12-1-2022

Estimating sewage flow rate in Jefferson County, Kentucky, using machine learning for wastewater-based epidemiology applications

Dhiraj Kanneganti

Lauren E. Reinersman

Rochelle H. Holm
University of Louisville, rochelle.holm@louisville.edu

Ted Smith

Original Publication Information
Dhiraj Kanneganti, Lauren E. Reinersman, Rochelle H. Holm, Ted Smith; Estimating sewage flow rate in Jefferson County, Kentucky, using machine learning for wastewater-based epidemiology applications. *Water Supply* 1 December 2022; 22 (12): 8434–8439. doi: https://doi.org/10.2166/ws.2022.395

Water Supply

Water Supply Vol 00 No 0, 1 doi: 10.2166/ws.2022.395

Estimating sewage flow rate in Jefferson County, Kentucky using machine learning for wastewater-based epidemiology applications

Dhiraj Kanneganti[a], Lauren E. Reinersman[b], Rochelle H. Holm [b,*] and Ted Smith[b]

[a] duPont Manual High School, 120 West Lee St., Louisville, KY 40208, United States
[b] Christina Lee Brown Envirome Institute, School of Medicine, University of Louisville, 302 E. Muhammad Ali Blvd., Louisville, KY 40202, United States
*Corresponding author. E-mail: rochelle.holm@lousiville.edu

RHH, 0000-0001-8849-1390

ABSTRACT

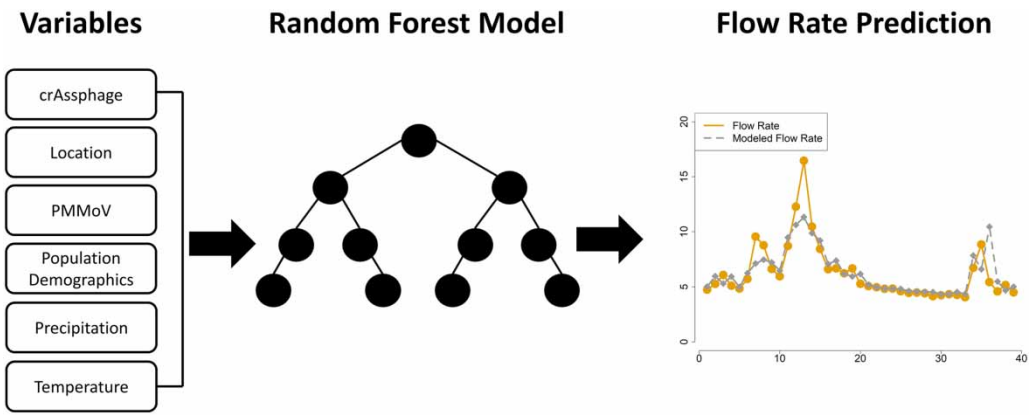
Direct measurement of the flow rate in sanitary sewer lines is not always feasible and is an important parameter for the normalization of data used in wastewater-based epidemiology applications. Machine learning to estimate past wastewater influent flow rates supporting public health applications has not been studied. The aim of this study was to assess wastewater treatment plant influent flow rates when compared with weather data and to retrospectively estimate flow rates in Louisville, Kentucky (USA), based on other data types using machine learning. A random forest model was trained using a range of variables, such as feces-related indicators, weather data that could be associated with dilution in sewage systems, and area demographics. The developed algorithm successfully estimated the flow rate with an accuracy of 91.7%, although it did not perform as well with short-term (1-day) high flow rates. This study suggests using variables such as precipitation (mm/day) and population size are more important for wastewater flow estimation. The fecal indicator concentration (cross-assembly phage and pepper mild mottle virus) was less important. Our study challenges currently accepted opinions by showing the important public health potential application of artificial intelligence in wastewater treatment plant flow rate estimation for wastewater-based epidemiological applications.

Key words: COVID-19, flow, model, random forest model, sewer, wastewater-based epidemiology

HIGHLIGHTS

- Machine learning to estimate wastewater influent flow rates has not been studied for wastewater-based epidemiology applications.
- Five wastewater treatment plants in Louisville, KY, USA, were studied to provide training and testing data sets of measured flow.
- The random forest algorithm to estimate past flow rate had an 91.7% accuracy.
- Artificial intelligence has potential applications in wastewater-based epidemiology.

