Environmental Science
Water Research & Technology

ROYAL SOCIETY OF CHEMISTRY

PAPER

Cite this: *Environ. Sci.: Water Res. Technol.*, 2023, **9**, 1403

Prewhitening and normalization help detect a strong cross-correlation between daily wastewater SARS-CoV-2 RNA abundance and COVID-19 cases in a community†

Min Ki Jeon, Bo Li, Doris Yoong Wen Di and Tao Yan *

Received 17th December 2022,
Accepted 2nd March 2023

DOI: 10.1039/d2ew00951j

rsc.li/es-water

Wastewater surveillance is a promising technology for real-time tracking and even early detection of COVID-19 infections in a community. Although correlation analysis between wastewater surveillance data and the daily clinical COVID-19 case numbers has been frequently conducted, the importance of stationarity of the time series data has not been well addressed. In this study, we demonstrated that strong yet spurious correlation could arise from non-stationary time series data in wastewater surveillance. Data prewhitening to remove trends by the first differences of values between two consecutive times helped to reveal distinct cross-correlation patterns between daily clinical case numbers and daily wastewater SARS-CoV-2 RNA abundance during a lockdown period in 2020 in Honolulu, Hawaii. Normalization of wastewater SARS-CoV-2 RNA concentration by the endogenous fecal viral markers in the same samples significantly improved the cross-correlation, and the best correlation was detected at a two-day lag of the daily clinical case numbers. The detection of a significant correlation between the daily wastewater SARS-CoV-2 RNA abundance and the clinical case numbers also suggests that disease burden fluctuation in the community should not be excluded as a contributor to the often observed weekly cyclic patterns of clinical cases.

Water impact

Wastewater surveillance represents an emerging water technology with significant human health benefits. The study demonstrated that non-stationary time series data could lead to spurious correlation, highlighting the need for prewhitening. Normalization strategies could alleviate variations in sample collection and analyses, which is useful for detecting actual underlying relationships between wastewater surveillance data and clinical data.

1. Introduction

Since the outbreak of COVID-19 pandemic in late 2019,[1] wastewater surveillance has been explored as a new way to monitor the spread of SARS-CoV-2 in human communities. Many studies have shown the presence of SARS-CoV-2 viral particles or genomic RNA in bodily wastes, including feces,[2] urine,[3] and respiratory fluids[4,5] in both symptomatic and asymptotic patients. In particular, asymptomatic infections are now known to account for a large percentage of total COVID-19 infections,[6,7] and also shed SARS-CoV-2 virus in feces.[8,9] Since wastewater collects human wastes from all individuals in the wastewater service area and hence can provide comprehensive information on COVID-19 infection in the community, this enables a unique advantage of wastewater surveillance in that it can potentially capture the "actual" infection rates, including the asymptomatic or mildly symptomatic patients in the community who are less likely to seek clinical testing.

Wastewater surveillance may also be able to provide real-time tracking and even early detection of infectious in a community. The sources of SARS-CoV-2 viral shedding to wastewater include mainly feces and partially saliva and sputum due to their high shedding probability and the possibility of entering the sewers.[10] It is known that COVID-19 infected patients start to shed SARS-CoV-2 virus in feces during the incubation period between the infection and the symptom onset[11] and the peak of SARS-CoV-2 viral

Department of Civil and Environmental Engineering, University of Hawaii at Manoa, 2540 Dole Street, 383 Holmes Hall, Honolulu, HI 96822, USA. E-mail: taoyan@hawaii.edu; Fax: 808 956 5014; Tel: 808 956 6024

† Electronic supplementary information (ESI) available. See DOI: https://doi.org/10.1039/d2ew00951j

concentration is reported generally at the beginning of the symptom onset.[12,13] In addition, model fitting results from a meta-data analysis study using experimental findings from various clinical studies have estimated the highest viral concentration at 0.34 days after symptom onset.[14]

Since data collected from both clinically confirmed COVID-19 cases and SARS-CoV-2 RNA abundance in wastewater are time series data, their relationships could be examined through time series data analyses including cross-correlation. Removing the trend or seasonality of time-series data sets to achieve stationarity is an important prerequisite to avoid spurious cross-correlation.[15] This transformation process is called "prewhitening" and it transforms time series data into stationary forms. One common prewhitening method is the single differencing of the time series data points by their first differences.[16] However, many of the wastewater surveillance studies that correlated SARS-CoV-2 RNA abundance in wastewater and COVID-19 cases in the community did not prewhiten the time series data.[17–20] As a result, the strong correlation coefficients observed could be attributed to trend or seasonality instead of the actual correlation of variation between the two types of data sets.

In this study, we re-analyzed previously collected time-series data on daily wastewater SARS-CoV-2 RNA abundance and clinical COVID-19 cases in a large metropolitan area to demonstrate the importance of prewhitening when conducting cross-correlation analysis. Both SARS-CoV-2 RNA concentration and its normalized abundance were subjected to time series cross-correlation analysis to determine any presence of lags between them and the corresponding daily clinical case numbers observed in the community. Also, normalization strategies of the wastewater data were compared to identify the best improvement of correlation with the daily clinical case numbers.

2. Materials and methods

2.1. Wastewater sampling, processing and molecular quantification

Wastewater sampling, processing, and RT-qPCR quantification of SARS-CoV-2 RNA and several fecal RNA viruses in wastewater were previously described in detail in Li *et al.*,[21] which are briefly summarized in the following (Tables S1 and S2†). The two largest wastewater treatment plants (WWTP), Sand Island (SI) and Honouliuli (HO) in the City and County of Honolulu, were selected to collect wastewater samples to represent the wastewater of the community. Untreated primary influent wastewater samples were collected by daily flow-adjusted composite sampling from the SI and HO WWTPs from August 27th, 2020 to October 4th, 2020 (*i.e.*, day 0 to 38, $n = 39$ for each WWTP). All daily wastewater samples were thoroughly mixed and aliquots were subsequently centrifuged to separate suspended solids and supernatant, which were referred to as solid and liquid fractions of the wastewater samples, respectively. The exogenous process control bovine coronavirus (BCoV) (Zoetis; Kalamazoo, MI, USA) was spiked into some solid and liquid subsamples to detect inhibition and assess recovery (four batches of samples, $n = 46$) (Table S2†). The liquid fraction was first treated by the polyethylene glycol (PEG) precipitation method[22] to concentrate and pellet viral particles in the liquid fraction. The precipitated pellets from the liquid fraction as well as the solid fraction were subjected to viral RNA extraction. All extracted viral RNA samples were reverse transcribed with random hexamer N6, and the produced cDNA samples were analyzed by qPCR assays targeting SARS-CoV-2 RNA (the N1 and N2 assays[23] and the E gene assay[24]) and fecal RNA viral surrogates (F+ RNA coliphages Group II (G2) and Group III (G3),[25] and pepper mild mottle virus (PMMoV)[26]).

