l), the increase in the crAssphage concentration only slightly having contributed to an increase in the SARS-CoV-2 modelled concentration. SARS-CoV-2 lower recoveries can also be observed at pH 6.8 – 7.6.

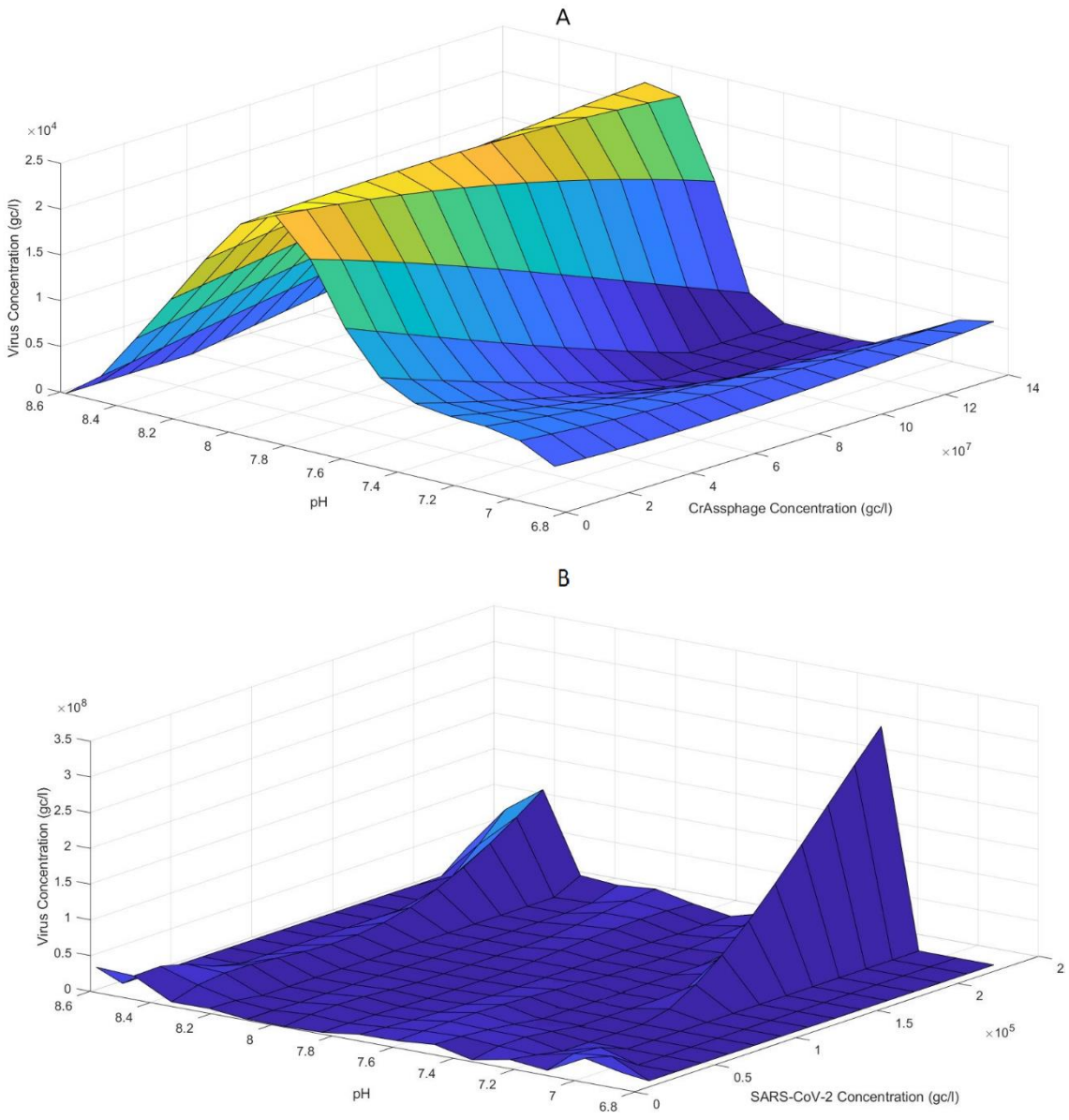

Figure 12. The Modelled Impact of pH and CrAssphage (or SARS-CoV-2) on SARS-CoV-2 (or CrAssphage) Concentration. *The modelling is based on a Sugeno adaptive neuro-fuzzy interference system (ANFIS), the sub-method being subtractive clustering followed by tuning with the training data, where the inputs are pH and crAssphage concentration (panel A) or pH and SARS-CoV-2 concentration (panel B), and the outputs are the SARS-CoV-2 concentration (A) or crAssphage concentration (B).*

CrAssphage projected concentrations demonstrated high values at pH 8.2 – 8.4, and significantly higher values at pH 7 – 7.35 and with an increase in the SARS-CoV-2 concentration, also increased the projected crAssphage concentration (Figure 12).

The pH conditions correlating with higher SARS-CoV-2 and crAssphage concentrations (Figure 12) coincided with those from previous models (Figure 5; Figure 6; Figure 9).

The modelled impact of turbidity or electrical conductivity and one virus concentration on the second virus concentration can be found in Figure S27 And Figure S28. Common observed trends are that at turbidity values of >150 NTU and an increasing crAssphage/SARS-CoV-2 concentration, also increased the modelled SARS-CoV-2/crAssphage concentration.

Additionally, at electrical conductivities of 1100 – 1800 µS/cm, the modelled SARS-CoV-2 concentrations showed significantly higher values regardless of the crAssphage concentration (Figure S28). Whereas the crAssphage model demonstrated that the optimal electrical conductivity for higher virus concentrations were at values of >1400 µS/cm, the increase in the SARS-CoV-2 concentration only slightly having correlated with the increase in the crAssphage concentration. The described optimal electrical conductivity values were consistent with the previous model in Figure 8.

3.4 Total and Infected Population Estimates

The data required for the population estimation with the EMCLT and the MCBA were collected from literature or measured experimentally in the case of the volumetric flow rate. These were recorded in the Table 7. It was not possible to find a standard deviation for crAssphage and NoVGII faecal shedding, hence for these 2 viruses, the standard deviation for the EMCLT infected population estimate was not calculated.

When a reliable mean value for a parameter could not be found, but a median value was available, the mean concentration used for modelling was selected based on the median value. Since the EMCLT and the employed MCBA assumes symmetrical normal distributions, it is expected that the mean and median values are close to each other. This was the case for the selected mean concentration for crAssphage and SARS-CoV-2.

Table 7. Mathematical Modelling Parameters

	Virus	Units	Selected Concentration of Mean	Selected Range of Concentration	Standard Deviation (σ)	Reference
Ammonium amount discharged daily by a person		$[M_{NH4-N}]$ = mg/d/capita	6000 mg/d/capita	4000-8000 mg/d/capita	N/A	(Zheng et al., 2017)
Phosphate amount disposed daily by a person		$[M_{PO4-P}]$ = mg/d/capita	850 mg/d/capita	594-1106 mg/d/capita	N/A	(Sun et al., 2017)
Volumetric Flow rate		[F] = L/day	1920000 L/day	1824000-2072235 L/day	96000 L/day (5% of the reading)	Measured experimentally
Faecal load		[α] = g/day/capita	149 g/day/capita	51-796 g/day/capita	95 g/day/capita	(Rose et al., 2015; Saththasivam et al., 2021)
Viral faecal shedding	CrAssphage	[β] = copies/g	8.1 log_{10} copies/g	4.1-10.1 log_{10} copies/g	Not available	(Park et al., 2020)
	SARS-CoV-2		6.9 log_{10} copies/g	4.3-10.81 log_{10} copies/g	7.47 log_{10} copies/g	(Anjos et al., 2022; Saththasivam et al., 2021)
	Influenza A Virus		4.5 log_{10} copies/g	3.69–7.9 log_{10} copies/g	4.8 log_{10} copies/g	(Chan et al., 2011, 2009)
	NoVGII		8.48 log_{10} copies/g	7.79–10.11 log_{10} copies/g	Not available	(Lee et al., 2007)

The table details the chemical parameters used throughout the modelling methods. The volumetric flow rate was measured experimentally on 2 occasions at the sampling manhole near the hospital, and the result was compared with the total water usage for the catchment. The "selected concentration of mean" was derived from the mean or median values when the former was not available.

An example of how exactly the MCBA was performed, is provided in the Appendix (Figure S6; Figure S7; Figure S8), specifically of the Approach 3 which also explains the Approach 1, 2 and the EMCLT.

The population at Ysbyty Gwynedd was estimated to be within the boundaries of 1500-3500 individuals, including the capacity of in-patients, commuting staff and population movement, such as visitors or out-patients. This range was used to evaluate the population estimate models.

The results of the simulations with the MCBA 1, 2 and 3 were integrated into figures alongside the EMCLT results for visual comparison (Figure S9, Figure S10, Figure S11, Figure S12).

Additionally, the mean, standard deviation, 99% confidence interval and the minimum degradation rate are provided for the estimated populations for each MCBA (Table S3, Table S4, Table S5, Table S6).

The EMCLT and MCBA 2 and 3 crAssphage-based estimates performed the best with 19.6%-21% of values being within the population limits of 1500-3000 individuals, whereas in 73%-75% of estimates, the total population number was underestimated (Figure S9). The MCBA 1 based on crAssphage wastewater concentration proved to be unsuitable for the total population estimation, whereas the ammonium and phosphate based estimates significantly overestimated the total population (Figure S9).

The probability distributions for each obtained population estimate are not given.

In the Appendix, it can also be found separately the EMCLT population estimate results for SARS-CoV-2, crAssphage, NoVGII and influenza A virus (Table S1). The standard deviation for the infected population estimate was calculated only for crAssphage and SARS-CoV-2.

Influenza A virus and NoVGII detection was not consistent throughout the study period. Therefore, it was not possible to create a reliable continuous trend for these two. Assuming the parameters of Table 7, the MCBA 1 proved not to be feasible in the case of influenza A due to the unrealistic estimation of the population number of 0-3 individuals given the detected concentrations (Table S5) in a volumetric flow rate of 1920000 ± 5% litres/day.

The MCBA 2 and 3 estimates were usually similar each other in the case of crAssphage, SARS-CoV-2 and influenza A. This is due to viral losses for each approach being well aligned. Surprisingly, although the EMCLT assumed a $\bar{\gamma}$ of 0, the estimates were within approximately ± 10% of the MCBA 2 and 3. In the case of norovirus GII, the calculated γ_{min} for MCBA 3 was not consistent for most of the estimates and jointly considering the EMCLT estimates for other viruses being higher than the MCBA estimates, it points out the flaw of the selected number of iterations which was no more than 5000 simulations (Supplementary Materials, “Monte-Carlo-Bayesian Approach Example”). However, for the γ_{min} calculation step, the number of simulations was even lower, 500 data points for MCBA 2 and 250 data points for MCBA 3. These values were selected based on the computational power and also because at 250-300 simulations, the distributions of the data points have normalised.

3.5 Comparison of Estimates with Clinical Data

3.5.1 SARS-CoV-2

It was not possible to obtain the data for the in-care clinical cases at Ysbyty Gwynedd. However, SARS-CoV-2 reported cases for the county Gwynedd were used instead (UKHSA, 2022a). The population in Gwynedd is 117,400 as of 2021 (ONS, 2022a). The population estimates of SARS-CoV-2 were divided by the population estimates based on crAssphage for each model and compared against the COVID-19 reported cases in Gwynedd divided by the total population of Gwynedd (Figure 13). It is not known if an increased or decreased disease prevalence can be expected in the hospital comparing to the community. On one side, healthcare professionals are at a higher risk of infection and a hospital is a hub for infected individuals or for those at a higher infection risk. On the other hand, extensive cleaning procedures are to be undertaken in a hospital, not all out-patients or visitors use toilet facilities and symptomatic individuals are more likely to undertake precautionary measures. However, correlation between the hospital and the community prevalence of COVID-19 is expected.

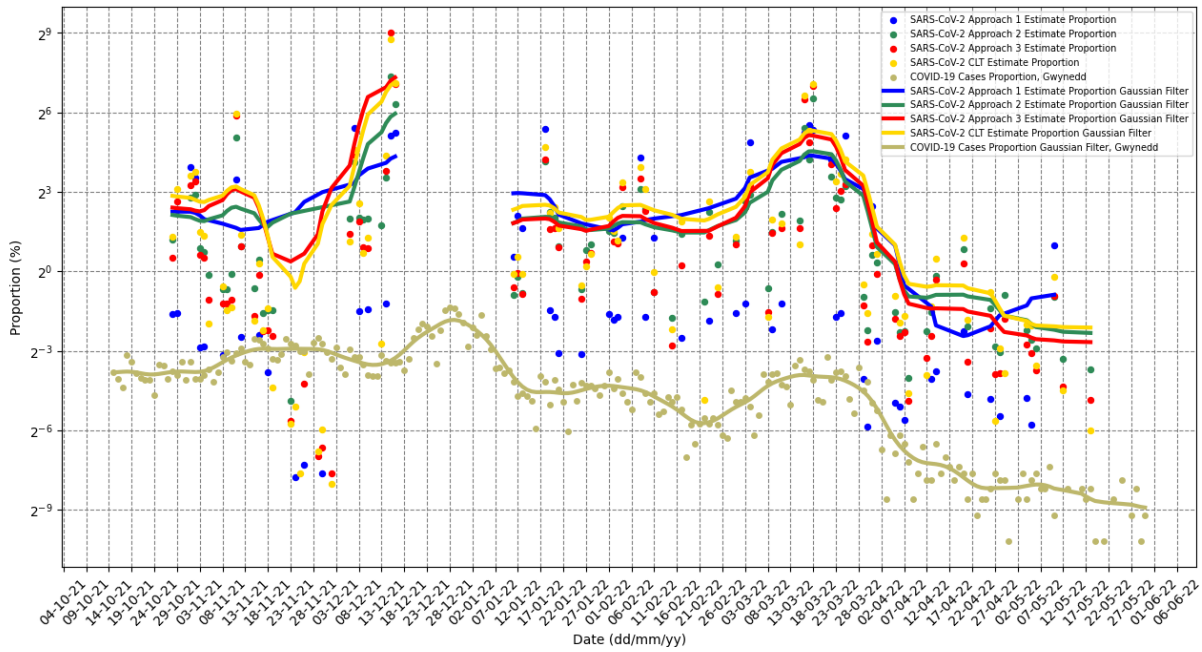

Figure 13. Proportion of estimated SARS-CoV-2 cases at Ysbyty Gwynedd against proportion of reported cases in Gwynedd. *The proportion of estimated SARS-CoV-2 cases at Ysbyty Gwynedd was calculated by dividing the SARS-CoV-2 population estimates by the crAssphage population estimates, while the proportion of reported cases in Gwynedd was calculated by dividing the number of cases reported clinically by the total population of Gwynedd, both being subsequently multiplied by 100. A Gaussian filter is applied with a sigma value of 3.*

In Figure S13, the comparison from Figure 13 is demonstrated with smaller values for the sigma argument in the Gaussian filtered plots or a lower “noise reducing” effect. Upon visual examination of the trends, a significant similarity can be observed, however it should be noted that the y-axis scale is exponential. Even if the reported cases were adjusted for the ratios of confirmed to unreported cases of SARS-CoV-2 as established previously to be between 1:11 and 1:15 (not shown) (95% CI between 1:4.2 and 1:17.5) (McMahan et al., 2021), the model still overestimated the proportion of people infected with SARS-CoV-2 in the hospital catchment comparing to the whole county.

In order to quantify the correlation observed in Figure 13, the proportions were converted to percent changes, and the pairwise Pearson correlation of the estimate-based percent versus reported-based changes was calculated (Table 8).

Table 8. Pairwise Pearson Correlation of SARS-CoV-2 Estimated Cases at Ysbyty Gwynedd against Reported Cases in Gwynedd Proportions Percent Changes

sigma	MCBA 1	MCBA 2	MCBA 3	EMCLT
0	-0.098	-0.042	0.066	0.035
1†	0.111	0.166	0.051	0.081
2†	0.309	0.321	0.274	0.246
3†	0.379	0.426	0.336	0.315
1*	0.225	0.286	0.187	0.190
2*	0.605	0.525	0.377	0.331
3*	0.750	0.694	0.556	0.496
5*	0.970	0.906	0.832	0.796

*Sigma was applied before percent conversion - *. Sigma was applied before and after percent conversion - †.*

The percentage changes were also plotted for visual comparison (Figure 14). Additionally, Figure S14 and Figure S15 show comparisons of percent changes with varying sigma argument values as described in Table 8. With an increasing sigma argument, especially when applied before and after percent changes calculation, also increased the pairwise correlation.

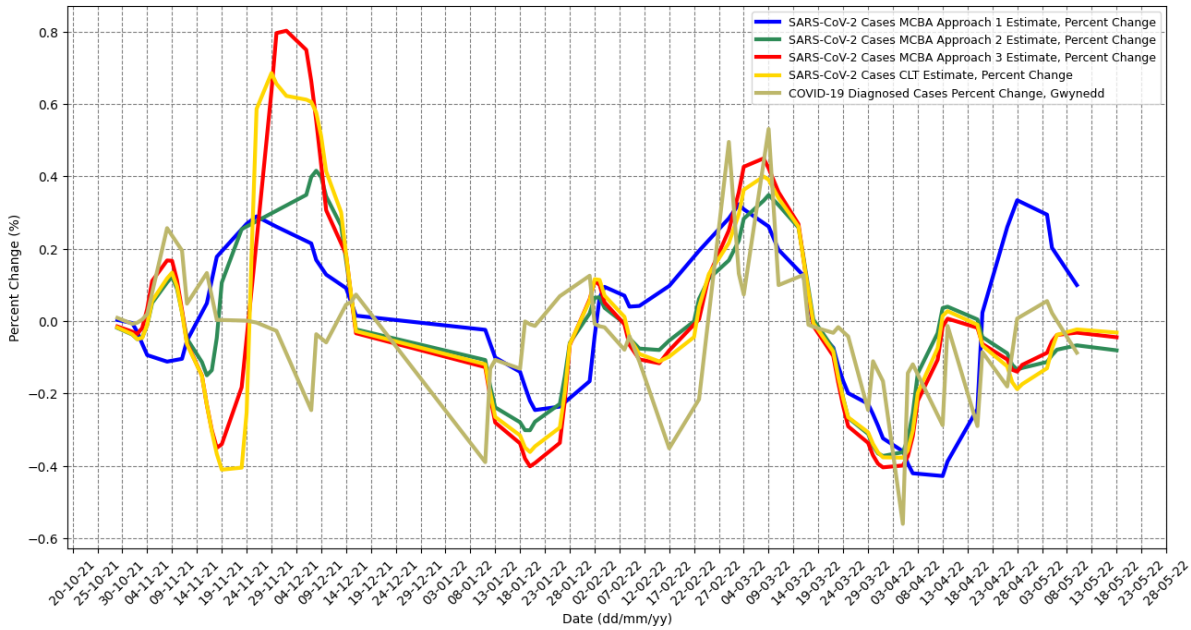

Figure 14. Estimated Cases at Ysbyty Gwynedd against Reported Cases in Gwynedd Proportions Percent Changes. *Initially the proportions are calculated by dividing the SARS-CoV-2 population estimates by the crAssphage population estimates for each of the modelling method: the Monte-Carlo-Bayesian approach (MCBA) 1, 2, 3 and the equation model following the Central Limit Theorem premise (CLT), and another proportion is calculated separately by dividing the SARS-CoV-2 reported cases in Gwynedd by the total population of Gwynedd. The proportions are computed with a Gaussian filter of a sigma argument of 3 only once before the conversion to percent changes. Then the percent changes were calculated for each proportion separately.*

Although the population estimates proportion were much higher than the reported cases proportion (Figure 13), the percent changes (Figure 14) demonstrated that the changes in the reported cases did indeed follow a similar trend as the estimated cases based on wastewater virus concentration.

The results of DTW can be observed in Figure 15, the method being applied on the estimated and reported cases proportions. Although the diagonal sections demonstrate an identity between the compared groups, the shifts towards either of the axes demonstrated that there lacks significant and consistent trend similarity and predictive capacity in terms of infected population prediction. However, the reason is that a Gaussian filter was not applied, hence the estimate trends have inconsistent significant variations (Figure S13), not allowing for proper quantification of the trends' similarity.

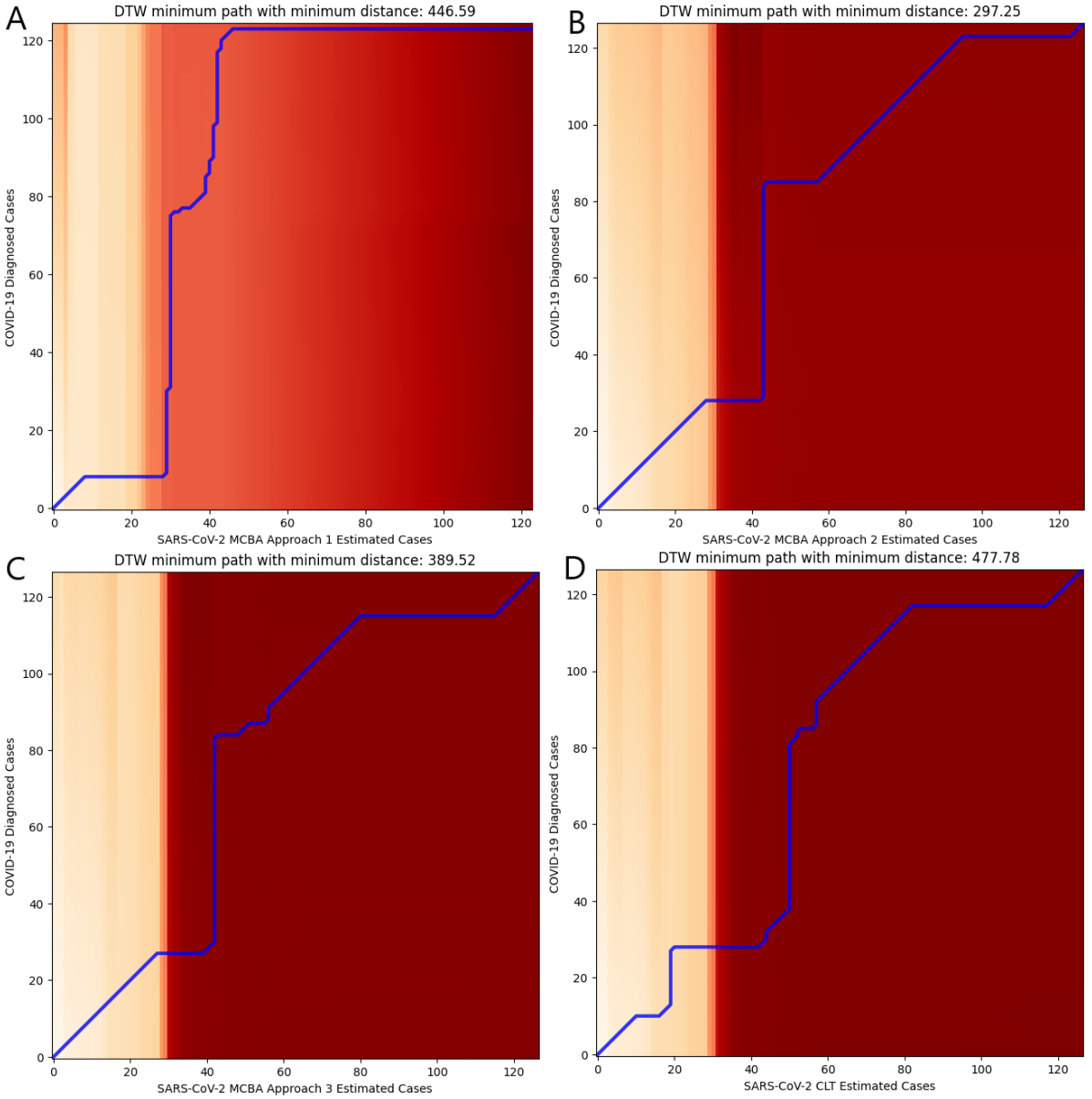

Figure 15. Dynamic Time Warping of SARS-CoV-2 Estimated Cases Proportion at Ysbyty Gwynedd against Reported Cases Proportion in Gwynedd. *The SARS-CoV-2 infected population estimates proportion at Ysbyty Gwynedd were compared against the SARS-CoV-2 reported population proportion in Gwynedd by Dynamic Time Warping (DTW). The estimates proportion were calculated by dividing the SARS-CoV-2 infected population estimate by crAssphage infected population estimate. The reported cases proportion was calculated by dividing the SARS-CoV-2 cases in Gwynedd by the total population of Gwynedd. The x and y axis represent the matrix cost, where the minimum path with minimum distance is computed as the sum of the absolute differences. A match is described by a diagonal warping path, hence indicating temporal alignment between the two series. Each of the Monte-Carlo-Bayesian approach (MCBA) and the equation model following the Central Limit Theorem premise (CLT) is compared independently against the COVID-19 reported cases in Gwynedd. A left-hand or a right-hand deviation from the diagonal indicates a leading or lagging relationship. For example, a right-hand shift towards the Estimated Cases Axes indicate that the proportion of estimated cases has increased sooner or has increased proportionally more than the proportion of reported cases in Gwynedd. In panel A, it can be observed a significant left-hand shift towards the "COVID-19 Reported Cases" axis and then a right-hand shift, whereas no diagonal path is present, this indicating no alignment between the trends. For panel C and D, although there is a significant portion of the path being diagonal which shows trend alignment, the cost matrix is 389.52 and 477.78, which is close to panel A, 446.59, indicating extensive adjustment by the DTW model in order to fit the trends. Therefore, it is unlikely that panel C and D have identified significant trend similarity or a lagging/leading relationship. The best result was yielded by panel B with the lowest cost matrix and the highest proportion of the diagonal path.*

In Figure 15, the SARS-CoV-2 MCBA 2 estimates proportion demonstrated the highest trend similarity with the COVID-19 reported cases proportion. However, a left-hand shift was present, indicating that the reported cases proportion has increased sooner or proportionally more, followed by trend alignment where the estimated cases trend was lagging the reported cases trend. This then follows by a right-hand shift, reducing the lagging distance of the estimated cases trend. In conclusion, the results demonstrate insufficient trend similarity and identity between the compared datasets. A significant limitation is that the clinical SARS-CoV-2 cases were bigger datasets containing data for almost each day between the study period, while the estimate cases datasets contained data for several days in a week. The lack of alignment is not necessarily due to the lack of WBE predictive capability of clinical cases but can be due to a much higher uncertainty in terms of viral losses and a lower accuracy in estimating the number of cases. Additionally, a factor is also that the clinically reported case numbers are not completely representative of the disease prevalence and do underestimate the spread within the catchment, whereas WBE would remove this limitation. However, this would be indicated by a right-hand shift towards the estimated cases axes (Figure 15) rather than the reported axes, demonstrating an inaccuracy in the employed infected population estimation models.

Another DTW comparison was between only the estimated and reported cases only on the same date (Figure S16); however, this adjustment did not improve the temporal alignment as quantified by the “minimum path” for any of the methods but the SARS-CoV-2 MCBA 1.

3.5.2 Influenza A

The weekly diagnosed influenza A cases in Wales as determined based on specimens submitted for virological testing for hospital patients and non-sentinel GPs, were retrieved from public resources (Table S7) (PHW, 2022). Influenza A estimates were cumulated by week, as well as the crAssphage estimates, after which the influenza A virus weekly estimates were divided by the crAssphage weekly estimates and compared against the weekly influenza A diagnosed cases divided by the total Wales population, the population in Wales being 3,107,500 as of 2021 (ONS, 2022b).

There is a significant discrepancy between the estimated Influenza A cases proportion and the diagnosed one (Table 9), however for the week 15/11/2021 – 05/12/2021, the estimated proportion reaches the range of 0.87% – 11.44%, a more realistic estimate which coincides with the CDC estimate of the prevalence of symptomatic flu illness of 3% – 11% (CDC, 2022b).

Table 9. Influenza A virus Estimated Cases Proportion at Ysbyty Gwynedd against Influenza A Diagnosed Cases Proportion in Wales

Week by Start Date	MCBA 1 (%)	MCBA 2 (%)	MCBA 3 (%)	EMCLT (%)	Diagnosed Cases Proportion (%)
08/11/2021	89.996	61.704	2273.896	2026.124	0.022
15/11/2021	29.239	11.442	4.003	4.318	0.046
22/11/2021	0.873	N/A	2.370	0.615	0.045
29/11/2021	1.011	N/A	0.117	0.178	0.098
06/12/2021	79.941	78.022	95.027	89.131	0.164
13/12/2021	97.560	67.796	77.487	62.470	0.130

The estimated cases proportion at the hospital Ysbyty Gwynedd was calculated by dividing the Influenza A estimated cases by the crAssphage population estimates whereas the proportion for Wales was calculated by dividing the number of clinically diagnosed cases in Wales by the number of tested specimens.

4. Discussion

4.1 Total Population Estimation

The projected total population numbers with crAssphage, ammonium and phosphate concentrations according to the EMCLT and MCBA demonstrated high variations conflicting with the expected population limits of 1500-3000 individuals (Figure S9). Generally, the ammonium and phosphate estimates were significantly above the higher limit, likely indicating that the ammonium and phosphate contribution per capita was underestimated. It resulted from a higher concentration discharge than 6 g/day/capita for ammonium and approximately 1 g/day/capita for phosphate. However, in previous research, the ammonium-based population estimates agreed with the corresponding population numbers (Zheng et al., 2017; Been et al., 2014). A significant difference is that Zheng et al. (2017) and Been et al. (2014) performed sampling of influent water at a WWTP, whilst in this hospital-based study sampling was performed at a nearby manhole. This factor is suggested to be important because at the hospital, there is likely to be an increased use of disinfectants containing quaternary ammonium compounds, especially during the SARS-CoV-2 pandemic (Hora et al., 2020). This would increase the total ammonium concentration unproportionally to the total population. In contrast, at a WWTP, unique events of increased ammonium disposal are likely to have a smaller effect on the population number predictability due to being diluted with a higher volume of wastewater. Therefore, ammonium did not prove to be suitable to estimate the population number within the hospital catchment. Additionally, urine-derived ammonium release increases with urea hydrolysis, and depending on the amount of urease, it can take up to 2-4 weeks for complete hydrolysis (Münch & Winker, 2011). Therefore, the difference of the amount of hydrolysed urea among processed wastewater samples is a source of error for total population estimates. It is therefore recommended that future estimates measure total soluble N in wastewater (i.e. ammonium, urea etc) to better estimate N loading rate. For the same reason, phosphate-based population estimates are not accurate either since phosphate is a complexing agent used in disinfectants and cleaning products (Yangxin, Jin & Bayly, 2008; Köhler, 2006). This conclusion is supported by the significant similarity in the trends of ammonium- and phosphate-based population estimates and by the fact that the population numbers fluctuated by 5000 – 10000 individuals, an unrealistic change for a catchment with an expected population of 1500 – 3000 individuals (Figure S9). Also, since these two trends mostly do not agree with the crAssphage-based trends (Figure S9), and assuming that the crAssphage losses rate was mostly consistent throughout the whole study period, it again demonstrates that the ammonium and phosphate fluctuations are not proportional exclusively to the population number. However, in this case, ammonium and phosphate concentrations can be evaluated for other purposes, for example for assessing virus degradation or normalising virological data (Hutchison et al., 2022).

The MCBA 1 is not suitable for crAssphage-based population estimation in the current setting. The EMCLT and MCBA 2 and 3 crAssphage-based estimates trends were within the proximity

of each other for most of the study period (Table S2; Table S6) despite the fact that the MCBA 2 and 3 assumed viral losses of >97% while the EMCLT had a $\bar{\gamma}$ of 0. It is unclear the reason of this similarity at such significantly different degradation rates between the MCBA and the EMCLT. Normalisation of the crAssphage-based estimates by the coefficient corresponding to the number of individuals shedding crAssphage in faeces did not improve the outcome since the trend overestimated and underestimated over periods of time of similar length. The obtained minimum viral losses rate of >97% is a realistic estimate considering that the ammonium-sulphate based concentration method had a mean recovery rate of 6% (Farkas et al., 2022a). Previous research has found a significant relationship between the served population and the crAssphage load, but no studies have estimated the exact population number based on crAssphage (Wilder et al., 2021). However, it should be noted that methods for crAssphage recovery from wastewater have not been optimised and their recovery is subject to considerable uncertainty. Potential solutions to solve the inaccuracy of the model are the following. Firstly, the crAssphage wastewater concentration is to be normalised with another human biomarker such as creatine, the pepper mild mottle virus (PMMoV) or the human mitochondrial gene NADH dehydrogenase subunit 5 (Hutchison et al., 2022). Secondly, the distribution of the crAssphage faecal shedding rate should be repeatedly evaluated in more detail for normality. In case it is a non-normal distribution, the distribution pattern should be integrated into the MCBA by artificially and proportionally increasing the incidence of pseudo-random number generation within specific intervals while decreasing for other intervals.

4.2 Wastewater Physico-chemical Properties Effect on SARS-CoV-2 and CrAssphage Concentration

Although previous experimental and ANFIS-based research on coronaviruses demonstrated that the viral RNA remains stable in the pH range of 7.1 – 7.4 (Casanova et al., 2009; Amoah et al., 2022), the employed ANFIS models demonstrated that the most optimal pH conditions for SARS-CoV-2 are in the range of 7.6 – 8.5 (Figure 5; Figure 6; Figure 9; Figure 12). This finding coincides with the pH range of 7.5 – 9.0 at which SARS-CoV-2 structural proteins human ACE2 and the S-protein have been shown to be stable (Xie et al., 2022). This difference is likely to originate from different chemical composition of wastewater as sampled from a hospital in this study or from the WWTP in Amoah et al. (2022) study. Similarly, for crAssphage, experimental data on five phage viruses from the same *Podoviridae* family demonstrated the optimal pH for stability to be in the range of 5 – 10 (Hamdi et al., 2016), corresponding with the ANFIS modelling indicating the most optimal pH conditions to be 6.8 – 7.8 and 8.15 – 8.5 (Figure 5; Figure 6; Figure 9; Figure 12). While it can be argued that the ANFIS model has narrowed the optimal pH range for crAssphage, it should be noted that the pH of wastewater is relatively stable (Table 4). Therefore, there was no sufficient data to train the model at pH values of <6.5 and >8.5, however, these values are unlikely for wastewater. The pH values which correlated with a higher concentration for both viruses corresponded with the mean

wastewater pH (Table 4), hence wastewater pH at the detected values does not contribute to SARS-CoV-2 or crAssphage instability, and their subsequent degradation. Therefore, it was also not possible to link the pH value to particular cleaning products. Factoring in the impact of a wastewater dilution factor on the cleaning product potential to degrade the pathogen or its genetic material requires more research. Since their pH vary, direct measurement of markers, such as total chlorine, might be more practical in a wastewater surveillance setting, and will remove uncertainties originating from using pH as a proxy indicator (Sridhar et al., 2022; CDC, 2022a).

Electrical conductivity can be interpreted as a good indicator of wastewater salinity, which electrostatically affects the binding capacity of viruses to charged surfaces, such as organic matter, a higher salt content promoting hydrophobic interactions (Farkas et al., 2022b; Liu et al., 2017). However, other indirect effects of salts are the change of pH and the formation of precipitates that adsorb viruses (Lukasik et al., 2000). Additionally, electrical conductivity is also an indicator of the dissolved organic matter (DOM) in wastewater, which may competitively inhibit virus sorption to polar sites. Therefore, it is predicted that with an increasing electrical conductivity, viruses detach from solid particles, hence increasing virus recovery from the liquid component of wastewater. That was observed for SARS-CoV-2 and crAssphage, EC correlating with higher virus recovery, starting at values of approximately 750 µS/cm as inferred from multiple ANFIS plots (Figure 8; Figure 9; Figure S28).

Some research reports the opposite, that salinity levels are negatively correlated with virus recovery from the liquid environment and positively correlated with virus sorption due to viral particles detaching from porous material as a result of salinity reduction (Zhang, Zabarankin & Prigiobbe, 2019). However, the model provided by Zhang, Zabarankin & Prigiobbe (2019) lacks sufficient experimental data. Moreover, different viruses have different charge distribution on the virion surface (i.e. positive and negatively charged domains), hence are likely to interact differently. For example, the charge distribution of the SARS-CoV-2 intact virion is heterogenous with the stalk part of the spike protein being net negatively charged while the apical part, particularly the receptor binding domain, being net positively charged (Adamczyk, Batys & Barbasz, 2021; Jones et al., 2022). The overall net charge is expected to be negative due to the isoelectric point being below pH 7 (Jones et al., 2022). Since the solid matter within raw wastewater contains organic matter, which is mostly negatively charged due to carboxylic and phenolic groups, the SARS-CoV-2 negative net charge will result in repulsion against particulate organic matter (Park et al., 2018). When positively charged groups will be present on particulate organic matter or other surfaces, SARS-CoV-2 virion attachment to them may be inhibited by salinity and DOM as demonstrated by the EC positive correlation with virus recovery. However, the wastewater matrix is complex, and it is unclear how the positive charge of the apical part of the SARS-CoV-2 spike protein would affect these electrostatic interactions. Moreover, even more uncertainty is introduced when the viral capsid is damaged or the nucleic acid is unprotected, which changes the surface charge (Jones et al., 2022).

In the case of influenza A virus, the surface net charge is positive, particularly due to haemagglutinin on its surface (Kobayashi & Suzuki, 2012; Arinaminpathy & Grenfell, 2010). In contrast, NoVGII has a surface net negative charge at environmental pH (da Silva et al., 2011; Mertens & Velev, 2015). Therefore, influenza A virus is likely to be attracted to negatively charged surfaces and receptors (e.g., wastewater organic matter), which is contrary to the NoVGII or SARS-CoV-2 electro-static interaction (Adamczyk, Batys & Barbasz, 2021). Unfortunately, it was not possible to train an ANFIS for influenza A virus given its much lower detection rate; however, other statistical analysis, such as principal component analysis, would make that possible. For modelling different viruses' sorption interactions, more research is required about the wastewater solid fraction composition and electric charge, and whether it is composed mostly of negatively charged organic matter.

The positive relationship between EC and SARS-CoV-2 and crAssphage virus recovery is also consistent with previous research that has shown that different wastewater concentration methods with and without pre-treatment give different viral recoveries. NaCl pre-treatment enhanced viral recovery from the liquid part of wastewater which was attributed to Na^+ and Cl^- additional competition for the polar binding sites on the surface of solid polar particles (Farkas et al., 2022b). However, in the experiment, NaCl was added to achieve a concentration of approximately 1.2% (Farkas et al., 2022b), whereas the NaCl concentration of >95% wastewater samples wastewater was <0.1% (<1500 µS/cm) as calculated by converting electrical conductivity into a NaCl concentration. Therefore, the salinity of the wastewater samples at the measured values (Table 7) is likely to have a smaller effect on virus recovery from the liquid part of wastewater comparing to the study from Farkas et al. (2022).

The positive association between ammonium concentration and viral recovery (Figure 5; Figure 7; Figure 10) is consistent with another ANFIS-based study (Keshavarzi et al., 2017). The phosphate concentration demonstrated a stronger association with the crAssphage concentration comparing to SARS-CoV-2 recovery due to the association of crAssphage with faecal matter and of faecal matter with orthophosphate (Figure 6; Figure 7; Figure 11) (Farkas et al., 2023; Chapuis-Lardy et al., 2004). In a larger population as demonstrated by higher ammonium or phosphate concentrations, increases the potential for a higher load of SARS-CoV-2 or crAssphage in wastewater (Keshavarzi et al., 2017).

In relation with the correlation between crAssphage and turbidity (Figure 8; Figure S27), previous research indicates weak to no correlation between crAssphage concentration and turbidity or total suspended solids of influent wastewater samples (Jennings et al., 2020; Cuevas-Ferrando et al., 2022).

4.3 Models Validation

The EMCLT and Monte-Carlo simulations have been previously trialled to estimate the SARS-CoV-2 infected population based on WBE data (Hasan et al., 2021; Vallejo et al., 2022), but little to no research is available for norovirus GII, influenza A virus and crAssphage.