GRAPHICAL ABSTRACT



INTRODUCTION

On March 11, 2020, the World Health Organization (2020) declared the spread of coronavirus disease 2019 (COVID-19) as a global pandemic. Conventionally, wastewater-based epidemiology (WBE) for severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) relied on the flow rate of wastewater treatment plant influent as a critical variable for standardized data reporting (McClary-Gutierrez *et al.* 2021). While these facilities have flow measurement systems for regulatory sampling (United States Environmental Protection Agency 2017), in-network sanitary sewer line upstream locations, such as streetline manholes, often lack equipment and access (McClary-Gutierrez *et al.* 2021), requiring the modelling and estimation of flow rate. The conventional approach for adjusting the level of SARS-CoV-2 RNA for dilution due to stormwater and other factors is to divide it by flow rate. Alternatively, when the flow rate is not available, imperfect, and variable measures of other fecal indicators, such as cross-assembly phage (crAssphage) and pepper mild mottle virus (PMMoV), have been used (Holm *et al.* 2022a, 2022b). A more comprehensive strategy would involve multiple inputs, including biological, meteorological, seasonal, and geographical data. Such a complex array of candidate inputs can be used to train a machine-learning algorithm aiding the flow rate estimation. Applications of machine learning to estimate wastewater influent flow rates at sampling locations supporting public health applications have not been reported previously.

The aim of this study was to assess: (1) wastewater treatment plant influent flow rates when compared with meteorological data [minimum or maximum daily temperature (°C); or precipitation (mm/day)]; and (2) to retrospectively estimate flow rates at wastewater treatment plants with a range of other data types using machine learning.

METHODS

Study site

Jefferson County, Kentucky (USA), contains five wastewater treatment plants (Table 1) that cover approximately 97% of the county population (Holm *et al.* 2022b). The Morris Forman Water Quality Treatment Plant is the only facility that is a combined sewer system, meaning that it combines wastewater and stormwater into the same piped network and is particularly susceptible to changes in influent flow rates due to precipitation. The other four facilities have separate piping networks for stormwater and wastewater.

Data

Daily influent flow rate data were provided by the wastewater utility, Louisville/Jefferson County Metropolitan Sewer District, from January 1, 2019, to December 31, 2021. Temperature and precipitation data georeferenced for each wastewater treatment plant from August 1, 2020 through June 16, 2021, at 15-minute resolution, were provided by a commercial service, Tomorrow.io (Boston, Massachusetts, USA), which uses the wireless network infrastructure to collect weather data. The crAssphage (copies/mL) and PMMoV (copies/mL) concentrations were obtained from Holm *et al.* (2022a, 2022b). Population, income, and race/ethnicity data for each wastewater treatment plant area were obtained from the United States Census Bureau (2020).

Table 1 | Characteristics of the studied wastewater treatment plants and associated areas

Water Quality Treatment Plant	Combined sewer	Income ($)[a]	Population[a]	Area (km^2)	2019 Mean flow rate (MGD)	2020 Mean flow rate (MGD)	2021 Mean flow rate (MGD)
Cedar Creek	No	76,606	55,928	80	5	6	6
Derek R. Guthrie	No	53,577	295,910	332	45	49	37
Floyds Fork	No	113,699	32,460	88	4	4	3
Hite Creek	No	106,769	31,269	67	5	5	5
Morris Forman	Yes	54,138	349,850	280	89	81	97

MGD, Million Gallons per Day.

[a]Based on 2018 United States Census Bureau American Community Survey (ACS). Income is mean median household.

Model

The random forest machine learning algorithm considered variables of crAssphage concentration (copies/mL), PMMoV concentration (copies/mL), site air temperature (°C) at 12:00 PM (midnight), location, site precipitation (mm/day), as well as area population size, income, and race/ethnicity. The three main software packages used to create the random forest model were NumPy, Pandas, and Scikit-learn (McKinney 2010; Pedregosa *et al.* 2011; Harris *et al.* 2020). NumPy is a software library that allows the user to store data in arrays, which can then be manipulated using pandas. Scikit-learn was used to import the default random forest model with no preset hyperparameters into the local Python development environment. The random forest model was constructed using the scikit-learn package (Kensert *et al.* 2018). The training group was randomly generated to contain 80% of the data from August 18, 2020 through June 16, 2021, excluding data from April 23 to May 31, 2021, and the testing group contained the remaining 20%. The groups were chosen randomly using the scikit-learn function. The following hyperparameters were optimized to best analyze the wastewater treatment plant data: maximum depth, maximum number of features, minimum number of samples per leaf, minimum number of samples per split, and number of estimators. The period of April 23 to May 31, 2021 was used to compare measured and estimated flow rate of the studied wastewater treatment plants.