2.2. Data analysis

All data analyses used both log-transformed SARS-CoV-2 RNA concentration data determined by the three qPCR assays (*i.e.*, log N1, log N2, and log E) and their abundances normalized by the three fecal viral indicators (*i.e.*, log (N1/G2), log (N2/G2), log (E/G2), log (N1/G3), log (N2/G3), log (E/G3), log (N1/PMMoV), log (N2/PMMoV) and log (E/PMMoV)). Daily new COVID-19 case counts for Honolulu were sourced from the local COVID-19 dashboard of the Disease Outbreak Control Division at the State of Hawaii Department of Health. The study only used publicly available data at the population level, and thus required no IRB review. Cross-correlation was used to examine the time-lagged association between new clinical COVID-19 cases in the community and wastewater SARS-CoV-2 RNA abundance. A prewhitening process was applied to all time series data, including COVID-19 clinical case numbers and the wastewater SARS-CoV-2 RNA abundance, to remove trends. The wastewater SARS-CoV-2 RNA abundance data were prewhitened by log transformation followed by the first differences, and the clinical case data were prewhitened by the first differences. The original and the prewhitened data were tested for normality by using Shapiro–Wilk test.[27] Mann–Kendall test[28,29] was used for the assessment of trend significance before and after the prewhitening to verify the successful removal of trends.

Cross-correlation of original and prewhitened SARS-CoV-2 RNA concentration and their normalized abundance in liquid or solid fractions and the daily new clinical COVID-19 cases were analyzed for the SI and HO WWTPs separately. The cross-correlation functions (CCF) function in the R environment was used with a maximum lag of six days. A positive lag indicates that the SARS-CoV-2 RNA concentration or normalized abundance was leading the clinical cases. Positive coefficients indicate a positive relationship between the SARS-CoV-2 RNA concentration or normalized abundance and clinical cases.

Results of the cross-correlation analysis were visualized by heatmaps and boxplots. Additionally, correlation coefficients from the cross-correlation analyses were compared with respect to different normalization strategies by using p-values obtained from pairwise t-test and were adjusted by the Benjamini and Hochberg correction[30] to determine which

normalization strategy showed the best improvement of cross-correlation coefficients. All statistical analyses and data visualization were conducted in R 4.2.1[31] by using the packages *tidyverse* 1.3.1,[32] *ggpubr* 0.4.0,[33] *scales* 1.1.1,[34] and *rstatix* 0.7.0.[35]

3. Results

3.1. Importance of prewhitening on cross-correlation

Our previous study[21] detected downward trends for both daily clinical COVID-19 case numbers in the community and SARS-CoV-2 RNA abundances (both with or without normalization by fecal RNA viral markers) in the wastewater samples. The observed downward trends were the result of a public health lockdown implemented to counter the COVID-19 outbreak on the Island of Oahu. Therefore, the time series data of raw wastewater SARS-CoV-2 RNA concentration and its normalized relative abundance, as well as the daily fluctuation of clinical case numbers, all need to be prewhitened in order to be detrended before cross-correlation analysis. The Shapiro–Wilk test showed that only 25 out of 49 of the original data (p = 0.083 ± 0.102) were normally distributed, but all of the prewhitened data (p = 0.394 ± 0.244) were normally distributed (Table S3†). The Mann–Kendall test results confirmed the removal of trends from both SARS-CoV-2 RNA abundance data in the wastewater samples and the daily clinical case data by the prewhitening process (Table S3†).

The prewhitened SI and HO WWTPs time series SARS-CoV-2 RNA concentration data were first compared with the prewhitened daily clinical case numbers, which showed only weak correlations (either positive or negative) (Fig. 1). The only significant correlation was observed at a two-day lag of the prewhitened daily clinical case numbers (x_{t+2}) falling behind the prewhitened daily wastewater SARS-CoV-2 RNA concentrations of log N1 (r = 0.38, p = 0.019) in the liquid fraction of HO WWTP (Fig. 1E). No significant correlations were found from any other lags for the two WWTPs, as indicated by the boxplots falling under the 95% confidence level (Fig. 1B and F).

To compare, cross-correlation of the original non-stationary time series SARS-CoV-2 data was also performed to illustrate the potential for spurious correlation (Fig. 1C, D, G and H). Both time series concentration data (liquid and solid fractions) from SI (Fig. 1C and D) and HO (Fig. 1G and H) WWTPs showed all positive correlation coefficients and the majority of the cross-correlation analyses showed statistically significant correlations with p-values less than 0.05 (SI: 21 out of 42 analyses; HO: 27 out of 42 analyses) with the original daily clinical case numbers (x_{t+h}, h = lag number). Because the normality assumption of cross-correlation analysis was not met, these high positive correlation coefficients are considered spurious and false positive.

3.2. Impact of normalized abundance on cross-correlation

The concentrations of SARS-CoV-2 RNA measured from the samples of the two WWTPs are expected to be impacted by various processes during wastewater sampling and sample processing, including total fecal discharge in the area, sewer