The evidence presented here suggests that the MCBA and the EMCLT overestimated the infected population proportions at Ysbyty Gwynedd, however, on-site data is required to confirm the results (Figure 13). The overestimation persists even if the reported number of cases are multiplied by a factor of 11 or 15 to account for the unreported number of SARS-CoV-2 cases (McMahan et al., 2021). A previous study of the EMCLT demonstrated that the model provided a realistic estimation of the SARS-CoV-2 infected population but also underpredicted the number when the infection cases were lower (Saththasivam et al., 2021). However, the reasons for other research models' underestimation are different than the SARS-CoV-2 infections overestimation of this study. This is due to the fact that previous research compared the estimated cases directly with the confirmed cases, which was not possible for this study. The hospital infected proportions were compared with the county infected proportions, which have different demographics and epidemiological dynamics. Possible explanations for the overestimation are an increased incidence of SARS-CoV-2 within the hospital catchment, underestimation of the total population based on crAssphage wastewater concentration, or inaccurate parameters set for the MCBA. Although, the percentage changes correlated well between the comparison groups (Table 8; Figure 14). Additionally, this points to the suitability of crAssphage data to normalise the SARS-CoV-2 data. Two other studies that used a susceptible-exposed-infectious recovered (SEIR) model have performed significantly better than the employed EMCLT and the MCBAs, finding a similarity ratio of 10:9 and 1:1 between the estimated cases at a WWTP and the adjusted for under-reporting (McMahan et al., 2022, 2021).

Previous research identified the predictive capacity of WBE using standard methods such as a visual inspection of the trends or Pearson and Spearman correlations (Olesen, Imakaev & Duvallet, 2021). DTW was not successful at validating the previously observed similarity of the trends in Figure 13 or at identifying the capacity of WBE in predicting an increase or a decrease in the number of infections due to not being possible to apply the Gaussian function which resulted in significant oscillations of the trends (Figure S13). However, the lead time of wastewater-based COVID-19 measurements against the case counts was previously estimated to be between 2 to 14 days depending on the setting (Zhang et al., 2022; Olesen, Imakaev & Duvallet, 2021).

The influenza A estimated prevalence for week 29/11/2021 was close to the Wales clinical cases proportion for week 06/12/2021, and between 15/11/2021 and 05/12/2021 (Table 9), it coincided with the CDC estimate of influenza A prevalence in the community of 3 – 11% (CDC, 2022b). Besides the positivity rate for Wales and the CDC estimates, sentinel swabbing in the UK found that the influenza A positivity rate reached as high as 4% for the week of 08/11/2021 which matches the MCBA 3 and EMCLT estimated proportions of 4% and 4.3%, respectively, for the week of 15/11/2021 (Table 9) (UKHSA, 2022c). Other similarities were observed between the estimated and sentinel swabbing positivity rates or the hospital admission rates in UK; however, the leading time was not consistent. Despite previous research describing the WBE capability to identify influenza A outbreaks (Wolfe et al., 2022),

the short period of time of influenza A virus positive detection (Table S7) was not sufficient to infer firm conclusions about the estimated-reported trends of this study.

The detection rate of NoVGII (Table 5) was consistent with the results of a systematic review reporting an 82.1% norovirus detection in wastewater (Huang et al., 2022), and the detection between 13/12/2021 and 04/03/2022 in multiple intervals of 3-4 days alongside the MCBA 1 estimates of 1 – 3 infected individuals is consistent with the norovirus infection period of up to 3 days (Table S3) (Mayo Clinic, 2022b). On the other hand, the MCBA 2, 3 and the EMCLT provided estimates (Table S1; Table S3) and would be treated as a health emergency, which was not the case, proving that these models are unsuitable for NoVGII. Nonetheless, the increase in the detected norovirus concentration and the MCBA 1 estimated cases between 29/03/2022 and 14/04/2022 indicated an outbreak at Ysbyty Gwynedd, but no outbreak was reported publicly specifically at the hospital. However, for the week commencing 14/03/2022, a significant peak was observed in the weekly reported norovirus cases in England based on laboratory reports and a peak of norovirus outbreaks was observed at hospitals in England as reported to the Hospital Norovirus Outbreak Reporting System (UKHSA, 2023). While the increase in laboratory reports and hospitals outbreaks in England does not demonstrate that an outbreak occurred at Ysbyty Gwynedd, it could provide basis for the assumption of the outbreak being possible based on national trends. Additionally, previous research has successfully used WBE to predict NoVGII outbreaks 2-3 weeks in advance (Hellmér et al., 2014). However, similarly to influenza A virus estimates, longer surveillance time for NoVGII detection is required.

5. Study Limitations

The initial sampling was performed as 12 individual 2-hour composite samples, followed by one 24-hour composite sample taken daily. The study accounted for this transition by using the daily mean values for the viruses, ammonium and phosphate concentrations, however, a thorough comparison between the two periods was not explored. Since sampling was done at the maintenance hole situated directly outside the hospital, it is likely that the fast-stream effluent was not homogenised sufficiently, and the sampled wastewater was not representative of the whole volume of wastewater passing the sampling point. This may have contributed to significant trend fluctuations, especially for low-abundance viruses, such as influenza A virus.

The parameters of the ANFIS model were selected as the most optimal based on the size of the dataset, which also yielded a high error rate in terms of the RMSE and the predictive capability. Perfect correlation could have been achieved by overfitting the dataset, but that would not provide any conclusions on the paired influence of two wastewater parameters on virus concentration. The solution is to decrease the similarity cluster influence range that would allow for more model membership relationships to be built, hence more accurate predictions to be made, providing there is one. However, due to a limit in the computational power and due to the size of the dataset that was not possible. Another limitation of the ANFIS model is that it explored exclusively the paired impact of two wastewater parameters on the virus concentration. Since it encompassed only the operator “and”, it was not exploring the “or” relationships when one of the parameters had an impact while the second one did not. Additionally, other statistical approaches exploring the impact of chemical parameters on virus concentrations were not tested, which otherwise, would have put the described impact into a more quantifiable and statistically significant perspective.

An important data analysis tool was the Gaussian filter. Its function was to reduce fluctuations to allow an easier comparison of the two trends. The limitation is that without this function, most of the similarities cannot be easily observed or quantified due to high variations in trends and outliers. The study tested a range of parameters for this function as to observe the corresponding change and the effect the function had in influencing the similarity. Although the Gaussian filter is efficient in reducing the trends jumps or “noise” of a variable, hence allowing to easier observe the trend over a time period, a too large reducing effect, manifested by choosing a higher *sigma* argument, may result in skewed conclusions, requiring a complementary statistical analysis to confirm the conclusions. Additionally, the need to apply this filter points out to inconsistencies of the methodology, particularly on the data analysis side such as the lack of direct comparison groups (no clinical data within the catchment) and potential inaccuracy or unsuitability of the MCBA or the EMCLT. However, the inconsistency could originate on the laboratory and study design sides such as the unsuitability of crAssphage as a human biomarker or the requirement to normalise virus concentrations with other human biomarkers.

A significant limitation is that the study resorted to comparing the population estimates with the reported cases of a much larger area by converting the estimated and the reported cases numbers into proportions corresponding to the total population. Additionally, there was an inconsistent influenza A virus detection, and it was not possible to observe the trends over a sufficiently large period. Samples were also negative when the number of influenza A cases has peaked in Wales, however, this may have resulted from an uneven disease spread distribution, e.g., many cases being concentrated in urban areas.

For random number generation in the context of the MCBA, the sampling was done from a uniform distribution. While under the CLT assumption, the simulated virus concentration would result in a normal distribution regardless of the initial distribution of parameters for a sufficiently large population, the distributions of the parameters, especially that of virus and faecal shedding, should be incorporated in the random number generation process to ensure accuracy even at low population numbers. Another significant limitation is that the EMCLT assumption of a normal virus concentration distribution for sufficiently large populations was not tested. In future studies, it can be addressed through Monte-Carlo simulations of normal and non-normal distributions for the parameters related to that large population, or through a sensitivity analysis.

Theoretically, crAssphage is the most suitable human biomarker among those investigated here. We ascribe this to it exclusively being related to its presence in human faecal matter, while ammonium and phosphate are less suitable as human biomarkers. It is not possible to differentiate the endogenous ammonium or phosphate contribution from exogenous sources. Moreover, the exogenous contribution, such as disinfectants discharge, is not likely to be consistent over time.

The estimated cases trend was explored with DTW for the potential to be leading for the reported cases trend. However, the wastewater RNA concentration is leading only when the number of cases is surging, after which it becomes lagging because individuals continue faecal shedding following recovery from the infection.

6. Conclusion

The study explored several approaches for the estimation of the total and infected population numbers to improve the outcome of wastewater surveillance in a near-source context (hospital). Additionally, several parameters of the wastewater matrix were studied for their impact on viruses based on previous research and ANFIS modelling.

Ammonium and phosphate were not good total population estimators due to significant fluctuations and unknown exogenous human discharge sources. Even if a more representative chemical discharge mean value was available, the fluctuations of the concentration would not allow for an adequate population number estimation. A range depending on the time of the day would be more suitable, but it is unknown whether it will be consistent over longer periods of time. CrAssphage was a more successful population estimator, but consistent sampling, model adjustment and concentration normalisation are essential to increase the accuracy of the model to a more acceptable success rate. Other human biomarkers should be explored to cross-validate with crAssphage.

The ANFIS model assessed the impact of the physico-chemical parameters on SARS-CoV-2 and crAssphage recoveries. Ammonium and phosphate concentrations had a directly proportional modelled correlation on viruses' concentrations likely due to ammonium and phosphate one of the main sources being urine, ammonium being better correlated with SARS-CoV-2 and phosphate with crAssphage. Most of the wastewater samples were within the pH range where the modelled recoveries were the highest, hence viral losses are likely to be due to other factors than pH. Additionally, the ANFIS modelled impact demonstrated that electrical conductivity correlates with virus concentration. This results from DOM and dissolved salts, such as NaCl, competitively inhibiting virus sorption, hence sample pre-concentration treatment with salts may increase virus recovery from turbid wastewater. Turbidity had a more notable impact on crAssphage recovery than on SARS-CoV-2. Although the model was based on finding clusters of similarity, other complementary statistics exploring linear and non-linear correlations, such as Spearman pairwise correlation, are essential to confirm the found impact of the physico-chemical parameters, and to subsequently establish an equation relationship.

CrAssphage, SARS-CoV-2 and NoVGII were detected at a relatively high rate and concentration in wastewater. Therefore, a large dataset was available for the modelling of the population number infected with crAssphage, SARS-CoV-2 and norovirus GII. At the same time, influenza A virus was detected sporadically, lacking trends to infer any conclusions.

The modelling results achieved for NoVGII and SARS-CoV-2 are promising. The MCBA 1 was suitable for NoVGII and influenza A virus, while the MCBA 2, 3 and the EMCLT was suitable for SARS-CoV-2 and crAssphage. The MCBA 1 has consistently calculated a norovirus infected population of 1-3 individuals in intervals of 3-4 days from an expected daily dynamic population of >1500 individuals. It was deemed to be an accurate estimation due to the norovirus infection period of up to 3 days. The model also identified a norovirus outbreak at

the hospital providing realistic estimates (5 – 40 individuals) but no clinical data was available to validate the results. For SARS-CoV-2 estimates, the MCBA 2, 3 and the EMCLT has provided realistic estimates. For validation, the infected estimates proportion cases were compared to the reported proportion observed from county-scale data, which showed that the estimated proportions were significantly overestimated. When converted to percentage changes, there was a significant correlation between the trends of the estimated and reported SARS-CoV-2 cases proportions. Consequently, the estimated cases should be compared directly with the reported cases rather than their total population proportions.

The capability to estimate the total and the infected population number is essential for the progression of WBE, and this study has provided insights into the effect of wastewater physico-chemical parameters and the usefulness of the EMCLT and different MCBAs. The use of infected population proportions to compare catchments of different sizes yielded inaccurate results due to an inaccurate crAssphage-based total population estimation. Due to the difficulty to count (dynamic population) and obtain the active number of infectious cases within a hospital catchment, future infected population models research should focus on larger catchments and perform direct comparisons with the number of reported cases.

7. Future Directions

To improve future results of the ANFIS model, the influence range of the cluster should be reduced, hence achieving smaller and more numerous clusters and, subsequently, a higher number of membership functions, allowing for a higher degree of freedom as to how the input parameters influence the output which is the virus concentration. However, this approach would require larger datasets and splitting the dataset into training and testing dataset, e.g., with a 80-20 ratio, as to be able to validate the model and to avoid model overfitting to the training data.

For the exploration of the wastewater parameters influence on virus recovery, besides the ANFIS modelling, monotonic and non-monotonic non-linear correlative relationships are necessary, such as the Spearman and Pearson pairwise and/or the mutual information and maximal integration correlations. The measurement of additional wastewater physico-chemical properties may provide more information on the potential influences on virus recovery (Li et al., 2022b; Amoah et al., 2022; Maal-Bared et al., 2022). For example, total or free chlorine is associated with virus inactivation and degradation and can be measured with a N,N-diethyl-p-phenylenediamine method (CDC, 2022a; Qiao et al., 2022). Other research has linked biochemical oxygen demand with an increased SARS-CoV-2 RNA concentration, likely due to the presence of organic matter (Nasseri et al., 2021). This can be measured by monitoring the dissolved oxygen either with a probe or with the Winkler method (APHA, 1992). Wastewater also contains extraction and PCR inhibitors such as metal ions which can be measured with inductivity-coupled plasma mass-spectrometry, therefore, providing an initial correlative assessment of metal ions concentration impact on virus recovery (Schrader et al., 2012; Yakimovich & Alekseev, 2018). Lastly, it would also be useful to measure total dissolved N (as a better measure of N loading rate) and also humic substances (qPCR inhibition).

To account for viral losses associated with laboratory processes, that varies among samples, and for the purpose of quality control, the wastewater can be spiked with control non-human viruses upon arrival at the laboratory or at the start of the analysis. The spiking virus should be surrogate to the target virus in terms of genome structure (e.g., ssRNA, dsRNA, ssDNA, dsDNA) and the virion structure (e.g., enveloped, non-enveloped, spherical, helical, polyhedral) (Farkas et al., 2020). Previous research has successfully used murine mengovirus, murine norovirus, and porcine respiratory and reproductive syndrome virus, but it is not limited to these viruses (Farkas et al., 2018, 2021).

Besides crAssphage and ammonium-nitrogen, other human biomarkers are suggested to be evaluated, including pepper mild mottle virus (PMMoV), human mitochondrial gene NADH dehydrogenase subunit 5 and creatinine (Maal-Bared et al., 2022; Hutchison et al., 2022). All three were previously demonstrated to improve the correlation of surveyed viruses in wastewater, particularly SARS-CoV-2, and clinical data. Furthermore, normalisation by these biomarkers have improved the WBE leading correlation, the lead time being 1 – 4 days for

PMMoV and, respectively, up to 5 days for creatinine and mitochondrial normalisation (Maal-Bared et al., 2022; Hutchison et al., 2022). Although creatinine may be not a suitable biomarker when sampling is performed at a WWTP due to wastewater long travel distances and a short half-life of approximately 24 hours (Thai et al., 2014), it might be a suitable biomarker when the sampling is performed near the source such as at a manhole near a hospital.

Before re-evaluating the EMCLT and the MCBAs, parameters such as faecal load per capita, virus faecal shedding rate or the volumetric flow rate should be researched in more detail in terms of their values range and distribution. This study has assumed a normal distribution which is likely to be inaccurate. Therefore, if present, the non-normal distribution should be integrated into MCBAs as explained in "4.3 Models Validation".

An additional infection number model discussed parallelly to the main modelling work was SEIR, which as explained in "4.3 Models Validation" performed significantly better than MCBA or EMCLT with a similarity ratio of 10:9 and 1:1 between estimated and adjusted reported cases (McMahan et al., 2022, 2021). The advantage of SEIR is that it organises the population into 4 subcategories: susceptible, exposed, infectious and recovered individuals (McMahan et al., 2021). It also uses Monte-Carlo simulations to predict a potential distribution of the mass rate of virus gc/day in wastewater over time in days depending on active infectious cases (McMahan et al., 2021). Therefore, the predicted numbers are based on a continuous trend taking into account the previous epidemiological situation and the virus mass rate in wastewater according to its probability as calculated with Monte-Carlo simulations, which is unlike this study MCBAs where each individual quantitative detection of viruses in wastewater is an independent event from previous detections. SEIR may provide an advantage in a residential area, however, since it relies on the susceptible, exposed, infectious or recovered number of individuals, the MCBA may be more appropriate for a hospital setting due to the number of people being more dynamic.

8. Bibliography

Adamczyk, Z., Batys, P. & Barbasz, J. (2021) SARS-CoV-2 virion physicochemical characteristics pertinent to abiotic substrate attachment. *Current Opinion in Colloid and Interface Science*.55 p.101466. doi:10.1016/j.cocis.2021.101466.

Ahmed, W., Bivins, A., Simpson, S.L., Smith, W.J.M., Metcalfe, S., McMinn, B., Symonds, E.M. & Korajkic, A. (2021) Comparative analysis of rapid concentration methods for the recovery of SARS-CoV-2 and quantification of human enteric viruses and a sewage-associated marker gene in untreated wastewater. *Science of the Total Environment*. 799, 149386. Doi:10.1016/j.scitotenv.2021.149386.

Ahmed, W., Payyappat, S., Cassidy, M., Besley, C. & Power, K. (2018) Novel crAssphage marker genes ascertain sewage pollution in a recreational lake receiving urban stormwater runoff. *Water Research*. 145, 769–778. Doi:10.1016/j.watres.2018.08.049.

Alexander, G.C. & Stevens, R.J. (1976) Per capita phosphorus loading from domestic sewage. *Water Research*. 10 (9), 757–764. Doi:10.1016/0043-1354(76)90093-2.

Amoah, I.D., Abunama, T., Awolusi, O.O., Pillay, L., Pillay, K., Kumari, S. & Bux, F. (2022) Effect of selected wastewater characteristics on estimation of SARS-CoV-2 viral load in wastewater. *Environmental Research*. 203, 111877. Doi:10.1016/j.envres.2021.111877.

Anjos, D., Fiaccadori, F.S., Servian, C. do P., da Fonseca, S.G., Guilarde, A.O., Borges, M.A.S.B., Franco, F.C., Ribeiro, B.M. & Souza, M. (2022) SARS-CoV-2 loads in urine, sera and stool specimens in association with clinical features of COVID-19 patients. *Journal of Clinical Virology Plus*. 2 (1), 100059. Doi:10.1016/j.jcvp.2021.100059.

APHA (1992) Standard methods for the examination of water and wastewater. *American Public Health Association*. 18. https://archive.epa.gov/water/archive/web/html/vms52.html.

Arena, C., Amoros, J.P., Vaillant, V., Balay, K., Chikhi-Brachet, R., Varesi, L., Arrighi, J., Blanchon, T., Carrat, F., Hanslik, T. & Falchi, A. (2012) Simultaneous investigation of influenza and enteric viruses in the stools of adult patients consulting in general practice for acute diarrhea. *Virology Journal*. 9, 116. Doi:10.1186/1743-422X-9-116.

Arinaminpathy, N. & Grenfell, B. (2010) Dynamics of Glycoprotein Charge in the Evolutionary History of Human Influenza J.-P. Vartanian (ed.). *PloS ONE*. 5 (12), e15674. Doi:10.1371/journal.pone.0015674.

Barclay, L., Park, G.W., Vega, E., Hall, A., Parashar, U., Vinjé, J. & Lopman, B. (2014) Infection control for norovirus. *Clinical Microbiology and Infection*.20 (8) pp.731–740. Doi:10.1111/1469-0691.12674.

Been, F., Rossi, L., Ort, C., Rudaz, S., Delémont, O. & Esseiva, P. (2014) Population normalization with ammonium in wastewater-based epidemiology: Application to illicit drug monitoring. *Environmental Science and Technology*. 48 (14), 8162–8169. Doi:10.1021/es5008388.

bioMérieux (2022) *NucliSENS easyMAG*. 2022. https://www.biomerieux.co.uk/product/nuclisensr-easymagr [Accessed: 21 August 2022].

Blümel, J., Burger, R., Drosten, C., Gröner, A., Gürtler, L., Heiden, M., Hildebrandt, M., Jansen, B., Klamm, H., Montag-Lessing, T., Offergeld, R., Pauli, G., Seitz, R., Schlenkrich, U., Schottstedt, V., Willkommen, H., von König, C.H.W. & Schweiger, B. (2009) Influenza virus. *Transfusion Medicine and Hemotherapy*.36 (1) pp.32–39. Doi:10.1159/000197314.

Cao, H., Tsai, F.T.-C. & Rusch, K.A. (2010) Salinity and Soluble Organic Matter on Virus Sorption in Sand and Soil Columns. *Ground Water*. 48 (1), 42–52. Doi:10.1111/j.1745-6584.2009.00645.x.

Casanova, L., Rutala, W.A., Weber, D.J. & Sobsey, M.D. (2009) Survival of surrogate coronaviruses in water. *Water Research*. 43 (7), 1893–1898. Doi:10.1016/j.watres.2009.02.002.

CDC (2022a) Chlorine Residual Testing. *Centers for Disease Control and Prevention*. https://www.cdc.gov/safewater/publications_pages/chlorineresidual.pdf.

CDC (2018) *History of 1918 Flu Pandemic*. 21 January 2018. Centers for Disease Control and Prevention. https://www.cdc.gov/flu/pandemic-resources/1918-commemoration/1918-pandemic-history.htm [Accessed: 21 August 2022].

CDC (2022b) *Key Facts About Influenza (Flu)*. 25 August 2022. Centers for Disease Control and Prevention. https://www.cdc.gov/flu/about/keyfacts.htm [Accessed: 19 September 2022].

CDC (2021) *Norovirus Worldwide*. 5 March 2021. Centers for Disease Control and Prevention. https://www.cdc.gov/norovirus/trends-outbreaks/worldwide.html [Accessed: 21 August 2022].

Chan, M.C.W., Lee, N., Chan, P.K.S., Leung, T.F. & Sung, J.J.Y. (2009) Fecal detection of influenza A virus in patients with concurrent respiratory and gastrointestinal symptoms. *Journal of Clinical Virology*. 45 (3), 208–211. Doi:10.1016/j.jcv.2009.06.011.

Chan, M.C.W., Lee, N., Chan, P.K.S., To, K.F., Wong, R.Y.K., Ho, W.S., Ngai, K.L.K. & Sung, J.J.Y. (2011) Seasonal influenza a virus in feces of hospitalized adults. *Emerging Infectious Diseases*. 17 (11), 2038–2042. Doi:10.3201/eid1711.110205.

Chang, H.-J., Huang, K.-C. & Wu, C.-H. (2006) Determination of Sample Size in Using Central Limit Theorem for Weibull Distribution. *Information and Management Sciences*. 17 (3), 31–46.

Chapuis-Lardy, L., Fiorini, J., Toth, J. & Dou, Z. (2004) Phosphorus concentration and solubility in dairy feces: Variability and affecting factors. *Journal of Dairy Science*. 87 (12), 4334–4341. Doi:10.3168/jds.S0022-0302(04)73579-1.

Cooper, B.S., Evans, S., Jafari, Y., Pham, T.M., Mo, Y., Lim, C., Pritchard, M.G., Pople, D., Hall, V., Stimson, J., Eyre, D.W., Read, J.M., Donnelly, C.A., Horby, P., Watson, C., Funk, S., Robotham, J. V. & Knight, G.M. (2023) The burden and dynamics of hospital-acquired SARS-CoV-2 in England. *Nature*. 623 (7985), 132–138. doi:10.1038/s41586-023-06634-z.

Corpuz, M.V.A., Buonerba, A., Vigliotta, G., Zarra, T., Ballesteros, F., Campiglia, P., Belgiorno, V., Korshin, G. & Naddeo, V. (2020) Viruses in wastewater: occurrence, abundance and detection methods. *Science of the Total Environment*. 745, 140910. Doi:10.1016/j.scitotenv.2020.140910.

Cuevas-Ferrando, E., Pérez-Cataluña, A., Falcó, I., Randazzo, W. & Sánchez, G. (2022) Monitoring Human Viral Pathogens Reveals Potential Hazard for Treated Wastewater Discharge or Reuse. *Frontiers in Microbiology*. 13, 779. Doi:10.3389/fmicb.2022.836193.

Cupisti, A. & Gallieni, M. (2018) Urinary phosphorus excretion: Not what we have believed it to be? *Clinical Journal of the American Society of Nephrology*.13 (7) pp.973–974. Doi:10.2215/CJN.06260518.

Diemert, S. & Yan, T. (2020) Municipal wastewater surveillance revealed a high community disease burden of a rarely reported and possibly subclinical salmonella enterica serovar derby strain. *Applied and Environmental Microbiology*. 86 (17). doi:10.1128/AEM.00814-20.

EMCDDA (2022) *Frequently-asked questions (FAQ): wastewater-based epidemiology and drugs*. 2022. European Monitoring Centre for Drugs and Drug Addiction. https://www.emcdda.europa.eu/publications/topic-overviews/content/wastewater-faq_en [Accessed: 25 November 2022].

Farkas, K., Adriaenssens, E.M., Walker, D.I., McDonald, J.E., Malham, S.K. & Jones, D.L. (2019) Critical Evaluation of CrAssphage as a Molecular Marker for Human-Derived Wastewater Contamination in the Aquatic Environment. *Food and Environmental Virology*. 11 (2), 113–119. Doi:10.1007/s12560-019-09369-1.

Farkas, K., Hillary, L.S., Malham, S.K., McDonald, J.E. & Jones, D.L. (2020) Wastewater and public health: the potential of wastewater surveillance for monitoring COVID-19. *Current Opinion in Environmental Science and Health*.17 pp.14–20. Doi:10.1016/j.coesh.2020.06.001.

Farkas, K., Hillary, L.S., Thorpe, J., Walker, D.I., Lowther, J.A., McDonald, J.E., Malham, S.K. & Jones, D.L. (2021) Concentration and quantification of sars-cov-2 rna in wastewater using polyethylene glycol-based concentration and qrt-pcr. *Methods and Protocols*. 4 (1), 1–9. Doi:10.3390/MPS4010017.

Farkas, K., Marshall, M., Cooper, D., McDonald, J.E., Malham, S.K., Peters, D.E., Maloney, J.D. & Jones, D.L. (2018) Seasonal and diurnal surveillance of treated and untreated wastewater for human enteric viruses. *Environmental Science and Pollution Research*. 25 (33), 33391–33401. Doi:10.1007/s11356-018-3261-y.

Farkas, K., Pântea, I., Woodhall, N., Williams, D., Lambert-Slosarska, K., Williams, R. & Jones, D.L. (2023) *Diurnal changes in pathogenic and indicator virus concentrations in wastewater*.

Farkas, K., Pellett, C., Alex-Sanders, N., Bridgman, M.T.P., Corbishley, A., Grimsley, J.M.S., Kasprzyk-Hordern, B., Kevill, J.L., Pântea, I., Richardson-O'Neill, I.S., Lambert-Slosarska, K., Woodhall, N. & Jones, D.L. (2022a) Comparative Assessment of Filtration- and Precipitation-Based Methods for the Concentration of SARS-CoV-2 and Other Viruses from Wastewater S.P. Faucher (ed.). *Microbiology Spectrum*. 10 (4). Doi:10.1128/spectrum.01102-22.

Farkas, K., Peters, D.E., McDonald, J.E., de Rougemont, A., Malham, S.K. & Jones, D.L. (2017) Evaluation of Two Triplex One-Step qRT-PCR Assays for the Quantification of Human Enteric Viruses in Environmental Samples. *Food and Environmental Virology*. 9 (3), 342–349. Doi:10.1007/s12560-017-9293-5.

Farkas, K., Williams, R., Alex-Sanders, N., Grimsley, J.M.S., Pântea, I., Wade, M.J., Woodhall, N. & Jones, D.L. (2022b) *Wastewater-based monitoring of SARS-CoV-2 at UK airports and its potential role in international public health surveillance*.

Forés, E., Bofill-Mas, S., Itarte, M., Martínez-Puchol, S., Hundesa, A., Calvo, M., Borrego, C.M., Corominas, L.L., Girones, R. & Rusiñol, M. (2021) Evaluation of two rapid ultrafiltration-based methods for SARS-CoV-2 concentration from wastewater. *Science of the Total Environment*. 768, 144786. Doi:10.1016/j.scitotenv.2020.144786.

Franses, P.H. & Wiemann, T. (2020) Intertemporal Similarity of Economic Time Series: An Application of Dynamic Time Warping. *Computational Economics*. 56 (1), 59–75. Doi:10.1007/s10614-020-09986-0.

Fumian, T.M., Fioretti, J.M., Lun, J.H., dos Santos, I.A.L., White, P.A. & Miagostovich, M.P. (2019) Detection of norovirus epidemic genotypes in raw sewage using next generation sequencing. *Environment International*. 123, 282–291. Doi:10.1016/j.envint.2018.11.054.

García-Aljaro, C., Ballesté, E., Muniesa, M. & Jofre, J. (2017) Determination of crAssphage in water samples and applicability for tracking human faecal pollution. *Microbial Biotechnology*. 10 (6), 1775–1780. Doi:10.1111/1751-7915.12841.

Guo, M., Tao, W., Flavell, R.A. & Zhu, S. (2021) Potential intestinal infection and faecal–oral transmission of SARS-CoV-2. *Nature Reviews Gastroenterology and Hepatology*.18 (4) pp.269–283. Doi:10.1038/s41575-021-00416-6.

Gupta, S., Parker, J., Smits, S., Underwood, J. & Dolwani, S. (2020) Persistent viral shedding of SARS-CoV-2 in faeces – a rapid review. *Colorectal Disease*. 22 (6), 611–620. Doi:10.1111/codi.15138.

Hamdi, S., Rousseau, G.M., Labrie, S.J., Kourda, R.S., Tremblay, D.M., Moineau, S. & Slama, K.B. (2016) Characterization of Five Podoviridae Phages Infecting Citrobacter freundii. *Frontiers in Microbiology*. 7 (JUN), 1023. Doi:10.3389/fmicb.2016.01023.

Hamming, I., Timens, W., Bulthuis, M., Lely, A., Navis, G. & van Goor, H. (2004) Tissue distribution of ACE2 protein, the functional receptor for SARS coronavirus. A first step in understanding SARS pathogenesis. *The Journal of Pathology*. 203 (2), 631–637. Doi:10.1002/path.1570.

Han, M.S., Seong, M.-W., Heo, E.Y., Park, J.H., Kim, N., Shin, S., Cho, S.I., Park, S.S. & Choi, E.H. (2020) Sequential Analysis of Viral Load in a Neonate and Her Mother Infected With Severe Acute Respiratory Syndrome Coronavirus 2. *Clinical Infectious Diseases*. 71 (16), 2236–2239. Doi:10.1093/cid/ciaa447.

Hasan, S.W., Ibrahim, Y., Daou, M., Kannout, H., Jan, N., Lopes, A., Alsafar, H. & Yousef, A.F. (2021) Detection and quantification of SARS-CoV-2 RNA in wastewater and treated effluents: Surveillance of COVID-19 epidemic in the United Arab Emirates. *Science of the Total Environment*. 764, 142929. Doi:10.1016/j.scitotenv.2020.142929.

Hassan, E. & Baldridge, M.T. (2019) Norovirus encounters in the gut: multifaceted interactions and disease outcomes. *Mucosal Immunology*.12 (6) pp.1259–1267. Doi:10.1038/s41385-019-0199-4.

He, X., Lau, E.H.Y., Wu, P., Deng, X., Wang, J., et al. (2020) Temporal dynamics in viral shedding and transmissibility of COVID-19. *Nature Medicine*. 26 (5), 672–675. doi:10.1038/s41591-020-0869-5.

Hellmér, M., Paxéus, N., Magnius, L., Enache, L., Arnholm, B., Johansson, A., Bergström, T. & Norder, H. (2014) Detection of pathogenic viruses in sewage provided early warnings of hepatitis A virus and norovirus outbreaks. *Applied and Environmental Microbiology*. 80 (21), 6771–6781. Doi:10.1128/AEM.01981-14.

Hirose, R., Daidoji, T., Naito, Y., Watanabe, Y., Arai, Y., Oda, T., Konishi, H., Yamawaki, M., Itoh, Y. & Nakaya, T. (2016) Long-term detection of seasonal influenza RNA in faeces and intestine. *Clinical Microbiology and Infection*. 22 (9), 813.e1-813.e7. doi:10.1016/j.cmi.2016.06.015.

Honap, T.P., Sankaranarayanan, K., Schnorr, S.L., Ozga, A.T., Warinner, C. & Lewis, C.M. (2020) Biogeographic study of human gut-associated crAssphage suggests impacts from industrialization and recent expansion. *PloS ONE*. 15 (1). Doi:10.1371/journal.pone.0226930.

Hora, P.I., Pati, S.G., McNamara, P.J. & Arnold, W.A. (2020) Increased Use of Quaternary Ammonium Compounds during the SARS-CoV-2 Pandemic and Beyond: Consideration of Environmental Implications. *Environmental Science and Technology Letters*.7 (9) pp.622–631. Doi:10.1021/acs.estlett.0c00437.

Huang, Y., Zhou, N., Zhang, S., Yi, Y., Han, Y., Liu, M., Han, Y., Shi, N., Yang, L., Wang, Q., Cui, T. & Jin, H. (2022) Norovirus detection in wastewater and its correlation with human gastroenteritis: a

systematic review and meta-analysis. *Environmental Science and Pollution Research*.29 (16) pp.22829–22842. Doi:10.1007/s11356-021-18202-x.

Huhti, L., Szakal, E.D., Puustinen, L., Salminen, M., Huhtala, H., Valve, O., Blazevic, V. & Vesikari, T. (2011) Norovirus GII-4 causes a more severe gastroenteritis than other noroviruses in young children. *Journal of Infectious Diseases*. 203 (10), 1442–1444. Doi:10.1093/infdis/jir039.

Hutchison, J.M., Li, Z., Chang, C.-N., Hiripitiyage, Y., Wittman, M. & Sturm, B.S.M. (2022) Improving correlation of wastewater SARS-CoV-2 gene copy numbers with COVID-19 public health cases using readily available biomarkers. *FEMS Microbes*. 3. Doi:10.1093/femsmc/xtac010.

IBM (2020) *Monte Carlo Simulation*. 24 August 2020. https://www.ibm.com/uk-en/cloud/learn/monte-carlo-simulation [Accessed: 21 August 2022].

Jennings, W.C., Gálvez-Arango, E., Prieto, A.L. & Boehm, A.B. (2020) CrAssphage for fecal source tracking in Chile: Covariation with norovirus, HF183, and bacterial indicators. *Water Research X*. 9. Doi:10.1016/j.wroa.2020.100071.

Jones, D.L., Baluja, M.Q., Graham, D.W., Corbishley, A., McDonald, J.E., Malham, S.K., Hillary, L.S., Connor, T.R., Gaze, W.H., Moura, I.B., Wilcox, M.H. & Farkas, K. (2020) Shedding of SARS-CoV-2 in feces and urine and its potential role in person-to-person transmission and the environment-based spread of COVID-19. *Science of the Total Environment*. 749, 141364. Doi:10.1016/j.scitotenv.2020.141364.

Jones, D. L., Grimsley, J. M. S., Kevill, J. L., Williams, R., Pellett, C., Lambert-Slosarska, K., Singer, A. C., Williams, G. B., Bargiela, R., Brown, R. W., Wade, M. J., & Farkas, K. (2022). Critical Evaluation of Different Passive Sampler Materials and Approaches for the Recovery of SARS-CoV-2, Faecal-Indicator Viruses and Bacteria from Wastewater. *Water (Switzerland)*, *14*(21), 3568. https://doi.org/10.3390/w14213568

Joung, M.J., Mangat, C., Mejia, E., Nagasawa, A., Nichani, A., Perez-Iratxeta, C., Peterson, S. & Champredon, D. (2023) Coupling wastewater-based epidemiological surveillance and modelling of SARS-COV-2/COVID-19: Practical applications at the Public Health Agency of Canada. *Canada Communicable Disease Report*. 49 (5), 166–174. doi:10.14745/ccdr.v49i05a01.

Karst, S.M. & Wobus, C.E. (2015) A Working Model of How Noroviruses Infect the Intestine. *PloS Pathogens*. 11 (2). Doi:10.1371/journal.ppat.1004626.

Kenton, W. (2022) *Monte Carlo Simulation*. 11 August 2022. Investopedia. https://www.investopedia.com/terms/m/montecarlosimulation.asp [Accessed: 21 August 2022].

Keshavarzi, A., Sarmadian, F., Shiri, J., Iqbal, M., Tirado-Corbalá, R. & Omran, E.S.E. (2017) Application of ANFIS-based subtractive clustering algorithm in soil Cation Exchange Capacity estimation using soil and remotely sensed data. *Measurement: Journal of the International Measurement Confederation*. 95, 173–180. Doi:10.1016/j.measurement.2016.10.010.

Kevill, J.L., Pellett, C., Farkas, K., Brown, M.R., Bassano, I., Denise, H., McDonald, J.E., Malham, S.K., Porter, J., Warren, J., Evens, N.P., Paterson, S., Singer, A.C. & Jones, D.L. (2022) A comparison of precipitation and filtration-based SARS-CoV-2 recovery methods and the influence of temperature, turbidity, and surfactant load in urban wastewater. *Science of the Total Environment*. 808, 151916. Doi:10.1016/j.scitotenv.2021.151916.

al Khatib, H.A., Coyle, P. v., al Maslamani, M.A., al Thani, A.A., Pathan, S.A. & Yassine, H.M. (2021) Molecular and biological characterization of influenza A viruses isolated from human fecal samples. *Infection, Genetics and Evolution*. 93, 104972. Doi:10.1016/j.meegid.2021.104972.

Klapsa, D., Wilton, T., Zealand, A., Bujaki, E., Saxentoff, E., et al. (2022) Sustained detection of type 2 poliovirus in London sewage between February and July, 2022, by enhanced environmental surveillance. *The Lancet*. 400 (10362), 1531–1538. doi:10.1016/S0140-6736(22)01804-9.

Kobayashi, Y. & Suzuki, Y. (2012) Compensatory Evolution of Net-Charge in Influenza A Virus Hemagglutinin K. Pyrc (ed.). *PloS ONE*. 7 (7), e40422. Doi:10.1371/journal.pone.0040422.

Köhler, J. (2006) Detergent Phosphates: An EU Policy Assessment. *Journal of Business Chemistry*. 3 (2). https://ssrn.com/abstract=934705.

Kongprajug, A., Mongkolsuk, S. & Sirikanchana, K. (2019) CrAssphage as a Potential Human Sewage Marker for Microbial Source Tracking in Southeast Asia. *Environmental Science and Technology Letters*. 6 (3), 156–164. Doi:10.1021/acs.estlett.9b00041.

Lai, C.C., Wang, Y.H., Wu, C.Y., Hung, C.H., Jiang, D.D.S. & Wu, F.T. (2013) A norovirus outbreak in a nursing home: Norovirus shedding time associated with age. *Journal of Clinical Virology*. 56 (2), 96–101. Doi:10.1016/j.jcv.2012.10.011.

Lamers, M.M., Beumer, J., Vaart, J. van der, Knoops, K., Puschhof, J., et al. (2020) SARS-CoV-2 productively infects human gut enterocytes. *Science*. 369 (6499), 50–54. Doi:10.1126/science.abc1669.

LaMorte, W.W. (2016) *Central Limit Theorem*. 24 July 2016. Boston University School of Public Health. https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_probability/BS704_Probability12.html [Accessed: 21 August 2022].

Lee, N., Chan, M.C.W., Wong, B., Choi, K.W., Sin, W., Lui, G., Chan, P.K.S., Lai, R.W.M., Cockram, C.S., Sung, J.J.Y. & Leung, W.K. (2007) Fecal viral concentration and diarrhea in norovirus gastroenteritis. *Emerging Infectious Diseases*. 13 (9), 1399–1401. Doi:10.3201/eid1309.061535.

LennTech (2022) *Phosphorous removal from wastewater*. 2022. https://www.lenntech.com/phosphorous-removal.htm [Accessed: 25 November 2022].