Data analysis

Data analysis for yearly flow data was performed using Minitab Statistical Software (version 21.1.0.0; Minitab, LLC, State College, PA, USA). Plots were produced using R Studio (version 1.4.1106; R Core Team, Vienna, Austria).

Ethics

The University of Louisville Institutional Review Board classified this project as non-human subject research (reference #:717950).

RESULTS AND DISCUSSION

The daily flow rate (millions of gallons per day; MGD) at each of the five treatment plants from 2019 to 2021 is shown in Figure 1. For our study period during the COVID-19 pandemic, Spearman correlation coefficients were found for each treatment plant, relating the flow rate with minimum and maximum daily temperature (°C), and with precipitation (mm/day). There was no good predictive value of minimum or maximum daily temperature and flow rate; weak correlation coefficients were observed across the treatment plants and a p-value < 0.05, except for two cases [flow rate and minimum daily

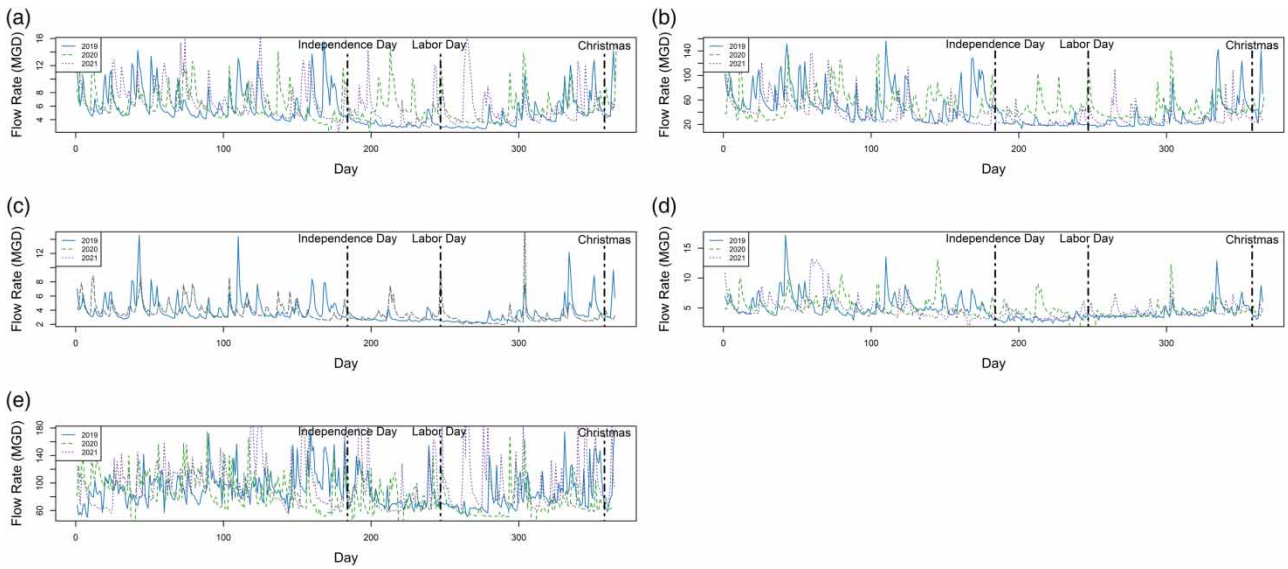


Figure 1 | Treatment plant flow rate by year, with referenced holidays. (a) Cedar Creek Water Quality Treatment Plant; (b) Derek R. Guthrie Water Quality Treatment Plant; (c) Floyds Fork Water Quality Treatment Plant; (d) Hites Creek Water Quality Treatment Plant; and (e) Morris Forman Water Quality Treatment Plant.

temperature at the Floyds Fork Water Quality Treatment Plant ($p = 0.063$) and Morris Forman Water Quality Treatment Plant ($p = 0.16$)].

For precipitation also, there was no good predictive value for flow rate; weak correlation coefficients were similarly observed across the treatment plants. Despite this weak relationship, the *p*-values for flow rate and precipitation were significant for each of the five treatment plants ($p < 0.05$). This common finding held despite that our five treatment plants were across a mixture of combined and separate sanitary sewer lines.