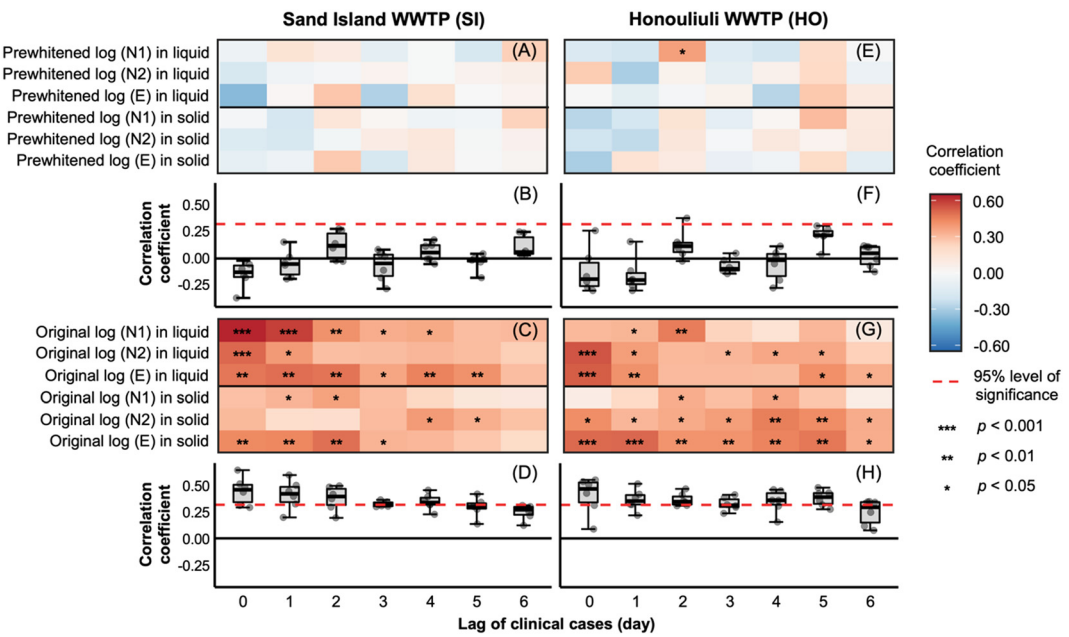

Fig. 1 Cross-correlation between with and without prewhitening by the first differences of daily new clinical COVID-19 case numbers and measured SARS-CoV-2 RNA concentration (log 10 transformed) in wastewater samples from Sand Island (prewhitened: A and B; non-prewhitened: C and D) and Honouliuli (prewhitened: E and F; non-prewhitened: G and H). Red dashed lines represent a 95% level of significance and the p-value of the correlation less than 0.05 are displayed as asterisks. The middle, upper, and lower lines in the box of the boxplot represent the median, 25th, and 75th percentiles, respectively, and the whiskers represent the largest and smallest values outside of the interquartile range.

collection to the WWTPs, wastewater viral precipitation, and molecular quantification steps. The resulting variations could be potentially mitigated by normalizing data to various endogenous fecal viral indicators (*e.g.,* log N1/G2).[21] The normalized daily wastewater SARS-CoV-2 RNA data were also prewhitened and then analyzed *via* cross-correlation with the prewhitened daily clinical case numbers. For wastewater samples from the SI WWTP, the normalization strategy produced significantly different cross-correlation patterns against different time lags (Fig. 2) than those without normalization (Fig. 1).

The most obvious improvement in the cross-correlation coefficients was observed at a two-day lag (x_{t+2}, Fig. 2B, D and F), with a range of r = −0.03-0.45 (0.23 ± 0.13). The average cross-correlation coefficients were increased from 0.12 to 0.33, 0.19, and 0.16 when G2, G3, and PMMoV were used for normalization, respectively, which showed an average of 0.11 ± 0.09 increase than those without normalization (Fig. 2B, D and F). Among all combinations, the best correlation coefficient (r = 0.45, p = 0.004) was observed between the log (E/G2) in the liquid fraction and the daily clinical case numbers. The normalization strategy showed a higher improvement of correlation coefficients in liquid fraction (Δr = 0.14) than in the solid fraction (Δr = 0.07) when compared to the raw data.

At the two-day lag, statistically significant correlations were observed more frequently with the normalized SARS-CoV-2 RNA abundance data than with the raw data. For example, at the SI WWTP, both log (N1/G2) and log (E/G2) showed statistically significant correlation coefficients in the liquid (r = 0.38 (p = 0.017) and r = 0.45 (p = 0.004), respectively) and solid fractions (r = 0.32 (p = 0.048) and r = 0.38 (p = 0.018), respectively). In contrast, there was no statistically significant correlation between clinical cases and raw wastewater SARS-CoV-2 RNA concentration data (*i.e.*, without normalization) at SI WWTP (Fig. 1A). G3 normalization showed two statistically significant correlations from log (N1/G3) and log (E/G3) (r = 0.33 (p = 0.043) and r = 0.33 (p = 0.044), respectively) in the liquid fractions. PMMoV normalization showed only one statistically significant correlation coefficient from the log (E/PMMoV) in the solid fraction (r = 0.33, p = 0.043).

For the HO WWTP, the largest correlation coefficients from the cross-correlation between the normalized SARS-CoV-2 RNA abundance and the daily clinical case numbers were also observed at a two-day lag (x_{t+2}, Fig. 3B and D), which is similar to the results of SI WWTP. The average cross-correlation coefficients were increased from 0.13 to 0.21 and

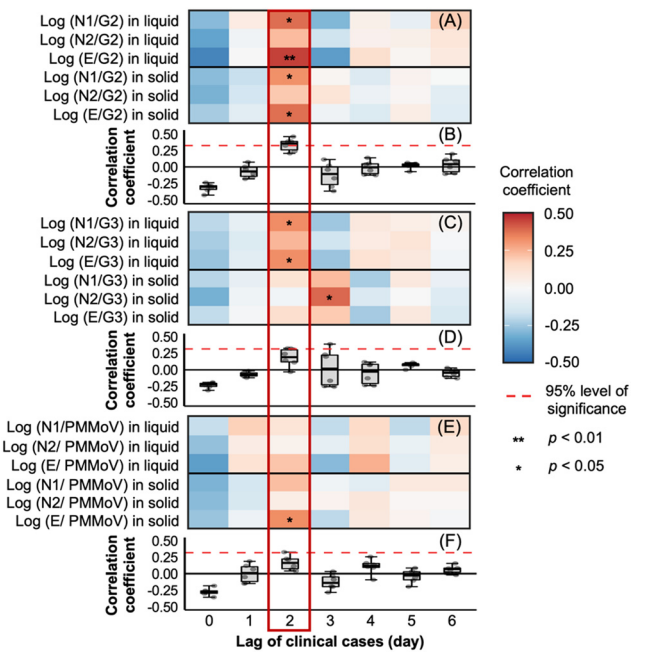

Fig. 2 Cross-correlation between the prewhitened COVID-19 new case numbers and the prewhitened SARS-CoV-2 RNA normalized abundance in wastewater samples from the SI WWTP. The normalized abundance was calculated by dividing SARS-CoV-2 RNA abundance by F+ RNA coliphage group II (A and B), group III (C and D), and PMMoV (E and F). All normalized abundances were transformed into log forms. Red dashed lines represent a 95% level of significance and the p-value of the correlation less than 0.05 are displayed as asterisks. The middle, upper, and lower lines in the box of the boxplot represent the median, 25th, and 75th percentiles, respectively, and the whiskers represent the largest and smallest values outside of the interquartile range.