Li, X., Kulandaivelu, J., Guo, Y., Zhang, S., Shi, J., O'Brien, J., Arora, S., Kumar, M., Sherchan, S.P., Honda, R., Jackson, G., Luby, S.P. & Jiang, G. (2022a) SARS-CoV-2 shedding sources in wastewater and implications for wastewater-based epidemiology. *Journal of Hazardous Materials*. 432, 128667. Doi:10.1016/j.jhazmat.2022.128667.

Li, X., Zhang, S., Shi, J., Luby, S.P. & Jiang, G. (2021) Uncertainties in estimating SARS-CoV-2 prevalence by wastewater-based epidemiology. *Chemical Engineering Journal*.415 p.129039. doi:10.1016/j.cej.2021.129039.

Li, Z.H., Wang, J.X., Lu, M., Zhang, T., Wang, X.C., Li, W.W. & Yu, H.Q. (2022b) Hospital sewage treatment facilities witness the fighting against the COVID-19 pandemic. *Journal of Environmental Management*. 309, 114728. Doi:10.1016/j.jenvman.2022.114728.

Liu, Z., Wickramasinghe, S. R., & Qian, X. (2017). Ion-specificity in protein binding and recovery for the responsive hydrophobic poly(vinylcaprolactam) ligand. *RSC Advances*, *7*(58), 36351–36360. https://doi.org/10.1039/c7ra06022j

Lukasik, J., Scott, T. M., Andryshak, D., & Farrah, S. R. (2000). Influence of salts on virus adsorption to microporous filters. *Applied and Environmental Microbiology*, *66*(7), 2914–2920. https://doi.org/10.1128/AEM.66.7.2914-2920.2000

Maal-Bared, R., Qiu, Y., Li, Q., Gao, T., Hrudey, S.E., Bhavanam, S., Ruecker, N.J., Ellehoj, E., Lee, B.E. & Pang, X. (2022) Does normalization of SARS-CoV-2 concentrations by Pepper Mild Mottle Virus improve correlations and lead time between wastewater surveillance and clinical data in Alberta (Canada): comparing twelve SARS-CoV-2 normalization approaches. *Science of the Total Environment*. 856, 158964. Doi:10.1016/j.scitotenv.2022.158964.

Madadizadeh, F., Asar, M.E. & Hosseini, M. (2015) Common statistical mistakes in descriptive statistics reports of normal and non-normal variables in biomedical sciences research. *Iranian Journal of Public Health*.44 (11) pp.1557–1558.

Maugeri, A., Barchitta, M., Battiato, S. & Agodi, A. (2020) Estimation of unreported novel coronavirus (Sars-coV-2) infections from reported deaths: A susceptible–exposed–infectious–recovered–dead model. *Journal of Clinical Medicine*. 9 (5). Doi:10.3390/jcm9051350.

Mayo Clinic (2021) *H1N1 flu (swine flu)*. 2021. https://www.mayoclinic.org/diseases-conditions/swine-flu/symptoms-causes/syc-20378103 [Accessed: 25 November 2022].

Mayo Clinic (2022a) *Influenza (flu)*. 15 October 2022. https://www.mayoclinic.org/diseases-conditions/flu/symptoms-causes/syc-20351719 [Accessed: 11 January 2023].

Mayo Clinic (2022b) *Norovirus infection*. 4 March 2022. https://www.mayoclinic.org/diseases-conditions/norovirus/symptoms-causes/syc-20355296 [Accessed: 19 September 2022].

McMahan, C.S., Lewis, D., Deaver, J.A., Dean, D., Rennert, L., Kalbaugh, C.A., Shi, L., Kriebel, D., Graves, D., Popat, S.C., Karanfil, T. & Freedman, D.L. (2022) Predicting COVID-19 Infected Individuals in a Defined Population from Wastewater RNA Data. *ACS ES and T Water*. Doi:10.1021/acsestwater.2c00105.

McMahan, C.S., Self, S., Rennert, L., Kalbaugh, C., Kriebel, D., Graves, D., Colby, C., Deaver, J.A., Popat, S.C., Karanfil, T. & Freedman, D.L. (2021) COVID-19 wastewater epidemiology: a model to estimate infected populations. *The Lancet Planetary Health*. 5 (12), e874–e881. Doi:10.1016/S2542-5196(21)00230-8.

MedlinePlus (2022) *Urine 24-hour volume*. 2022. https://medlineplus.gov/ency/article/003425.htm [Accessed: 21 August 2022].

Mertens, B.S. & Velev, O.D. (2015) Characterization and control of surfactant-mediated Norovirus interactions. *Soft Matter*. 11 (44), 8621–8631. Doi:10.1039/c5sm01778e.

Mertz, D. & Speicher, D. (2020) *C. difficile: The Role of Colonization and Phages*. 2020. Insitute for Infectious Disease Research. https://iidr.mcmaster.ca/c-difficile-the-role-of-colonization-and-phages/ [Accessed: 21 August 2022].

Minodier, L., Charrel, R.N., Ceccaldi, P.E., van der Werf, S., Blanchon, T., Hanslik, T. & Falchi, A. (2015) Prevalence of gastrointestinal symptoms in patients with influenza, clinical significance, and pathophysiology of human influenza viruses in faecal samples: What do we know? *Virology Journal*.12 (1) p.215. doi:10.1186/s12985-015-0448-4.

Morvan, M., Lojacomo, A., Souque, C., Wade, M., Hoffmann, T., Pouwels, K., Singer, A., Bunce, J., Engeli, A., Grimsley, J., O'Reilly, K. & Danon, L. (2022) Estimating SARS-CoV-2 prevalence from large-scale wastewater surveillance: insights from combined analysis of 44 sites in England. *International Journal of Infectious Diseases*. 116, S24. doi:10.1016/j.ijid.2021.12.057.

Mulvaney, R.L. (2018) Nitrogen-Inorganic Forms. In: *Methods of Soil Analysis, Part 3: Chemical Methods*. Wiley. Pp. 1123–1184. Doi:10.2136/sssabookser5.3.c38.

Münch, E. von & Winker, M. (2011) *Technology review of urine diversion components*. May 2011. https://www.susana.org/_resources/documents/default/2-875-giz2011-en-technology-review-urine-diversion.pdf [Accessed: 21 August 2022].

Murphy, J. & Riley, J.P. (1962) A modified single solution method for the determination of phosphate in natural waters. *Analytica Chimica Acta*. 27 (C), 31–36. doi:10.1016/S0003-2670(00)88444-5.

Murray, C.J.L. & Piot, P. (2021) The Potential Future of the COVID-19 Pandemic: Will SARS-CoV-2 Become a Recurrent Seasonal Infection? *JAMA - Journal of the American Medical Association*.325 (13) pp.1249–1250. doi:10.1001/jama.2021.2828.

Nasseri, S., Yavarian, J., Baghani, A.N., Azad, T.M., Nejati, A., Nabizadeh, R., Hadi, M., Jandaghi, N.Z.S., Vakili, B., Vaghefi, S.K.A., Baghban, M., Yousefi, S., Nazmara, S. & Alimohammadi, M. (2021) The presence of SARS-CoV-2 in raw and treated wastewater in 3 cities of Iran: Tehran, Qom and Anzali during coronavirus disease 2019 (COVID-19) outbreak. *Journal of Environmental Health Science and Engineering*. 19 (1), 573–584. doi:10.1007/s40201-021-00629-6.

NHS (2022) *Symptoms of coronavirus (COVID-19)*. 2022. National Health Service. https://www.nhs.uk/conditions/coronavirus-covid-19/symptoms/ [Accessed: 11 January 2023].

Nourbakhsh, S., Fazil, A., Li, M., Mangat, C.S., Peterson, S.W., Daigle, J., Langner, S., Shurgold, J., D'Aoust, P., Delatolla, R., Mercier, E., Pang, X., Lee, B.E., Stuart, R., Wijayasri, S. & Champredon, D. (2022) A wastewater-based epidemic model for SARS-CoV-2 with application to three Canadian cities. *Epidemics*. 39, 100560. doi:10.1016/j.epidem.2022.100560.

Oberle, W. (2015) *Monte Carlo Simulations: Number of Iterations and Accuracy*. https://apps.dtic.mil/sti/pdfs/ADA621501.pdf.

O'Keeffe, J. (2021) Wastewater-based epidemiology: current uses and future opportunities as a public health surveillance tool. *Environmental Health Review*. 64 (3), 44–52. doi:10.5864/d2021-015.

Olesen, S.W., Imakaev, M. & Duvallet, C. (2021) Making waves: Defining the lead time of wastewater-based epidemiology for COVID-19. *Water Research*.202 p.117433. doi:10.1016/j.watres.2021.117433.

ONS (2022a) *How the population changed in Gwynedd: Census 2021*. 22 June 2022. Office for National Statistics. https://www.ons.gov.uk/visualisations/censuspopulationchange/W06000002/ [Accessed: 21 August 2022].

ONS (2022b) *Population and household estimates, Wales: Census 2021*. 28 June 2022. Office for National Statistics. https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/bulletins/populationandhouseholdestimateswales/census2021 [Accessed: 21 August 2022].

Park, G.W., Ng, T.F.F., Freeland, A.L., Marconi, V.C., Boom, J.A., Staat, M.A., Montmayeur, A.M., Browne, H., Narayanan, J., Payne, D.C., Cardemil, C. v., Treffiletti, A. & Vinjé, J. (2020) CrAssphage as a novel tool to detect human fecal contamination on environmental surfaces and hands. *Emerging Infectious Diseases*. 26 (8), 1731–1739. doi:10.3201/eid2608.200346.

Park, J., Cho, K.H., Lee, E., Lee, S. & Cho, J. (2018) Sorption of pharmaceuticals to soil organic matter in a constructed wetland by electrostatic interaction. *Science of the Total Environment*. 635, 1345–1350. doi:10.1016/j.scitotenv.2018.04.212.

Peccia, J., Zulli, A., Brackney, D.E., Grubaugh, N.D., Kaplan, E.H., Casanovas-Massana, A., Ko, A.I., Malik, A.A., Wang, D., Wang, M., Warren, J.L., Weinberger, D.M., Arnold, W. & Omer, S.B. (2020)

Measurement of SARS-CoV-2 RNA in wastewater tracks community infection dynamics. *Nature Biotechnology*. 38 (10), 1164–1167. doi:10.1038/s41587-020-0684-z.

Peng, L., Liu, J., Xu, W., Luo, Q., Chen, D., Lei, Z., Huang, Z., Li, X., Deng, K., Lin, B. & Gao, Z. (2020) SARS-CoV-2 can be detected in urine, blood, anal swabs, and oropharyngeal swabs specimens. *Journal of Medical Virology*. 92 (9), 1676–1680. doi:10.1002/jmv.25936.

PHW (2022) *Weekly Influenza and Acute Respiratory Infection Report*. 2022. Public Health Wales. https://phw.nhs.wales/topics/immunisation-and-vaccines/fluvaccine/weekly-influenza-and-acute-respiratory-infection-report/ [Accessed: 21 August 2022].

Qiao, Z., Ye, Y., Szczuka, A., Harrison, K.R., Dodd, M.C. & Wigginton, K.R. (2022) Reactivity of Viral Nucleic Acids with Chlorine and the Impact of Virus Encapsidation. *Environmental Science and Technology*. 56 (1), 218–227. doi:10.1021/acs.est.1c04239.

Quinlan, B. (2015) *Dimensional Analysis: How Many Monte Carlo Simulations Should I Run?* 26 January 2015. https://blog.3dcs.com/dimensional-analysis-how-many-monte-carlo-simulations-should-i-run [Accessed: 19 September 2022].

Riedel, S. (2006) Crossing the Species Barrier: The Threat of an Avian Influenza Pandemic. *Baylor University Medical Center Proceedings*. 19 (1), 16–20. doi:10.1080/08998280.2006.11928118.

Rose, C., Parker, A., Jefferson, B. & Cartmell, E. (2015) The Characterization of Feces and Urine: A Review of the Literature to Inform Advanced Treatment Technology. *Critical Reviews in Environmental Science and Technology*. 45 (17), 1827–1879. doi:10.1080/10643389.2014.1000761.

Sala-Comorera, L., Reynolds, L.J., Martin, N.A., Pascual-Benito, M., Stephens, J.H., Nolan, T.M., Gitto, A., O'Hare, G.M.P., O'Sullivan, J.J., García-Aljaro, C. & Meijer, W.G. (2021) crAssphage as a human molecular marker to evaluate temporal and spatial variability in faecal contamination of urban marine bathing waters. *Science of the Total Environment*. 789, 147828. doi:10.1016/j.scitotenv.2021.147828.

Saththasivam, J., El-Malah, S.S., Gomez, T.A., Jabbar, K.A., Remanan, R., et al. (2021) COVID-19 (SARS-CoV-2) outbreak monitoring using wastewater-based epidemiology in Qatar. *Science of the Total Environment*. 774, 145608. doi:10.1016/j.scitotenv.2021.145608.

Schmitz, B.W., Innes, G.K., Prasek, S.M., Betancourt, W.Q., Stark, E.R., Foster, A.R., Abraham, A.G., Gerba, C.P. & Pepper, I.L. (2021) Enumerating asymptomatic COVID-19 cases and estimating SARS-CoV-2 fecal shedding rates via wastewater-based epidemiology. *Science of the Total Environment*. 801, 149794. doi:10.1016/j.scitotenv.2021.149794.

Schrader, C., Schielke, A., Ellerbroek, L. & Johne, R. (2012) PCR inhibito–s - occurrence, properties and removal. *Journal of Applied Microbiology*. 113 (5), 1014–1026. doi:10.1111/j.1365-2672.2012.05384.x.

SciPy v1.9.0 Manual (2022) *scipy.ndimage.gaussian_filter1d*. 2022. https://docs.scipy.org/doc/scipy/reference/generated/scipy.ndimage.gaussian_filter1d.html [Accessed: 21 August 2022].

Scudellari, M. (2021) How the coronavirus infects cel–s - and why Delta is so dangerous. *Nature*. 595 (7869), 640–644. doi:10.1038/d41586-021-02039-y.

Seitz, S.R., Leon, J.S., Schwab, K.J., Lyon, G.M., Dowd, M., McDaniels, M., Abdulhafid, G., Fernandez, M.L., Lindesmith, L.C., Baric, R.S. & Moe, C.L. (2011) Norovirus infectivity in humans and persistence in water. *Applied and Environmental Microbiology*. 77 (19), 6884–6888. doi:10.1128/AEM.05806-11.

Short, K.R., Richard, M., Verhagen, J.H., van Riel, D., Schrauwen, E.J.A., van den Brand, J.M.A., Mänz, B., Bodewes, R. & Herfst, S. (2015) One health, multiple challenges: The inter-species transmission of influenza A virus. *One Health*. 1, 1–13. doi:10.1016/j.onehlt.2015.03.001.

Shrestha, S., Yoshinaga, E., Chapagain, S.K., Mohan, G., Gasparatos, A. & Fukushi, K. (2021) Wastewater-Based Epidemiology for Cost-Effective Mass Surveillance of COVID-19 in Low- and Middle-Income Countries: Challenges and Opportunities. *Water*. 13 (20), 2897. doi:10.3390/w13202897.

da Silva, A.K., Kavanagh, O. v., Estes, M.K. & Elimelech, M. (2011) Adsorption and aggregation properties of norovirus GI and GII virus-like particles demonstrate differing responses to solution chemistry. *Environmental Science and Technology*. 45 (2), 520–526. doi:10.1021/es102368d.

Sims, N. & Kasprzyk-Hordern, B. (2020) Future perspectives of wastewater-based epidemiology: Monitoring infectious disease spread and resistance to the community level. *Environment International*.139 p.105689. doi:10.1016/j.envint.2020.105689.

Singer, A.C., Thompson, J.R., Filho, C.R.M., Street, R., Li, X., Castiglioni, S. & Thomas, K. V. (2023) A world of wastewater-based epidemiology. *Nature Water*. 1 (5), 408–415. doi:10.1038/s44221-023-00083-8.

Sonnenwald, F., Shuttleworth, J., Bailey, O., Williams, M., Frankland, J., Rhead, B., Mark, O., Wade, M.J. & Guymer, I. (2023) Quantifying Mixing in Sewer Networks for Source Localization. *Journal of Environmental Engineering*. 149 (5), 04023019. doi:10.1061/joeedu.eeeng-7134.

Sridhar, J., Parit, R., Boopalakrishnan, G., Rexliene, M.J., Praveen, R. & Viswananathan, B. (2022) Importance of wastewater-based epidemiology for detecting and monitoring SARS-CoV-2. *Case Studies in Chemical and Environmental Engineering*. 6, 100241. doi:10.1016/j.cscee.2022.100241.

Stachler, E., Kelty, C., Sivaganesan, M., Li, X., Bibby, K. & Shanks, O.C. (2017) Quantitative CrAssphage PCR Assays for Human Fecal Pollution Measurement. *Environmental Science and Technology*. 51 (16), 9146–9154. doi:10.1021/acs.est.7b02703.

Sun, Q., Bertrand, K.A., Franke, A.A., Rosner, B., Curhan, G.C. & Willett, W.C. (2017) Reproducibility of urinary biomarkers in multiple 24-h urine samples. *American Journal of Clinical Nutrition*. 105 (1), 159–168. doi:10.3945/ajcn.116.139758.

Thai, P.K., O'Brien, J., Jiang, G., Gernjak, W., Yuan, Z., Eaglesham, G. & Mueller, J.F. (2014) Degradability of creatinine under sewer conditions affects its potential to be used as biomarker in sewage epidemiology. *Water Research*. 55, 272–279. doi:10.1016/j.watres.2014.02.035.

To, K.K.W., Chan, K.-H., Li, I.W.S., Tsang, T.-Y., Tse, H., Chan, J.F.W., Hung, I.F.N., Lai, S.-T., Leung, C.-W., Kwan, Y.-W., Lau, Y.-L., Ng, T.-K., Cheng, V.C.C., Peiris, J.S.M. & Yuen, K.-Y. (2010) Viral load in patients infected with pandemic H1N1 2009 influenza A virus. *Journal of Medical Virology*. 82 (1), 1–7. doi:10.1002/jmv.21664.

UKHSA (2022a) *Coronavirus (COVID-19) in the UK*. 2022. UK Health Security Agency. https://coronavirus.data.gov.uk/ [Accessed: 21 August 2022].

UKHSA (2023) *National norovirus and rotavirus report, week 49 report: data up to week 47 (27 November 2022)*. 5 January 2023. UK Health Security Agency. https://www.gov.uk/government/statistics/national-norovirus-and-rotavirus-surveillance-reports-2022-to-2023-season/national-norovirus-and-rotavirus-report-week-49-report-data-up-to-week-47-27-november-2022#acknowledgements [Accessed: 11 January 2023].

UKHSA (2022b) *Norovirus outbreaks increasing in England*. 3 March 2022. UK Health Security Agency. https://www.gov.uk/government/news/norovirus-outbreaks-increasing-in-england-1 [Accessed: 21 August 2022].

UKHSA (2022c) *Weekly national Influenza and COVID-19 surveillance report: Weekly national Influenza and COVID-19 surveillance report Week 14 report (up to week 13 data) 7 April 2022*. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1067092/Weekly_Flu_and_COVID-19_report_w14.pdf.

Vallejo, J.A., Trigo-Tasende, N., Rumbo-Feal, S., Conde-Pérez, K., López-Oriona, Á., et al. (2022) Modeling the number of people infected with SARS-COV-2 from wastewater viral load in Northwest Spain. *Science of the Total Environment*. 811, 152334. doi:10.1016/j.scitotenv.2021.152334.

Vaselli, N.M., Setiabudi, W., Subramaniam, K., Adams, E.R., Turtle, L., et al. (2021) Investigation of SARS-CoV-2 faecal shedding in the community: a prospective household cohort study (COVID-LIV) in the UK. *BMC Infectious Diseases*. 21 (1), 784. doi:10.1186/s12879-021-06443-7.

Weiner, I.D. & Verlander, J.W. (2013) Renal ammonia metabolism and transport. *Comprehensive Physiology*. 3 (1), 201–220. doi:10.1002/cphy.c120010.

Welsh Government (2023) *Wastewater monitoring reports: coronavirus*. 9 January 2023. https://www.gov.wales/wastewater-monitoring-reports-coronavirus [Accessed: 11 January 2023].

WHO (2022) *WHO Coronavirus (COVID-19) Dashboard*. 2022. World Health Organization . https://covid19.who.int/ [Accessed: 25 November 2022].

Wilder, M.L., Middleton, F., Larsen, D.A., Du, Q., Fenty, A., Zeng, T., Insaf, T., Kilaru, P., Collins, M., Kmush, B. & Green, H.C. (2021) Co-quantification of crAssphage increases confidence in wastewater-based epidemiology for SARS-CoV-2 in low prevalence areas. *Water Research X*. 11, 100100. doi:10.1016/j.wroa.2021.100100.

Wolfe, M.K., Duong, D., Bakker, K.M., Ammerman, M., Mortenson, L., Hughes, B., Arts, P., Lauring, A.S., Fitzsimmons, W.J., Bendall, E., Hwang, C.E., Martin, E.T., White, B.J., Boehm, A.B. & Wigginton, K.R. (2022) Wastewater-Based Detection of Two Influenza Outbreaks. *Environmental Science and Technology Letters*. 9 (8), 687–692. doi:10.1021/acs.estlett.2c00350.

Wade, M.J., Lo Jacomo, A., Armenise, E., Brown, M.R., Bunce, J.T., et al. (2022) Understanding and managing uncertainty and variability for wastewater monitoring beyond the pandemic: Lessons learned from the United Kingdom national COVID-19 surveillance programmes. *Journal of Hazardous Materials*. 424, 127456. doi:10.1016/j.jhazmat.2021.127456.

Westbury, C.F. (2010) Bayes' rule for clinicians: An introduction. *Frontiers in Psychology*. 1 (NOV). doi:10.3389/fpsyg.2010.00192.

Wölfel, R., Corman, V.M., Guggemos, W., Seilmaier, M., Zange, S., Müller, M.A., Niemeyer, D., Jones, T.C., Vollmar, P., Rothe, C., Hoelscher, M., Bleicker, T., Brünink, S., Schneider, J., Ehmann, R., Zwirglmaier, K., Drosten, C. & Wendtner, C. (2020) Virological assessment of hospitalized patients with COVID-2019. *Nature*. 581 (7809), 465–469. doi:10.1038/s41586-020-2196-x.

Xie, Y., Guo, W., Lopez-Hernadez, A., Teng, S. & Li, L. (2022) The pH Effects on SARS-CoV and SARS-CoV-2 Spike Proteins in the Process of Binding to hACE2. *Pathogens*. 11 (2), 238. doi:10.3390/pathogens11020238.

Xie, Z.W., Gl, Y., Zhao, L. & Huang, D.M. (2008) A tentatice approach to pollutant discharge coefficient of urban residential quarters-in the cities of Kunming, Dali and Luquan. *Yunan Geographic Environment Research*. 20 (2), 119–123.

Yakimovich, P. v. & Alekseev, A. v. (2018) Analysis of Wastewater by ICP-MS. *Metallurgist*. 62 (1–2), 3–9. doi:10.1007/s11015-018-0617-y.

Yangxin, Y., Jin, Z. & Bayly, A.E. (2008) Development of Surfactants and Builders in Detergent Formulations. *Chinese Journal of Chemical Engineering*. 16 (4), 517–527. http://www.paper.edu.cn/scholar/showpdf/NUT2IN0IMTj0cxeQh.

Yoon, J.G., Yoon, J., Song, J.Y., Yoon, S.Y., Lim, C.S., Seong, H., Noh, J.Y., Cheong, H.J. & Kim, W.J. (2020) Clinical significance of a high SARS-CoV-2 viral load in the Saliva. *Journal of Korean Medical Science*. 35 (20). doi:10.3346/JKMS.2020.35.E195.

Zang, R., Castro, M.F.G., McCune, B.T., Zeng, Q., Rothlauf, P.W., Sonnek, N.M., Liu, Z., Brulois, K.F., Wang, X., Greenberg, H.B., Diamond, M.S., Ciorba, M.A., Whelan, S.P.J. & Ding, S. (2020) TMPRSS2 and TMPRSS4 promote SARS-CoV-2 infection of human small intestinal enterocytes. *Science Immunology*. 5 (47). doi:10.1126/sciimmunol.abc3582.

Zhang, D., Duran, S.S.F., Lim, W.Y.S., Tan, C.K.I., Cheong, W.C.D., Suwardi, A. & Loh, X.J. (2022) SARS-CoV-2 in wastewater: From detection to evaluation. *Materials Today Advances*.13 p.100211. doi:10.1016/j.mtadv.2022.100211.

Zhang, D., Zabarankin, M. & Prigiobbe, V. (2019) Modeling salinity-dependent transport of viruses in porous media. *Advances in Water Resources*. 127, 252–263. doi:10.1016/j.advwatres.2019.03.017.

Zhang, H., Kang, Z., Gong, H., Xu, D., Wang, J., Li, Z., Li, Z., Cui, X., Xiao, J., Zhan, J., Meng, T., Zhou, W., Liu, J. & Xu, H. (2020) Digestive system is a potential route of COVID-19: An analysis of single-cell coexpression pattern of key proteins in viral entry process. *Gut*. 69 (6), 1010–1018. doi:10.1136/gutjnl-2020-320953.

Zhang, Y., Cen, M., Hu, M., Du, L., Hu, W., Kim, J.J. & Dai, N. (2021) Prevalence and Persistent Shedding of Fecal SARS-CoV-2 RNA in Patients With COVID-19 Infection: A Systematic Review and Meta-analysis. *Clinical and Translational Gastroenterology*. 12 (4), e00343. doi:10.14309/ctg.0000000000000343.

Zheng, Q. da, Lin, J.G., Pei, W., Guo, M.X., Wang, Z. & Wang, D.G. (2017) Estimating nicotine consumption in eight cities using sewage epidemiology based on ammonia nitrogen equivalent population. *Science of the Total Environment*. 590–591, 226–232. doi:10.1016/j.scitotenv.2017.02.214.

Zhu, H., Li, H.Z., Ye, J.F. & Fu, W. (2010) Coefficients of major pollutants in domestic sewage in Shanghai. *Zhongguo Huanjing Kexue/China Environmental Science*. 30 (1), 37–41.

9. Appendix: Supplementary Materials

```
import random
from openpyxl import load_workbook
import numpy
import pandas as pd
import statistics
import scipy.stats as st


'''
1. Estimates N based on an average concentration.
2. Generates an N_dataset based on the N calculated in step 1, where the elements are 0.005xN, 0.01xN,
0.015xN up to 2xN.
3. For each N from the N_dataset, creates a Monte-Carlo full simulation.
   Assumes a degradation between 0% and 100% or [0;1].
4. Checks each experimentally measured concentration within an acceptable deviation against the
concentrations generated
    for each of the N from step 3. Subsequently calculates what is the minimum N and the maximum N.
5. Also, per each measured concentration, calculates and exports the probability for each N.
'''


def n_estimate():
    test_concentration_gc_L = 918513  # to be used only once
    '''
    test_concentration_gc_L = 23005960  # to be used only once
    test_concentration_gc_L = 23368  # to be used only once
    test_concentration_gc_L = 4000  # to be used only once
    test_concentration_gc_L = 918513  # to be used only once
    '''
    flow_l_day = 1920000
    mean_alpha = 149
    mean_beta = 301995172  # for Norovirus GII
    '''
    mean_beta = 125892541  # for CrAssphage
    mean_beta = 32000  # for Influenza A
    mean_beta = 8010000  # for SARS-CoV-2
    mean_beta = 301995172  # for Norovirus GII
    '''
    mean_gamma = 0
    n_estimate_local = (test_concentration_gc_L * flow_l_day) / (mean_alpha * mean_beta * (1 - mean_gamma))
    return n_estimate_local


def get_virus_concentration_list():
    from openpyxl import load_workbook
    workbook_v6 = load_workbook(filename="Xres_c1.xlsx", read_only=True)  # filename="Xres_c1.xlsx" for
influenza and norovirus
    sheet_v6 = workbook_v6.active
    labels = []
    '''
    for row in sheet_v6.iter_rows(min_row=1, max_row=1):
        for i in row:
            labels.append(i.value)
    '''
    virus_concentration_dataset = []
    # min_col=42, max_col=42  #  for SARS_CoV-2
    # min_col=36, max_col=36  #  for CrAssphage
    # min_col=18, max_col=18  #  for Influenza A and Norovirus GII
    # min_row=2, max_row=353  #  for Influenza A
    # min_row=354, max_row=466  #  for Norovirus GII
    for value in sheet_v6.iter_rows(min_row=354, max_row=466, min_col=18, max_col=18, values_only=True):
        for i in value:
            if i >= 0:
                virus_concentration_dataset.append(i*666.6666)  # *666.6666 only for CrAssphage, Norovirus
GII, Influenza A
            else:
                virus_concentration_dataset.append(1)
    sample_date_dataset = []
    # column 5 for flu and norovirus
    # column 8 for crass and sars-cov-2
    for value in sheet_v6.iter_rows(min_row=354, max_row=466, min_col=5, max_col=5, values_only=True):
        for i in value:
            date_only = i.date()
            sample_date_dataset.append(date_only)
    '''
    grab_comp_dataset = []
    for value in sheet_v6.iter_rows(min_row=2, min_col=12, max_col=12, values_only=True):
        for i in value:
            grab_comp_dataset.append(i)
     '''
    return virus_concentration_dataset, sample_date_dataset


def monte_carlo_concentration_list_for_n(n):
    monte_carlo_concentration_list = []
```

```
    for i in range(2000):  # choose the number of simulations
        concentration = 0
        for j in range(n):
            alpha_random_dataset = random.randint(51, 796)
            beta_random_dataset = random.randint(61659500, 12882495517)  # for Norovirus GII
            '''
            beta_random_dataset = random.randint(12589, 12589254117)  # for CrAssphage
            beta_random_dataset = random.randint(4900, 80000000)  # for Influenza A
            beta_random_dataset = random.randint(20000, 65300000000)  # for SARS-CoV-2
            beta_random_dataset = random.randint(61659500, 12882495517)  # for Norovirus GII
            '''
            gamma_random_dataset = random.uniform(0, 1)
            concentration_temp = (alpha_random_dataset * beta_random_dataset * (1 - gamma_random_dataset))
            concentration += concentration_temp
        flow_random_dataset = random.randint(1824000, 2072235)
        concentration = concentration / flow_random_dataset
        monte_carlo_concentration_list.append(concentration)
    return monte_carlo_concentration_list


def determine_range_of_n_by_a_measured_concentration(measured_concentration):
    deviation_min = measured_concentration - acceptable_deviation
    deviation_max = measured_concentration + acceptable_deviation
    in_n_number_of_concentrations_within_measured_interval = []
    i = 0
    for key in monte_carlo_concentration_dictionary_by_n:
        temp_concentrations_list = monte_carlo_concentration_dictionary_by_n[key]
        in_n_number_of_concentrations_within_measured_interval.append(0)
        for j in range(len(temp_concentrations_list)):
            if deviation_min <= temp_concentrations_list[j] <= deviation_max:
                in_n_number_of_concentrations_within_measured_interval[i] += 1
        i += 1
    del i
    del temp_concentrations_list
    return in_n_number_of_concentrations_within_measured_interval


def mean_by_date(concentration_list_local, date_list_local):
    concentration_mean_cumulated_by_date_temporary = []
    for a in range(len(date_list_local)):
        if date_list_local[a] not in concentration_mean_cumulated_by_date_temporary:
            concentration_mean_cumulated_by_date_temporary.append([date_list_local[a]][0])

    concentration_mean_cumulated_by_date = []
    for a in range(len(concentration_mean_cumulated_by_date_temporary)):
        concentration_mean_cumulated_by_date.append([concentration_mean_cumulated_by_date_temporary[a], 0])

    counter_list = []
    for a in range(len(concentration_mean_cumulated_by_date)):
        counter = 0
        for b in range(len(date_list_local)):
            if concentration_mean_cumulated_by_date[a][0] == date_list_local[b]:
                counter += 1
                if concentration_list_local[b] >= 0:
                    concentration_mean_cumulated_by_date[a][1] = concentration_mean_cumulated_by_date[a][1]
+ \
                                                                 concentration_list_local[b]
        counter_list.append(counter)

    concentration_mean_by_date = []
    concentration_mean_dates = []
    for a in range(len(concentration_mean_cumulated_by_date)):
        concentration_mean_by_date.append(concentration_mean_cumulated_by_date[a][1])
        concentration_mean_dates.append(concentration_mean_cumulated_by_date[a][0])
    for a in range(len(counter_list)):
        concentration_mean_by_date[a] = concentration_mean_by_date[a] / counter_list[a]
    return concentration_mean_by_date, concentration_mean_dates


n_estimate = n_estimate()
N_dataset = []
print(n_estimate)
i = 0
while i <= 0.1:  # can be changed to <=4 when there is no N for higher concentration
    i += 0.001
    N_dataset.append(round(n_estimate * i))
del i


original_length = len(N_dataset)
N_dataset = list(dict.fromkeys(N_dataset))
while original_length > len(N_dataset):
    N_dataset.append(N_dataset[-1] + 1)
N_dataset = list(dict.fromkeys(N_dataset))
while original_length > len(N_dataset):
    N_dataset.append(N_dataset[-1] + 4)
N_dataset = list(dict.fromkeys(N_dataset))
while original_length > len(N_dataset):
    N_dataset.append(N_dataset[-1] + 3)
del original_length
```

```
monte_carlo_concentration_dictionary_by_n = {}
for i in range(len(N_dataset)):
    temp_concentration_list = monte_carlo_concentration_list_for_n(N_dataset[i])
    monte_carlo_concentration_dictionary_by_n[N_dataset[i]] = temp_concentration_list
    print(N_dataset[i])
del temp_concentration_list

measured_concentration_list, date_list = get_virus_concentration_list()
# conversion into average by date
measured_concentration_list, date_list = mean_by_date(measured_concentration_list, date_list)

acceptable_deviation = 10000   # for Norovirus GII
'''
acceptable_deviation = 100000   # for CrAssphage
acceptable_deviation = 100   # for Influenza A
acceptable_deviation = 10000   # for SARS-CoV-2
acceptable_deviation = 10000   # for Norovirus GII
'''

n_minimum_list = []
n_maximum_list = []
n_by_probability_list = []
mean_n_list = []
variance_list = []
standard_deviation_list = []
confidence_interval_list = []

for i in measured_concentration_list:
    in_n_number_of_concentrations_within_measured_interval =
determine_range_of_n_by_a_measured_concentration(i)
    # the values produced in in_n_number_of_concentrations_within_measured_interval corresponds with
N_dataset
    for j in range(len(in_n_number_of_concentrations_within_measured_interval)):
        if in_n_number_of_concentrations_within_measured_interval[j] > 0:
            minimum_n = N_dataset[j]
            break
        else:
            minimum_n = 'Min does not exist'
    for k in reversed(in_n_number_of_concentrations_within_measured_interval):
        if k > 0:
            index = len(in_n_number_of_concentrations_within_measured_interval) -
in_n_number_of_concentrations_within_measured_interval[::-1].index(k) - 1
            maximum_n = N_dataset[index]
            break
        else:
            maximum_n = 'Max does not exist'
    n_minimum_list.append(minimum_n)
    n_maximum_list.append(maximum_n)

    #  A new list is created from the N_dataset that corresponds with the number of occurrences of the
measured concentration in the interval
    all_n = []
    for i in range(len(N_dataset)):
        for k in range(in_n_number_of_concentrations_within_measured_interval[i]):
            all_n.append(N_dataset[i])

    if len(all_n) == 0:
        mean_n_list.append('N/A')
        variance_list.append('N/A')
        standard_deviation_list.append('N/A')
        confidence_interval_list.append('N/A')
    else:
        #  Calculate Mean
        mean_for_all_n = sum(all_n)/len(all_n)
        mean_n_list.append(round(mean_for_all_n, 2))
        #  Calculate Variance
        #  Calculate SD
        if len(all_n) <= 1:
            variance_list.append("N/A")
            standard_deviation_list.append("N/A")
        else:
            variance = statistics.variance(all_n)
            variance_list.append(round(variance, 2))
            standard_deviation = statistics.stdev(all_n)
            standard_deviation_list.append(round(standard_deviation, 2))
        #  Calculate a 99% Confidence Interval
        confidence_interval = st.norm.interval(alpha=0.99, loc=numpy.mean(all_n), scale=st.sem(all_n))
        confidence_interval = (round(confidence_interval[0], 2), round(confidence_interval[1], 2))
        confidence_interval_list.append(confidence_interval)
    total_number_of_points = 0
    for j in range(len(in_n_number_of_concentrations_within_measured_interval)):
        total_number_of_points += in_n_number_of_concentrations_within_measured_interval[j]
    if total_number_of_points == 0:
        total_number_of_points = 1  # to prevent a division by zero event
    for j in range(len(in_n_number_of_concentrations_within_measured_interval)):
        in_n_number_of_concentrations_within_measured_interval[j] =
round((in_n_number_of_concentrations_within_measured_interval[j]/total_number_of_points)*100, 2)

    #  append only N with a probability >0
    index_list = []
```

```
    N_intermediate = []
    in_n_number_of_concentrations_within_measured_interval_intermediate = []
    for j in range(len(N_dataset)):
        if in_n_number_of_concentrations_within_measured_interval[j] > 0:

in_n_number_of_concentrations_within_measured_interval_intermediate.append(in_n_number_of_concentrations_wi
thin_measured_interval[j])
            N_intermediate.append(N_dataset[j])
    n_by_probability_list.append([in_n_number_of_concentrations_within_measured_interval_intermediate,
N_intermediate])


intermediate_dict = {'Date': date_list,
                     'Measured Concentration (log10 gc/L)': numpy.log10(measured_concentration_list),
                     'Minimum N': n_minimum_list,
                     'Maximum N': n_maximum_list,
                     'Mean N': mean_n_list,
                     'Variance (σ^2)': variance_list,
                     'Standard Deviation (σ)': standard_deviation_list,
                     'Confidence Interval (CI, 99%)': confidence_interval_list,
                     'N dataset by probability': n_by_probability_list,
                     'Minimum Viral Losses': 0}
df = pd.DataFrame(intermediate_dict)
df.to_excel('export_dataframe_approach1.xlsx', index=False, header=True)
```