A machine learning algorithm was developed for three treatment plants from April 23 to May 31, 2021 (Figure 2) to estimate the daily flow rate (measured in MGD). The model accuracy was 91.7%. At Cedar Creek Water Quality Treatment Plant, the mean absolute difference was 0.57 MGD and the mean percent difference was 8.8%. For the Derek R. Guthrie Water Quality Treatment Plant, the model had a mean difference of 4.9 MGD and a mean percent difference of 8.2% between the actual and measured flow values. At the Floyds Fork Water Quality Treatment Plant, the mean absolute difference was 0.47 MGD and the mean percent difference was 11.1%.

Feature importance is the weight of the random forest model assigned to each of the input variables. Each variable affects the modelled flow rate differently, with more important variables having higher feature importance (Figure 3). Precipitation

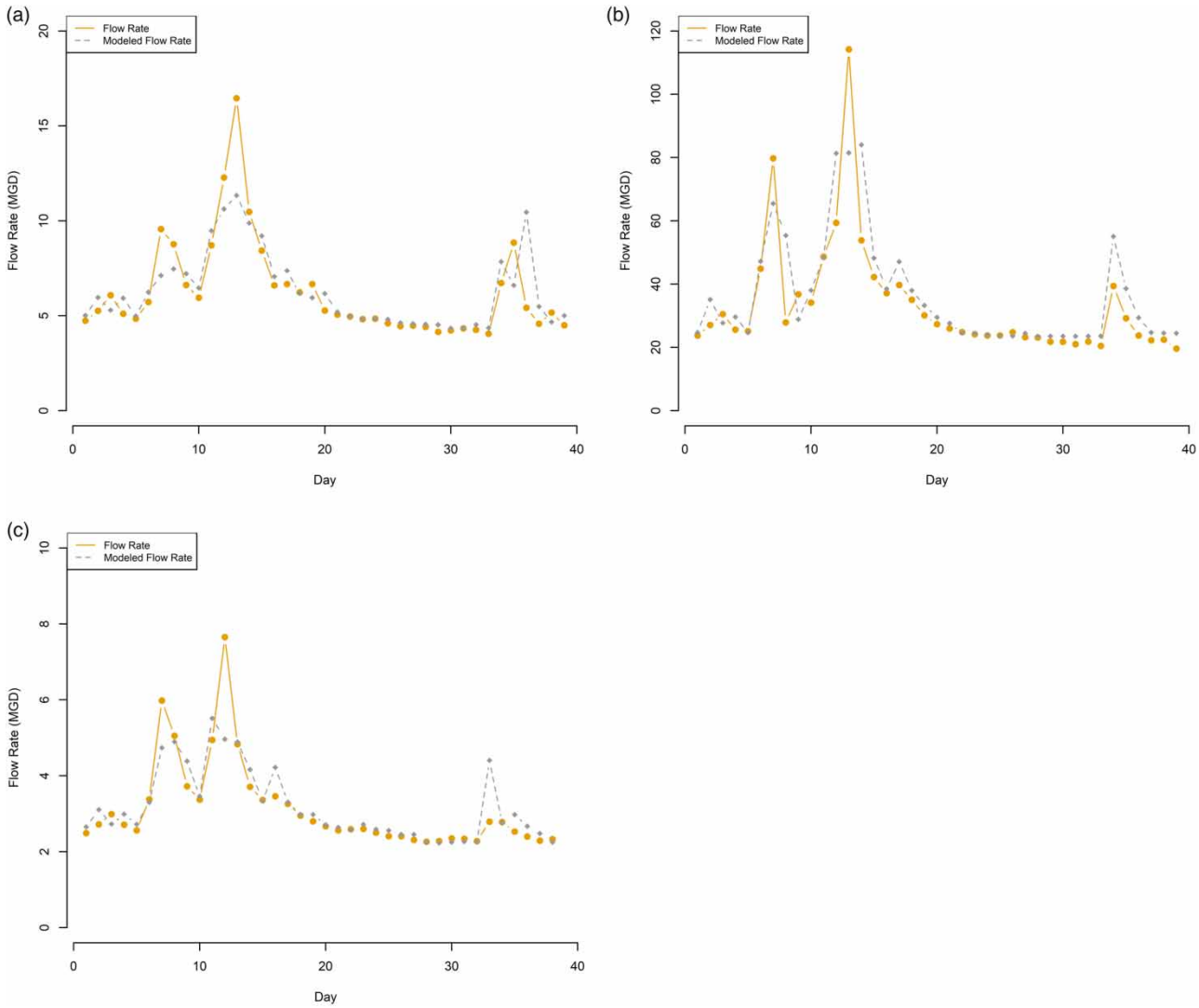


Figure 2 | Actual and estimated flow rate (millions of gallons per day) from the random forest algorithm. (a) Cedar Creek Water Quality Treatment Plant; (b) Derek R. Guthrie Water Quality Treatment Plant; and (c) Floyds Fork Water Quality Treatment Plant.

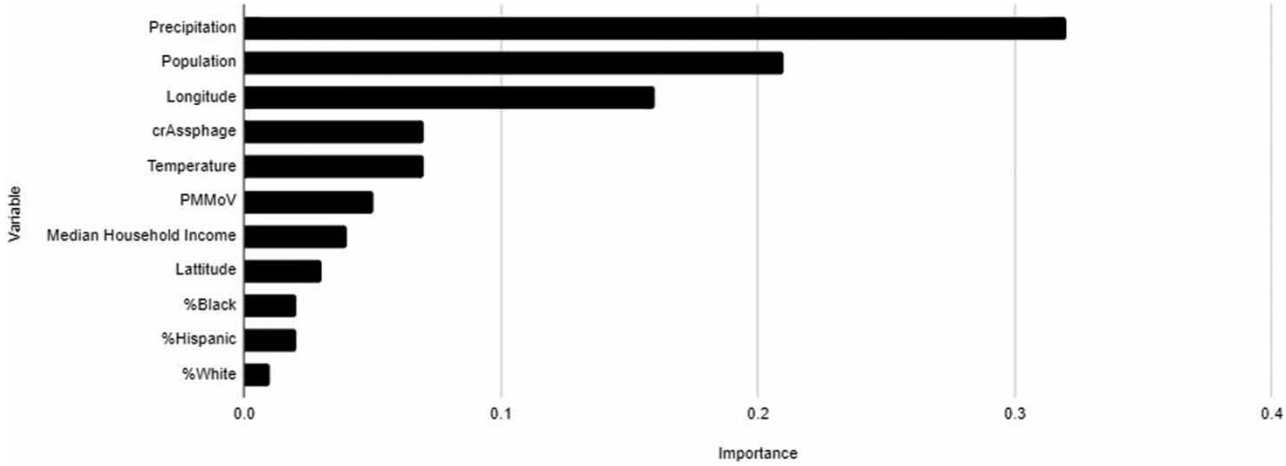


Figure 3 | Feature importance in wastewater flow rate random forest model.