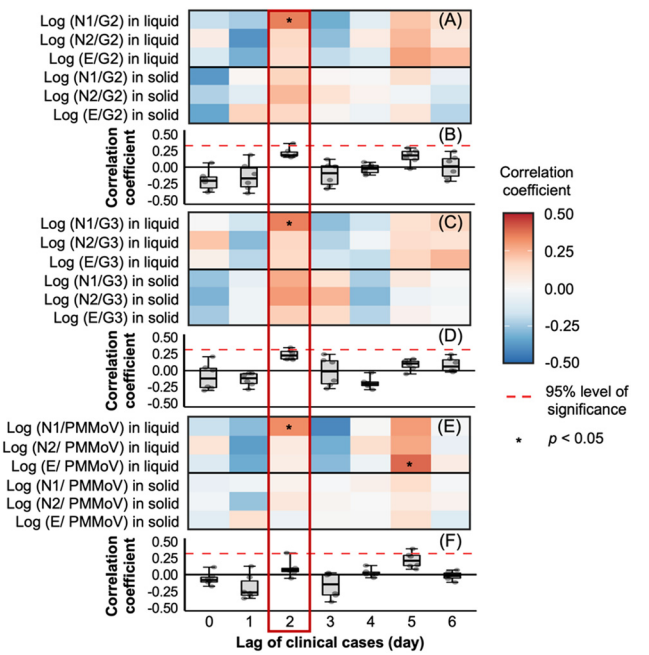

Fig. 3 Cross-correlation between the prewhitened COVID-19 new case numbers and the prewhitened SARS-CoV-2 RNA normalized abundance in wastewater samples from the HO WWTP. The normalized abundance was calculated by dividing SARS-CoV-2 RNA abundance by F+ RNA coliphage group II (A and B), group III (C and D), and PMMoV (E and F). All normalized abundances were transformed into log forms. Red dashed lines represent a 95% level of significance and the p-value of the correlation less than 0.05 are displayed as asterisks. The middle, upper, and lower lines in the box of the boxplot represent the median, 25th, and 75th percentiles, respectively, and the whiskers represent the largest and smallest values outside of the interquartile range.

0.24 for G2 and G3 normalizations, respectively, which showed an average of 0.10 ± 0.02 increase than those without normalization (Fig. 3B and D). The best correlation coefficient among all three normalizations was observed between the log (N1/G2) in the liquid fraction and the daily clinical case numbers (r = 0.35, p = 0.029). Similar to the SI WWTP, the normalization strategy showed a larger improvement of correlation coefficients in liquid fractions (Δr = 0.07) than in the solid fractions (Δr = 0.03) when compared to the raw data.

At the two-day lag, all normalized forms of log N1 in the liquid fractions (log (N1/G2): r = 0.35, p = 0.029; log (N1/G3): r = 0.35, p = 0.030; log (N1/PMMoV): r = 0.33, p = 0.043) showed statistically significant correlations between the daily clinical cases and the normalized SARS-CoV-2 RNA abundance at the HO WWTP (Fig. 3A, C and E). Although no statistically significant correlation was observed from solid fractions from all normalization strategies (r = −0.05-0.27, 0.16 ± 0.11), all correlation coefficients observed from the solid fractions were increased by normalization with G2 and G3.

The correlation coefficients of all SARS-CoV-2 marker genes from both liquid and solid fractions of HO WWTP decreased when they were normalized with PMMoV; average correlation coefficients decreased from 0.13 to 0.09 at the two-day lag (Fig. 3E). Interestingly, PMMoV normalized SARS-CoV-2 RNA abundances (for all three gene markers) in the liquid fraction showed large correlation coefficients (r = 0.32 ± 0.06) at a five-day lag of daily clinical case numbers (Fig. 3F), and a statistically significant correlation was observed for log (E/PMMoV) (r = 0.39, p = 0.014).

3.3. Comparison of different normalization strategies for cross-correlation analysis

The correlation coefficients between the daily clinical case numbers and SARS-CoV-2 RNA abundance at a two-day lag of clinical cases were compared with respect to the normalization strategies by using a pairwise t-test (Fig. 4). Among the three different endogenous fecal viral RNA controls used, only normalizing the data with G2 (r = 0.33 ± 0.10, p = 0.002) showed a statistically significant improvement of correlation coefficients in comparison to that using the raw data (r = 0.12 ± 0.13). While G3 (r = 0.19 ± 0.14, p = 0.396) and PMMoV (r = 0.16 ± 0.11, p = 0.198) mildly improved the correlation (Fig. 4A), the improvement was not statistically significant. Furthermore, the G2 normalization showed a significantly larger average correlation coefficient than both G3 (p = 0.040) and PMMoV (p = 0.014) normalizations.

For the HO WWTP, the G3 (r = 0.24 ± 0.08, p = 0.032) normalization method was the only strategy that significantly improved the correlation coefficients (Fig. 4B). G2 (r = 0.21 ± 0.08) also increased the mean values of correlation coefficients from the raw data (r = 0.13 ± 0.14), although the improvement was only marginally significant (p = 0.058). Both G2 (p = 0.028) and G3 (p = 0.028) normalizations resulted in better correlation than the PMMoV normalization

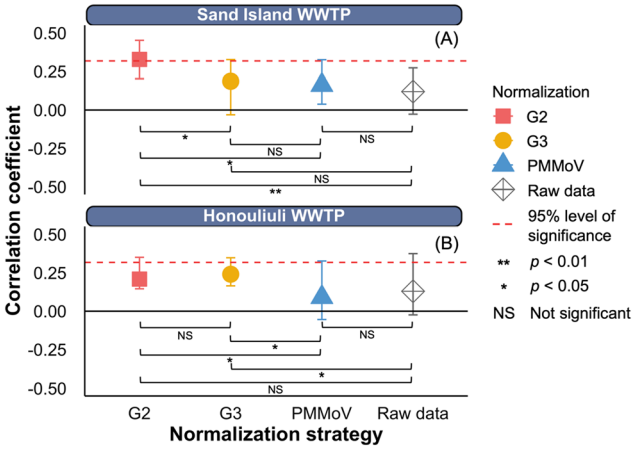

Fig. 4 Cross-correlation between both prewhitened daily new clinical COVID-19 case numbers in Honolulu County and normalized SARS-CoV-2 RNA concentration (log 10 transformed) by F+ RNA coliphage group II, group III, and PMMoV in wastewater samples from Sand Island (A) and Honouliuli (B) at a two-day lag of clinical cases. Red dashed lines represent a 95% level of significance and the whiskers represent the largest and the smallest values.

(r = 0.09 ± 0.13), which actually showed a lower average correlation coefficient than the raw data.