Figure S1. Monte-Carlo-Bayesian Approach 1 Code. The code was developed in the Python programming language. It allows for Monte-Carlo-Bayesian simulations, specifically the approach 1 as explained in the description provided within the code.

```
import random
from openpyxl import load_workbook
import numpy
import pandas as pd
import statistics
import scipy.stats as st


'''
0. An average concentration is calculated from all the measured concentrations
1. Estimates N based on an average concentration.
2. Generates an N_dataset based on the N calculated in step 1, where the elements are 0.005xN, 0.01xN,
0.015xN up to 2xN.
3. Calculates the minimal degradation rate (gamma minimum)
4. For each N from the N_dataset, creates a Monte-Carlo full simulation.
   Assumes the degradation between the calculated gamma minimum and 100% or [gamma_minimum; 1] for all the
next operations.
5. Checks each experimentally measured concentration within an acceptable deviation against the
concentrations generated
    for each of the N from step 3. Subsequently calculates what is the minimum N and the maximum N.
6. Also, per each measured concentration, calculates and exports the probability for each N.
'''


def n_estimate():
    test_concentration_gc_L = 918513  # to be used only once for Norovirus GII minimum gamma estimation
    '''
    test_concentration_gc_L = 23005960  # to be used only once
    test_concentration_gc_L = 23368  # to be used only once
    test_concentration_gc_L = 4000  # to be used only once
    test_concentration_gc_L = 918513  # to be used only once
    '''
    flow_l_day = 1920000
    mean_alpha = 149
    mean_beta = 301995172  # for Norovirus GII
    '''
    mean_beta = 125892541  # for CrAssphage
    mean_beta = 32000  # for Influenza A
    mean_beta = 8010000  # for SARS-CoV-2
    mean_beta = 301995172  # for Norovirus GII
    '''
    mean_gamma = 0
    n_estimate_local = (test_concentration_gc_L * flow_l_day) / (mean_alpha * mean_beta * (1 - mean_gamma))
    return n_estimate_local


def get_virus_concentration_list():
    from openpyxl import load_workbook
    workbook_v6 = load_workbook(filename="Xres_cl.xlsx", read_only=True)  # filename="Xres_cl.xlsx" for
influenza and norovirus
    sheet_v6 = workbook_v6.active
    labels = []
    '''
    for row in sheet_v6.iter_rows(min_row=1, max_row=1):
        for i in row:
            labels.append(i.value)
    '''
    virus_concentration_dataset = []
    # min_col=42, max_col=42  #  for SARS_CoV-2
    # min_col=36, max_col=36  #  for CrAssphage
    # min_col=18, max_col=18  #  for Influenza A and Norovirus GII
    # min_row=2, max_row=353  #  for Influenza A
    # min_row=354, max_row=466  #  for Norovirus GII
    for value in sheet_v6.iter_rows(min_row=354, max_row=466, min_col=18, max_col=18, values_only=True):
        for i in value:
            if i >= 0:
                virus_concentration_dataset.append(i*666.6666)  # *666.6666 only for CrAssphage, Norovirus
GII, Influenza A
            else:
                virus_concentration_dataset.append(1)
    sample_date_dataset = []
    # column 5 for flu and norovirus
    # column 8 for crass and sars-cov-2
    for value in sheet_v6.iter_rows(min_row=354, max_row=466, min_col=5, max_col=5, values_only=True):
        for i in value:
            date_only = i.date()
            sample_date_dataset.append(date_only)
    '''
    grab_comp_dataset = []
    for value in sheet_v6.iter_rows(min_row=2, min_col=12, max_col=12, values_only=True):
        for i in value:
            grab_comp_dataset.append(i)
     '''
    return virus_concentration_dataset, sample_date_dataset


def determine_average_gamma(n):
    monte_carlo_concentration_list = []
    for i in range(500):  # choose the number of simulations
        concentration = 0
```

```
        for j in range(n):
            alpha_random_dataset = random.randint(51, 796)
            beta_random_dataset = random.randint(61659500, 12882495517)  # for Norovirus GII
            '''
            beta_random_dataset = random.randint(12589, 12589254117)  # for CrAssphage
            beta_random_dataset = random.randint(4900, 80000000)  # for Influenza A
            beta_random_dataset = random.randint(20000, 65300000000)  # for SARS-CoV-2
            beta_random_dataset = random.randint(61659500, 12882495517)  # for Norovirus GII
            '''
            gamma_random_dataset = random.uniform(0, 1)
            concentration_temp = (alpha_random_dataset * beta_random_dataset * (1 - gamma_random_dataset))
            concentration += concentration_temp
        flow_random_dataset = random.randint(1824000, 2072235)
        concentration = concentration / flow_random_dataset
        monte_carlo_concentration_list.append(concentration)

        test_concentration_gc_L = 918513  # to be used only once for Norovirus GII minimum gamma estimation
        '''
        test_concentration_gc_L = 23005960  # to be used only once for CrAssphage minimum gamma estimation
        test_concentration_gc_L = 23368  # to be used only once for SARS-CoV-2 minimum gamma estimation
        test_concentration_gc_L = 7966  # to be used only once for Influenza A minimum gamma estimation
        test_concentration_gc_L = 918513  # to be used only once for Norovirus GII minimum gamma estimation
        '''
        calculated_average_gamma = 1 - (test_concentration_gc_L/min(monte_carlo_concentration_list))

    return calculated_average_gamma


def monte_carlo_concentration_list_for_n(n, minimum_gamma):
    monte_carlo_concentration_list = []
    for i in range(2000):  # choose the number of simulations
        concentration = 0
        for j in range(n):
            alpha_random_dataset = random.randint(51, 796)
            beta_random_dataset = random.randint(61659500, 12882495517)  # for Norovirus GII
            '''
            beta_random_dataset = random.randint(12589, 12589254117)  # for CrAssphage
            beta_random_dataset = random.randint(4900, 80000000)  # for Influenza A
            beta_random_dataset = random.randint(20000, 65300000000)  # for SARS-CoV-2
            beta_random_dataset = random.randint(61659500, 12882495517)  # for Norovirus GII
            '''
            gamma_random_dataset = random.uniform(minimum_gamma, 1)
            concentration_temp = (alpha_random_dataset * beta_random_dataset * (1 - gamma_random_dataset))
            concentration += concentration_temp
        flow_random_dataset = random.randint(1824000, 2072235)
        concentration = concentration / flow_random_dataset
        monte_carlo_concentration_list.append(concentration)
    return monte_carlo_concentration_list


def determine_range_of_n_by_a_measured_concentration(measured_concentration):
    deviation_min = measured_concentration - acceptable_deviation
    deviation_max = measured_concentration + acceptable_deviation
    in_n_number_of_concentrations_within_measured_interval = []
    i = 0
    for key in monte_carlo_concentration_dictionary_by_n:
        temp_concentrations_list = monte_carlo_concentration_dictionary_by_n[key]
        in_n_number_of_concentrations_within_measured_interval.append(0)
        for j in range(len(temp_concentrations_list)):
            if deviation_min <= temp_concentrations_list[j] <= deviation_max:
                in_n_number_of_concentrations_within_measured_interval[i] += 1
        i += 1
    del i
    del temp_concentrations_list
    return in_n_number_of_concentrations_within_measured_interval


def mean_by_date(concentration_list_local, date_list_local):
    concentration_mean_cumulated_by_date_temporary = []
    for a in range(len(date_list_local)):
        if date_list_local[a] not in concentration_mean_cumulated_by_date_temporary:
            concentration_mean_cumulated_by_date_temporary.append([date_list_local[a]][0])

    concentration_mean_cumulated_by_date = []
    for a in range(len(concentration_mean_cumulated_by_date_temporary)):
        concentration_mean_cumulated_by_date.append([concentration_mean_cumulated_by_date_temporary[a], 0])

    counter_list = []
    for a in range(len(concentration_mean_cumulated_by_date)):
        counter = 0
        for b in range(len(date_list_local)):
            if concentration_mean_cumulated_by_date[a][0] == date_list_local[b]:
                counter += 1
                if concentration_list_local[b] >= 0:
                    concentration_mean_cumulated_by_date[a][1] = concentration_mean_cumulated_by_date[a][1]
+ \
                                                                 concentration_list_local[b]
        counter_list.append(counter)

    concentration_mean_by_date = []
```

```
    concentration_mean_dates = []
    for a in range(len(concentration_mean_cumulated_by_date)):
        concentration_mean_by_date.append(concentration_mean_cumulated_by_date[a][1])
        concentration_mean_dates.append(concentration_mean_cumulated_by_date[a][0])
    for a in range(len(counter_list)):
        concentration_mean_by_date[a] = concentration_mean_by_date[a] / counter_list[a]
    return concentration_mean_by_date, concentration_mean_dates


n_estimate = n_estimate()
N_dataset = []
print(n_estimate)
i = 0
while i <= 8:  # can be changed to <=4 when there is no N for higher concentration
    i += 0.025
    N_dataset.append(round(n_estimate * i))
del i

i = 10
while i <= 250:
    N_dataset.append(320 + i)
    i += 15
del i
average_minimum_gamma = determine_average_gamma(round(n_estimate))

original_length = len(N_dataset)
N_dataset = list(dict.fromkeys(N_dataset))
while original_length > len(N_dataset):
    N_dataset.append(N_dataset[-1] + 1)
N_dataset = list(dict.fromkeys(N_dataset))
while original_length > len(N_dataset):
    N_dataset.append(N_dataset[-1] + 4)
N_dataset = list(dict.fromkeys(N_dataset))
while original_length > len(N_dataset):
    N_dataset.append(N_dataset[-1] + 3)
del original_length
monte_carlo_concentration_dictionary_by_n = {}
for i in range(len(N_dataset)):
    temp_concentration_list = monte_carlo_concentration_list_for_n(N_dataset[i], average_minimum_gamma)
    monte_carlo_concentration_dictionary_by_n[N_dataset[i]] = temp_concentration_list
    print(N_dataset[i])
del temp_concentration_list

measured_concentration_list, date_list = get_virus_concentration_list()
# conversion into average by date
measured_concentration_list, date_list = mean_by_date(measured_concentration_list, date_list)

acceptable_deviation = 10000   # for Norovirus GII
'''
acceptable_deviation = 50000   # for CrAssphage
acceptable_deviation = 100   # for Influenza A
acceptable_deviation = 2000   # for SARS-CoV-2
acceptable_deviation = 10000   # for Norovirus GII
'''

n_minimum_list = []
n_maximum_list = []
n_by_probability_list = []
mean_n_list = []
variance_list = []
standard_deviation_list = []
confidence_interval_list = []

for i in measured_concentration_list:
    in_n_number_of_concentrations_within_measured_interval =
determine_range_of_n_by_a_measured_concentration(i)
    # the values produced in in_n_number_of_concentrations_within_measured_interval corresponds with
N_dataset
    for j in range(len(in_n_number_of_concentrations_within_measured_interval)):
        if in_n_number_of_concentrations_within_measured_interval[j] > 0:
            minimum_n = N_dataset[j]
            break
        else:
            minimum_n = 'Min does not exist'
    for k in reversed(in_n_number_of_concentrations_within_measured_interval):
        if k > 0:
            index = len(in_n_number_of_concentrations_within_measured_interval) -
in_n_number_of_concentrations_within_measured_interval[::-1].index(k) - 1
            maximum_n = N_dataset[index]
            break
        else:
            maximum_n = 'Max does not exist'
    n_minimum_list.append(minimum_n)
    n_maximum_list.append(maximum_n)

    #  A new list is created from the N_dataset that corresponds with the number of occurrences of the
measured concentration in the interval
    all_n = []
    for i in range(len(N_dataset)):
        for k in range(in_n_number_of_concentrations_within_measured_interval[i]):
```

```
            all_n.append(N_dataset[i])

    if len(all_n) == 0:
        mean_n_list.append('N/A')
        variance_list.append('N/A')
        standard_deviation_list.append('N/A')
        confidence_interval_list.append('N/A')
    else:
        #  Calculate Mean
        mean_for_all_n = sum(all_n)/len(all_n)
        mean_n_list.append(round(mean_for_all_n, 2))
        #  Calculate Variance
        #  Calculate SD
        if len(all_n) <= 1:
            variance_list.append("N/A")
            standard_deviation_list.append("N/A")
        else:
            variance = statistics.variance(all_n)
            variance_list.append(round(variance, 2))
            standard_deviation = statistics.stdev(all_n)
            standard_deviation_list.append(round(standard_deviation, 2))
        #  Calculate a 99% Confidence Interval
        confidence_interval = st.norm.interval(alpha=0.99, loc=numpy.mean(all_n), scale=st.sem(all_n))
        confidence_interval = (round(confidence_interval[0], 2), round(confidence_interval[1], 2))
        confidence_interval_list.append(confidence_interval)
    total_number_of_points = 0
    for j in range(len(in_n_number_of_concentrations_within_measured_interval)):
        total_number_of_points += in_n_number_of_concentrations_within_measured_interval[j]
    if total_number_of_points == 0:
        total_number_of_points = 1  # to prevent a division by zero event
    for j in range(len(in_n_number_of_concentrations_within_measured_interval)):
        in_n_number_of_concentrations_within_measured_interval[j] =
round((in_n_number_of_concentrations_within_measured_interval[j]/total_number_of_points)*100, 2)

    #  append only N with a probability >0
    index_list = []
    N_intermediate = []
    in_n_number_of_concentrations_within_measured_interval_intermediate = []
    for j in range(len(N_dataset)):
        if in_n_number_of_concentrations_within_measured_interval[j] > 0:

in_n_number_of_concentrations_within_measured_interval_intermediate.append(in_n_number_of_concentrations_wi
thin_measured_interval[j])
            N_intermediate.append(N_dataset[j])
    n_by_probability_list.append([in_n_number_of_concentrations_within_measured_interval_intermediate,
N_intermediate])


intermediate_dict = {'Date': date_list,
                     'Measured Concentration (log10 gc/L)': numpy.log10(measured_concentration_list),
                     'Minimum N': n_minimum_list,
                     'Maximum N': n_maximum_list,
                     'Mean N': mean_n_list,
                     'Variance (σ^2)': variance_list,
                     'Standard Deviation (σ)': standard_deviation_list,
                     'Confidence Interval (CI, 99%)': confidence_interval_list,
                     'N dataset by probability': n_by_probability_list,
                     'Minimum Viral Losses': average_minimum_gamma}
df = pd.DataFrame(intermediate_dict)
df.to_excel('export_dataframe_approach2.xlsx', index=False, header=True)
```

Figure S2. Monte-Carlo-Bayesian Approach 2 Code. The code was developed in the Python programming language. It allows for Monte-Carlo-Bayesian simulations, specifically the approach 2 as explained in the description provided within the code.

```
import random
from openpyxl import load_workbook
import numpy
import scipy.stats as st
import pandas as pd
import statistics


'''
All the following operations are performed individually for each of the measured concentration:

1. Estimates N based on the measured concentration.
2. Generates an N_dataset based on the N calculated in step 1, where the elements are 0.005xN, 0.01xN,
0.015xN up to 2xN.
3. Calculates the minimal degradation rate (gamma minimum)
4. For each N from the N_dataset, creates a Monte-Carlo full simulation.
   Assumes the degradation between the calculated gamma minimum and 100% or [gamma_minimum; 1] for all the
next operations.
5. Checks measured concentration within an acceptable deviation against the concentrations generated
    for each of the N from step 3. Subsequently calculates what is the minimum N and the maximum N.
6. Also, calculates and exports the probability for each N.
'''


def n_estimate_function(concentration):
    flow_l_day = 1920000
    mean_alpha = 149
    mean_beta = 301995172  # for Norovirus GII
    '''
    mean_beta = 125892541  # for CrAssphage
    mean_beta = 32000  # for Influenza A
    mean_beta = 8010000  # for SARS-CoV-2
    mean_beta = 301995172  # for Norovirus GII
    '''
    mean_gamma = 0
    n_estimate_local = (concentration * flow_l_day) / (mean_alpha * mean_beta * (1 - mean_gamma))
    return n_estimate_local


def get_virus_data_list():
    from openpyxl import load_workbook
    workbook_v6 = load_workbook(filename="Xres_cl.xlsx", read_only=True)  # filename="Xres_cl.xlsx for
influenza and norovirus
    sheet_v6 = workbook_v6.active
    labels = []
    '''
    for row in sheet_v6.iter_rows(min_row=1, max_row=1):
        for i in row:
            labels.append(i.value)
    '''
    virus_concentration_dataset = []
    # min_col=42, max_col=42  #  for SARS_CoV-2
    # min_col=36, max_col=36  #  for CrAssphage
    # min_col=18, max_col=18  #  for Influenza A and Norovirus GII
    # min_row=2, max_row=353  #  for Influenza A
    # min_row=354, max_row=466  #  for Norovirus GII
    for value in sheet_v6.iter_rows(min_row=354, max_row=466, min_col=18, max_col=18, values_only=True):
        for i in value:
            if i >= 0:
                virus_concentration_dataset.append(i*666.6666)  # *666.6666 only for CrAssphage, Norovirus
GII, Influenza A
            else:
                virus_concentration_dataset.append(1)
    sample_date_dataset = []
    # column 5 for flu and norovirus
    # column 8 for crass and sars-cov-2
    for value in sheet_v6.iter_rows(min_row=354, max_row=466, min_col=5, max_col=5, values_only=True):
        for i in value:
            date_only = i.date()
            sample_date_dataset.append(date_only)
    '''
    grab_comp_dataset = []
    for value in sheet_v6.iter_rows(min_row=2, min_col=12, max_col=12, values_only=True):
        for i in value:
            grab_comp_dataset.append(i)
     '''
    return virus_concentration_dataset, sample_date_dataset


def determine_average_gamma(n, concentration_variable):
    monte_carlo_concentration_list = []
    for i in range(250):  # choose the number of simulations
        concentration = 0
        for j in range(n):
            alpha_random_dataset = random.randint(51, 796)
            beta_random_dataset = random.randint(61659500, 12882495517)  # for Norovirus GII
            '''
            beta_random_dataset = random.randint(12589, 12589254117)  # for CrAssphage
            beta_random_dataset = random.randint(4900, 80000000)  # for Influenza A
            beta_random_dataset = random.randint(20000, 65300000000)  # for SARS-CoV-2
```

```
            beta_random_dataset = random.randint(61659500, 12882495517)  # for Norovirus GII
            '''
            gamma_random_dataset = random.uniform(0, 1)
            concentration_temp = (alpha_random_dataset * beta_random_dataset * (1 - gamma_random_dataset))
            concentration += concentration_temp
        flow_random_dataset = random.randint(1824000, 2072235)
        concentration = concentration / flow_random_dataset
        monte_carlo_concentration_list.append(concentration)
        if min(monte_carlo_concentration_list) <= 0:
            calculated_minimum_gamma = 0
        else:
            calculated_minimum_gamma = 1 - (concentration_variable/min(monte_carlo_concentration_list))
    return calculated_minimum_gamma


def monte_carlo_concentration_list_for_n(n, minimum_gamma_local):
    monte_carlo_concentration_list = []
    for i in range(700):  # choose the number of simulations
        concentration = 0
        for j in range(n):
            alpha_random_dataset = random.randint(51, 796)
            beta_random_dataset = random.randint(61659500, 12882495517)  # for Norovirus GII
            '''
            beta_random_dataset = random.randint(12589, 12589254117)  # for CrAssphage
            beta_random_dataset = random.randint(4900, 80000000)  # for Influenza A
            beta_random_dataset = random.randint(20000, 65300000000)  # for SARS-CoV-2
            beta_random_dataset = random.randint(61659500, 12882495517)  # for Norovirus GII
            '''
            gamma_random_dataset = random.uniform(minimum_gamma_local, 1)
            concentration_temp = (alpha_random_dataset * beta_random_dataset * (1 - gamma_random_dataset))
            concentration += concentration_temp
        flow_random_dataset = random.randint(1824000, 2072235)
        concentration = concentration / flow_random_dataset
        monte_carlo_concentration_list.append(concentration)
    return monte_carlo_concentration_list


def determine_range_of_n_by_a_measured_concentration(concentration_local):
    deviation_min = concentration_local - acceptable_deviation
    deviation_max = concentration_local + acceptable_deviation
    in_n_number_of_concentrations_within_measured_interval_local = []
    i = 0
    for key in monte_carlo_concentration_dictionary_by_n:
        temp_concentrations_list = monte_carlo_concentration_dictionary_by_n[key]
        in_n_number_of_concentrations_within_measured_interval_local.append(0)
        for j in range(len(temp_concentrations_list)):
            if deviation_min <= temp_concentrations_list[j] <= deviation_max:
                in_n_number_of_concentrations_within_measured_interval_local[i] += 1
        i += 1
    return in_n_number_of_concentrations_within_measured_interval_local


def mean_by_date(concentration_list_local, date_list_local):
    concentration_mean_cumulated_by_date_temporary = []
    for a in range(len(date_list_local)):
        if date_list_local[a] not in concentration_mean_cumulated_by_date_temporary:
            concentration_mean_cumulated_by_date_temporary.append([date_list_local[a]][0])

    concentration_mean_cumulated_by_date = []
    for a in range(len(concentration_mean_cumulated_by_date_temporary)):
        concentration_mean_cumulated_by_date.append([concentration_mean_cumulated_by_date_temporary[a], 0])

    counter_list = []
    for a in range(len(concentration_mean_cumulated_by_date)):
        counter = 0
        for b in range(len(date_list_local)):
            if concentration_mean_cumulated_by_date[a][0] == date_list_local[b]:
                counter += 1
                if concentration_list_local[b] >= 0:
                    concentration_mean_cumulated_by_date[a][1] = concentration_mean_cumulated_by_date[a][1]
+ \
                                                                 concentration_list_local[b]
        counter_list.append(counter)

    concentration_mean_by_date = []
    concentration_mean_dates = []
    for a in range(len(concentration_mean_cumulated_by_date)):
        concentration_mean_by_date.append(concentration_mean_cumulated_by_date[a][1])
        concentration_mean_dates.append(concentration_mean_cumulated_by_date[a][0])
    for a in range(len(counter_list)):
        concentration_mean_by_date[a] = concentration_mean_by_date[a] / counter_list[a]
    return concentration_mean_by_date, concentration_mean_dates


measured_concentration_list, date_list = get_virus_data_list()
# conversion into average by date
measured_concentration_list, date_list = mean_by_date(measured_concentration_list, date_list)

print(measured_concentration_list)
# choose the acceptable deviation from the measured concentration
```

```
acceptable_deviation = 10000   # for Norovirus GII
'''
acceptable_deviation = 10000   # for CrAssphage
acceptable_deviation = 100   # for Influenza A
acceptable_deviation = 10000   # for SARS-CoV-2
acceptable_deviation = 10000   # for Norovirus GII
'''
n_minimum_list = []
n_maximum_list = []
gamma_minimum_list = []
n_estimate_list = []
n_by_probability_list = []
n_by_occurrence_list = []
mean_n_list = []
variance_list = []
standard_deviation_list = []
confidence_interval_list = []

for measured_concentration in measured_concentration_list:
    n_estimate = n_estimate_function(measured_concentration)
    minimum_gamma = determine_average_gamma(round(n_estimate), measured_concentration)
    if minimum_gamma < 0:
        minimum_gamma = 0
    n_estimate_list.append(n_estimate)
    gamma_minimum_list.append(minimum_gamma)

    if n_estimate == 0:
        n_estimate = 1  # this is to append an n =! 0 so that it does not give errors later
    N_dataset = []
    print(n_estimate)
    i = 0
    while i <= 2:
        i += 0.05  # choose the fractionation of the population
        if round(n_estimate * i) == 0:
            N_dataset.append(round(i*100))  # this is to append an n =! 0 so that it does not give errors
later
        else:
            N_dataset.append(round(n_estimate * i))
    del i

    original_length = len(N_dataset)
    N_dataset = list(dict.fromkeys(N_dataset))
    while original_length > len(N_dataset):
        N_dataset.append(N_dataset[-1] + 1)
    N_dataset = list(dict.fromkeys(N_dataset))
    while original_length > len(N_dataset):
        N_dataset.append(N_dataset[-1] + 4)
    N_dataset = list(dict.fromkeys(N_dataset))
    while original_length > len(N_dataset):
        N_dataset.append(N_dataset[-1] + 3)
    del original_length

    monte_carlo_concentration_dictionary_by_n = {}
    for i in range(len(N_dataset)):
        temp_concentration_list = monte_carlo_concentration_list_for_n(N_dataset[i], minimum_gamma)
        monte_carlo_concentration_dictionary_by_n[N_dataset[i]] = temp_concentration_list
    del temp_concentration_list
    in_n_number_of_concentrations_within_measured_interval =
determine_range_of_n_by_a_measured_concentration(measured_concentration)
    # the values produced in in_n_number_of_concentrations_within_measured_interval corresponds with
N_dataset

    for j in range(len(in_n_number_of_concentrations_within_measured_interval)):
        if in_n_number_of_concentrations_within_measured_interval[j] > 0:
            minimum_n = N_dataset[j]
            break
        else:
            minimum_n = 'Min does not exist'
    for k in reversed(in_n_number_of_concentrations_within_measured_interval):
        if k > 0:
            index = len(in_n_number_of_concentrations_within_measured_interval) -
in_n_number_of_concentrations_within_measured_interval[::-1].index(k) - 1
            maximum_n = N_dataset[index]
            break
        else:
            maximum_n = 'Max does not exist'

    n_minimum_list.append(minimum_n)
    n_maximum_list.append(maximum_n)

    #  delete all N that had a 0 probability
    index_list = []
    for i in range(len(N_dataset)):
        if in_n_number_of_concentrations_within_measured_interval[i] == 0:
            index_list.append(i)
    for i in reversed(index_list):
        del in_n_number_of_concentrations_within_measured_interval[i]
        del N_dataset[i]

    #  A new list is created from the N_dataset that corresponds with the number of occurrences of the
```

```
measured concentration in the interval
    all_n = []
    for i in range(len(N_dataset)):
        for k in range(in_n_number_of_concentrations_within_measured_interval[i]):
            all_n.append(N_dataset[i])

    if len(all_n) == 0:
        mean_n_list.append('N/A')
        variance_list.append('N/A')
        standard_deviation_list.append('N/A')
        confidence_interval_list.append('N/A')
    else:
        #  Calculate Mean
        mean_for_all_n = sum(all_n)/len(all_n)
        mean_n_list.append(round(mean_for_all_n, 2))
        #  Calculate Variance
        #  Calculate SD
        if len(all_n) <= 1:
            variance_list.append("N/A")
            standard_deviation_list.append("N/A")
        else:
            variance = statistics.variance(all_n)
            variance_list.append(round(variance, 2))
            standard_deviation = statistics.stdev(all_n)
            standard_deviation_list.append(round(standard_deviation, 2))
        #  Calculate a 99% Confidence Interval
        confidence_interval = st.norm.interval(alpha=0.99, loc=numpy.mean(all_n), scale=st.sem(all_n))
        confidence_interval = (round(confidence_interval[0], 2), round(confidence_interval[1], 2))
        confidence_interval_list.append(confidence_interval)

    total_number_of_points = 0
    for i in range(len(in_n_number_of_concentrations_within_measured_interval)):
        total_number_of_points += in_n_number_of_concentrations_within_measured_interval[i]
    if total_number_of_points == 0:
        total_number_of_points = 1  # to prevent a division by zero error
    for i in range(len(in_n_number_of_concentrations_within_measured_interval)):
        in_n_number_of_concentrations_within_measured_interval[i] =
round((in_n_number_of_concentrations_within_measured_interval[i]/total_number_of_points)*100, 2)
    n_by_probability_list.append([in_n_number_of_concentrations_within_measured_interval, N_dataset])


intermediate_dict = {'Date': date_list,
                     'Measured Concentration (log10 gc/L)': numpy.log10(measured_concentration_list),
                     'Minimum Viral Losses': gamma_minimum_list,
                     'N Estimate (losses = 0)': n_estimate_list,
                     'Minimum N': n_minimum_list,
                     'Maximum N': n_maximum_list,
                     'Mean N': mean_n_list,
                     'Variance (σ^2)': variance_list,
                     'Standard Deviation (σ)': standard_deviation_list,
                     'Confidence Interval (CI, 99%)': confidence_interval_list,
                     'N by probability': n_by_probability_list}
df = pd.DataFrame(intermediate_dict)
df.to_excel('export_dataframe_approach3.xlsx', index=False, header=True)
```

Figure S3. Monte-Carlo-Bayesian Approach 3 Code. The code was developed in the Python programming language. It allows for Monte-Carlo-Bayesian simulations, specifically the approach 3 as explained in the description provided within the code. Additionally, an intermediate step during the code execution is the estimation of the population according to the Central Limit Theorem.

```
from openpyxl import load_workbook
import matplotlib.pyplot as plot
import numpy as np
import pandas
from dtw import accelerated_dtw

workbook = load_workbook(filename="covid_estimates.xlsx", read_only=True)
sheet = workbook.active

# Collect the MCBA Approach 1 population estimates
date_dataset1 = []
for value in sheet.iter_rows(min_row=2, max_row=93, min_col=1, max_col=1, values_only=True):
    for i in value:
        date_only = i.date()
        date_dataset1.append(date_only)

virus1_dataset = []
for value in sheet.iter_rows(min_row=2, max_row=93, min_col=12, max_col=12, values_only=True):
    for i in value:
        virus1_dataset.append(i)

# removes empty data
index_list = []
for i in range(len(virus1_dataset)):
    index_list.append(i)
index_list.reverse()
for i in index_list:
    if virus1_dataset[i] is None:
        del virus1_dataset[i]
        del date_dataset1[i]

# Collect the MCBA Approach 2 population estimates
date_dataset2 = []
for value in sheet.iter_rows(min_row=2, max_row=93, min_col=1, max_col=1, values_only=True):
    for i in value:
        date_only = i.date()
        date_dataset2.append(date_only)

virus2_dataset = []
for value in sheet.iter_rows(min_row=2, max_row=93, min_col=13, max_col=13, values_only=True):
    for i in value:
        virus2_dataset.append(i)

# removes empty data
index_list = []
for i in range(len(virus2_dataset)):
    index_list.append(i)
index_list.reverse()
for i in index_list:
    if virus2_dataset[i] is None:
        del virus2_dataset[i]
        del date_dataset2[i]

# Collect the MCBA Approach 3 population estimates
date_dataset3 = []
for value in sheet.iter_rows(min_row=2, max_row=93, min_col=1, max_col=1, values_only=True):
    for i in value:
        date_only = i.date()
        date_dataset3.append(date_only)

virus3_dataset = []
for value in sheet.iter_rows(min_row=2, max_row=93, min_col=14, max_col=14, values_only=True):
    for i in value:
        virus3_dataset.append(i)

# removes empty data
index_list = []
for i in range(len(virus3_dataset)):
    index_list.append(i)
index_list.reverse()
for i in index_list:
    if virus3_dataset[i] is None:
        del virus3_dataset[i]
        del date_dataset3[i]

# Collect the CLT population estimates
date_dataset4 = []
for value in sheet.iter_rows(min_row=2, max_row=93, min_col=1, max_col=1, values_only=True):
    for i in value:
        date_only = i.date()
        date_dataset4.append(date_only)

virus4_dataset = []
for value in sheet.iter_rows(min_row=2, max_row=93, min_col=15, max_col=15, values_only=True):
    for i in value:
        virus4_dataset.append(i)

# removes empty data
index_list = []
for i in range(len(virus4_dataset)):
```

```
    index_list.append(i)
index_list.reverse()
for i in index_list:
    if virus4_dataset[i] is None:
        del virus4_dataset[i]
        del date_dataset4[i]

# Collect COVID-19 diagnosed cases in Gwynedd
date_dataset5 = []
for value in sheet.iter_rows(min_row=2, max_row=230, min_col=16, max_col=16, values_only=True):
    for i in value:
        date_only = i.date()
        date_dataset5.append(date_only)

virus5_dataset = []
for value in sheet.iter_rows(min_row=2, max_row=230, min_col=18, max_col=18, values_only=True):
    for i in value:
        virus5_dataset.append(i)

# removes empty data
index_list = []
for i in range(len(virus5_dataset)):
    index_list.append(i)
index_list.reverse()
for i in index_list:
    if virus5_dataset[i] is None:
        del virus5_dataset[i]
        del date_dataset5[i]

# Select the data only after 10th of Jan 2022
date_dataset1 = date_dataset1[20:]
virus1_dataset = virus1_dataset[20:]
date_dataset2 = date_dataset2[27:]
virus2_dataset = virus2_dataset[27:]
date_dataset3 = date_dataset3[30:]
virus3_dataset = virus3_dataset[30:]
date_dataset4 = date_dataset4[33:]
virus4_dataset = virus4_dataset[33:]
date_dataset5 = date_dataset5[10:135]
virus5_dataset = virus5_dataset[10:135]

intermediate_dict1 = {'Date': date_dataset1,
                      'Virus1': virus1_dataset}
intermediate_dict2 = {'Date': date_dataset2,
                      'Virus2': virus2_dataset}
intermediate_dict3 = {'Date': date_dataset3,
                      'Virus3': virus3_dataset}
intermediate_dict4 = {'Date': date_dataset4,
                      'Virus4': virus4_dataset}
intermediate_dict5 = {'Date': date_dataset5,
                      'Virus5': virus5_dataset}

df1 = pandas.DataFrame.from_dict(intermediate_dict1, orient='index')
df1 = df1.transpose()
df2 = pandas.DataFrame.from_dict(intermediate_dict2, orient='index')
df2 = df2.transpose()
df3 = pandas.DataFrame.from_dict(intermediate_dict3, orient='index')
df3 = df3.transpose()
df4 = pandas.DataFrame.from_dict(intermediate_dict4, orient='index')
df4 = df4.transpose()
df5 = pandas.DataFrame.from_dict(intermediate_dict5)

df_1 = pandas.merge(df1, df5, on='Date', how='outer')
df_2 = pandas.merge(df2, df5, on='Date', how='outer')
df_3 = pandas.merge(df3, df5, on='Date', how='outer')
df_4 = pandas.merge(df4, df5, on='Date', how='outer')

# removes the rows containing empty data: the diagnosed cases are never empty, but only the estimated cases
# to be included at the user's discretion for comparison

df_1 = df_1.dropna()
df_2 = df_2.dropna()
df_3 = df_3.dropna()
df_4 = df_4.dropna()


df_1['Virus5'] = df_1['Virus5'].astype(float)
df_1['Virus1'] = df_1['Virus1'].astype(float)
df_2['Virus5'] = df_2['Virus5'].astype(float)
df_2['Virus2'] = df_2['Virus2'].astype(float)
df_3['Virus5'] = df_3['Virus5'].astype(float)
df_3['Virus3'] = df_3['Virus3'].astype(float)
df_4['Virus5'] = df_4['Virus5'].astype(float)
df_4['Virus4'] = df_4['Virus4'].astype(float)

'''
# For MCBA Approach1
d1 = df_1['Virus1'].interpolate().values
d2 = df_1['Virus5'].interpolate().values
# For MCBA Approach2
```

```
d1 = df_2['Virus2'].interpolate().values
d2 = df_2['Virus5'].interpolate().values
# For MCBA Approach3
d1 = df_3['Virus3'].interpolate().values
d2 = df_3['Virus5'].interpolate().values
# For CLT
d1 = df_4['Virus4'].interpolate().values
d2 = df_4['Virus5'].interpolate().values
'''
# For CLT
d1 = df_4['Virus4'].interpolate().values
d2 = df_4['Virus5'].interpolate().values
d, cost_matrix, acc_cost_matrix, path = accelerated_dtw(d1,d2, dist='euclidean')
plot.imshow(acc_cost_matrix.T, origin='lower', cmap='OrRd', interpolation='nearest')
plot.plot(path[0], path[1], 'b', linewidth=3, alpha=0.8)
'''
plot.xlabel('SARS-CoV-2 MCBA Approach 1 Estimated Cases')
plot.xlabel('SARS-CoV-2 MCBA Approach 2 Estimated Cases')
plot.xlabel('SARS-CoV-2 MCBA Approach 3 Estimated Cases')
plot.xlabel('SARS-CoV-2 CLT Estimated Cases')
'''
plot.xlabel('SARS-CoV-2 CLT Estimated Cases')
plot.ylabel('COVID-19 Diagnosed Cases')
plot.title(f'DTW minimum path with minimum distance: {np.round(d,2)}')
plot.show()
```

Figure S4. Dynamic Time Warping Code for Comparison of Clinical and Estimated Data. The code was developed in the Python programming language.

```
def pol_trendline_plot(time_variable, dependent_variable, label):
    time_variable = mdates.date2num(time_variable)
    polynomialOrder = 10
    fittedParameters_A = numpy.poly1d(numpy.polyfit(time_variable, dependent_variable,
polynomialOrder))
    print('Fitted Parameters A:', fittedParameters_A)
    modelPredictions_A = numpy.polyval(fittedParameters_A, time_variable)
    absError_A = modelPredictions_A - dependent_variable
    SE_A = numpy.square(absError_A)  # squared errors A
    MSE_A = numpy.mean(SE_A)  # mean squared errors A
    RMSE_A = numpy.sqrt(MSE_A)  # Root Mean Squared Error, RMSE A
    Rsquared_A = 1.0 - (numpy.var(absError_A) / numpy.var(dependent_variable))
    print('RMSE A:', RMSE_A)
    print('R-squared A:', Rsquared_A)
    print(r2_score(dependent_variable, fittedParameters_A(time_variable)))
    xModel_A = numpy.linspace(min(time_variable), max(time_variable))
    yModel_A = numpy.polyval(fittedParameters_A, xModel_A)
    ax.plot(xModel_A, yModel_A, linewidth=3, label=label)
```

Figure S5. Polynomial Trendline Function. The code was developed in the Python programming language. It details a function required for the calculation of a polynomial trendline. The function requires three variables as arguments, two of which are of an array data type, and the third a string type.

Monte-Carlo-Bayesian Approach Example:

1. For demonstration purposes, a mean value is calculated for the CrAssphage measured concentration, this being 27930540 or 7.45 $\log_{10}$ gc/l.

2. N is estimated using the mean values according to Table 7 and the equation:

$$N = \frac{C_{RNA}\bar{F}}{\bar{\alpha}\bar{\beta}(1-\bar{\gamma})}.$$

   Also, this N estimate would be the result as per the EMCLT.

3. N estimate is 2859 individuals.

4. The simulation is performed assuming viral losses between [0; 1] or 0-100%. The number of Monte-Carlo simulations in this case is 1000 data points. The histogram (Figure S6) is obtained by separating the range between the minimum and maximum simulated concentrations in 30 intervals of equal length.

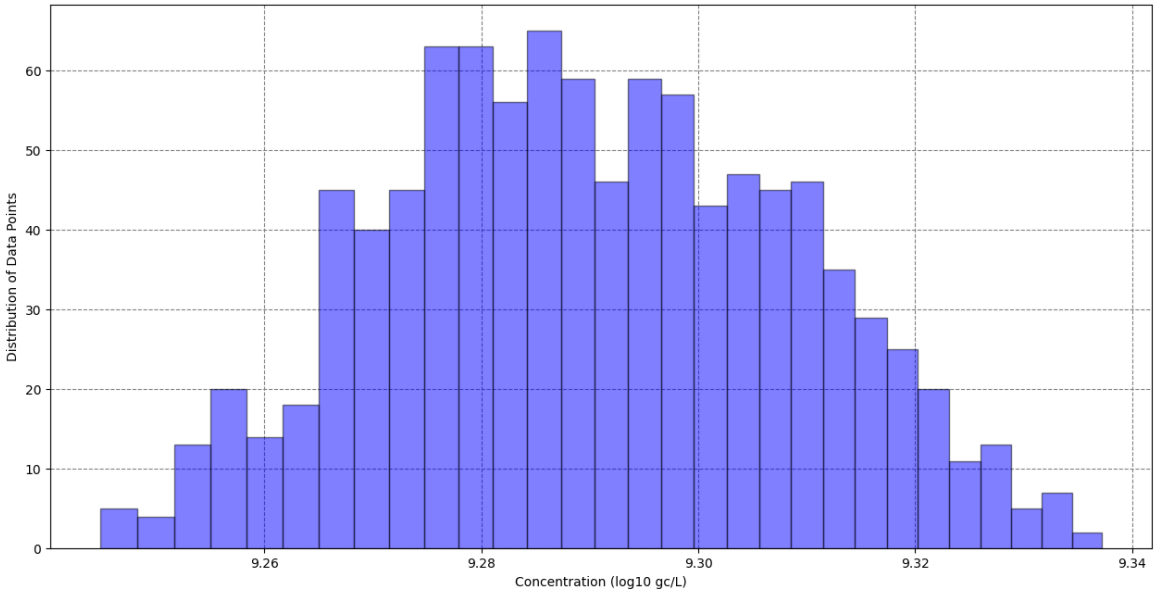

Figure S6. Monte-Carlo full simulation of crAssphage concentrations in wastewater. The probability of crAssphage concentrations in wastewater according to a Monte-Carlo full simulation as expected for a population number of 2859 individuals. The number of simulated data-points is 1000. The number of intervals is 30 and the distribution was calculated by dividing the number of data points fitting within one interval by the total number of data points.

5. A degradation range of 0% to 100% is unlikely in a field setting, hence the obtained histogram in the previous step will be used to estimate the minimum value of the degradation rate. This is done with the formula

$$\gamma_{min} = 1 - (C_{measured}/C_{estimated\_min}),$$

   where γ_{min} – minimum degradation rate, $C_{measured}$ – the measured concentration in step 1, and $C_{estimated\_min}$ – the minimum simulated concentration from step 4.

6. Since

$$C_{estimated\_min} = 9.24 \log_{10} gc/L,$$

then

$$(1 - \gamma_{min}) = 27930540 \text{ gc/L} \div 1737800828 \text{ gc/L},$$

and

$$\gamma_{min} = 0.9839.$$

Therefore, the range of the degradation rate is deducted to be [0.9839; 1], corresponding to viral losses of 98.39%-100%.