had the greatest impact on the model, with a normalized feature importance of 0.32. This was consistent with our earlier findings of flow rates when compared with a single meteorological data type, but was surprising, as the model was used only on treatment plants having separate pipe networks of stormwater and sewage. The dependence on precipitation may indicate a key deficiency in the model for replication at other locations; for treatment plants that experience less precipitation, the model may be less effective. While other variables, such as population size and temperature, were also deemed important, losing access to quantifiable and regular precipitation data as a key variable may lead to a lack of estimation quality.

The random forest model accurately estimated the flow rates across the three treatment plants with varying population sizes and flow rates. However, at Derek R. Guthrie Water Quality Treatment Plant, the model unreliably calculated flow values over several days when there was a temporary 1-day high flow rate on May 5, 2021, the model under calculated that day by nearly 33 MGD (29% variance) and the +/− 1 day variance was 20–30 MGD.

The concentrations of the two fecal indicators, crAssphage and PMMoV, were not highly weighted by the model, indicating their lower importance when estimating the flow rate. This finding may be due to the large variance in the data (Holm *et al.* 2022), and therefore less useful for estimating retrospective flow rates. However, in situations where precipitation is no longer a usable variable, owing to a drier climate or georeferenced precipitation data not being available, the use of crAssphage and PMMoV concentrations may provide some useful information.