The overall results indicate that the normalization of SARS-CoV-2 RNA abundance improved the cross-correlations with the daily clinical case numbers. G2 normalization showed the largest improvement of cross-correlations in the SI WWTP samples, while G3 normalization resulted in the largest improvement of the cross-correlations in the HO WWTP samples. When considering the liquid fractions only, the G2 normalization of log N1 from SI (r = 0.38, p = 0.017) and HO (r = 0.35, p = 0.029) WWTPs showed both significant correlations with the daily clinical case numbers.

4. Discussion

In our previous study,[21] we observed simultaneous downward trends between SARS-CoV-2 RNA abundance (both with and without normalization by fecal viral markers) in wastewater samples from the SI and HO WWTPs and the daily clinical COVID-19 case numbers during a COVID-19 public health lockdown. This is congruent with previous observations where increases in wastewater SARS-CoV-2 RNA concentration corresponded with rapidly expanding COVID-19 outbreaks.[17,36–39] The fine-scale temporal dynamics revealed by the daily sampling also detected significant intra-day fluctuation of the wastewater SARS-CoV-2 RNA abundance, even within the same weeks. Many factors could have contributed to the observed intra-day fluctuation, including errors in wastewater sampling and sample analysis, variations in viral shedding by infected individuals, and daily fluctuations in disease burden in the community. Since similar trends were detected in the two replicate WWTPs,

over multiple weeks, and regardless of normalization strategies, the former two (*i.e.*, sampling and analysis errors and variation in viral shedding) are unlikely to explain the observations entirely.

Cross-correlation analysis of the time series data of wastewater SARS-CoV-2 RNA abundance and clinical case numbers could be used to infer potential association and determine if daily fluctuations in disease burden in the community contributed to the observed intra-day fluctuation. Many wastewater surveillance studies have compared wastewater SARS-CoV-2 RNA abundance with clinical case data in the community through correlation analysis.[17–20] However, few previous studies have conducted prewhitening treatment to achieve stationarity of the time series data. Stationarity in time series data indicates consistency of the distribution (mean and variance) over time,[40] and non-stationary time series data can often lead to spurious outcomes in correlation analysis. This was clearly demonstrated when we observed significant spurious cross-correlations with the original data that contained trends (Fig. 1C, D, G and H). Similar phenomenon may explain reports of wastewater surveillance showing strong correlations to clinical cases, especially when the studies were conducted during a period when COVID-19 clinical cases were continuously increasing or decreasing.[36,41,42] Therefore, prewhitening the data for wastewater surveillance to meet the stationarity requirement of cross-correlation analysis must be practiced in order to identify the actual association between data sets.

After prewhitening the data, the cross-correlation coefficients decreased significantly, and the overall patterns with respect to the time lag also changed drastically (Fig. 1). For example, at zero-time lag, cross-correlation using the original data detected the best positive correlation, whereas the prewhitened data actually detected some negative correlations. The low levels of correlation detected are not entirely unexpected, considering the extraordinary complexity involved in collecting the wastewater SARS-CoV-2 RNA data and associate variations. Many factors, including varying fecal discharge by infected individuals, dilution and fluctuation during transportation in sanitary sewers, and wastewater sample collection and processing, could have contributed to the variations in SARS-CoV-2 RNA abundance in wastewater samples. The molecular quantification processes could also introduce additional variation to the results; for example, RNA recovery during sample extraction could have different efficiencies, and subsequent reverse transcription and qPCR quantification could introduce additional biases.

The incorporation of normalization strategies and the resulting relative abundance of SARS-CoV-2 RNA in the wastewater samples led to the identification of a two-day lag showing the best correlation (Fig. 2 and 3). Given the complex and multi-step process required for quantifying SARS-CoV-2 RNA in wastewater, the use of endogenous viral RNA control for global normalization may be important to reduce the variations during the analysis and enable statistical comparison. Amongst the three fecal RNA viruses tested as endogenous controls in this study, normalization by G2 provided the most significant improvement in correlation between wastewater SARS-CoV-2 RNA abundance and clinical new cases. Cole *et al.* found that G2 had the highest proportion among the total F+ RNA groups in WWTP samples (51.9%) and G2 was found more in human-impacted wastes than G3.[43] This supports our results of G2 normalization of SARS-CoV-2 RNA abundance having higher correlation coefficients compared to G3. On the other hand, PMMoV only provided marginal improvement in correlation. This difference could be attributed to their respective sources in human feces where G2 and G3 are inherently linked with fecal coliforms while PMMoV is subjected to dietary variation in pepper consumption. Some previous wastewater surveillance studies that used PMMoV for the SARS-CoV-2 abundance normalization also reported that the PMMoV did not improve the correlation with the clinical cases.[44–46] Other biomarkers and chemical indicators for population normalization are also recently considered, such as paraxanthine,[47] cross-assembly phage,[48,49] human RNase P,[50] total nitrogen and phosphate.[51] Therefore, more studies related to improving the normalization methods in the wastewater surveillance field are required to more efficiently reduce the variations.

The significant cross-correlation between the normalized abundance of daily wastewater SARS-CoV-2 RNA and new clinical cases in the community is highly intriguing. Since the onset of COVID-19 pandemic, weekly intra-day oscillations in new clinical case numbers have been widely observed in communities across the globe.[52] One school of thought is that these weekly intra-day oscillations are primarily a reflection of diagnostic and reporting biases,[53,54] while a competing theory is that this could be caused by actual disease transmission dynamics due to weekly behavior patterns.[55–57] The strong correlation observed in this study between the intra-day fluctuation and weekly oscillation of wastewater SARS-CoV-2 RNA abundance and clinical case numbers suggests that the observed weekly oscillation of clinical cases may be indeed a true reflection of the disease burden dynamics in the community, in addition to contributions from clinical sampling and reporting biases and errors.