This can be read in the following way: "Assuming a population estimate of 2859 individuals and a measured concentration of CrAssphage in wastewater of 7.45 $\log_{10}$ gc/L, the estimated viral losses were simulated to be in the range of 98.39%-100%".

7. The Monte-Carlo full simulation is repeated by using the γ_{min} variable value obtained in step 6. In this case, 5000 simulations are performed (Figure S7).

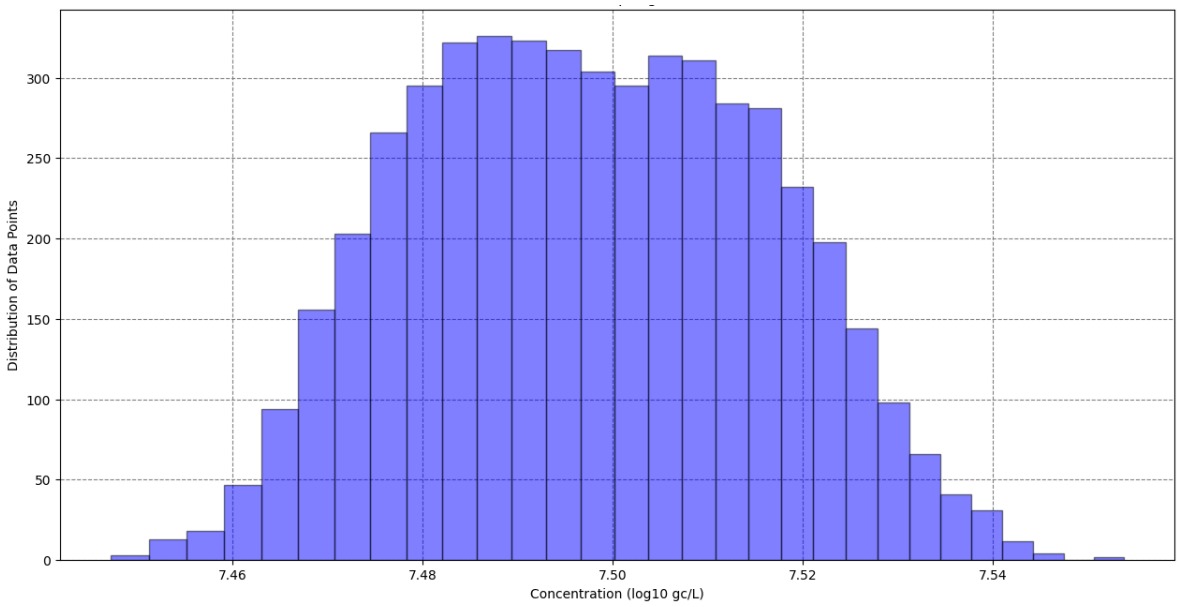

Figure S7. Monte-Carlo full simulation of crAssphage concentration in wastewater. The probability of crAssphage concentrations in wastewater according to a Monte-Carlo full simulation as expected for a population number of 2859 individuals and assuming viral losses between 98.39-100%. The number of simulated data-points is 5000. The number of intervals is 30 and the distribution was calculated by dividing the number of data points fitting within one interval by the total number of data points.

8. Now, using the N estimated in step 2 and 3, a dataset of N values will be obtained which corresponds to an arithmetic progression of 0.05xN: 0.05N, 0.1N, 0.15N… 1.9N,

1.95N, 2N. And Monte-Carlo simulations in number of 1000 data points will be performed for each of the values from the N dataset.

9. An acceptable deviation is chosen from the measured concentration. For demonstration purposes, the chosen acceptable deviation is 6 $\log_{10}$ gc/L.

10. Within the range of the measured concentration ± acceptable deviation, it is calculated how many simulated concentration data points fit within the range, and it is divided by the total number of data points. This calculation is performed for each individual N value from the N dataset obtained in step 8.

    A histogram is plotted illustrating the distribution of the number of Monte-Carlo data points fitting within the acceptable deviation from the measured concentration according to the estimated population number.

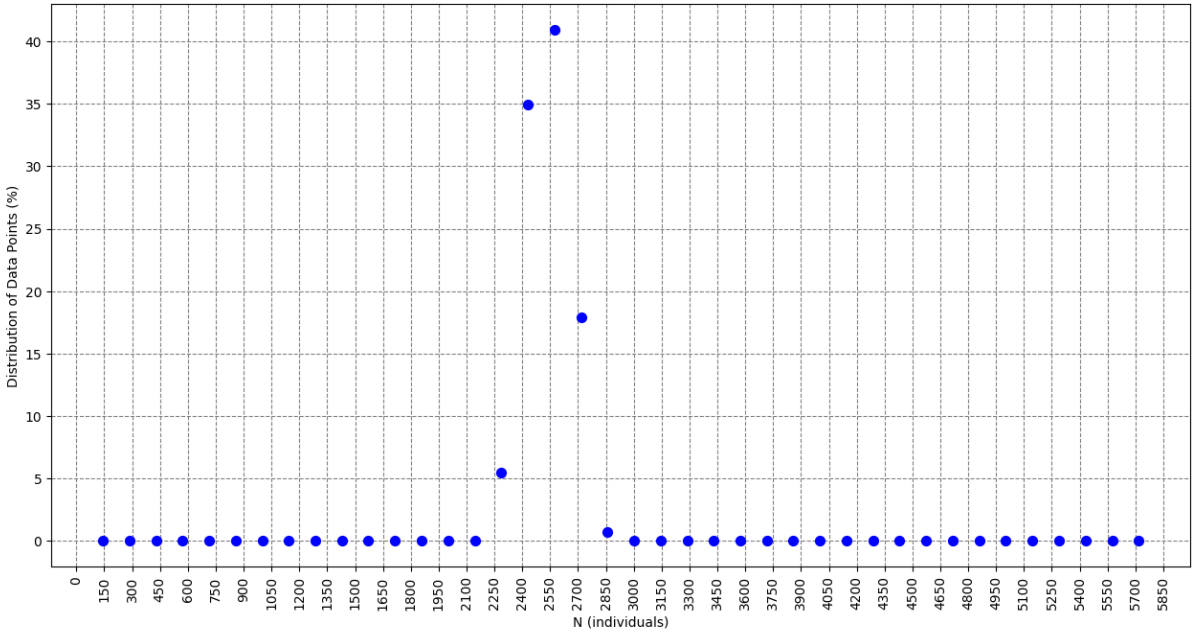

Figure S8. Probability of each population estimate according to the measured crAssphage concentration. The probability of each number of individuals of a population to result in a measured concentration of crAssphage in wastewater according to a full Monte-Carlo simulation while the probability was calculated through Bayes' rule.

In conclusion, Figure S8 depicts the expected density, distribution or probability for each of the estimated population numbers, assuming the measured crAssphage concentration, thereby indicating the simulated distribution of the population infected with CrAssphage.

Table S1. Population Estimates According to the Central Limit Theorem

Date	SARS-CoV-2			Influenza A			Norovirus GII		CrAssphage	
	Measured Concentration (gc/L)	Population Estimate	SD	Measured Concentration (gc/L)	Population Estimate	SD	Measured Concentration (gc/L)	Population Estimate	Measured Concentration (gc/L)	Population Estimate
2021-10-28	1968.2	3.2	8	0.0	0.0	0	N/A	N/A	1246483.9	127.6
2021-10-29	6042.3	9.7	14	0.0	0.0	0	N/A	N/A	1102212.4	112.8
2021-11-01	27336.3	44	29	0.0	0.0	0	N/A	N/A	3497303.0	358.0
2021-11-02	40882.5	65.8	36	0.0	0.0	0	N/A	N/A	4757886.2	487.0
2021-11-03	7133.6	11.5	15	0.0	0.0	0	N/A	N/A	3984320.1	407.8
2021-11-04	6929.2	11.1	15	0.0	0.0	0	N/A	N/A	4308999.6	441.1
2021-11-05	423.2	0.7	4	0.0	0.0	0	N/A	N/A	2580834.2	264.2
2021-11-08	2169.3	3.5	8	0.0	0.0	0	N/A	N/A	5067248.4	518.7
2021-11-09	962.7	1.5	6	524.4	211.2	37	N/A	N/A	4143788.5	424.1
2021-11-10	633.0	1	4	3629.5	1461.5	118	N/A	N/A	2563706.4	262.4
2021-11-11	200290.8	322.2	81	327.5	131.9	29	N/A	N/A	5072683.9	519.2
2021-11-12	3705.9	6	11	67854.9	27324.1	1424	N/A	N/A	2265422.6	231.9
2021-11-15	659.2	1.1	5	41.2	16.6	10	N/A	N/A	3755780.7	384.4
2021-11-16	1669.9	2.7	7	0.0	0.0	0	N/A	N/A	2150719.1	220.1
2021-11-17	1057.8	1.7	6	0.0	0.0	0	N/A	N/A	7726449.9	790.9
2021-11-18	2004.3	3.2	8	0.0	0.0	0	N/A	N/A	8192900.0	838.6
2021-11-19	186.0	0.3	2	0.0	0.0	0	N/A	N/A	6036388.1	617.9
2021-11-23	990.8	1.6	6	0.0	0.0	0	N/A	N/A	83201203.3	8516.2
2021-11-24	2737.2	4.4	9	412.3	166.0	32	N/A	N/A	146432181.7	14988.3
2021-11-25	731.1	1.2	5	51.6	20.8	11	N/A	N/A	229164943.8	23456.5
2021-11-26	8196.2	13.2	16	289.7	116.7	27	N/A	N/A	106413637.4	10892.1
2021-11-29	829.0	1.3	5	0.0	0.0	0	N/A	N/A	145771815.4	14920.7
2021-11-30	1363.1	2.2	7	60.8	24.5	12	N/A	N/A	134253098.1	13741.7
2021-12-01	0.0	0	0	0.0	0.0	0	N/A	N/A	73779722.6	7551.8
2021-12-02	473.0	0.8	4	0.0	0.0	0	N/A	N/A	194349167.2	19892.9
2021-12-06	941.2	1.5	5	0.0	0.0	0	N/A	N/A	680372.2	69.6

2021-12-07	11532.8	18.6	19	130.2	52.4	18	N/A	N/A	756665.5	77.4
2021-12-08	4595.3	7.4	12	613.3	247.0	40	N/A	N/A	1215579.3	124.4
2021-12-09	782.4	1.3	5	55.5	22.4	12	N/A	N/A	754810.5	77.3
2021-12-10	1357.0	2.2	7	19.1	7.7	7	N/A	N/A	884271.0	90.5
2021-12-13	29.3	0	1	48.0	19.3	11	111136.8	4.7	302215.5	30.9
2021-12-14	7326.7	11.8	15	0.0	0.0	0	119962.8	5.1	553417.4	56.6
2021-12-15	318014.5	511.6	103	0.0	0.0	0	1202461.0	51.3	1147019.3	117.4
2021-12-16	78992.7	127.1	50	0.0	0.0	0	1558950.5	66.5	911191.0	93.3
2022-01-11	27526.0	44.3	30	0.0	0.0	0	205598.1	8.8	45783688.8	4686.3
2022-01-12	14020.0	22.6	21	0.0	0.0	0	24840.2	1.1	15113998.5	1547.0
2022-01-13	12466.0	20.1	20	0.0	0.0	0	1783510.8	76.1	20907117.9	2140.0
2022-01-18	11823.0	19	19	0.0	0.0	0	0.0	0.0	725453.3	74.3
2022-01-19	2646.0	4.3	9	0.0	0.0	0	0.0	0.0	895906.6	91.7
2022-01-20	5253.0	8.5	13	0.0	0.0	0	0.0	0.0	1525019.8	156.1
2022-01-21	9246.0	14.9	17	0.0	0.0	0	0.0	0.0	4673452.9	478.4
2022-01-26	2413.0	3.9	9	0.0	0.0	0	0.0	0.0	5525519.4	565.6
2022-01-27	1133.0	1.8	6	0.0	0.0	0	0.0	0.0	1557313.2	159.4
2022-01-28	1073.0	1.7	6	0.0	0.0	0	0.0	0.0	1061726.6	108.7
2022-02-01	2650.0	4.3	9	0.0	0.0	0	0.0	0.0	1019553.2	104.4
2022-02-02	2550.0	4.1	9	0.0	0.0	0	0.0	0.0	1384519.9	141.7
2022-02-03	2190.0	3.5	8	0.0	0.0	0	0.0	0.0	1541526.5	157.8
2022-02-04	208273.0	335.1	83	0.0	0.0	0	0.0	0.0	31698656.8	3244.6
2022-02-08	25203.0	40.5	28	0.0	0.0	0	N/A	N/A	2594086.4	265.5
2022-02-09	8764.0	14.1	17	0.0	0.0	0	N/A	N/A	1602686.5	164.0
2022-02-11	17483.0	28.1	23	0.0	0.0	0	N/A	N/A	27651390.6	2830.3
2022-02-15	1270.0	2	6	0.0	0.0	0	N/A	N/A	9143865.8	935.9
2022-02-17	5031.0	8.1	13	0.0	0.0	0	N/A	N/A	2156286.5	220.7
2022-02-18	0.0	0	0	0.0	0.0	0	N/A	N/A	1614233.2	165.2
2022-02-22	106.0	0.2	2	0.0	0.0	0	N/A	N/A	4863752.8	497.8
2022-02-23	5750.0	9.3	13	0.0	0.0	0	N/A	N/A	1450666.5	148.5
2022-02-24	0.0	0	0	0.0	0.0	0	0.0	0.0	1169366.5	119.7

2022-02-25	906.0	1.5	5	0.0	0.0	0	122716.1	5.2	2171439.8	222.3
2022-03-01	1736.0	2.8	7	0.0	0.0	0	59047.9	2.5	1096066.6	112.2
2022-03-02	0.0	0	0	0.0	0.0	0	58427.1	2.5	2933093.0	300.2
2022-03-03	3123.0	5	10	0.0	0.0	0	36657.5	1.6	684619.9	70.1
2022-03-04	14146.0	22.8	21	0.0	0.0	0	24188.7	1.0	1651213.2	169.0
2022-03-08	746.0	1.2	5	0.0	0.0	0	N/A	N/A	3849226.3	394.0
2022-03-09	4730.0	7.6	12	0.0	0.0	0	N/A	N/A	1922359.8	196.8
2022-03-10	0.0	0	0	0.0	0.0	0	N/A	N/A	903059.9	92.4
2022-03-11	1670.0	2.7	7	0.0	0.0	0	N/A	N/A	742799.9	76.0
2022-03-15	630.0	1	4	0.0	0.0	0	0.0	0.0	492653.3	50.4
2022-03-16	62190.0	100	45	0.0	0.0	0	0.0	0.0	985953.2	100.9
2022-03-17	12533.0	20.2	20	0.0	0.0	0	0.0	0.0	503666.6	51.6
2022-03-18	73130.0	117.6	48	0.0	0.0	0	0.0	0.0	864446.6	88.5
2022-03-22	27487.0	44.2	29	N/A	N/A	N/A	N/A	N/A	2127233.1	217.7
2022-03-23	8927.0	14.4	17	N/A	N/A	N/A	N/A	N/A	1343739.9	137.5
2022-03-24	8436.0	13.6	16	N/A	N/A	N/A	N/A	N/A	1033799.9	105.8
2022-03-25	11464.0	18.4	19	N/A	N/A	N/A	N/A	N/A	961719.9	98.4
2022-03-29	4553.0	7.3	12	0.0	0.0	0	2321199.5	99.0	10008139.0	1024.4
2022-03-30	8113.0	13.1	16	0.0	0.0	0	1128173.4	48.1	38348282.8	3925.2
2022-03-31	20356.0	32.7	25	0.0	0.0	0	3213769.9	137.1	11901705.5	1218.2
2022-04-01	3123.0	5	10	0.0	0.0	0	1913305.9	81.6	3127659.7	320.1
2022-04-05	7026.0	11.3	15	0.0	0.0	0	21925716.5	935.6	20813371.3	2130.4
2022-04-06	3626.0	5.8	11	0.0	0.0	0	11823313.2	504.5	21972197.8	2249.0
2022-04-07	6187.0	10	14	0.0	0.0	0	27150820.0	1158.5	31023070.2	3175.4
2022-04-08	1102.0	1.8	6	0.0	0.0	0	2117679.1	90.4	42132175.8	4312.5
2022-04-12	481.0	0.8	4	N/A	N/A	N/A	48278.2	2.1	11271818.9	1153.7
2022-04-13	2652.0	4.3	9	N/A	N/A	N/A	13506.9	0.6	10440165.6	1068.6
2022-04-14	7009.0	11.3	15	N/A	N/A	N/A	53110.7	2.3	7969752.5	815.8
2022-04-20	3789.0	6.1	11	N/A	N/A	N/A	N/A	N/A	2489379.8	254.8
2022-04-21	2926.0	4.7	10	N/A	N/A	N/A	N/A	N/A	16458005.0	1684.6
2022-04-22	0.0	0	0	N/A	N/A	N/A	N/A	N/A	4596766.2	470.5

2022-04-26	7082.0	11.4	15	N/A	N/A	N/A	N/A	N/A	19172384.7	1962.4
2022-04-27	203.0	0.3	3	N/A	N/A	N/A	N/A	N/A	15894451.7	1626.9
2022-04-28	2414.0	3.9	9	N/A	N/A	N/A	N/A	N/A	28770783.8	2944.9
2022-04-29	181.0	0.3	2	N/A	N/A	N/A	N/A	N/A	4081399.6	417.8
2022-05-04	2787.0	4.5	9	N/A	N/A	N/A	N/A	N/A	17770931.6	1819.0
2022-05-05	5855.0	9.4	14	N/A	N/A	N/A	N/A	N/A	37692782.9	3858.1
2022-05-06	1077.0	1.7	6	N/A	N/A	N/A	N/A	N/A	19765658.0	2023.1
2022-05-10	18126.0	29.2	24	N/A	N/A	N/A	N/A	N/A	32479383.4	3324.5
2022-05-11	0.0	0	0	N/A	N/A	N/A	N/A	N/A	4755912.9	486.8
2022-05-12	675.0	1.1	5	N/A	N/A	N/A	N/A	N/A	23905257.6	2446.9
2022-05-13	0.0	0	0	N/A	N/A	N/A	N/A	N/A	29649157.0	3034.8
2022-05-17	0.0	0	0	N/A	N/A	N/A	N/A	N/A	20081224.7	2055.4
2022-05-18	291.0	0.5	3	N/A	N/A	N/A	N/A	N/A	29634163.7	3033.2
2022-05-19	0.0	0	0	N/A	N/A	N/A	N/A	N/A	33347396.7	3413.3
2022-05-20	0.0	0	0	N/A	N/A	N/A	N/A	N/A	23488151.0	2404.2

The table depicts population estimates for each virus target according to the Central Limit Theorem. The standard deviation (SD) is provided only for SARS-CoV-2 and influenza A because it was possible to find all the required data only for these two targets. When the measured concentration displays "0.0" it indicates that no virus genome was detected in the wastewater samples. N/A – not applicable and indicates that the wastewater sample with the corresponding date was not analysed for that virus target.

Table S2. Total Population Estimates

Date	Population estimate (individuals) based on the wastewater concentration of						
	Ammonium	Phosphate	CrAssphage, MCBA Approach 1	CrAssphage, MCBA Approach 2	CrAssphage, MCBA Approach 3	CrAssphage, MCBA Approach 2 Adjusted	CrAssphage, EMCLT
2021-10-28	8318.7	6329.2	3.1	118.0	96.2	330.5	127.6
2021-10-29	4019.6	2865.9	3.0	118.0	77.8	330.5	112.8
2021-11-01	4721.6	5880.5	6.6	319.5	291.7	895.0	358.0
2021-11-02	7501.1	7047.5	8.6	439.8	416.2	1232.0	487.0
2021-11-03	5407.5	4732.2	7.3	366.5	345.2	1026.7	407.8
2021-11-04	5556.9	4881.1	7.1	397.6	366.2	1113.6	441.1
2021-11-05	6476.8	4557.2	4.7	236.7	207.8	663.0	264.2
2021-11-08	N/A	N/A	9.0	467.6	431.9	1309.7	518.7
2021-11-09	N/A	N/A	7.3	379.1	357.6	1062.0	424.1
2021-11-10	11347.2	6121.4	4.8	236.1	215.6	661.4	262.4
2021-11-11	10114.8	5446.6	9.0	467.5	439.9	1309.6	519.2
2021-11-12	7982.9	6000.9	5.6	216.3	177.2	605.7	231.9
2021-11-15	1886.4	7841.9	6.8	349.6	320.3	979.2	384.4
2021-11-16	2056.3	6076.2	5.3	191.6	171.4	536.7	220.1
2021-11-17	2348.8	11054.7	13.0	709.8	667.3	1988.3	790.9
2021-11-18	13208.7	8254.4	14.0	759.0	761.0	2126.0	838.6
2021-11-19	16488.2	10597.1	9.8	557.2	546.6	1560.8	617.9
2021-11-23	8692.8	18352.9	122.7	7044.7	7665.0	19733.0	8516.2
2021-11-24	10682.3	9808.3	215.7	N/A	N/A	N/A	14988.3
2021-11-25	11159.5	10393.1	339.5	N/A	N/A	N/A	23456.5
2021-11-26	11097.6	9450.9	158.1	N/A	9803.0	N/A	10892.1
2021-11-29	13200.5	12547.8	212.8	N/A	13429.0	N/A	14920.7
2021-11-30	9785.3	9856.7	197.6	N/A	11680.0	N/A	13741.7
2021-12-01	4266.7	5729.9	108.1	6735.8	6797.0	18867.8	7551.8
2021-12-02	5638.4	10977.9	287.6	N/A	19893.0	N/A	19892.9
2021-12-06	15048.0	8903.5	2.3	59.0	44.2	165.3	69.6

2021-12-07	9248.5	5409.9	2.4	59.0	50.1	165.3	77.4
2021-12-08	6511.2	4329.4	2.8	118.0	96.5	330.5	124.4
2021-12-09	8086.4	6368.0	2.4	59.0	53.8	165.3	77.3
2021-12-10	6286.9	6189.2	2.7	64.2	62.6	179.9	90.5
2021-12-13	11876.8	9140.7	2.1	59.0	19.1	165.3	30.9
2021-12-14	15080.0	7254.2	2.3	59.0	38.2	165.3	56.6
2021-12-15	10820.9	10004.5	3.0	118.0	84.4	330.5	117.4
2021-12-16	8899.2	7201.9	2.8	79.4	68.0	222.5	93.3
2022-01-11	12080.0	12468.7	68.9	4207.0	4159.0	11784.3	4686.3
2022-01-12	24156.8	9103.1	23.2	1388.2	1375.1	3888.5	1547.0
2022-01-13	22051.2	10074.4	32.2	1922.6	1926.0	5385.4	2140.0
2022-01-18	16889.6	9871.1	2.4	59.0	52.9	165.3	74.3
2022-01-19	13862.4	7589.6	2.8	69.7	67.5	195.3	91.7
2022-01-20	11827.2	10029.2	3.3	130.3	121.2	365.0	156.1
2022-01-21	12220.8	7386.4	8.6	430.2	389.0	1205.0	478.4
2022-01-26	N/A	N/A	8.8	507.9	489.8	1422.7	565.6
2022-01-27	N/A	N/A	3.2	139.2	119.1	389.9	159.4
2022-01-28	N/A	N/A	3.0	118.0	83.8	330.5	108.7
2022-02-01	10582.4	11497.4	3.0	118.0	76.0	330.5	104.4
2022-02-02	17164.8	9148.2	3.6	118.0	95.4	330.5	141.7
2022-02-03	14313.6	8628.7	3.3	133.9	119.9	375.1	157.8
2022-02-04	17990.4	12333.2	46.1	2905.0	2920.0	8137.2	3244.6
2022-02-08	14902.4	9758.1	5.1	238.1	213.2	666.9	265.5
2022-02-09	11254.4	10932.7	3.3	156.2	127.1	437.6	164.0
2022-02-11	13699.2	10752.0	41.3	2543.5	2500.2	7124.8	2830.3
2022-02-15	13827.2	14569.4	14.5	847.2	837.1	2373.2	935.9
2022-02-17	8192.0	2349.2	5.7	193.7	171.0	542.5	220.7
2022-02-18	10956.8	7838.1	3.3	158.2	132.2	443.1	165.2
2022-02-22	13824.0	9374.1	8.3	446.3	424.8	1250.2	497.8
2022-02-23	10243.2	7386.4	3.6	118.4	107.8	331.7	148.5

2022-02-24	7664.0	6347.3	2.8	118.0	88.5	330.5	119.7
2022-02-25	8659.2	5240.5	5.8	195.6	179.2	547.9	222.3
2022-03-01	12464.0	10300.2	3.0	118.0	87.0	330.5	112.2
2022-03-02	15820.8	9351.5	5.9	274.3	247.2	768.4	300.2
2022-03-03	12595.2	11136.0	2.3	59.0	48.2	165.3	70.1
2022-03-04	14288.0	8673.9	3.5	167.1	133.9	467.9	169.0
2022-03-08	16860.8	16150.6	6.9	355.3	313.3	995.3	394.0
2022-03-09	14691.2	9351.5	4.6	177.0	147.2	495.8	196.8
2022-03-10	9353.6	9148.2	2.7	71.6	63.2	200.7	92.4
2022-03-11	15152.0	10910.1	2.3	59.0	45.2	165.3	76.0
2022-03-15	12633.6	7634.8	2.2	59.0	33.6	165.3	50.4
2022-03-16	11107.2	7228.2	2.9	115.5	75.4	323.6	100.9
2022-03-17	13568.0	7567.1	2.2	59.0	33.0	165.3	51.6
2022-03-18	3718.4	8357.6	2.6	62.8	63.6	175.8	88.5
2022-03-22	6166.4	7137.9	5.3	186.9	170.6	523.6	217.7
2022-03-23	6643.2	7770.4	3.3	118.2	100.7	331.2	137.5
2022-03-24	4764.8	11858.8	3.0	118.0	71.9	330.5	105.8
2022-03-25	5635.2	8719.1	2.9	108.7	72.6	304.4	98.4
2022-03-29	22454.4	10797.2	16.5	922.7	870.9	2584.7	1024.4
2022-03-30	11500.8	8425.4	57.5	3523.4	3729.0	9869.5	3925.2
2022-03-31	11792.0	9464.5	18.1	1094.0	1062.3	3064.4	1218.2
2022-04-01	11830.4	7928.5	6.1	291.9	237.0	817.6	320.1
2022-04-05	15929.6	5376.0	31.4	1924.1	1890.8	5389.7	2130.4
2022-04-06	22195.2	5872.9	34.6	2030.2	2062.0	5686.9	2249.0
2022-04-07	19788.8	8583.5	48.4	2850.0	2884.5	7983.3	3175.4
2022-04-08	19574.4	12852.7	64.8	3910.3	3881.0	10953.3	4312.5
2022-04-12	16483.2	19764.7	17.5	1040.8	980.6	2915.5	1153.7
2022-04-13	12854.4	3907.8	16.6	962.2	944.0	2695.1	1068.6
2022-04-14	13612.8	8651.3	13.7	736.8	698.4	2063.7	815.8
2022-04-20	13510.4	7521.9	4.8	234.7	201.5	657.3	254.8

2022-04-21	13814.4	7183.1	24.6	1512.3	1541.2	4236.2	1684.6
2022-04-22	20140.8	13281.9	7.9	423.8	377.2	1187.1	470.5
2022-04-26	11376.0	7860.7	28.4	1769.7	1962.0	4957.0	1962.4
2022-04-27	13804.8	8176.9	25.9	1459.1	1464.2	4087.0	1626.9
2022-04-28	8112.0	4133.6	43.7	2620.9	2601.2	7341.5	2944.9
2022-04-29	11836.8	7431.5	7.6	379.2	343.4	1062.2	417.8
2022-05-04	21852.8	16060.2	27.8	1627.8	1614.3	4559.8	1819.0
2022-05-05	15801.6	10548.7	55.9	3473.7	3472.0	9730.4	3858.1
2022-05-06	17158.4	12378.4	30.0	1828.2	1745.3	5121.1	2023.1
2022-05-10	10051.2	12039.5	50.2	2992.2	2992.0	8381.5	3324.5
2022-05-11	14246.4	11768.5	8.6	439.6	416.8	1231.4	486.8
2022-05-12	13971.2	9396.7	36.9	2218.2	2243.0	6213.4	2446.9
2022-05-13	14220.8	12468.7	45.2	2718.6	2731.3	7615.1	3034.8
2022-05-17	11782.4	11587.8	30.3	1839.7	1832.8	5153.3	2055.4
2022-05-18	9843.2	6912.0	45.1	2713.1	2882.0	7599.8	3033.2
2022-05-19	13779.2	9306.4	49.0	3078.3	3072.0	8622.6	3413.3
2022-05-20	14518.4	7838.1	35.8	2163.4	2144.0	6059.9	2404.2

Total population number according to the wastewater ammonium and phosphate concentration, and according to the Monte-Carlo-Bayesian Approach (MCBA) and the equation model following the Central Limit Theorem premise (EMCLT) for crAssphage-based estimates. The MCBA Approach 2 adjustment consists in the assumption that only 35.7% of people shed CrAssphage in faeces.

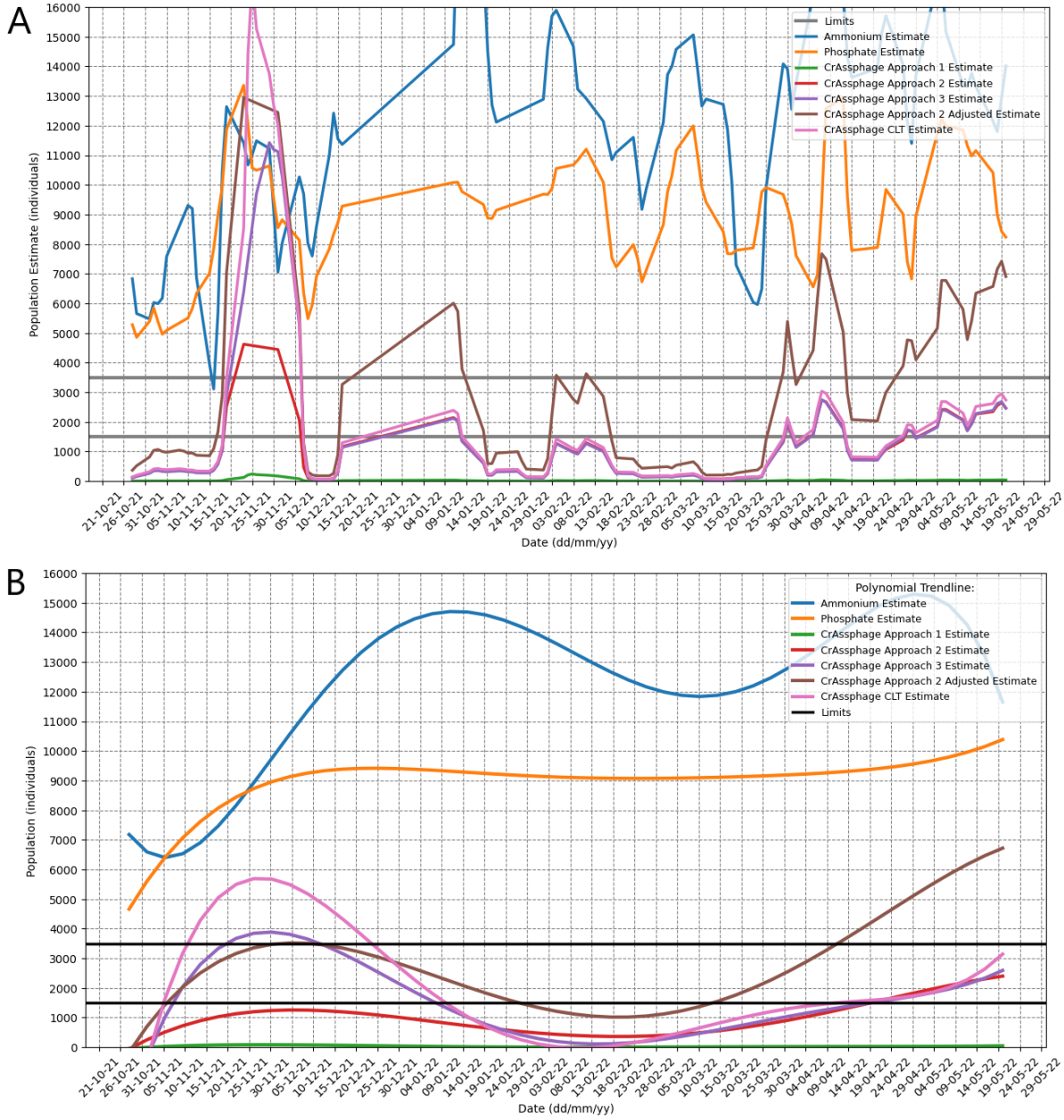

Figure S9. Total Population Estimates comparing to the Population Expected Limits Total population estimates based on ammonium, phosphate, and crAssphage concentrations. *Panel A: non-filtered data plot. Panel B: polynomial trendline of a 10-th order for each population estimate. The limits (black) are the expected numbers for the population within the catchment area. The MCBA Approach 2 adjustment consists in the assumption that only 35.7% of people shed CrAssphage in faeces.*

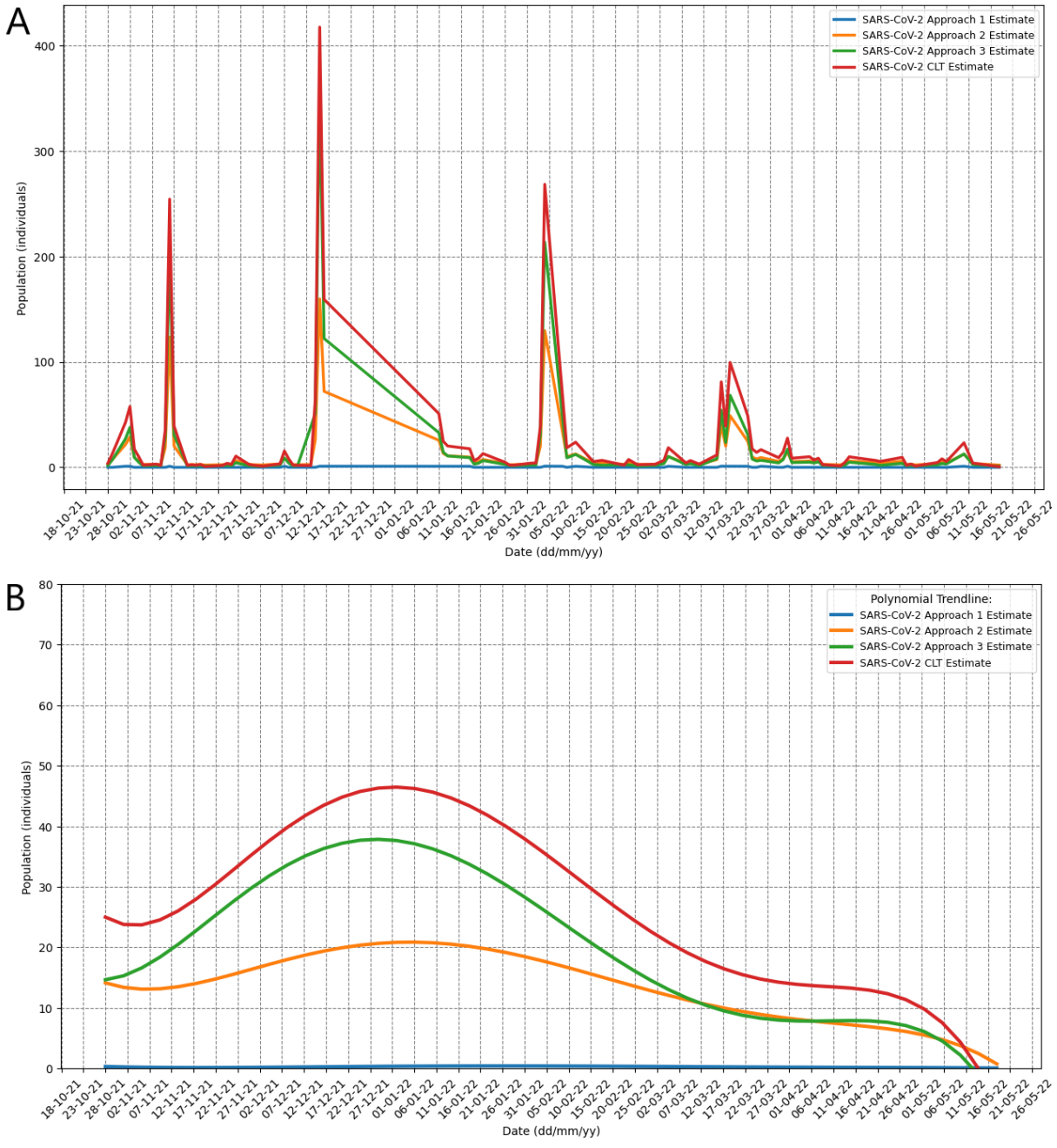

Figure S10. SARS-CoV-2 Infected Population Estimates. Estimates of the population infected with SARS-CoV-2 according to the Monte-Carlo-Bayesian approach 1, 2 and 3, and the equation model following the Central Limit Theorem premise (CLT). Panel A: non-filtered data plot. Panel B: polynomial trendline of a 10-th order for each estimate type.

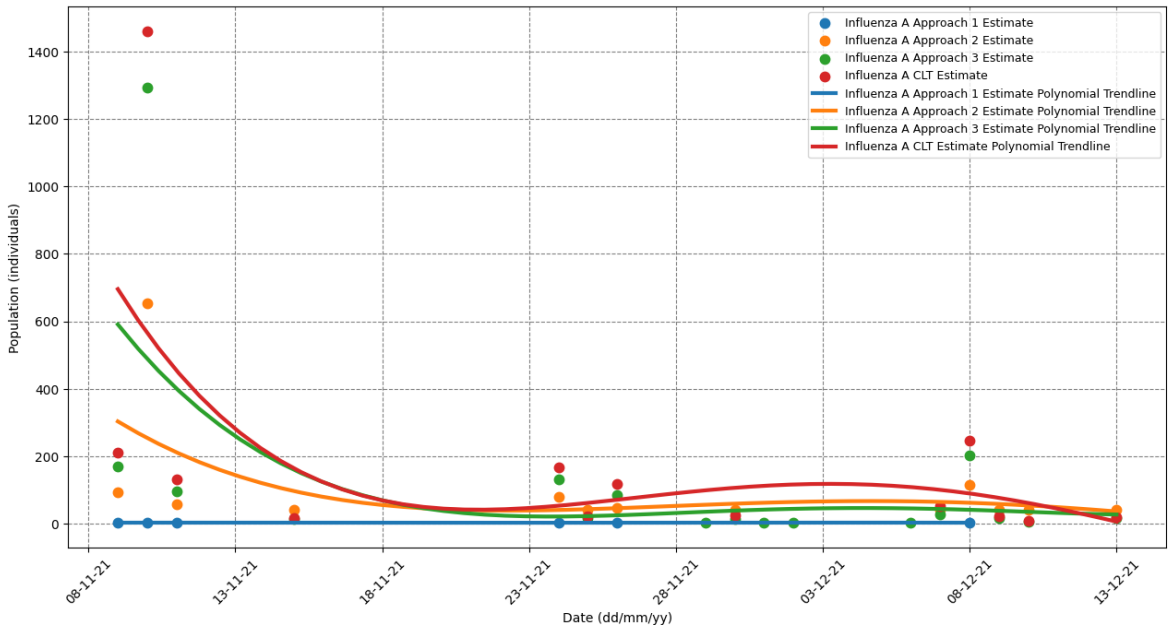

Figure S11. Influenza A Infected Population Estimates. Estimates of the population infected with influenza A according to the Monte-Carlo-Bayesian approach 1, 2 and 3, and the equation model following the Central Limit Theorem premise (CLT). Panel A: non-filtered scatter plot. Panel B: polynomial trendline of a 10-th order for each estimate type.