FUTURE RESEARCH

While this research provides preliminary insight into the potential application of machine learning for wastewater-based epidemiology in settings where the flow rate is unavailable, there are a few key areas that require further investigation. First, alternative machine learning models can be used. Algorithms, such as artificial neural networks, recurrent neural networks, and support vector machines, may produce different results. Second, while the random forest model may have proven to be effective in this study, the dataset used to train the model was limited to a temporal scale and pandemic conditions may have influenced the flow rates. Finally, the application of machine learning models to wastewater sampling locations upstream of treatment plants, mainly streetline manholes, may be most useful in areas that routinely lack in-place flow-rate measuring equipment.

CONCLUSIONS

Wastewater systems exhibit a regular flux of flow rates over time. The use of a machine learning model to retrospectively estimate the flow rate resulted in an algorithm with an accuracy of 91.7%. For future machine learning applications to estimate wastewater flow rates, the present study suggests prioritizing including variables of site precipitation (mm/day) and population size, but fecal indicators (crAssphage and PMMoV) were found to be less important. A model limitation is that a single high flow event could affect three days of estimation, and there may be some high flow rate temporal resolution

boundaries. Our study challenges currently accepted opinions by showing the potential application of artificial intelligence in wastewater treatment plant flow rate for wastewater-based epidemiology applications.

ACKNOWLEDGEMENTS

We thank the Louisville/Jefferson County Metropolitan Sewer District for their valuable collaboration for the wastewater sample collection. We would also like to thank Dr Andrew Karem for insight into creating the machine learning model.

FUNDING

This work was supported by a contract from the Louisville-Jefferson County Metro Government as a component of the Coronavirus Aid, Relief, and Economic Security Act, as well as grants from the James Graham Brown Foundation and Owsley Brown II Family Foundation. The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

DATA AVAILABILITY STATEMENT

All relevant data are included in the paper or its Supplementary Information.

CONFLICT OF INTEREST

The authors declare there is no conflict.

REFERENCES

Harris, C., Millman, K. & van der Walt, S. *et al.* 2020 Array programming with NumPy. *Nature* **585**, 357–362. doi:10.1038/s41586-020-2649-2.

Holm, R. H., Nagarkar, M., Yeager, R., Talley, D., Chaney, A., Rai, J. P., Mukherjee, A., Rai, S. N., Bhatnagar, A. & Smith, T. 2022a Surveillance of RNase P, PMMoV, and CrAssphage in wastewater as indicators of human fecal concentration across urban sewer neighborhoods, Kentucky. *FEMS Microbes* **3**, xtac003. doi:10.1093/femsmc/xtac003.

Holm, R. H., Mukherjee, A., Rai, J. P., Yeager, R., Talley, D., Rai, S. N., Bhatnagar, A. & Smith, T. 2022b SARS-CoV-2 RNA abundance in wastewater as a function of distinct urban sewershed size. *Environmental Science: Water Research & Technology* **8**, 807–819. doi:10.1039/d1ew00672j.

Kensert, A., Alvarsson, J., Norinder, U. & Spjuth, O. 2018 Evaluating parameters for ligand-based modeling with random forest on sparse data sets. *Journal of Cheminformatics* **10** (1). doi:10.1186/s13321-018-0304-9.

McClary-Gutierrez, J. S., Aanderud, Z. T. & Al-faliti, M. *et al.* 2021 Standardizing data reporting in the research community to enhance the utility of open data for SARS-CoV-2 wastewater surveillance. *Environmental Science: Water Research & Technology* **7**, 1545–1551. doi:10.1039/D1EW00235J.

McKinney, W. 2010 Data structures for statistical computing in Python. In *Proceedings of the 9th Python in Science Conference*. Vol. 445, pp. 56–61.

Pedregosa, F., Varoquaux, G. & Gramfort, A. *et al.* 2011 Scikit-learn: machine learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830.

R Core Team 2019 *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

United States Census Bureau 2020 *2018 American Community Survey (ACS)*. U.S. Department of Commerce. Available from: https://www.census.gov/data/developers/data-sets/acs-5year.html (accessed 21 June 2021).

United States Environmental Protection Agency 2017 *Wastewater Sampling*. Available from: https://www.epa.gov/sites/default/files/2017-07/documents/wastewater_sampling306_af.r4.pdf (accessed 27 July 2022).

World Health Organization 2020 *WHO Director-General's Opening Remarks at the Media Briefing on COVID-19–11 March 2020*. Available from: https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19—11-march-2020 (accessed 27 July 2022).

First received 31 March 2022; accepted in revised form 11 November 2022. Available online 16 November 2022