Since the average turnaround for clinical testing during the study period was approximately one day, with the assumption of one day lag between symptom onset and specimen collection, the observed two-day lag in cross-correlation analysis indicates that the wastewater SARS-CoV-2 RNA abundance may be synchronizing with symptom development of new COVID-19 cases in the community. Studies at the early stage of the pandemic, which likely experienced clinical testing delays, have reported the detection of the SARS-CoV-2 RNA in wastewater about one week ahead of reported clinical cases in the communities.[17,37] Another study reported wastewater sludge SARS-CoV-2 RNA

concentration leading the cinical specimen collection by 0–2 days.[23] The apparent synchronous correspondence supports the possibility of using wastewater for early detection of viral transmission in communities, as viral shedding can start 3–5 days before the peak of symptom onset.[12]

In this study, both the solid and liquid fractions of the same wastewater samples were analyzed separately, and normalized SARS-CoV-2 RNA abundance data in both fractions showed similar cross-correlation patterns (Fig. 2 and 3). In the previous study,[21] the solid fraction contained a higher per mass concentration of SARS-CoV-2 RNA than the liquid fraction, while the normalized abundances between the two fractions were quite similar. The normalized abundance of SARS-CoV-2 RNA in the liquid fraction exhibited slightly stronger correlations with the clinical COVID-19 case numbers in the community than the normalized abundance of SARS-CoV-2 in the solid fraction. This could be attributed to the more complex matrix effects in the solid fraction than the liquid fraction, as indicated in our previous study, where lower recovery and higher variation of the spiked BCoV as exogenous control were observed in the solid fraction than in the liquid fraction.[21]

It is important to note that all three quantification assays showed similar cross-correlation patterns between normalized SARS-CoV-2 RNA abundance in wastewater and clinical case numbers in the community. While the SARS-CoV-2 RNA genome contains a single copy of N and E genes, our previously published study[21] and many other studies[37,58] have shown that different assays usually generate different abundance data, indicating that the molecular quantification processes have variations. Nevertheless, all three assays were able to reveal strong correlations at a two-day lag between normalized SARS-CoV-2 RNA abundance in the wastewater and community disease burden, with the N1 gene assay providing the highest correlation coefficients in both tested WWTPs (Fig. 2 and 3). Even though the E gene showed strong positive cross-correlation patterns at a two-day lag from both SI and HO WWTPs, it is considered the least specific PCR target for SARS-CoV-2 detection due to homologous sequence similarities with other coronaviruses.[59] Our previous paper[21] showed that the N2 gene assay had the least sensitivity, consequently, more wastewater surveillance studies for SARS-CoV-2 RNA detection are using the N1 assay and showed higher positivity rates compared to the E gene assay.[60,61] With the reasons above, we would recommend the N1 gene assay for future wastewater surveillance for SARS-CoV-2 RNA detection and G2 for wastewater normalization.

5. Conclusions

This study demonstrated the importance of prewhitening to remove the trends of the daily fluctuation of wastewater surveillance data and the clinical case numbers before cross-correlation analysis of the time series data sets to avoid spurious correlations. We also observed that normalization strategies to account for variations in the process are helpful in improving cross-correlation coefficients. Amongst the various normalization strategies, SARS-CoV-2 RNA abundances normalized with F+ RNA coliphage Group II provided the best correlation coefficients in this study. We observed that the N1 assay was showing the best correlation, while N2 showed less sensitivity and the E gene has been reported to be less specific. Although there were significant inherent variations in the data due to the complexity of the wastewater samples and the process, the daily clinical case numbers appeared to lag two days behind the SARS-CoV-2 RNA detection in wastewater based on multiple gene markers and multiple normalization strategies. This supports the notion that wastewater surveillance has the potential to provide earlier detection of SARS-CoV-2 in the community than clinical diagnosis. Most interestingly, the strong cross-correlation between the intra-day fluctuation and weekly oscillation of wastewater SARS-CoV-2 RNA abundance and clinical cases suggest that the observed weekly oscillation of clinical cases may indeed be a true reflection (at least partially) of the disease burden dynamics in the community, in addition to contributions from clinical sampling and reporting biases and errors, which requires further research to delineate their respective contributions.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. CBET-2027059.

References

1 P. Zhou, X.-L. Yang, X.-G. Wang, B. Hu, L. Zhang and W. Zhang, *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin, *Nature*, 2020, **579**(7798), 270–273.

2 Y. Chen, L. Chen, Q. Deng, G. Zhang, K. Wu and L. Ni, *et al.* The presence of SARS-CoV-2 RNA in the feces of COVID-19 patients, *J. Med. Virol.*, 2020, **92**(7), 833–840.

3 A. H. Kashi, J. De la Rosette, E. Amini, H. Abdi, M. Fallah-Karkan and M. Vaezjalali, Urinary viral shedding of COVID-19 and its clinical associations: a systematic review and meta-analysis of observational studies, *Urol. J.*, 2020, **17**(5), 433–441.

4 K. K.-W. Tor, O. T.-Y. Tsang, W.-S. Leung, A. R. Tam, T.-C. Wu and D. C. Lung, *et al.* Temporal profiles of viral load in posterior oropharyngeal saliva samples and serum antibody responses during infection by SARS-CoV-2: an observational cohort study, *Lancet Infect. Dis.*, 2020, **20**(5), 565–574.

5 A. L. Wyllie, J. Fournier, A. Casanovas-Massana, M. Campbell, M. Tokuyama and P. Vijayakumar, *et al.* Saliva or nasopharyngeal swab specimens for detection of SARS-CoV-2, *N. Engl. J. Med.*, 2020, **383**(13), 1283–1286.

6 Q.-X. Long, X.-J. Tang, Q.-L. Shi, Q. Li, H.-J. Deng and J. Yuan, *et al.* Clinical and immunological assessment of asymptomatic SARS-CoV-2 infections, *Nat. Med.*, 2020, **26**(8), 1200–1204.

7 D. Buitrago-Garcia, D. Egli-Gany, M. J. Counotte, S. Hossmann, H. Imeri and A. M. Ipekci, *et al.* Occurrence and transmission potential of asymptomatic and presymptomatic SARS-CoV-2 infections: A living systematic review and meta-analysis, *PLoS Med.*, 2020, **17**(9), e1003346.

8 B. W. Schmitz, G. K. Innes, S. M. Prasek, W. Q. Betancourt, E. R. Stark and A. R. Foster, *et al.* Enumerating asymptomatic COVID-19 cases and estimating SARS-CoV-2 fecal shedding rates via wastewater-based epidemiology, *Sci. Total Environ.*, 2021, **801**, 149794.

9 S.-K. Park, C.-W. Lee, D.-I. Park, H.-Y. Woo, H. S. Cheong and H. C. Shin, *et al.* Detection of SARS-CoV-2 in fecal samples from patients with asymptomatic and mild COVID-19 in Korea, *Clin. Gastroenterol. Hepatol.*, 2021, **19**(7), 1387–1394. e2.

10 X. Li, J. Kulandaivelu, Y. Guo, S. Zhang, J. Shi and J. O'Brien, *et al.* SARS-CoV-2 shedding sources in wastewater and implications for wastewater-based epidemiology, *J. Hazard. Mater.*, 2022, **432**, 128667.

11 X. He, E. H. Lau, P. Wu, X. Deng, J. Wang and X. Hao, *et al.* Temporal dynamics in viral shedding and transmissibility of COVID-19, *Nat. Med.*, 2020, **26**(5), 672–675.