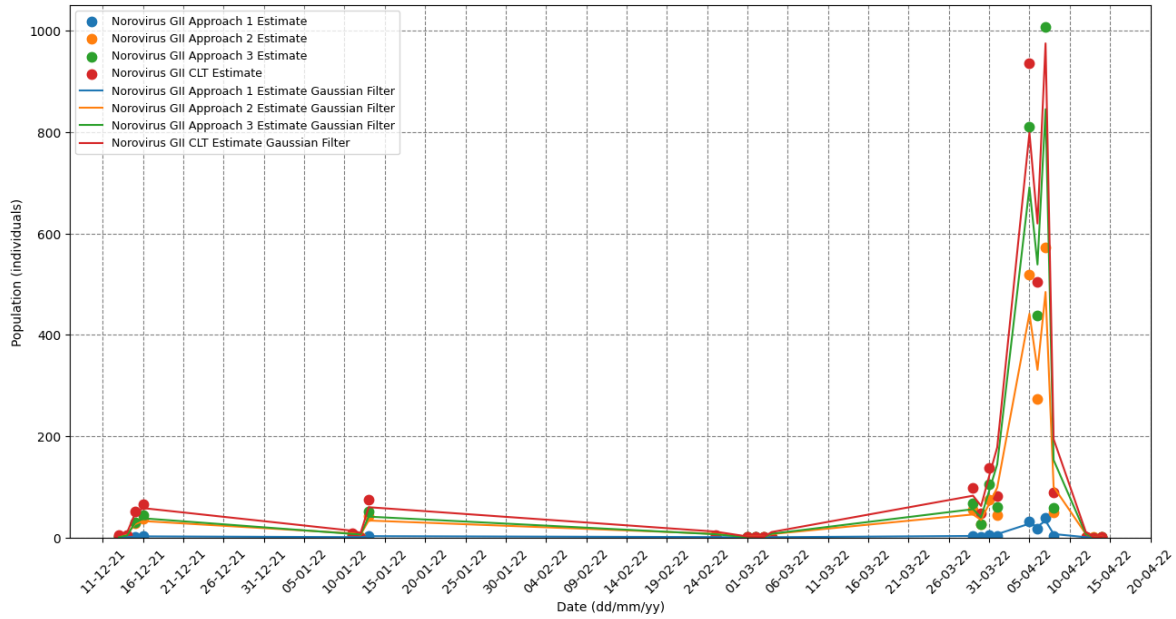

Figure S12. Norovirus GII Infected Population Estimates. Estimates of the population infected with norovirus GII according to the Monte-Carlo-Bayesian approach 1, 2 and 3, and the equation model following the Central Limit Theorem premise (CLT). A Gaussian filter was applied, sigma=0.5.

Table S3. Norovirus GII Monte-Carlo-Bayesian Simulations Results

Date		Approach 1			Approach 2				Approach 3			
	Concentration (log_{10} gc/L)	Mean N	SD (σ)	99% CI	Mean N	SD (σ)	99% CI	γ_{min}	Mean N	SD (σ)	99% CI	γ_{min}
2021-12-13	5.045857931	1.3	0.5	(1.12, 1.38)	4.0	2.0	(3.84, 4.18)	0.938	2.0	1.0	(1.7, 2.25)	0.781
2021-12-14	5.079046568	1.4	0.6	(1.18, 1.53)	4.1	2.0	(3.95, 4.29)	0.938	1.5	0.8	(1.25, 1.76)	0.605
2021-12-15	6.080071004	2.9	1.4	(2.39, 3.35)	28.8	5.8	(28.31, 29.27)	0.938	31.5	6.4	(30.12, 32.8)	0.944
2021-12-16	6.192832338	3.3	1.4	(2.85, 3.82)	37.5	6.8	(36.91, 38.05)	0.938	44.4	7.0	(42.75, 45.98)	0.949
2022-01-11	5.313019124	1.4	0.6	(1.23, 1.58)	6.0	2.7	(5.79, 6.25)	0.938	3.8	1.8	(3.41, 4.14)	0.891
2022-01-12	4.395155045	1.1	0.3	(1.0, 1.13)	1.8	0.9	(1.74, 1.89)	0.938	1.0	0.2	(0.96, 1.1)	0.000
2022-01-13	6.251275733	4.1	1.9	(3.52, 4.68)	42.7	7.3	(42.11, 43.32)	0.938	52.9	8.4	(50.93, 54.83)	0.952
2022-02-25	5.088901383	1.4	0.6	(1.2, 1.6)	4.2	2.0	(4.02, 4.37)	0.938	1.9	1.0	(1.63, 2.19)	0.749
2022-03-01	4.771204168	1.2	0.4	(1.05, 1.28)	2.7	1.4	(2.62, 2.85)	0.938	1.2	0.4	(1.04, 1.33)	0.550
2022-03-02	4.76661404	1.2	0.4	(1.05, 1.29)	2.7	1.4	(2.59, 2.82)	0.938	1.3	0.7	(1.0, 1.58)	0.000
2022-03-03	4.564163195	1.1	0.3	(1.01, 1.15)	2.1	1.1	(2.04, 2.23)	0.938	1.2	0.4	(0.97, 1.35)	0.000
2022-03-04	4.383613083	1.1	0.3	(1.0, 1.14)	1.8	1.0	(1.73, 1.89)	0.938	1.1	0.3	(0.95, 1.19)	0.000
2022-03-29	6.365712469	4.4	2.0	(3.73, 5.07)	54.8	8.6	(54.11, 55.57)	0.938	68.0	10.1	(65.12, 70.88)	0.951
2022-03-30	6.052375864	2.9	1.7	(2.31, 3.42)	27.4	5.9	(26.88, 27.92)	0.938	28.0	5.7	(26.76, 29.16)	0.940
2022-03-31	6.507014776	5.9	2.8	(4.88, 6.99)	75.3	10.0	(74.49, 76.16)	0.938	105.8	12.5	(101.85, 109.65)	0.957
2022-04-01	6.281784405	4.3	2.4	(3.47, 5.06)	45.5	7.4	(44.87, 46.11)	0.938	60.3	8.3	(58.3, 62.2)	0.953
2022-04-05	7.340953794	32.5	6.5	(30.13, 34.83)	518.3	32.6	(507.92, 528.73)	0.938	810.7	33.2	(782.13, 839.2)	0.962
2022-04-06	7.072739192	19.0	5.3	(17.22, 20.75)	273.6	19.3	(271.89, 275.21)	0.938	439.2	36.7	(416.24, 462.11)	0.960
2022-04-07	7.43378295	39.7	7.5	(37.41, 42.07)	572.3	6.4	(569.24, 575.38)	0.938	1006.8	61.5	(950.73, 1062.77)	0.962
2022-04-08	6.325860141	4.2	1.6	(3.67, 4.69)	50.1	8.1	(49.39, 50.77)	0.938	58.6	9.0	(56.24, 60.99)	0.950
2022-04-12	4.683751026	1.1	0.3	(1.02, 1.19)	2.5	1.3	(2.34, 2.56)	0.938	1.1	0.3	(0.96, 1.24)	0.000
2022-04-13	4.130556713	1.1	0.3	(1.0, 1.13)	1.5	0.8	(1.44, 1.56)	0.938	1.0	0.2	(0.97, 1.12)	0.000
2022-04-14	4.725182255	1.1	0.3	(1.02, 1.21)	2.6	1.3	(2.48, 2.71)	0.938	1.3	0.4	(1.02, 1.48)	0.029

Results of the Monte-Carlo-Bayesian simulations for the detected norovirus GII concentration in wastewater: the detected concentration, the mean of the estimates distribution, the standard deviation (SD) of the distribution and the 99% confidence interval (CI 99%). The γ_{min} for the approach 2 and 3 stands for the assumed minimum viral losses, which depends on the approach type and the measured concentration.

Table S4. SARS-CoV-2 Monte-Carlo-Bayesian Simulations Results

Date		Approach 1			Approach 2				Approach 3			
	Concentration (log10 gc/L)	Mean N	SD (σ)	99% CI	Mean N	SD (σ)	99% CI	γmin	Mean N	SD (σ)	99% CI	γmin
2021-10-28	3.294069227	0.01	0.07	(0.0, 0.01)	2.72	1.72	(2.67, 2.78)	0.99963	1.36	0.62	(1.29, 1.42)	0.99330
2021-10-29	3.78119869	0.01	0.09	(0.0, 0.01)	5.88	2.65	(5.8, 5.97)	0.99963	4.87	3.12	(4.78, 4.96)	0.99938
2021-11-01	4.436740262	1	0	(nan, nan)	22.25	5.33	(22.08, 22.42)	0.99963	27.85	8.19	(27.63, 28.08)	0.99971
2021-11-02	4.611537445	1	0	(nan, nan)	32.84	6.44	(32.63, 33.05)	0.99963	43.68	9.52	(43.37, 43.99)	0.99973
2021-11-03	3.853307014	0.01	0.1	(0.0, 0.01)	6.71	2.84	(6.62, 6.81)	0.99963	5.28	3.39	(5.19, 5.38)	0.99941
2021-11-04	3.840681008	0.01	0.1	(0.0, 0.01)	6.57	2.81	(6.48, 6.66)	0.99963	5.22	3.32	(5.12, 5.31)	0.99940
2021-11-05	2.62651145	0	0.07	(0.0, 0.01)	2.14	1.3	(2.09, 2.19)	0.99963	1	0	(nan, nan)	0.00000
2021-11-08	3.33632629	0.01	0.07	(0.0, 0.01)	2.92	1.78	(2.86, 2.98)	0.99963	1.86	1.09	(1.79, 1.94)	0.99735
2021-11-09	2.983475934	0	0.07	(0.0, 0.01)	2.37	1.49	(2.32, 2.43)	0.99963	1.54	0.81	(1.47, 1.61)	0.99599
2021-11-10	2.80140371	0	0.07	(0.0, 0.01)	2.23	1.36	(2.17, 2.28)	0.99963	1.03	0.16	(0.98, 1.07)	0.92432
2021-11-11	5.301660893	1	0	(nan, nan)	156.08	14.74	(155.6, 156.56)	0.99963	259.89	21.62	(258.49, 261.28)	0.99978
2021-11-12	3.568888674	0.01	0.07	(0.0, 0.01)	4.17	2.2	(4.09, 4.24)	0.99963	3.45	2.18	(3.37, 3.53)	0.99913
2021-11-15	2.818995237	0	0.07	(0.0, 0.01)	2.24	1.37	(2.18, 2.29)	0.99963	1	0	(nan, nan)	-0.89574
2021-11-16	3.222690465	0.01	0.07	(0.0, 0.01)	2.63	1.67	(2.57, 2.69)	0.99963	1.56	0.84	(1.48, 1.63)	0.99531
2021-11-17	3.024403563	0	0.07	(0.0, 0.01)	2.4	1.51	(2.35, 2.46)	0.99963	1.4	0.68	(1.33, 1.47)	0.99392
2021-11-18	3.301951892	0.01	0.07	(0.0, 0.01)	2.74	1.73	(2.69, 2.8)	0.99963	1.62	0.91	(1.55, 1.69)	0.99645
2021-11-19	2.269512944	0	0.07	(0.0, 0.01)	2.04	1.22	(1.98, 2.09)	0.99963	1	0	(nan, nan)	0.00000
2021-11-23	2.995964081	0	0.07	(0.0, 0.01)	2.38	1.5	(2.33, 2.44)	0.99963	1.55	0.8	(1.49, 1.61)	0.99619
2021-11-24	3.437310057	0.01	0.07	(0.0, 0.01)	3.41	1.93	(3.34, 3.47)	0.99963	2.13	1.28	(2.06, 2.21)	0.99809
2021-11-25	2.863983384	0	0.07	(0.0, 0.01)	2.27	1.39	(2.21, 2.32)	0.99963	1.03	0.16	(0.96, 1.09)	0.87653
2021-11-26	3.913612547	0.01	0.1	(0.0, 0.01)	7.52	3.02	(7.42, 7.62)	0.99963	5.21	3.3	(5.12, 5.3)	0.99937
2021-11-29	2.918554531	0	0.07	(0.0, 0.01)	2.31	1.43	(2.26, 2.37)	0.99963	1.06	0.24	(0.91, 1.2)	0.60566
2021-11-30	3.134524821	0.01	0.07	(0.0, 0.01)	2.52	1.59	(2.47, 2.58)	0.99963	1.16	0.42	(1.09, 1.23)	0.98236
2021-12-02	2.674861141	0	0.07	(0.0, 0.01)	2.16	1.31	(2.11, 2.22)	0.99963	1	0	(nan, nan)	0.75290
2021-12-06	2.973666537	0	0.07	(0.0, 0.01)	2.36	1.48	(2.31, 2.41)	0.99963	1.17	0.43	(1.1, 1.25)	0.97558
2021-12-07	4.061936016	1	0	(nan, nan)	10.14	3.55	(10.03, 10.26)	0.99963	10.6	5.79	(10.48, 10.72)	0.99965
2021-12-08	3.662309144	0.01	0.07	(0.0, 0.01)	4.82	2.38	(4.74, 4.9)	0.99963	3.52	2.27	(3.44, 3.61)	0.99910
2021-12-09	2.893438093	0	0.07	(0.0, 0.01)	2.3	1.42	(2.24, 2.35)	0.99963	1	0	(nan, nan)	0.45176
2021-12-10	3.132579848	0.01	0.07	(0.0, 0.01)	2.52	1.59	(2.46, 2.57)	0.99963	1.16	0.43	(1.1, 1.22)	0.98474

2021-12-13	1.467361417	0	0.07	(0.0, 0.01)	1.97	1.17	(1.92, 2.02)	0.99963	N/A	N/A	N/A	0.00000
2021-12-14	3.864906793	0.01	0.1	(0.0, 0.01)	6.85	2.86	(6.76, 6.95)	0.99963	5.33	3.37	(5.23, 5.42)	0.99941
2021-12-15	5.502446984	1.05	0.22	(0.92, 1.18)	193.6	4.99	(191.68, 195.52)	0.99963	438.17	29.12	(435.9, 440.43)	0.99979
2021-12-16	4.897586775	1.06	0.25	(0.9, 1.22)	62.29	9.08	(62.0, 62.58)	0.99963	92.18	12.67	(91.64, 92.73)	0.99976
2022-01-11	4.439743106	1	0	(nan, nan)	22.47	5.36	(22.29, 22.64)	0.99963	27.34	8.01	(27.12, 27.56)	0.99970
2022-01-12	4.146748014	1	0	(nan, nan)	12.14	3.97	(12.01, 12.27)	0.99963	13.14	6.24	(13.01, 13.27)	0.99967
2022-01-13	4.095727123	1	0	(nan, nan)	10.9	3.71	(10.78, 11.02)	0.99963	10.56	5.53	(10.45, 10.68)	0.99962
2022-01-18	4.07272769	1	0	(nan, nan)	10.37	3.6	(10.25, 10.49)	0.99963	9.93	5.41	(9.81, 10.04)	0.99961
2022-01-19	3.42258984	0.01	0.07	(0.0, 0.01)	3.34	1.91	(3.28, 3.41)	0.99963	2.05	1.21	(1.98, 2.12)	0.99792
2022-01-20	3.720407401	0.01	0.08	(0.0, 0.01)	5.3	2.51	(5.21, 5.38)	0.99963	3.74	2.43	(3.65, 3.83)	0.99915
2022-01-21	3.965953889	0.01	0.1	(0.0, 0.02)	8.35	3.18	(8.25, 8.46)	0.99963	7.26	4.52	(7.16, 7.37)	0.99955
2022-01-26	3.382557322	0.01	0.07	(0.0, 0.01)	3.15	1.85	(3.09, 3.22)	0.99963	2.41	1.53	(2.33, 2.49)	0.99849
2022-01-27	3.05422991	0	0.07	(0.0, 0.01)	2.43	1.53	(2.38, 2.49)	0.99963	1.54	0.85	(1.47, 1.62)	0.99572
2022-01-28	3.030599722	0	0.07	(0.0, 0.01)	2.41	1.51	(2.35, 2.46)	0.99963	1.34	0.64	(1.27, 1.41)	0.99288
2022-02-01	3.423245874	0.01	0.07	(0.0, 0.01)	3.35	1.91	(3.28, 3.41)	0.99963	2.44	1.53	(2.37, 2.52)	0.99854
2022-02-02	3.40654018	0.01	0.07	(0.0, 0.01)	3.25	1.87	(3.19, 3.32)	0.99963	2.08	1.29	(2.01, 2.16)	0.99794
2022-02-03	3.340444115	0.01	0.07	(0.0, 0.01)	2.94	1.78	(2.88, 3.0)	0.99963	2.51	1.59	(2.43, 2.59)	0.99868
2022-02-04	5.318632973	1.1	0.31	(0.92, 1.28)	161.54	14.64	(161.06, 162.02)	0.99963	267.59	21.75	(266.15, 269.03)	0.99978
2022-02-08	4.401452239	1	0	(nan, nan)	20.76	5.11	(20.59, 20.93)	0.99963	24.31	7.54	(24.11, 24.51)	0.99969
2022-02-09	3.942702369	0.01	0.1	(0.0, 0.02)	7.96	3.1	(7.86, 8.06)	0.99963	6.09	3.85	(5.99, 6.19)	0.99946
2022-02-11	4.242615958	1	0	(nan, nan)	14.73	4.24	(14.59, 14.87)	0.99963	14.56	6.09	(14.41, 14.71)	0.99963
2022-02-15	3.103803721	0	0.07	(0.0, 0.01)	2.48	1.56	(2.42, 2.53)	0.99963	1.2	0.44	(1.13, 1.26)	0.98503
2022-02-17	3.701654317	0.01	0.08	(0.0, 0.01)	5.14	2.47	(5.06, 5.22)	0.99963	1.98	1.19	(1.9, 2.05)	0.99723
2022-02-22	2.025305865	0	0.07	(0.0, 0.01)	2	1.19	(1.95, 2.06)	0.99963	N/A	N/A	N/A	0.00000
2022-02-23	3.759667845	0.01	0.09	(0.0, 0.01)	5.67	2.6	(5.58, 5.76)	0.99963	2.73	1.76	(2.65, 2.81)	0.99851
2022-02-25	2.957128198	0	0.07	(0.0, 0.01)	2.34	1.47	(2.29, 2.4)	0.99963	1	0	(nan, nan)	0.71114
2022-03-01	3.239549721	0.01	0.07	(0.0, 0.01)	2.65	1.68	(2.59, 2.7)	0.99963	1.75	0.99	(1.68, 1.82)	0.99706
2022-03-03	3.494571984	0.01	0.07	(0.0, 0.01)	3.7	2.03	(3.64, 3.77)	0.99963	2.59	1.63	(2.52, 2.67)	0.99867
2022-03-04	4.150633654	1	0	(nan, nan)	12.24	3.97	(12.11, 12.37)	0.99963	12.63	5.94	(12.5, 12.76)	0.99965
2022-03-08	2.872738827	0	0.07	(0.0, 0.01)	2.28	1.41	(2.23, 2.33)	0.99963	1.09	0.29	(1.02, 1.17)	0.95180
2022-03-09	3.674861141	0.01	0.08	(0.0, 0.01)	4.93	2.41	(4.85, 5.01)	0.99963	4.07	2.6	(3.98, 4.16)	0.99928
2022-03-11	3.222716471	0.01	0.07	(0.0, 0.01)	2.63	1.67	(2.57, 2.69)	0.99963	1.41	0.72	(1.34, 1.49)	0.99356
2022-03-15	2.799340549	0	0.07	(0.0, 0.01)	2.23	1.36	(2.17, 2.28)	0.99963	1.05	0.22	(0.99, 1.1)	0.93702

2022-03-16	4.793720557	1.1	0.3	(0.93, 1.26)	49.38	7.99	(49.12, 49.65)	0.99963	67.41	11.28	(66.97, 67.86)	0.99974
2022-03-17	4.09805504	1	0	(nan, nan)	10.94	3.71	(10.81, 11.06)	0.99963	9.67	5.1	(9.56, 9.78)	0.99958
2022-03-18	4.864095573	1.05	0.22	(0.92, 1.18)	57.66	8.81	(57.38, 57.95)	0.99963	81.81	12.07	(81.31, 82.32)	0.99974
2022-03-22	4.439127342	1	0	(nan, nan)	22.41	5.36	(22.23, 22.58)	0.99963	27.78	8.02	(27.56, 28.0)	0.99971
2022-03-23	3.950705535	0.01	0.1	(0.0, 0.02)	8.09	3.12	(7.99, 8.19)	0.99963	5.33	3.37	(5.24, 5.43)	0.99935
2022-03-24	3.926136571	0.01	0.1	(0.0, 0.01)	7.71	3.05	(7.61, 7.81)	0.99963	5.94	3.76	(5.84, 6.03)	0.99945
2022-03-25	4.059336177	1	0	(nan, nan)	10.09	3.55	(9.98, 10.21)	0.99963	7.1	4.04	(6.99, 7.2)	0.99945
2022-03-29	3.65829765	0.01	0.07	(0.0, 0.01)	4.79	2.37	(4.71, 4.86)	0.99963	3.55	2.26	(3.47, 3.63)	0.99913
2022-03-30	3.909181476	0.01	0.1	(0.0, 0.01)	7.47	3.01	(7.37, 7.57)	0.99963	5.87	3.71	(5.77, 5.97)	0.99945
2022-03-31	4.308692442	1	0	(nan, nan)	16.77	4.57	(16.62, 16.92)	0.99963	20.77	7.55	(20.59, 20.95)	0.99971
2022-04-01	3.494571984	0.01	0.07	(0.0, 0.01)	3.7	2.03	(3.64, 3.77)	0.99963	2.22	1.39	(2.15, 2.3)	0.99817
2022-04-05	3.846708145	0.01	0.1	(0.0, 0.01)	6.62	2.82	(6.53, 6.72)	0.99963	5.4	3.44	(5.31, 5.5)	0.99943
2022-04-06	3.5594278	0.01	0.07	(0.0, 0.01)	4.11	2.18	(4.03, 4.18)	0.99963	3.8	2.44	(3.71, 3.89)	0.99925
2022-04-07	3.791480116	0.01	0.09	(0.0, 0.01)	6	2.68	(5.91, 6.09)	0.99963	5.9	3.74	(5.81, 6.0)	0.99952
2022-04-08	3.042181595	0	0.07	(0.0, 0.01)	2.42	1.52	(2.37, 2.48)	0.99963	1.32	0.62	(1.25, 1.39)	0.99221
2022-04-12	2.682145076	0	0.07	(0.0, 0.01)	2.17	1.31	(2.11, 2.22)	0.99963	1.01	0.12	(0.98, 1.05)	0.90565
2022-04-13	3.42357352	0.01	0.07	(0.0, 0.01)	3.35	1.91	(3.28, 3.41)	0.99963	1.75	1.01	(1.67, 1.82)	0.99665
2022-04-14	3.84565606	0.01	0.1	(0.0, 0.01)	6.61	2.82	(6.52, 6.71)	0.99963	5.59	3.53	(5.49, 5.68)	0.99945
2022-04-20	3.578524605	0.01	0.07	(0.0, 0.01)	4.23	2.22	(4.16, 4.3)	0.99963	2.47	1.56	(2.39, 2.55)	0.99845
2022-04-21	3.466274322	0.01	0.07	(0.0, 0.01)	3.55	1.98	(3.49, 3.62)	0.99963	1.45	0.73	(1.39, 1.51)	0.99477
2022-04-26	3.850155922	0.01	0.1	(0.0, 0.01)	6.67	2.83	(6.58, 6.76)	0.99963	4.44	2.83	(4.35, 4.53)	0.99926
2022-04-27	2.307496038	0	0.07	(0.0, 0.01)	2.04	1.23	(1.99, 2.1)	0.99963	1	N/A	(nan, nan)	0.00000
2022-04-28	3.382737266	0.01	0.07	(0.0, 0.01)	3.16	1.85	(3.09, 3.22)	0.99963	1.79	1.05	(1.72, 1.86)	0.99712
2022-04-29	2.257678575	0	0.07	(0.0, 0.01)	2.03	1.22	(1.98, 2.09)	0.99963	1	0	(nan, nan)	0.00000
2022-05-04	3.445136969	0.01	0.07	(0.0, 0.01)	3.45	1.93	(3.38, 3.51)	0.99963	2.35	1.47	(2.28, 2.43)	0.99846
2022-05-05	3.767526899	0.01	0.09	(0.0, 0.01)	5.74	2.62	(5.65, 5.82)	0.99963	4.08	2.61	(3.99, 4.17)	0.99923
2022-05-06	3.032215703	0	0.07	(0.0, 0.01)	2.41	1.51	(2.36, 2.47)	0.99963	1.32	0.6	(1.25, 1.39)	0.99212
2022-05-10	4.258301976	1	0	(nan, nan)	15.2	4.29	(15.06, 15.34)	0.99963	15.79	6.32	(15.63, 15.94)	0.99965
2022-05-12	2.829303773	0	0.07	(0.0, 0.01)	2.24	1.38	(2.19, 2.3)	0.99963	1.1	0.33	(1.02, 1.18)	0.95139
2022-05-18	2.463892989	0	0.07	(0.0, 0.01)	2.08	1.25	(2.03, 2.13)	0.99963	1	0	(nan, nan)	0.00000

Results of the Monte-Carlo-Bayesian simulations for the detected SARS-CoV-2 concentration in wastewater: the detected concentration, the mean of the estimates distribution, the standard deviation (SD) of the distribution and the 99% confidence interval (CI 99%). The γ_{min} for the approach 2 and 3 stands for the assumed minimum viral losses, which depends on the approach type and the measured concentration.

Table S5. Influenza A Monte-Carlo-Bayesian Simulations Results

Date		Approach 1			Approach 2				Approach 3			
	Concentration (log_{10} gc/L)	Mean N	SD (σ)	99% CI	Mean N	SD (σ)	99% CI	γmin	Mean N	SD (σ)	99% CI	γmin
2021-11-09	2.719699446	2.12	0.35	(1.8, 2.45)	91.35	17.53	(89.86, 92.83)	0.99872	169.28	24.14	(168.15, 170.41)	0.99928
2021-11-10	3.559842769	2.62	0.81	(2.11, 3.14)	652.69	39.6	(649.23, 656.14)	0.99872	1295.3	67.07	(1287.62, 1302.88)	0.99936
2021-11-11	2.51527756	2.12	0.35	(1.8, 2.45)	57.53	20.3	(55.65, 59.4)	0.99872	95.87	20.03	(95.09, 96.65)	0.99920
2021-11-12	4.831580897	17.07	5.54	(13.26, 20.88)	N/A	N/A	N/A	0.99872	25503	899.2	(24997.23, 26008.11)	0.99939
2021-11-15	1.615131358	2	N/A	(nan, nan)	40	0	(nan, nan)	0.99872	12.82	7.77	(12.63, 13.0)	0.99861
2021-11-24	2.615248402	2.12	0.35	(1.8, 2.45)	79.91	8.33	(79.19, 80.64)	0.99872	131.83	22.91	(130.87, 132.78)	0.99927
2021-11-25	1.712462584	2	N/A	(nan, nan)	40	0	(nan, nan)	0.99872	16.67	9.98	(16.46, 16.88)	0.99888
2021-11-26	2.46199842	2.11	0.33	(1.82, 2.4)	45.68	14.17	(44.47, 46.88)	0.99872	83.85	19.59	(83.13, 84.58)	0.99919
2021-11-30	1.784163204	2	N/A	(nan, nan)	40	0	(nan, nan)	0.99872	13.67	8.22	(13.46, 13.88)	0.99852
2021-12-07	2.114685059	2	0	(nan, nan)	40	0	(nan, nan)	0.99872	27.36	12.99	(26.97, 27.74)	0.99886
2021-12-08	2.787672921	2.17	0.39	(1.88, 2.46)	114.23	16.47	(112.76, 115.7)	0.99872	201.46	25.94	(200.17, 202.75)	0.99929
2021-12-09	1.744640584	2	N/A	(nan, nan)	40	0	(nan, nan)	0.99872	14.81	8.94	(14.6, 15.02)	0.99869
2021-12-10	1.281285896	2	N/A	(nan, nan)	40	N/A	(nan, nan)	0.99872	6.32	4	(6.17, 6.46)	0.99732
2021-12-13	1.681140652	2	N/A	(nan, nan)	40	0	(nan, nan)	0.99872	14.8	8.88	(14.61, 15.0)	0.99876

Results of the Monte-Carlo-Bayesian simulations for the detected influenza A concentration in wastewater: the detected concentration, the mean of the estimates distribution, the standard deviation (SD) of the distribution and the 99% confidence interval (CI 99%). The γ_{min} for the approach 2 and 3 stands for the assumed minimum viral losses, which depends on the approach type and the measured concentration.

Table S6. CrAssphage Monte-Carlo-Bayesian Simulations Results

Date		Approach 1			Approach 2				Approach 3			
	Concentration (log_{10} gc/L)	Mean N	SD (σ)	99% CI	Mean N	SD (σ)	99% CI	γmin	Mean N	SD (σ)	99% CI	γmin
2021-10-28	6.095686665	3.08	1.64	(2.5, 3.66)	118	0	(nan, nan)	0.98409	96.24	11.99	(92.55, 99.93)	0.98105
2021-10-29	6.042265289	2.97	1.69	(2.41, 3.53)	118	0	(nan, nan)	0.98409	77.76	10.12	(74.55, 80.97)	0.97877
2021-11-01	6.543733258	6.56	2.58	(5.67, 7.46)	319.52	30.17	(312.98, 326.07)	0.98409	291.66	19.02	(282.99, 300.32)	0.98246
2021-11-02	6.677414049	8.55	3.26	(7.33, 9.78)	439.82	31.49	(433.71, 445.94)	0.98409	416.18	18	(406.29, 426.07)	0.98318
2021-11-03	6.60035422	7.32	2.92	(6.34, 8.3)	366.53	26.62	(361.06, 372.0)	0.98409	345.21	28.29	(325.74, 364.69)	0.98295
2021-11-04	6.634376451	7.08	2.41	(6.28, 7.89)	397.57	26.43	(391.79, 403.34)	0.98409	366.2	27.09	(350.6, 381.8)	0.98232
2021-11-05	6.411760103	4.74	2.76	(3.81, 5.68)	236.7	11.42	(234.66, 238.73)	0.98409	207.78	14.96	(200.36, 215.19)	0.98154
2021-11-08	6.704772193	9	2.99	(8.18, 9.82)	467.55	28.81	(461.57, 473.53)	0.98409	431.9	25.69	(417.1, 446.7)	0.98271
2021-11-09	6.617397578	7.26	3.02	(6.27, 8.25)	379.14	29.4	(373.11, 385.16)	0.98409	357.56	27.31	(345.49, 369.62)	0.98299
2021-11-10	6.408868289	4.77	2.63	(3.9, 5.64)	236.13	9.93	(234.37, 237.9)	0.98409	215.57	16.99	(209.32, 221.82)	0.98280
2021-11-11	6.705237804	8.95	3.02	(8.11, 9.8)	467.53	28.9	(461.51, 473.55)	0.98409	439.87	27.69	(425.0, 454.74)	0.98299
2021-11-12	6.355149235	5.61	2.73	(4.62, 6.59)	216.25	27.27	(209.19, 223.31)	0.98409	177.18	15.94	(170.04, 184.33)	0.98149
2021-11-15	6.57470023	6.84	2.15	(6.16, 7.51)	349.57	18.83	(345.87, 353.27)	0.98409	320.27	24.21	(308.04, 332.5)	0.98280
2021-11-16	6.332583696	5.33	2.41	(4.46, 6.2)	191.61	25.27	(186.01, 197.21)	0.98409	171.44	16.68	(164.73, 178.15)	0.98183
2021-11-17	6.887979993	13	4.04	(11.72, 14.28)	709.82	42.38	(700.82, 718.83)	0.98409	667.25	24.39	(651.54, 682.96)	0.98310
2021-11-18	6.913437655	14	4.93	(12.45, 15.55)	758.99	44.81	(749.12, 768.85)	0.98409	761	44.9	(717.29, 804.71)	0.98394
2021-11-19	6.780777155	9.84	3.32	(8.75, 10.92)	557.2	33.82	(550.18, 564.22)	0.98409	546.57	31.61	(529.59, 563.54)	0.98366
2021-11-23	7.920129608	122.72	13.06	(118.51, 126.92)	7044.7	89.58	(6911.45, 7177.88)	0.98409	7665	N/A	(nan, nan)	0.98454
2021-11-24	8.165636533	215.68	17.81	(209.99, 221.37)	N/A	N/A	N/A	0.98409	N/A	N/A	N/A	0.98459
2021-11-25	8.360148183	339.53	23.1	(331.35, 347.7)	N/A	N/A	N/A	0.98409	N/A	N/A	N/A	0.98460
2021-11-26	8.026997288	158.06	14.95	(152.77, 163.35)	N/A	N/A	N/A	0.98409	9803	0	(nan, nan)	0.98453
2021-11-29	8.163673562	212.84	14.05	(208.45, 217.23)	N/A	N/A	N/A	0.98409	13429	N/A	(nan, nan)	0.98465
2021-11-30	8.127924316	197.63	15.39	(192.08, 203.18)	N/A	N/A	N/A	0.98409	11680	N/A	(nan, nan)	0.98454
2021-12-01	7.867937018	108.08	13.24	(103.78, 112.38)	6735.8	234.6	(6683.59, 6787.99)	0.98409	6797	0	(nan, nan)	0.98444

2021-12-02	8.288582684	287.63	20.44	(281.2, 294.06)	N/A	N/A	N/A	0.98409	19893	N/A	(nan, nan)	0.98467
2021-12-06	5.832746531	2.34	0.96	(2.0, 2.68)	59	0	(nan, nan)	0.98409	44.22	7.56	(42.37, 46.06)	0.97686
2021-12-07	5.878903921	2.37	0.99	(2.0, 2.73)	59	0	(nan, nan)	0.98409	50.12	8.3	(47.89, 52.35)	0.97755
2021-12-08	6.084783304	2.82	1.48	(2.31, 3.33)	118	0	(nan, nan)	0.98409	96.54	12.96	(92.79, 100.3)	0.98101
2021-12-09	5.877837921	2.36	0.98	(2.0, 2.72)	59	0	(nan, nan)	0.98409	53.79	8.73	(51.84, 55.74)	0.97944
2021-12-10	5.946585394	2.67	1.34	(2.17, 3.17)	64.21	16.99	(56.7, 71.71)	0.98409	62.58	8.19	(60.33, 64.83)	0.97937
2021-12-13	5.480316771	2.05	0.38	(1.92, 2.17)	59	N/A	(nan, nan)	0.98409	19.1	5.02	(18.22, 19.97)	0.97515
2021-12-14	5.743052809	2.33	0.94	(2.0, 2.66)	59	0	(nan, nan)	0.98409	38.2	6.66	(36.75, 39.65)	0.97748
2021-12-15	6.059570718	2.98	1.59	(2.44, 3.52)	118	0	(nan, nan)	0.98409	84.39	11.99	(80.8, 87.98)	0.97981
2021-12-16	5.959609431	2.84	1.57	(2.27, 3.41)	79.42	28.62	(64.96, 93.88)	0.98409	67.95	9.79	(65.02, 70.88)	0.97994
2022-01-11	7.660710781	68.86	9.74	(65.6, 72.13)	4207	165.7	(4173.26, 4240.75)	0.98409	4159	224.6	(3869.79, 4448.21)	0.98429
2022-01-12	7.179379374	23.22	5.48	(21.5, 24.95)	1388.2	64.78	(1373.32, 1403.05)	0.98409	1375.1	64.53	(1319.7, 1430.52)	0.98376
2022-01-13	7.320294169	32.21	5.99	(30.24, 34.19)	1922.6	86.84	(1904.57, 1940.62)	0.98409	1926	0	(nan, nan)	0.98419
2022-01-18	5.860609437	2.38	1	(2.0, 2.75)	59	0	(nan, nan)	0.98409	52.94	8.03	(51.1, 54.79)	0.97949
2022-01-19	5.952262725	2.76	1.4	(2.23, 3.29)	69.73	23.11	(59.37, 80.09)	0.98409	67.54	8.76	(65.43, 69.64)	0.98030
2022-01-20	6.183275496	3.33	1.83	(2.69, 3.96)	130.29	24.13	(122.97, 137.62)	0.98409	121.16	13.49	(117.12, 125.2)	0.98133
2022-01-21	6.669637867	8.59	3.01	(7.57, 9.6)	430.18	28.2	(424.17, 436.2)	0.98409	389	24.98	(376.84, 401.16)	0.98217
2022-01-26	6.742373112	8.82	3.6	(7.52, 10.12)	507.92	31.64	(501.38, 514.47)	0.98409	489.77	37.45	(469.21, 510.34)	0.98334
2022-01-27	6.192375959	3.23	1.86	(2.59, 3.87)	139.2	28.53	(130.02, 148.39)	0.98409	119.1	12.37	(115.08, 123.11)	0.98071
2022-01-28	6.026012682	2.96	1.7	(2.38, 3.55)	118	0	(nan, nan)	0.98409	83.84	10.21	(81.06, 86.63)	0.98086
2022-02-01	6.008409905	3.02	1.67	(2.44, 3.6)	118	0	(nan, nan)	0.98409	76.03	9.54	(73.46, 78.61)	0.98079
2022-02-02	6.14129919	3.59	1.91	(2.9, 4.28)	118	0	(nan, nan)	0.98409	95.37	11.44	(91.36, 99.38)	0.97878
2022-02-03	6.187950999	3.31	1.87	(2.66, 3.96)	133.92	26.4	(125.35, 142.49)	0.98409	119.86	10.96	(116.36, 123.36)	0.98135
2022-02-04	7.50104086	46.14	7.23	(43.92, 48.37)	2905	128.6	(2878.53, 2931.41)	0.98409	2920	N/A	(nan, nan)	0.98419
2022-02-08	6.413984438	5.13	2.92	(4.1, 6.16)	238.09	13.18	(235.75, 240.43)	0.98409	213.21	17.87	(206.56, 219.85)	0.98205
2022-02-09	6.20484858	3.3	1.89	(2.63, 3.97)	156.23	28.38	(147.55, 164.9)	0.98409	127.05	13.15	(122.53, 131.58)	0.98124
2022-02-11	7.441716977	41.26	7.42	(38.73, 43.79)	2543.5	110.4	(2521.73, 2565.34)	0.98409	2500.2	115.5	(2378.76, 2621.57)	0.98435
2022-02-15	6.961129841	14.47	3.86	(13.3, 15.63)	847.23	49.4	(835.94, 858.52)	0.98409	837.11	54.44	(790.37, 883.85)	0.98393
2022-02-17	6.333706454	5.68	2.49	(4.8, 6.56)	193.69	26.35	(187.93, 199.45)	0.98409	171.03	13.74	(165.28, 176.77)	0.98142
2022-02-18	6.207966268	3.33	1.84	(2.72, 3.94)	158.19	27.7	(149.6, 166.78)	0.98409	132.2	13.07	(127.7, 136.7)	0.98174

2022-02-22	6.686971498	8.28	3.16	(7.21, 9.34)	446.32	30.73	(439.92, 452.72)	0.98409	424.79	25.39	(412.42, 437.15)	0.98327
2022-02-23	6.161567589	3.6	1.99	(2.88, 4.31)	118.4	4.83	(117.38, 119.42)	0.98409	107.75	12.02	(104.26, 111.23)	0.98023
2022-02-24	6.067950666	2.83	1.48	(2.34, 3.33)	118	0	(nan, nan)	0.98409	88.48	9.22	(85.66, 91.3)	0.98048
2022-02-25	6.33674779	5.78	2.61	(4.88, 6.69)	195.61	27.18	(189.56, 201.66)	0.98409	179.15	14.63	(173.6, 184.71)	0.98218
2022-03-01	6.039836927	2.97	1.69	(2.41, 3.53)	118	0	(nan, nan)	0.98409	86.95	12.2	(83.39, 90.51)	0.98114
2022-03-02	6.467325839	5.92	2.22	(5.21, 6.64)	274.33	27.92	(268.14, 280.52)	0.98409	247.2	17.92	(237.97, 256.43)	0.98258
2022-03-03	5.835449539	2.33	0.95	(2.0, 2.67)	59	0	(nan, nan)	0.98409	48.21	7.5	(46.41, 50.0)	0.97834
2022-03-04	6.217803143	3.45	1.81	(2.86, 4.04)	167.05	22.23	(160.76, 173.33)	0.98409	133.93	11.61	(129.97, 137.89)	0.98175
2022-03-08	6.585373442	6.89	2.28	(6.1, 7.68)	355.31	18.09	(351.83, 358.79)	0.98409	313.32	26.57	(298.73, 327.91)	0.98233
2022-03-09	6.283834678	4.59	2.19	(3.86, 5.33)	177	0	(nan, nan)	0.98409	147.18	15.63	(140.81, 153.54)	0.98095
2022-03-10	5.955716563	2.74	1.39	(2.22, 3.27)	71.64	24.65	(59.64, 83.64)	0.98409	63.22	8.67	(60.88, 65.56)	0.97872
2022-03-11	5.870871852	2.31	0.93	(1.97, 2.66)	59	0	(nan, nan)	0.98409	45.23	8.42	(42.97, 47.49)	0.97553
2022-03-15	5.692541382	2.15	0.66	(1.93, 2.38)	59	0	(nan, nan)	0.98409	33.64	6.28	(32.36, 34.92)	0.97754
2022-03-16	5.993856316	2.9	1.59	(2.32, 3.48)	115.54	12.04	(109.21, 121.87)	0.98409	75.42	9.62	(72.88, 77.96)	0.98049
2022-03-17	5.702143166	2.2	0.75	(1.95, 2.45)	59	0	(nan, nan)	0.98409	33.04	6.35	(31.65, 34.42)	0.97680
2022-03-18	5.936738161	2.6	1.3	(2.12, 3.09)	62.77	14.58	(57.29, 68.24)	0.98409	63.64	7.77	(61.66, 65.62)	0.97896
2022-03-22	6.327815086	5.3	2.08	(4.57, 6.04)	186.92	21.91	(182.34, 191.5)	0.98409	170.6	17.59	(163.99, 177.2)	0.98193
2022-03-23	6.128315202	3.3	1.83	(2.66, 3.94)	118.24	3.73	(117.63, 118.84)	0.98409	100.7	12.19	(96.78, 104.63)	0.98008
2022-03-24	6.014436484	2.95	1.63	(2.4, 3.5)	118	0	(nan, nan)	0.98409	71.9	9.08	(69.52, 74.27)	0.97881
2022-03-25	5.983048604	2.86	1.58	(2.28, 3.44)	108.68	22.1	(95.62, 121.75)	0.98409	72.57	9.08	(70.12, 75.02)	0.98008
2022-03-29	7.000353328	16.49	4.82	(15.0, 17.99)	922.74	50.23	(911.96, 933.53)	0.98409	870.91	60.51	(823.91, 917.91)	0.98353
2022-03-30	7.583745922	57.54	8.77	(54.88, 60.2)	3523.4	149.6	(3492.34, 3554.45)	0.98409	3729	160	(3522.89, 3935.11)	0.98440
2022-03-31	7.075609199	18.1	4.79	(16.36, 19.84)	1094	61.69	(1081.3, 1106.67)	0.98409	1062.3	61.49	(1009.54, 1115.13)	0.98355
2022-04-01	6.495219492	6.09	2.18	(5.35, 6.82)	291.87	14.37	(289.0, 294.74)	0.98409	236.95	24.06	(223.43, 250.48)	0.98121
2022-04-05	7.318342431	31.43	6.51	(29.15, 33.71)	1924.1	83.54	(1906.94, 1941.28)	0.98409	1890.8	101.9	(1759.48, 2022.02)	0.98404
2022-04-06	7.3418735	34.61	7.13	(32.37, 36.85)	2030.2	90.52	(2011.92, 2048.55)	0.98409	2062	129.9	(1868.81, 2255.19)	0.98441
2022-04-07	7.491684776	48.37	7.67	(46.09, 50.65)	2850	133	(2823.2, 2876.88)	0.98409	2884.5	64.91	(2816.24, 2952.76)	0.98451
2022-04-08	7.624613888	64.76	10.43	(61.14, 68.39)	3910.3	155.7	(3878.93, 3941.75)	0.98409	3881	0	(nan, nan)	0.98455
2022-04-12	7.051994001	17.47	4.86	(15.82, 19.13)	1040.8	49.61	(1030.62, 1051.02)	0.98409	980.6	66.49	(926.44, 1034.76)	0.98301
2022-04-13	7.018707388	16.57	4.96	(14.89, 18.25)	962.15	47.78	(952.54, 971.76)	0.98409	944	55.19	(885.96, 1002.04)	0.98403

2022-04-14	6.901444837	13.74	4.46	(12.28, 15.2)	736.75	37.99	(728.65, 744.85)	0.98409	698.38	36.57	(679.15, 717.6)	0.98351
2022-04-20	6.396091153	4.77	2.14	(4.05, 5.48)	234.67	7.65	(233.18, 236.16)	0.98409	201.45	16.57	(194.87, 208.04)	0.98173
2022-04-21	7.21637719	24.6	5.59	(22.88, 26.32)	1512.3	70.69	(1497.92, 1526.71)	0.98409	1541.2	56.7	(1495.02, 1587.38)	0.98418
2022-04-22	6.662452416	7.85	2.82	(6.96, 8.74)	423.8	25.51	(418.52, 429.08)	0.98409	377.17	25.99	(363.22, 391.13)	0.98208
2022-04-26	7.282676136	28.39	5.33	(26.64, 30.13)	1769.7	91.57	(1751.08, 1788.25)	0.98409	1962	N/A	(nan, nan)	0.98404
2022-04-27	7.201245552	25.89	6.49	(23.66, 28.13)	1459.1	72.21	(1442.35, 1475.76)	0.98409	1464.2	57.63	(1397.81, 1530.59)	0.98402
2022-04-28	7.458951693	43.7	7.67	(41.08, 46.32)	2620.9	114.5	(2597.66, 2644.14)	0.98409	2601.2	120.4	(2474.61, 2727.73)	0.98421
2022-04-29	6.610809117	7.61	2.98	(6.58, 8.63)	379.22	29.42	(372.91, 385.54)	0.98409	343.41	26.67	(330.66, 356.17)	0.98262
2022-05-04	7.249710194	27.77	5.27	(26.07, 29.46)	1627.8	76.71	(1611.26, 1644.42)	0.98409	1614.3	80.66	(1540.79, 1687.71)	0.98395
2022-05-05	7.576258203	55.92	9.77	(52.8, 59.04)	3473.7	145.5	(3442.93, 3504.56)	0.98409	3472	272.9	(2974.86, 3969.14)	0.98417
2022-05-06	7.295911277	30	6.17	(28.07, 31.93)	1828.2	77.76	(1810.52, 1845.93)	0.98409	1745.3	127.1	(1581.57, 1908.93)	0.98394
2022-05-10	7.511607776	50.15	8.39	(47.4, 52.89)	2992.2	122.7	(2968.08, 3016.28)	0.98409	2992	166	(2745.13, 3238.87)	0.98427
2022-05-11	6.677233888	8.55	3.26	(7.33, 9.78)	439.62	31.52	(433.45, 445.8)	0.98409	416.75	26.18	(402.99, 430.51)	0.98303
2022-05-12	7.378493428	36.88	7.5	(34.39, 39.38)	2218.2	93.18	(2197.77, 2238.63)	0.98409	2243	71.01	(2137.39, 2348.61)	0.98417
2022-05-13	7.47201235	45.18	7.1	(42.86, 47.5)	2718.6	110.9	(2696.13, 2741.03)	0.98409	2731.3	114.5	(2626.95, 2835.55)	0.98424
2022-05-17	7.302790195	30.27	6.4	(27.97, 32.58)	1839.7	81.69	(1823.31, 1856.17)	0.98409	1832.8	77.54	(1751.3, 1914.37)	0.98400
2022-05-18	7.471792676	45.05	7.02	(42.72, 47.38)	2713.1	110.8	(2689.83, 2736.43)	0.98409	2882	N/A	(nan, nan)	0.98435
2022-05-19	7.523061935	49.03	7.63	(46.71, 51.34)	3078.3	125.9	(3053.61, 3102.94)	0.98409	3072	241.8	(2631.53, 3512.47)	0.98425
2022-05-20	7.37084883	35.83	6.59	(33.74, 37.92)	2163.4	95.59	(2142.37, 2184.44)	0.98409	2144	90.33	(2049.01, 2238.99)	0.98423

Results of the Monte-Carlo-Bayesian simulations for the detected crAssphage concentration in wastewater: the detected concentration, the mean of the estimates distribution, the standard deviation (SD) of the distribution and the 99% confidence interval (CI 99%). The γ_{min} for the approach 2 and 3 stands for the assumed minimum viral losses, which depends on the approach type and the measured concentration.

Table S7. The number of influenza A cases in Wales and the detection of the virus in the wastewater sampled at Ysbyty Gwynedd.

Week by Start Date	Total Influenza A Cases	Total Specimens Tested	Influenza A Virus Detected at YG
04/10/2021	5	10194	N/A
11/10/2021	0	9412	N/A
18/10/2021	5	8311	N/A
25/10/2021	2	7537	No
01/11/2021	3	6990	No
08/11/2021	2	8753	Yes
15/11/2021	4	8678	Yes
22/11/2021	4	8819	Yes
29/11/2021	9	9094	Yes
06/12/2021	14	8486	Yes
13/12/2021	11	8416	Yes
20/12/2021	8	7468	No
27/12/2021	6	8134	No
03/01/2022	19	9001	No
10/01/2022	4	7451	No
17/01/2022	7	6049	No
24/01/2022	4	6391	No
31/01/2022	6	5593	No
07/02/2022	2	5865	No
14/02/2022	20	5997	No
21/02/2022	12	5497	No
28/02/2022	17	5879	No
07/03/2022	27	4188	No
14/03/2022	44	5864	No
21/03/2022	43	5899	No
28/03/2022	46	4509	No
04/04/2022	70	3946	No
11/04/2022	38	3180	N/A
18/04/2022	40	4180	N/A
25/04/2022	54	3114	N/A
02/05/2022	53	2566	N/A
09/05/2022	43	2868	N/A
16/05/2022	24	3175	N/A
23/05/2022	17	3056	N/A
30/05/2022	11	2318	N/A
06/06/2022	4	3135	N/A
13/06/2022	13	3701	N/A

The total influenza A cases numbers are based on the virus detection in specimens submitted for virological testing for hospital patients and non-sentinel GPs and in specimens from hospital patients submitted for RSV, Influenza A, and SARS-CoV-2 testing only in Wales. The table also presents whether the virus was detected in the wastewater sampled at Ysbyty Gwynedd (YG), yes standing for "detected", "no" for not detected and a zero concentration and "N/A" is for dates when a wastewater sample was not available.

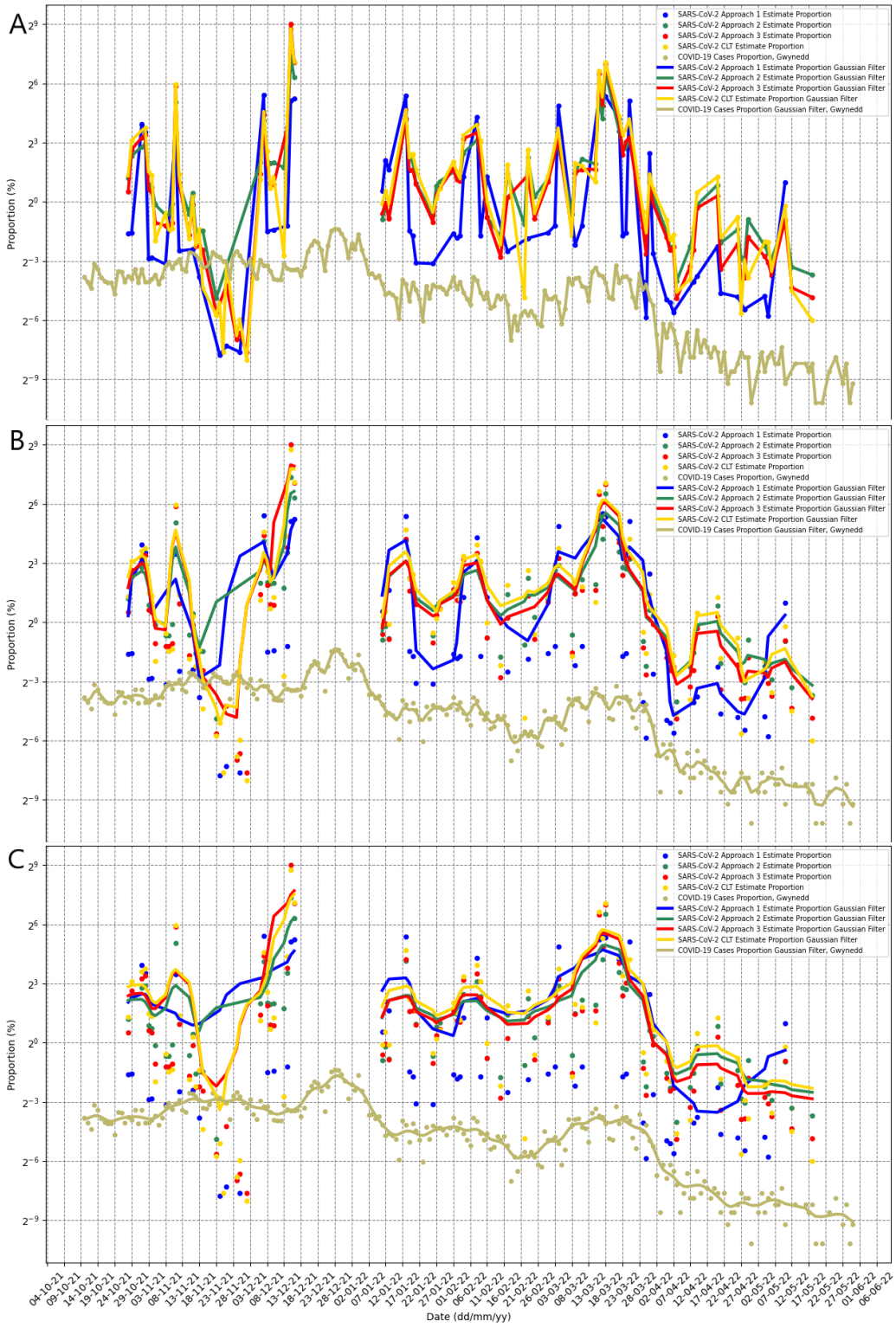

Figure S13. Comparison of SARS-CoV-2 Infected Population Estimate at Ysbyty Gwynedd Proportion and the Reported Number Proportion in Gwynedd. The proportion of SARS-CoV-2 estimates is compared against the proportion of SARS-CoV-2 cases in Gwynedd. The estimate proportion is calculated by dividing the SARS-CoV-2 estimate by the crAssphage estimate for each simulation method separately and multiplied by 100 (%). The proportion of SARS-CoV-2 cases in Gwynedd is calculated by dividing the SARS-CoV-2 reported cases in Gwynedd by the total Gwynedd population and multiplied by 100 (%). The obtained values are filtered with a Gaussian filter of varying sigma value: 0.5 (panel A), 1 (panel B) and 2 (panel C).

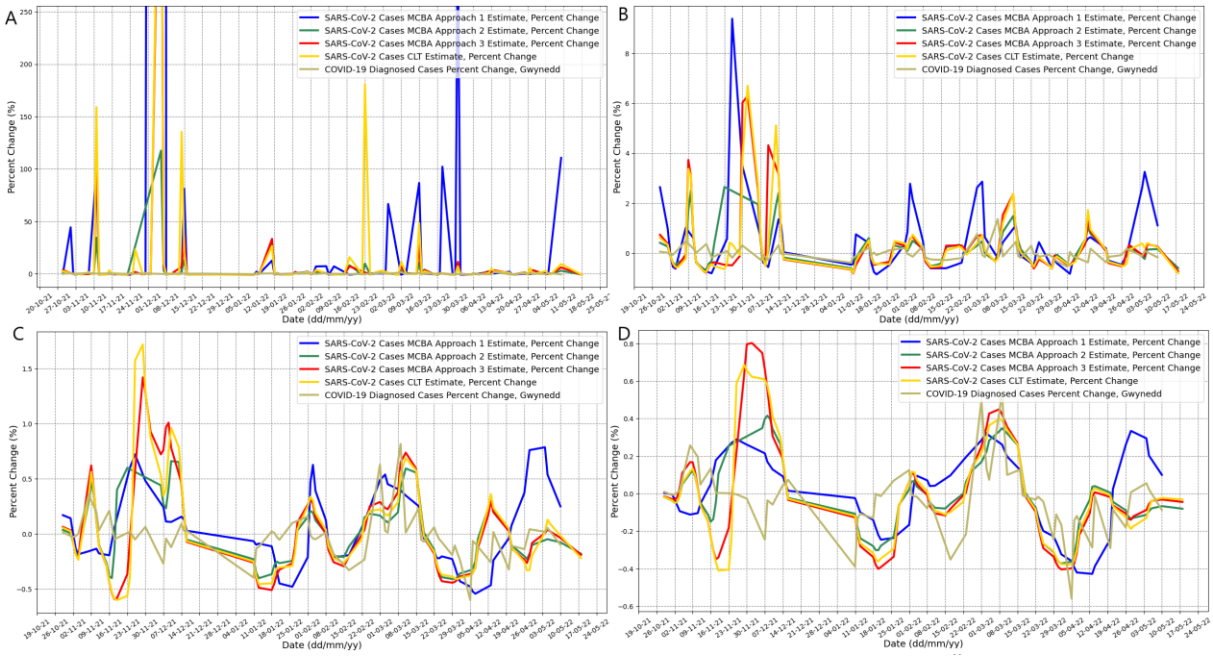

Figure S14. Comparison of Percent Changes of the Proportion of SARS-CoV-2 Infected Population Estimate and the Proportion of the Reported Cases Number. Initially the proportions are calculated by dividing the SARS-CoV-2 population estimates by the crAssphage population estimates in the case of the Monte-Carlo-Bayesian approach (MCBA) 1, 2, 3 and the equation model following the Central Limit Theorem premise (CLT) and separately by dividing the SARS-CoV-2 reported cases by the total population of Gwynedd. The proportions are computed with a Gaussian filter of varying sigma argument ***(only once before percent changes calculations)****: 0 (panel A), 1 (panel B), 2 (panel C) and 3 (panel D). Then the percent changes were calculated for each proportion separately.*

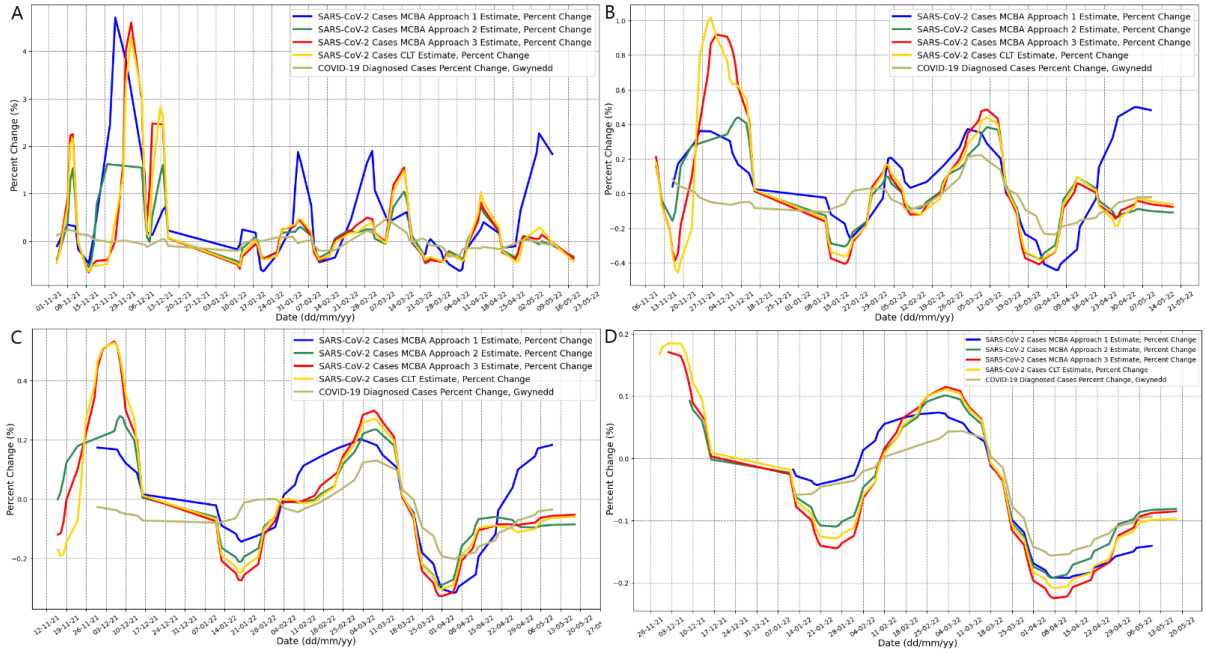

Figure S15. Comparison of Percent Changes of the Proportion of SARS-CoV-2 Infected Population Estimate and the Proportion of the Reported Cases Number. Initially the proportions are calculated by dividing the SARS-CoV-2 population estimates by the crAssphage population estimates in the case of the Monte-Carlo-Bayesian approach (MCBA) 1, 2, 3 and the equation model following the Central Limit Theorem premise (CLT) and separately by dividing the SARS-CoV-2 reported cases by the total population of Gwynedd. Then the percent changes were calculated for each proportion separately. The proportions are computed with a Gaussian filter of varying sigma argument: 1 (panel A), 2 (panel B), 3 (panel C) and 5 (panel D) ***before and after the conversion to percent changes****.*

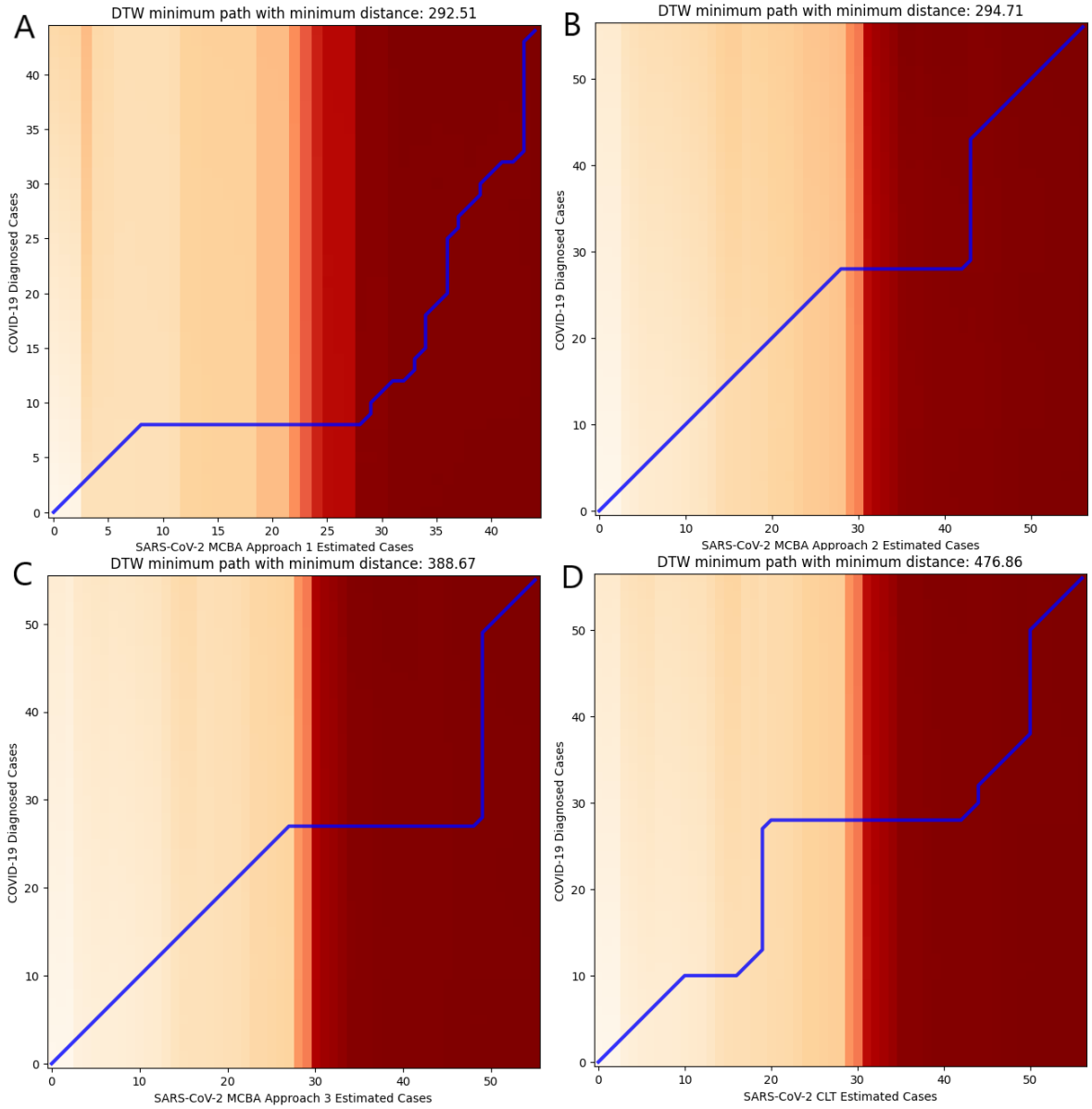

Figure S16. Dynamic Time Warping of SARS-CoV-2 Infected Population Estimates Proportion against the Reported Cases Numbers Proportion. The SARS-CoV-2 infected population estimates proportion at Ysbyty Gwynedd were compared against the SARS-CoV-2 reported population proportion in Gwynedd by Dynamic Time Warping (DTW), hence allowing for measuring the similarity between these two time series. The estimates proportion were calculated by dividing the SARS-CoV-2 infected population estimate by crAssphage infected population estimate. The reported cases proportion was calculated by dividing the SARS-CoV-2 cases in Gwynedd by the total population of Gwynedd. The x and y axis represent the matrix cost, where the minimum path with minimum distance is computed as the sum of the absolute differences. A match is described by a diagonal warping path, hence indicating temporal alignment between the two series. Each of the Monte-Carlo-Bayesian approach (MCBA) and the equation model following the Central Limit Theorem premise (CLT) is compared independently against the COVID-19 reported cases in Gwynedd. The series contain only the number of cases for the matching dates, whereas the non-matching dates were excluded from this comparison. A left-hand or a right-hand deviation from the diagonal indicates a leading or lagging relationship. For example, in panel A, it can be observed a significant right-hand shift of the diagonal towards the "SARS-CoV-2 MCBA Approach 1" axis, this indicating that the

proportion of estimated cases has increased sooner or has increased proportionally more than the proportion of reported cases in Gwynedd.

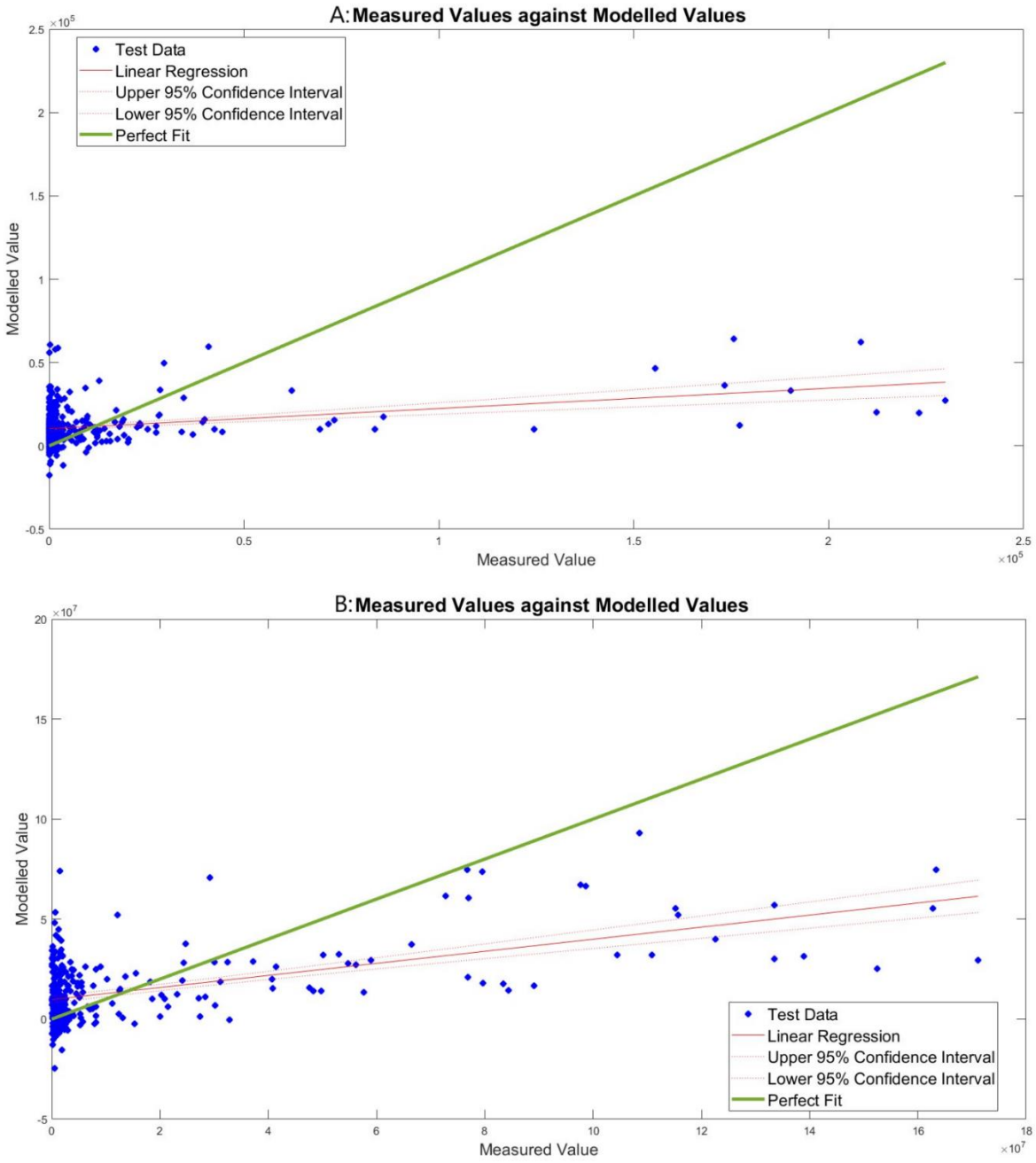

Figure S17. Correlation of Measured Virus Concentrations against Modelled Values based on the FIS pH and Ammonium Impact. The measured virus concentration is compared against the modelled virus concentration, the correlation being described by the linear regression (the R, R^2 and RMSE can be found in Table 6). A perfect fit is described by a diagonal line where the test data has identical values on both axes. The model is a Sugeno adaptive neuro-fuzzy interference system (ANFIS) where pH and ammonium are the input parameters and SARS-CoV-2 (A) or crAssphage (B) are the output parameters. The initial fuzzy interference system (FIS) is trained with subtractive clustering followed by tuning using input/output training data, which can also be described as additional fitting of the model to the training data. The figure depicts the final modelled data.

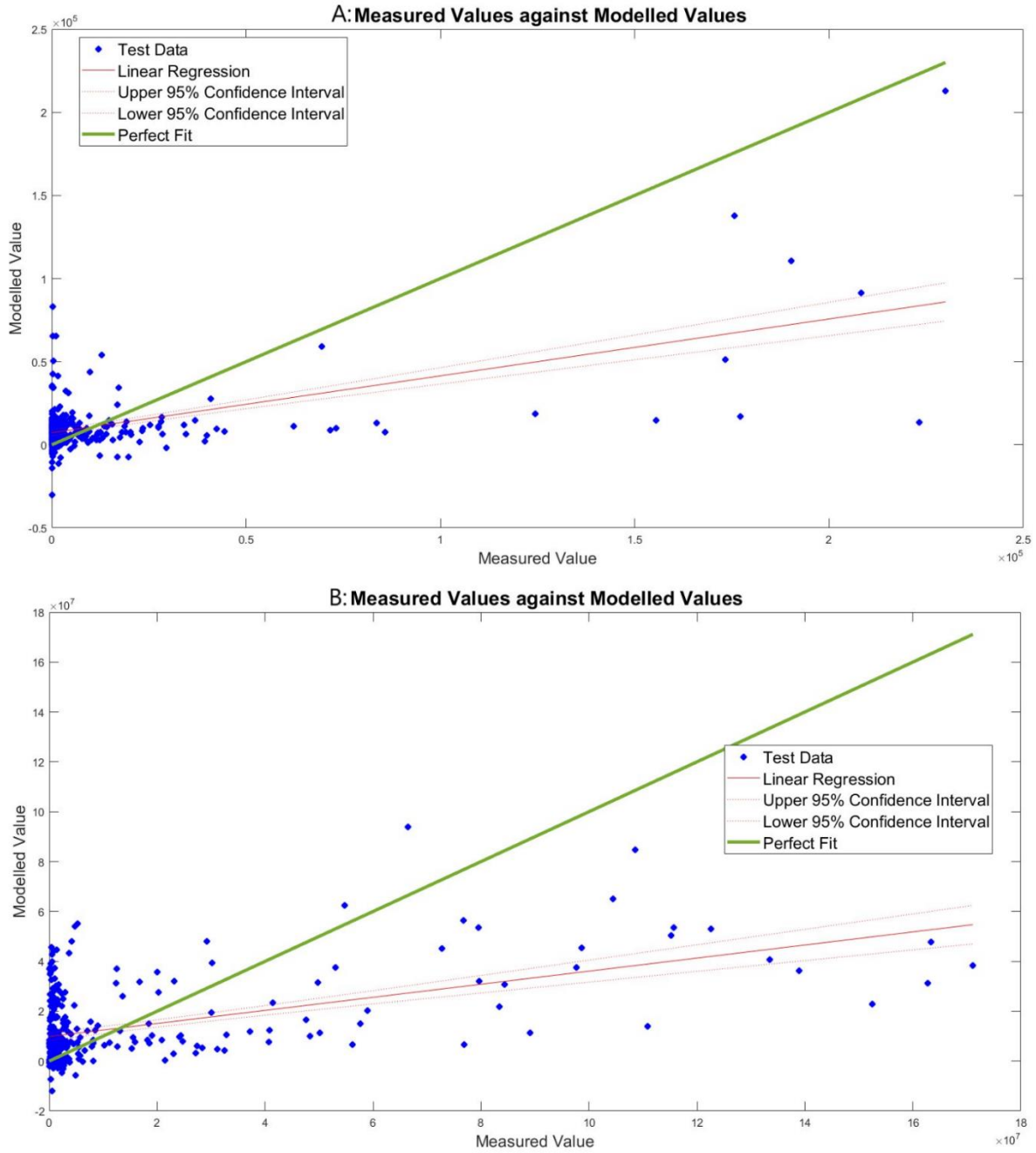

Figure S18. Correlation of Measured Virus Concentrations against Modelled Values based on the FIS pH and Phosphate Impact. The measured virus concentration is compared against the modelled virus concentration, the correlation being described by the linear regression (the R, R^2 and RMSE can be found in Table 6). A perfect fit is described by a diagonal line where the test data has identical values on both axes. The model is a Sugeno adaptive neuro-fuzzy interference system (ANFIS) where pH and phosphate are the input parameters and SARS-CoV-2 (A) or crAssphage (B) are the output parameters. The initial fuzzy interference system (FIS) is trained with subtractive clustering followed by tuning using input/output training data, which can also be described as additional fitting of the model to the training data. The figure depicts the final modelled data.

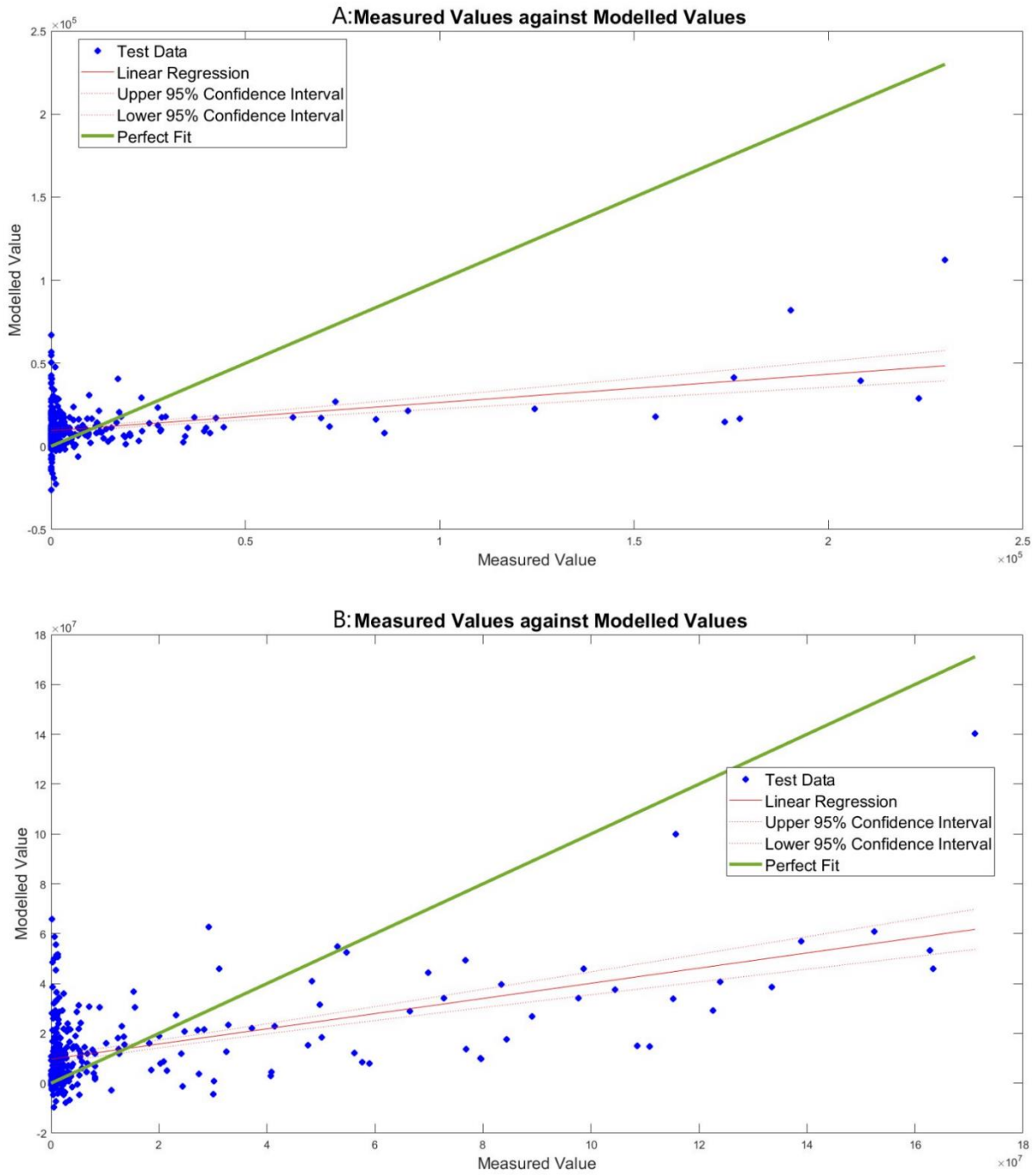

Figure S19. Correlation of Measured Virus Concentrations against Modelled Values based on the FIS Ammonium and Phosphate Impact. The measured virus concentration is compared against the modelled virus concentration, the correlation being described by the linear regression (the R, R^2 and RMSE can be found in Table 6). A perfect fit is described by a diagonal line where the test data has identical values on both axes. The model is a Sugeno adaptive neuro-fuzzy interference system (ANFIS) where ammonium and phosphate are the input parameters and SARS-CoV-2 (A) or crAssphage (B) are the output parameters. The initial fuzzy interference system (FIS) is trained with subtractive clustering followed by tuning using input/output training data, which can also be described as additional fitting of the model to the training data. The figure depicts the final modelled data.