12 D. L. Jones, M. Q. Baluja, D. W. Graham, A. Corbishley, J. E. McDonald and S. K. Malham, *et al.* 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, *Sci. Total Environ.*, 2020, **749**, 141364.

13 M. Cevik, M. Tate, O. Lloyd, A. E. Maraolo, J. Schafers and A. Ho, SARS-CoV-2, SARS-CoV, and MERS-CoV viral load dynamics, duration of viral shedding, and infectiousness: a systematic review and meta-analysis, *Lancet Microbe*, 2021, **2**(1), e13–e22.

14 F. Miura, M. Kitajima and R. Omori, Duration of SARS-CoV-2 viral shedding in faeces as a parameter for wastewater-based epidemiology: Re-analysis of patient data using a shedding dynamics model, *Sci. Total Environ.*, 2021, **769**, 144549.

15 J. D. Cryer and K.-S. Chan, *Time series analysis: with applications in R*, Springer, 2008.

16 M. Bayazit and B. Önöz, To prewhiten or not to prewhiten in trend analysis?, *Hydrol. Sci. J.*, 2007, **52**(4), 611–624.

17 J. Peccia, A. Zulli, D. E. Brackney, N. D. Grubaugh, E. H. Kaplan and A. Casanovas-Massana, *et al.* Measurement of SARS-CoV-2 RNA in wastewater tracks community infection dynamics, *Nat. Biotechnol.*, 2020, **38**(10), 1164–1167.

18 P. Aberi, R. Arabzadeh, H. Insam, R. Markt, M. Mayr and N. Kreuzinger, *et al.* Quest for optimal regression models in SARS-CoV-2 wastewater based epidemiology, *Int. J. Environ. Res. Public Health*, 2021, **18**(20), 10778.

19 L. R. Lara-Jacobo, G. Islam, J.-P. Desaulniers, A. E. Kirkwood and D. B. Simmons, Detection of SARS-CoV-2 Proteins in Wastewater Samples by Mass Spectrometry, *Environ. Sci. Technol.*, 2022, **56**(8), 5062–5070.

20 L. Li, L. Mazurowski, A. Dewan, M. Carine, L. Haak and T. C. Guarin, *et al.* Longitudinal monitoring of SARS-CoV-2 in wastewater using viral genetic markers and the estimation of unconfirmed COVID-19 cases, *Sci. Total Environ.*, 2022, **817**, 152958.

21 B. Li, D. Y. W. Di, P. Saingam, M. K. Jeon and T. Yan, Fine-scale temporal dynamics of SARS-CoV-2 RNA abundance in wastewater during a COVID-19 lockdown, *Water Res.*, 2021, **197**, 117093.

22 M. H. Hjelmsø, M. Hellmér, X. Fernandez-Cassi, N. Timoneda, O. Lukjancenko and M. Seidel, *et al.* Evaluation of methods for the concentration and extraction of viruses from sewage in the context of metagenomic sequencing, *PLoS One*, 2017, **12**(1), e0170199.

23 X. Lu, L. Wang, S. K. Sakthivel, B. Whitaker, J. Murray and S. Kamili, *et al.* US CDC real-time reverse transcription PCR panel for detection of severe acute respiratory syndrome coronavirus 2, *Emerging Infect. Dis.*, 2020, **26**(8), 1654.

24 V. M. Corman, O. Landt, M. Kaiser, R. Molenkamp, A. Meijer and D. K. Chu, *et al.* Detection of 2019 novel coronavirus(2019-nCoV) by real-time RT-PCR, *Eurosurveillance*, 2020, **25**(3), 2000045.

25 S. D. Friedman, E. M. Cooper, K. R. Calci and F. J. Genthner, Design and assessment of a real time reverse transcription-PCR method to genotype single-stranded RNA male-specific coliphages(Family Leviviridae), *J. Virol. Methods*, 2011, **173**(2), 196–202.

26 K. Rosario, E. M. Symonds, C. Sinigalliano, J. Stewart and M. Breitbart, Pepper mild mottle virus as an indicator of fecal pollution, *Appl. Environ. Microbiol.*, 2009, **75**(22), 7261–7267.

27 S. S. Shapiro and M. B. Wilk, An analysis of variance test for normality(complete samples), *Biometrika*, 1965, **52**(3/4), 591–611.

28 H. B. Mann, Nonparametric tests against trend, *Econometrica*, 1945, 245–259.

29 M. G. Kendall, *Rank correlation methods*, 1948.

30 Y. Benjamini and Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing, *J. R. Stat. Soc. Series B Stat. Methodol.*, 1995, **57**(1), 289–300.

31 R. C. Teami, *A language and environment for statistical computing*, Vienna, Austria, R Foundation for Statistical Computing, 2022.

32 H. Wickham, M. Averick, J. Bryan, W. Chang, L. D. A. McGowan and R. François, *et al.* Welcome to the Tidyverse, *J. Open Source Softw.*, 2019, **4**(43), 1686.

33 A. Kassambara, *ggpubr:'ggplot2'Based Publication Ready Plots, R package version 0.4. 0, Computer software*, **https://cran-r-project org/web/packages/ggpubr/indes html**, 2020.

34 H. Wickham and D. Seidel, *scales: Scale functions for visualization, R package version*, 2020, vol. 11, p. 678.

35 A. Kassambara, *rstatix: Pipe-friendly framework for basic statistical tests. R package version 06 0*, 2020.

36 E. Róka, B. Khayer, Z. Kis, L. B. Kovács, E. Schuler and N. Magyar, *et al.* Ahead of the second wave: early warning for COVID-19 by wastewater surveillance in Hungary, *Sci. Total Environ.*, 2021, **786**, 147398.

37 G. Medema, L. Heijnen, G. Elsinga, R. Italiaander and A. Brouwer, Presence of SARS-Coronavirus-2 RNA in sewage and correlation with reported COVID-19 prevalence in the early stage of the epidemic in the Netherlands, *Environ. Sci. Technol. Lett.*, 2020, **7**(7), 511–516.

38 W. Randazzo, P. Truchado, E. Cuevas-Ferrando, P. Simón, A. Allende and G. Sánchez, SARS-CoV-2 RNA in wastewater anticipated COVID-19 occurrence in a low prevalence area, *Water Res.*, 2020, **181**, 115942.

39 W. Ahmed, N. Angel, J. Edson, K. Bibby, A. Bivins and J. W. O'Brien, *et al.* First confirmed detection of SARS-CoV-2 in untreated wastewater in Australia: a proof of concept for the wastewater surveillance of COVID-19 in the community, *Sci. Total Environ.*, 2020, **728**, 138764.