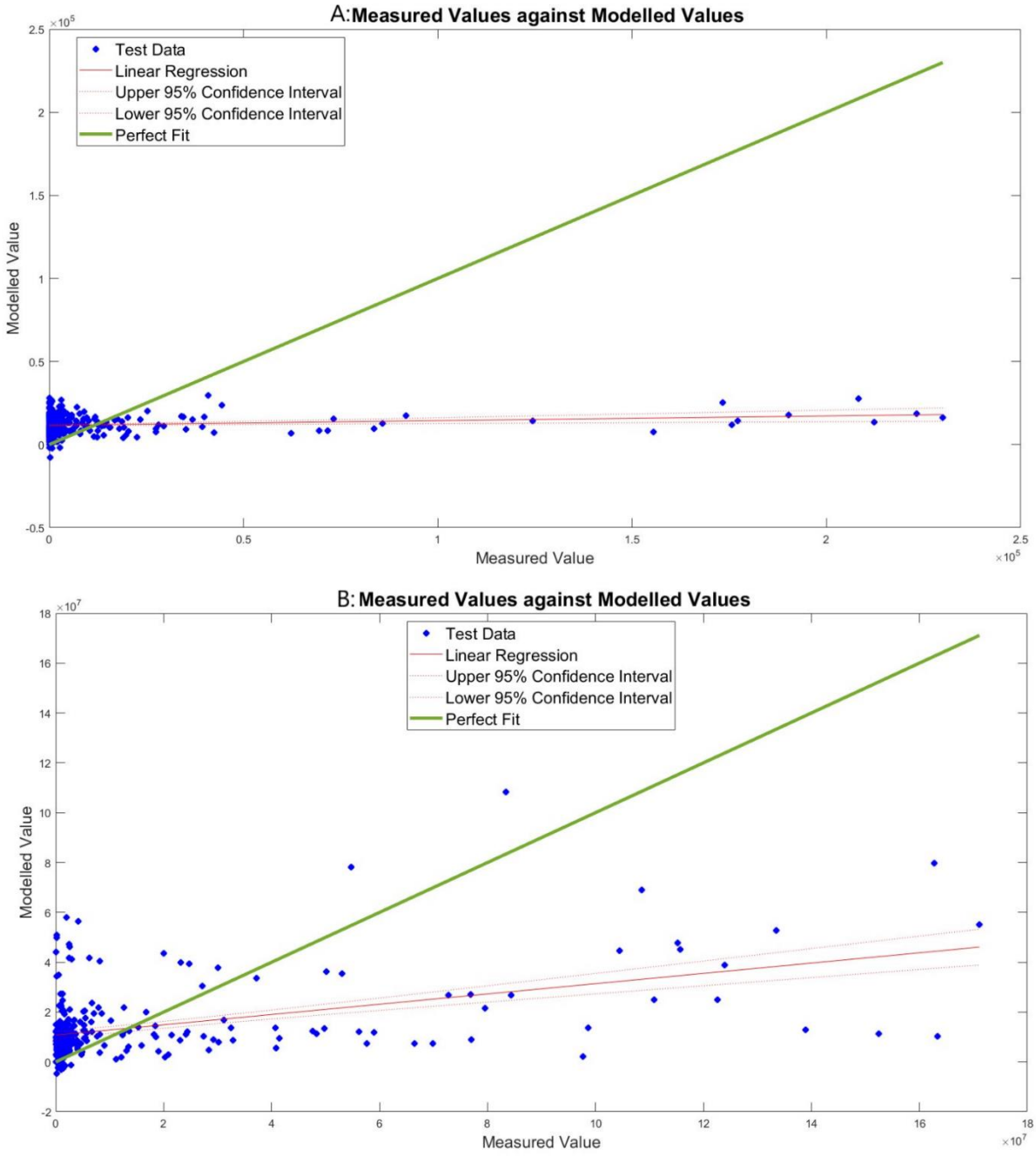

Figure S20. Correlation of Measured Virus Concentrations against Modelled Values based on the FIS Electrical conductivity and Turbidity Impact. The measured virus concentration is compared against the modelled virus concentration, the correlation being described by the linear regression (the R, R^2 and RMSE can be found in Table 6). A perfect fit is described by a diagonal line where the test data has identical values on both axes. The model is a Sugeno adaptive neuro-fuzzy interference system (ANFIS) where electrical conductivity and turbidity are the input parameters and SARS-CoV-2 (A) or crAssphage (B) are the output parameters. The initial fuzzy interference system (FIS) is trained with subtractive clustering followed by tuning using input/output training data, which can also be described as additional fitting of the model to the training data. The figure depicts the final modelled data.

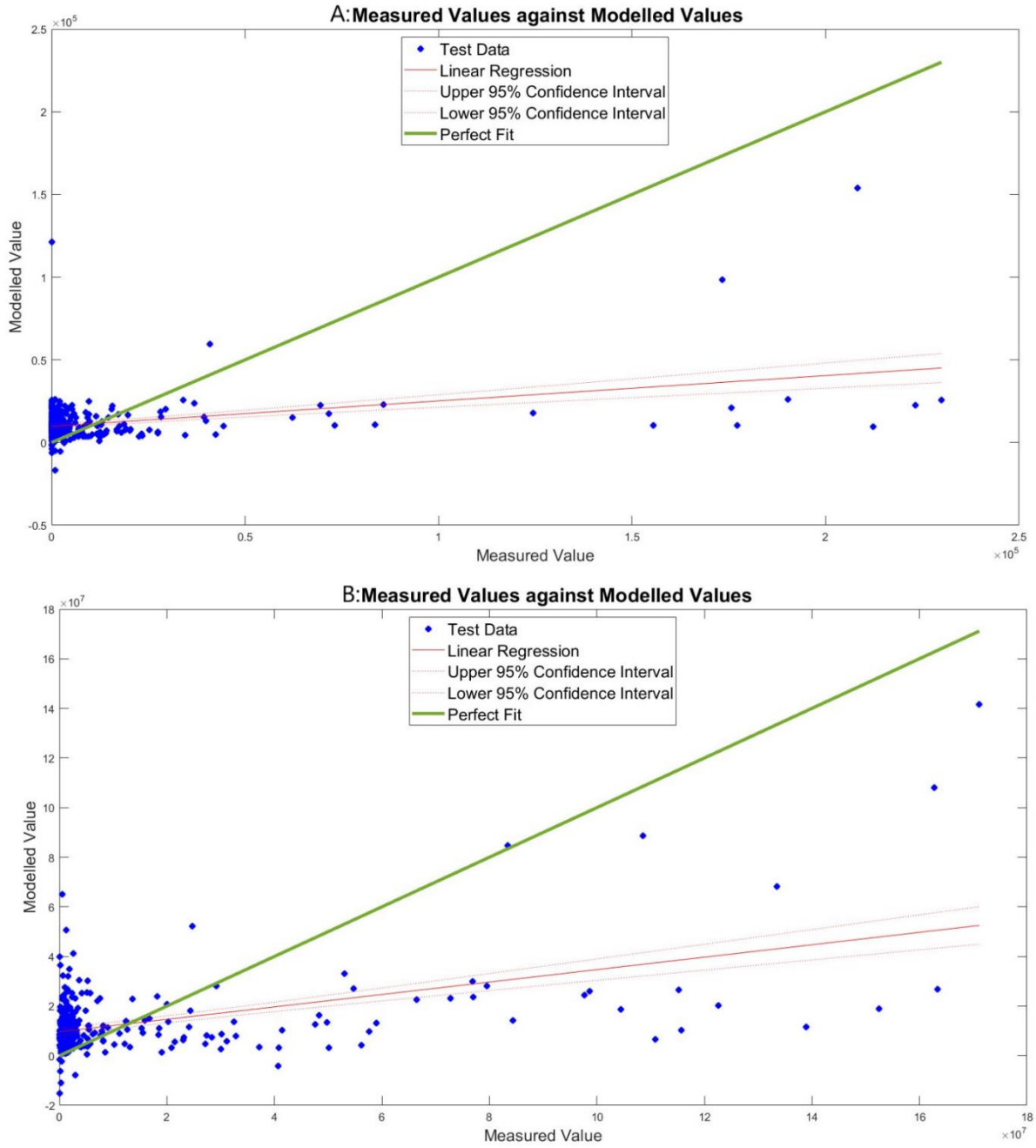

Figure S21. Correlation of Measured Virus Concentrations against Modelled Values based on the FIS Electrical conductivity and pH Impact. The measured virus concentration is compared against the modelled virus concentration, the correlation being described by the linear regression (the R, R^2 and RMSE can be found in Table 6). A perfect fit is described by a diagonal line where the test data has identical values on both axes. The model is a Sugeno adaptive neuro-fuzzy interference system (ANFIS) where electrical conductivity and pH are the input parameters and SARS-CoV-2 (A) or crAssphage (B) are the output parameters. The initial fuzzy interference system (FIS) is trained with subtractive clustering followed by tuning using input/output training data, which can also be described as additional fitting of the model to the training data. The figure depicts the final modelled data.

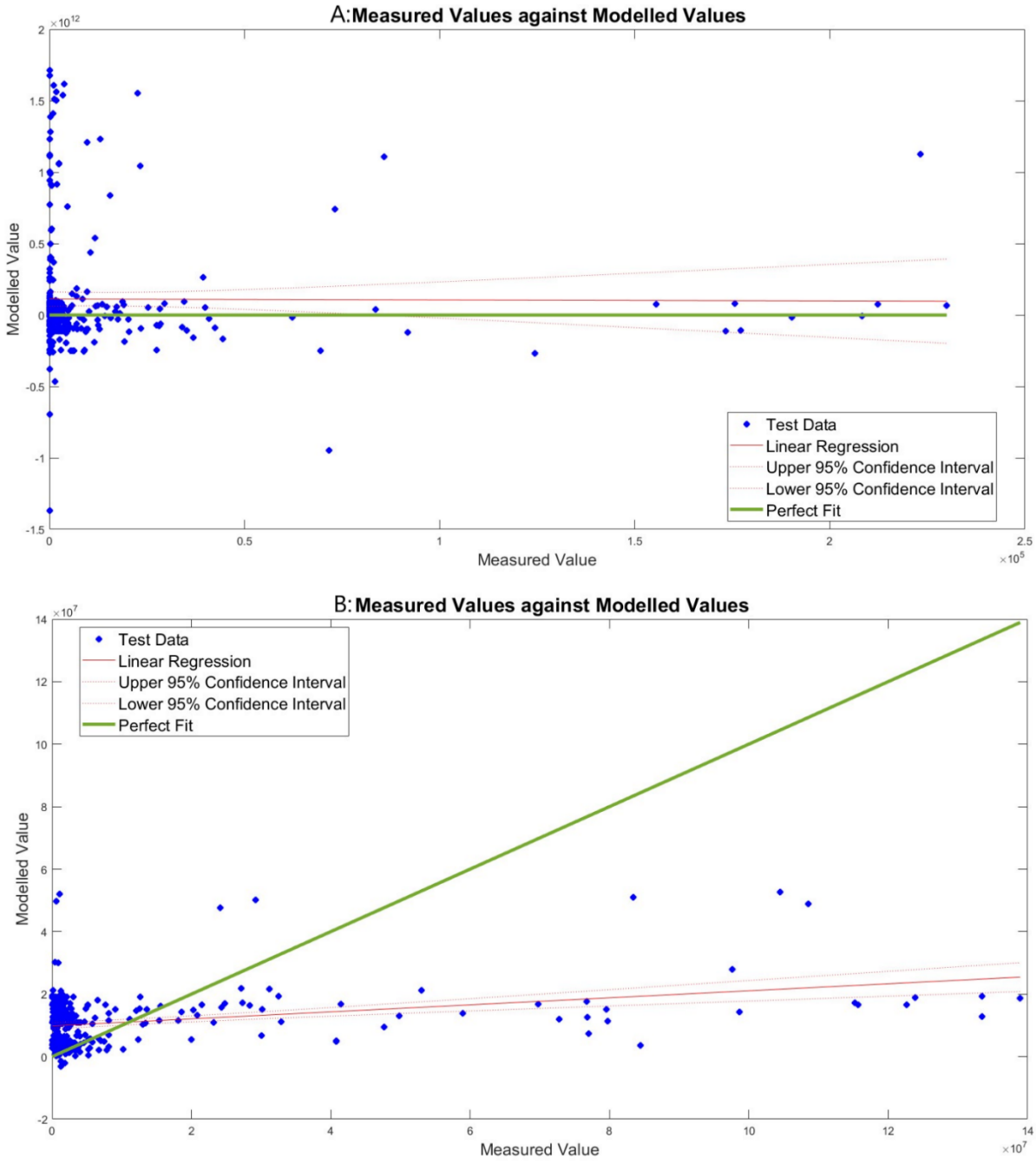

Figure S22. Correlation of Measured SARS-CoV-2 (or CrAssphage) Concentrations against Modelled Values based on the CrAssphage (or SARS-CoV-2) and Ammonium Impact. The measured virus concentration is compared against the modelled virus concentration, the correlation being described by the linear regression (the R, R^2 and RMSE can be found in Table 6). A perfect fit is described by a diagonal line where the test data has identical values on both axes. The model is a Sugeno adaptive neuro-fuzzy interference system (ANFIS) where crAssphage for panel A (or SARS-CoV-2 for panel B) and ammonium are the input parameters and SARS-CoV-2 (A) or crAssphage (B) are the output parameters. The initial fuzzy interference system (FIS) is trained with subtractive clustering followed by tuning using input/output training data, which can also be described as additional fitting of the model to the training data. The figure depicts the final modelled data.

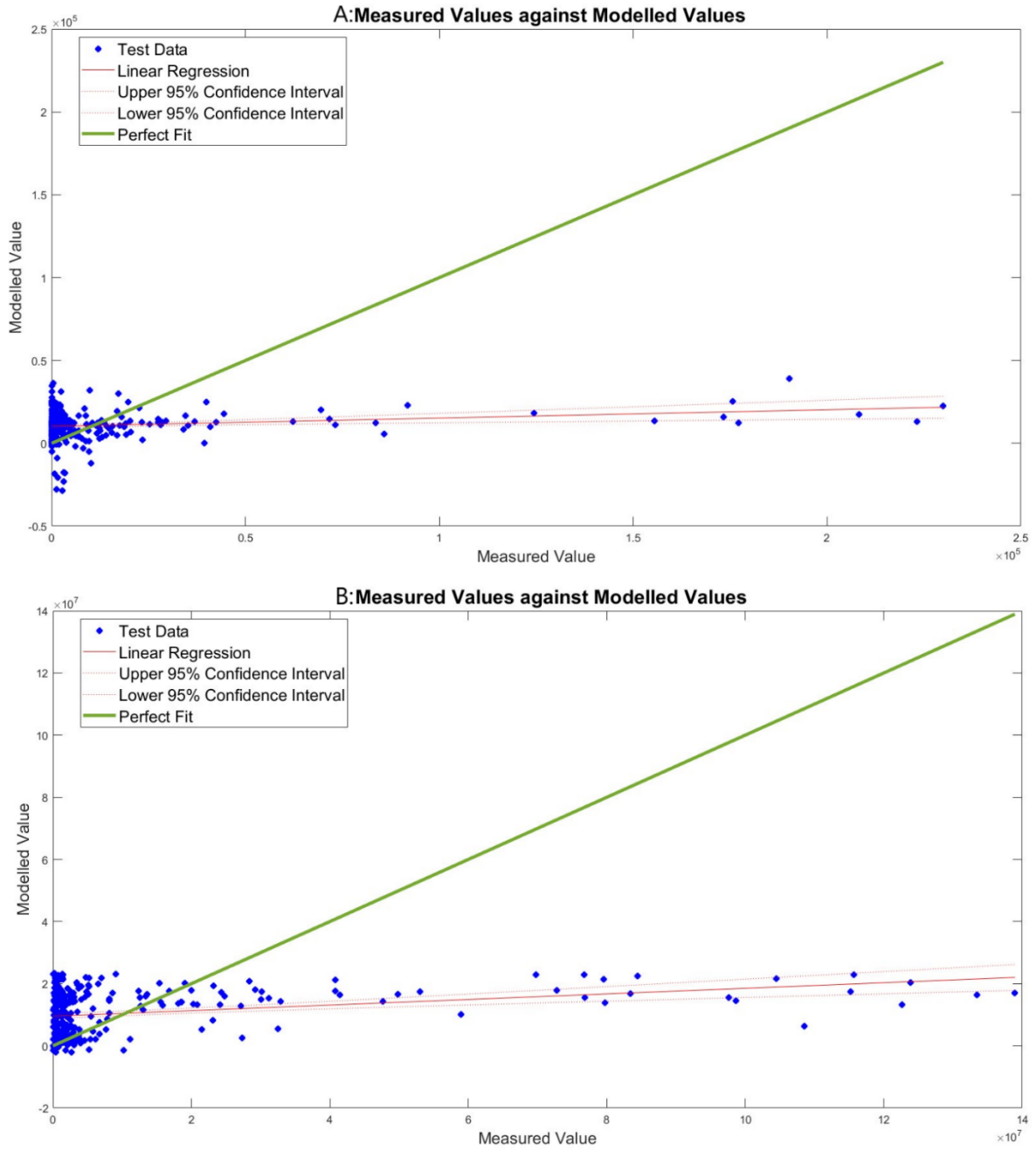

Figure S23. Correlation of Measured SARS-CoV-2 (or CrAssphage) Concentrations against Modelled Values based on the CrAssphage (or SARS-CoV-2) and Phosphate Impact. The measured virus concentration is compared against the modelled virus concentration, the correlation being described by the linear regression (the R, R^2 and RMSE can be found in Table 6). A perfect fit is described by a diagonal line where the test data has identical values on both axes. The model is a Sugeno adaptive neuro-fuzzy interference system (ANFIS) where crAssphage for panel A (or SARS-CoV-2 for panel B) and phosphate are the input parameters and SARS-CoV-2 (A) or crAssphage (B) are the output parameters. The initial fuzzy interference system (FIS) is trained with subtractive clustering followed by tuning using input/output training data, which can also be described as additional fitting of the model to the training data. The figure depicts the final modelled data.

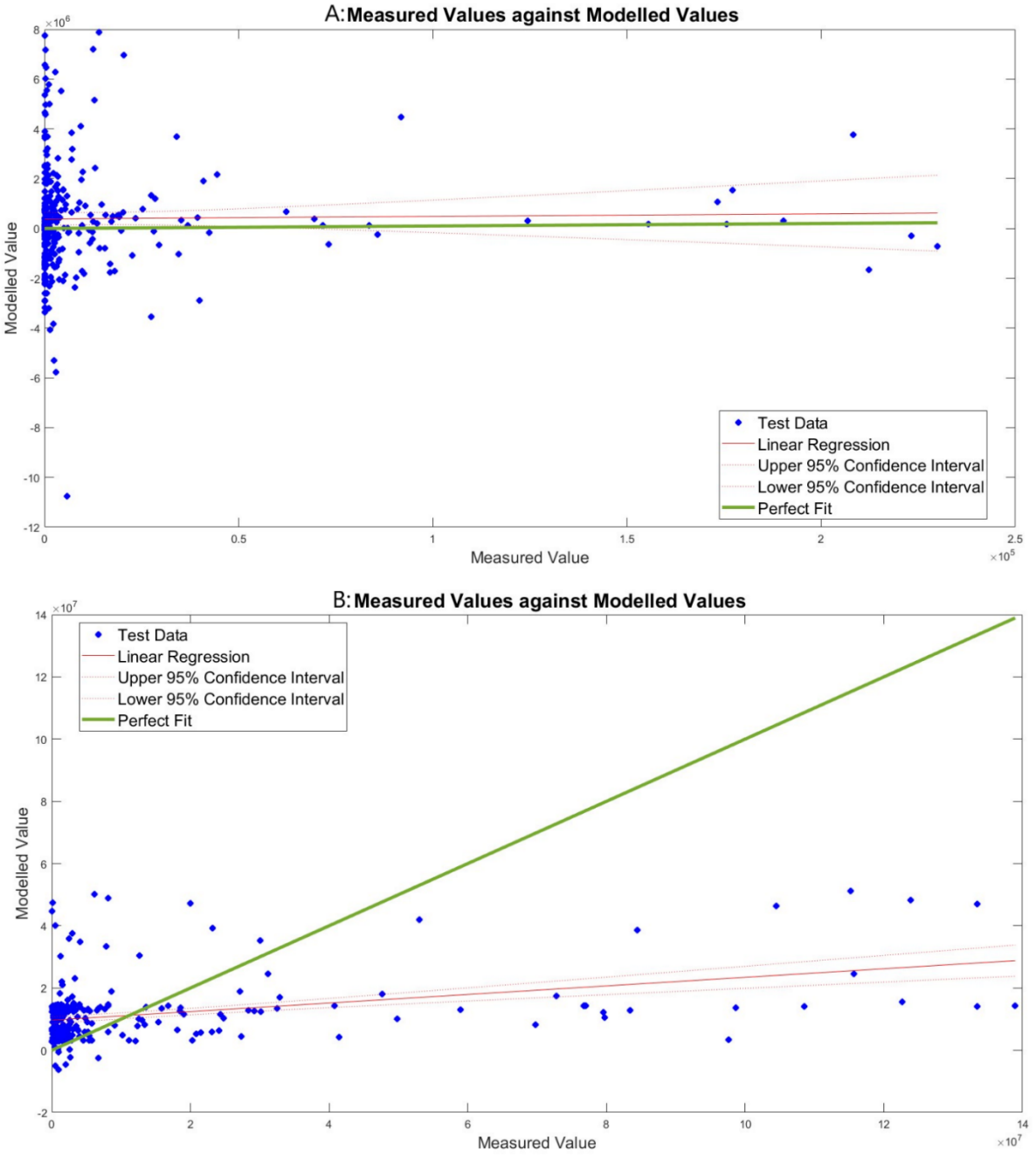

Figure S24. Correlation of Measured SARS-CoV-2 (or CrAssphage) Concentrations against Modelled Values based on the CrAssphage (or SARS-CoV-2) and Turbidity Impact. The measured virus concentration is compared against the modelled virus concentration, the correlation being described by the linear regression (the R, R^2 and RMSE can be found in Table 6). A perfect fit is described by a diagonal line where the test data has identical values on both axes. The model is a Sugeno adaptive neuro-fuzzy interference system (ANFIS) where crAssphage for panel A (or SARS-CoV-2 for panel B) and turbidity are the input parameters and SARS-CoV-2 (A) or crAssphage (B) are the output parameters. The initial fuzzy interference system (FIS) is trained with subtractive clustering followed by tuning using input/output training data, which can also be described as additional fitting of the model to the training data. The figure depicts the final modelled data.

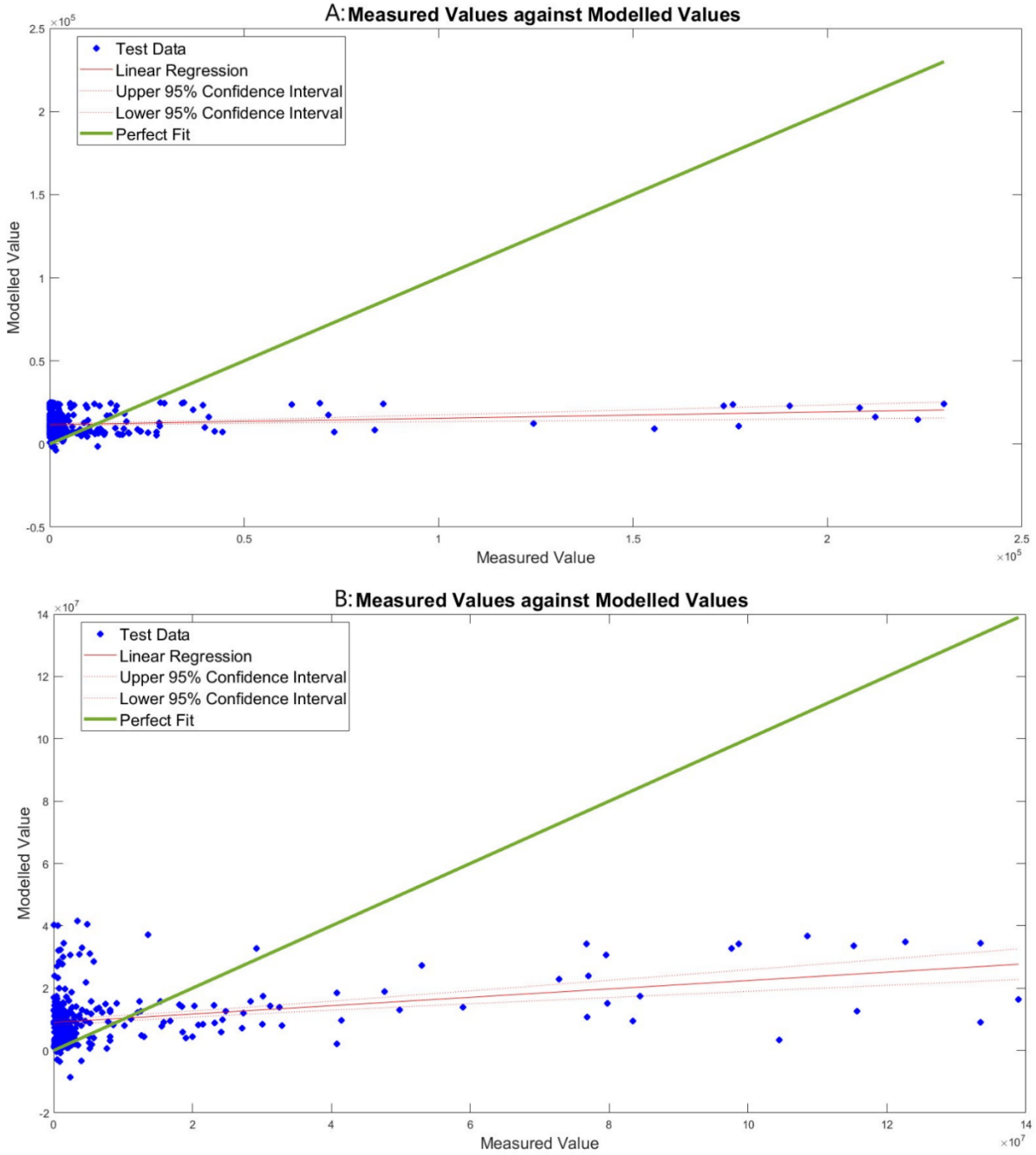

Figure S25. Correlation of Measured SARS-CoV-2 (or CrAssphage) Concentrations against Modelled Values based on the CrAssphage (or SARS-CoV-2) and pH Impact. The measured virus concentration is compared against the modelled virus concentration, the correlation being described by the linear regression (the R, R^2 and RMSE can be found in Table 6). A perfect fit is described by a diagonal line where the test data has identical values on both axes. The model is a Sugeno adaptive neuro-fuzzy interference system (ANFIS) where crAssphage for panel A (or SARS-CoV-2 for panel B) and pH are the input parameters and SARS-CoV-2 (A) or crAssphage (B) are the output parameters. The initial fuzzy interference system (FIS) is trained with subtractive clustering followed by tuning using input/output training data, which can also be described as additional fitting of the model to the training data. The figure depicts the final modelled data.

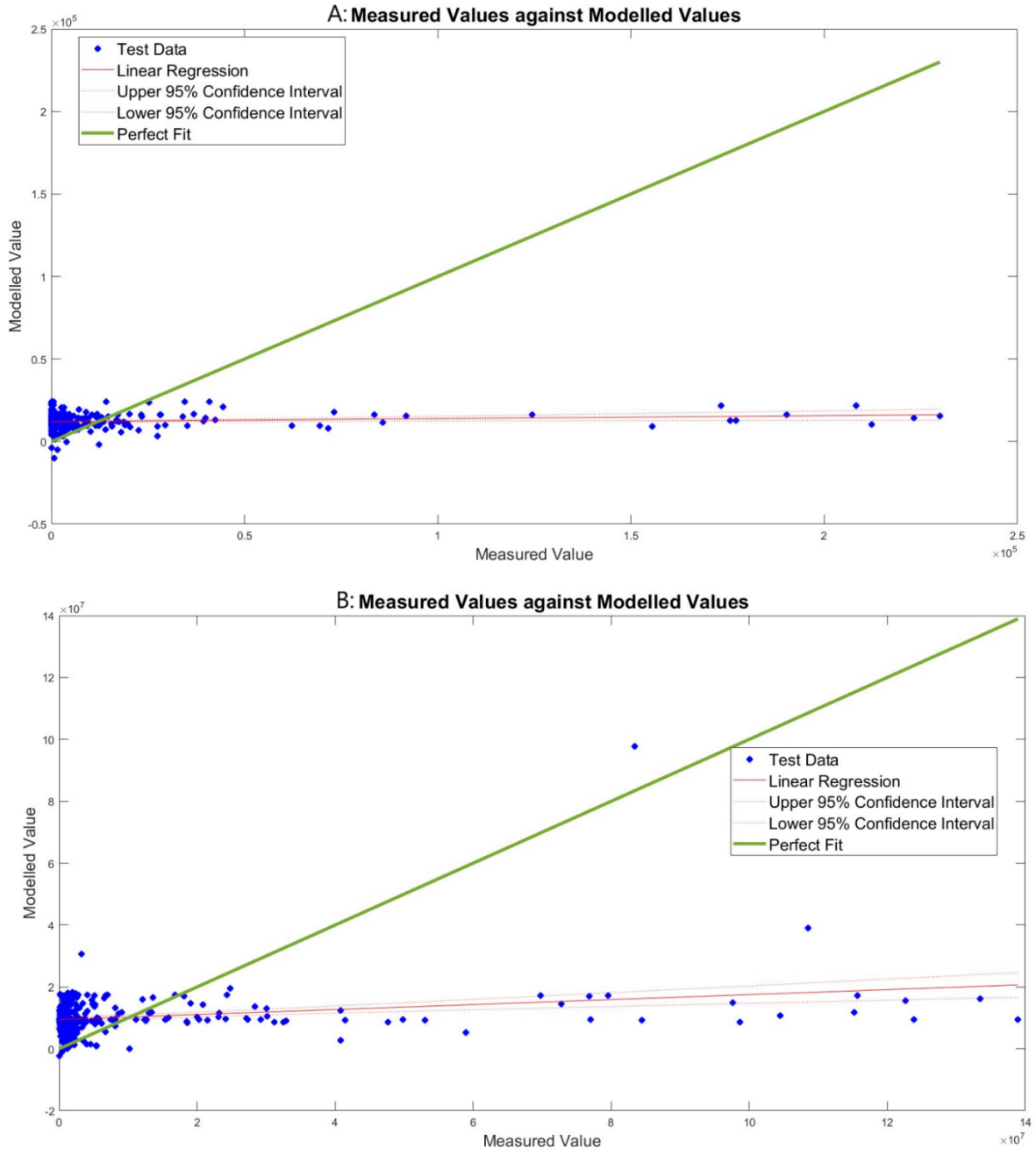

Figure S26. Correlation of Measured SARS-CoV-2 (or CrAssphage) Concentrations against Modelled Values based on the CrAssphage (or SARS-CoV-2) and Electrical conductivity Impact. The measured virus concentration is compared against the modelled virus concentration, the correlation being described by the linear regression (the R, R^2 and RMSE can be found in Table 6). A perfect fit is described by a diagonal line where the test data has identical values on both axes. The model is a Sugeno adaptive neuro-fuzzy interference system (ANFIS) where crAssphage for panel A (or SARS-CoV-2 for panel B) and electrical conductivity are the input parameters and SARS-CoV-2 (A) or crAssphage (B) are the output parameters. The initial fuzzy interference system (FIS) is trained with subtractive clustering followed by tuning using input/output training data, which can also be described as additional fitting of the model to the training data. The figure depicts the final modelled data.

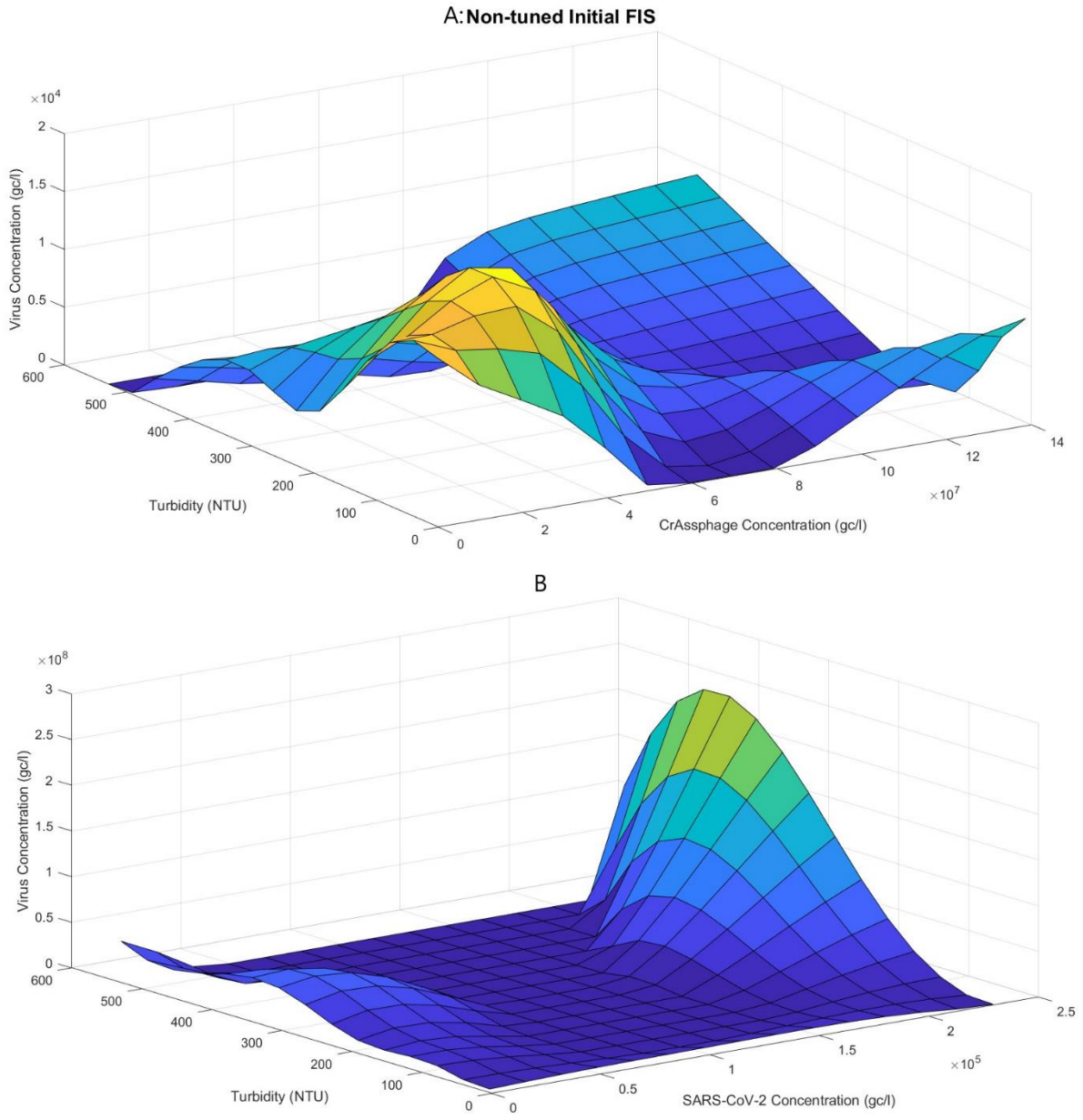

Figure S27. The Modelled Impact of Turbidity and CrAssphage (or SARS-CoV-2) on SARS-CoV-2 (or CrAssphage) Concentration. The modelling is based on a Sugeno adaptive neuro-fuzzy interference system (ANFIS), the sub-method being subtractive clustering followed by tuning with the training data, where the inputs are turbidity and crAssphage concentration (panel A) or turbidity and SARS-CoV-2 concentration (panel B), and the outputs are the SARS-CoV-2 concentration (A) or crAssphage concentration (B). Due to an unacceptable error of the tuned model A, the non-tuned initial FIS is presented for panel A.

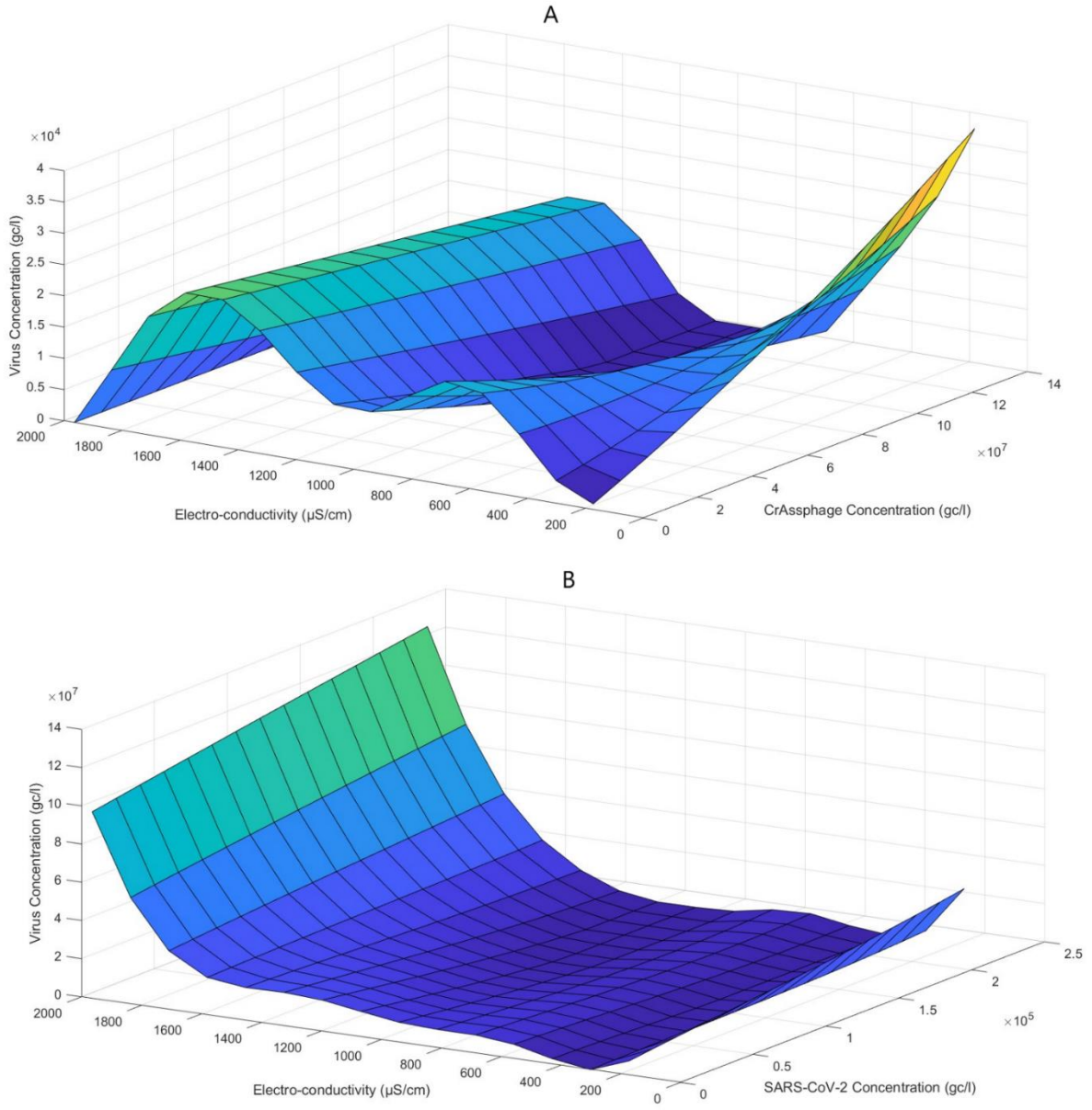

Figure S28. The Modelled Impact of Electrical conductivity and CrAssphage (or SARS-CoV-2) on SARS-CoV-2 (or CrAssphage) Concentration. The modelling is based on a Sugeno adaptive neuro-fuzzy interference system (ANFIS), the sub-method being subtractive clustering followed by tuning with the training data, where the inputs are electrical conductivity and crAssphage concentration (panel A) or electrical conductivity and SARS-CoV-2 concentration (panel B), and the outputs are the SARS-CoV-2 concentration (A) or crAssphage concentration (B).