40 D. Kwiatkowski, P. C. Phillips, P. Schmidt and Y. Shin, Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root?, *J. Econom.*, 1992, **54**(1–3), 159–178.

41 J. Sangsanont, S. Rattanakul, A. Kongprajug, N. Chyerochana, M. Sresung and N. Sriporatana, *et al.* SARS-CoV-2 RNA surveillance in large to small centralized wastewater treatment plants preceding the third COVID-19 resurgence in Bangkok, Thailand, *Sci. Total Environ.*, 2022, **809**, 151169.

42 J. Weidhaas, Z. T. Aanderud, D. K. Roper, J. VanDerslice, E. B. Gaddis and J. Ostermiller, *et al.* Correlation of SARS-CoV-2 RNA in wastewater with COVID-19 disease burden in sewersheds, *Sci. Total Environ.*, 2021, **775**, 145790.

43 D. Cole, S. C. Long and M. D. Sobsey, Evaluation of F+ RNA and DNA coliphages as source-specific indicators of fecal contamination in surface waters, *Appl. Environ. Microbiol.*, 2003, **69**(11), 6507–6514.

44 Y. Ai, A. Davis, D. Jones, S. Lemeshow, H. Tu and F. He, *et al.* Wastewater SARS-CoV-2 monitoring as a community-level COVID-19 trend tracker and variants in Ohio, United States, *Sci. Total Environ.*, 2021, **801**, 149757.

45 S. Feng, A. Roguet, J. S. McClary-Gutierrez, R. J. Newton, N. Kloczko and J. G. Meiman, *et al.* Evaluation of sampling, analysis, and normalization methods for SARS-CoV-2 concentrations in wastewater to assess COVID-19 burdens in Wisconsin communities, *ACS ES&T Water*, 2021, **1**(8), 1955–1965.

46 H. D. Greenwald, L. C. Kennedy, A. Hinkle, O. N. Whitney, V. B. Fan and A. Crits-Christoph, *et al.* Tools for interpretation of wastewater SARS-CoV-2 temporal and spatial trends demonstrated with data collected in the San Francisco Bay Area, *Water Res.: X*, 2021, **12**, 100111.

47 S.-Y. Hsu, M. Bayati, C. Li, H.-Y. Hsieh, A. Belenchia and J. Klutts, *et al.* Biomarkers Selection for Population Normalization in SARS-CoV-2 Wastewater-based Epidemiology, *Water Res.*, 2022, 118985.

48 A. Bivins, K. Crank, J. Greaves, D. North, Z. Wu and K. Bibby, Cross-assembly phage and pepper mild mottle virus as viral water quality monitoring tools—potential, research gaps, and way forward, *Curr. Opin. Environ. Sci. Health*, 2020, **16**, 54–61.

49 H. Green, M. Wilder, M. Collins, A. Fenty, K. Gentile and B. L. Kmush, *et al.* Quantification of SARS-CoV-2 and cross-assembly phage(crAssphage) from wastewater to monitor coronavirus transmission within communities, *MedRxiv*, 2020, preprint, DOI: **10.1101/2020.05.21.20109181**.

50 S. S. El-Malah, J. Saththasivam, K. A. Jabbar, K. Arun, T. A. Gomez and A. A. Ahmed, *et al.* Application of human RNase P normalization for the realistic estimation of SARS-CoV-2 viral load in wastewater: A perspective from Qatar wastewater surveillance, *Environ. Technol. Innovation*, 2022, **27**, 102775.

51 F. Isaksson, L. Lundy, A. Hedström, A. J. Székely and N. Mohamed, Evaluating the use of alternative normalization approaches on sars-cov-2 concentrations in wastewater: Experiences from two catchments in northern sweden, *Environments*, 2022, **9**(3), 39.

52 Q. Bukhari, Y. Jameel, J. M. Massaro, R. B. D'Agostino and S. Khan, Periodic oscillations in daily reported infections and deaths for coronavirus disease 2019, *JAMA Netw. Open*, 2020, **3**(8), e2017521-e.

53 P. L. Bragato, Assessment of the weekly fluctuations of the Covid-19 cases in Italy and worldwide, *Preprints*, 2020, DOI: **10.20944/preprints202005.0202.v1**.

54 A. Bergman, Y. Sella, P. Agre and A. Casadevall, Oscillations in US COVID-19 incidence and mortality data reflect diagnostic and reporting factors, *mSystems*, 2020, **5**(4), e00544-20.

55 W. T. Cecil, COVID-19: daily fluctuations, a weekly cycle, and a negative trend, *Am. J. Manag. Care*, 2020, **26**(7), 284–285.

56 M. Derakhshan, H. R. Ansarian and M. Ghomshei, Temporal variations in COVID-19: an epidemiological discussion with a practical application, *J. Int. Med. Res.*, 2021, **49**(8), 03000605211033208.

57 I. Ricon-Becker, R. Tarrasch, P. Blinder and S. Ben-Eliyahu, A seven-day cycle in COVID-19 infection and mortality rates: Are inter-generational social interactions on the weekends killing susceptible people, *medRxiv*, 2020, preprint, DOI: **10.1101/2020.05.03.20089508**.

58 A. K. Nalla, A. M. Casto, M.-L. W. Huang, G. A. Perchetti, R. Sampoleo and L. Shrestha, *et al.* Comparative performance of SARS-CoV-2 detection assays using seven different primer-probe sets and one assay kit, *J. Clin. Microbiol.*, 2020, **58**(6), e00557-20.

59 M. Artesi, S. Bontems, P. Göbbels, M. Franckh, P. Maes and R. Boreux, *et al.* A recurrent mutation at position 26340 of SARS-CoV-2 is associated with failure of the E gene quantitative reverse transcription-PCR utilized in a commercial dual-target diagnostic assay, *J. Clin. Microbiol.*, 2020, **58**(10), e01598-20.

60 M. Muenchhoff, H. Mairhofer, H. Nitschko, N. Grzimek-Koschewa, D. Hoffmann and A. Berger, *et al.* Multicentre comparison of quantitative PCR-based assays to detect SARS-CoV-2, Germany, March 2020, *Eurosurveillance*, 2020, **25**(24), 2001057.

61 A. Pérez-Cataluña, E. Cuevas-Ferrando, W. Randazzo, I. Falcó, A. Allende and G. Sánchez, Comparing analytical methods to detect SARS-CoV-2 in wastewater, *Sci. Total Environ.*, 2021, **758**, 143870.

