Using the right wastewater characteristics for early COVID-19 pandemic warning and forecast using deep machine-learning.

J.D. Therrien[1*], T. Maere[1], S. Halle[2], P. Dallaire[1,3], P.A. Vanrolleghem[1]

[1]modelEAU, Université Laval, Québec City, QC, G1V 0A7 Canada

[2]Thales Group, Québec City, QC, Canada

[3]SmartyfAI, Québec City, QC, Canada

**Corresponding author email:* jean-david.therrien.1@ulaval.ca

Highlights

- A deep ML model based on epidemiological and wastewater data was developed for 1-7 day forecasting of COVID-19 case counts and tested on Québec City's third wave.
- A novel method was used to characterize the added value of different wastewater characteristics for case prediction.
- Wastewater data was found to increase 5-7 day forecast accuracy by 40% compared to models using only epidemiological data.

Introduction

The goal of this project was to evaluate the potential of wastewater analyses to improve COVID-19 forecasts and to create a reusable model that can be used for monitoring the remaining of the COVID-19 pandemic, future pandemics, or any other population exposure to illness, drugs, and other agents.

Methodology

Context

The wastewater-based epidemiology (WBE) community now agrees on a proven lead time of four to six days for WBE over clinical tests (Bibby et al., 2021). Unfortunately, the composition of wastewater is not constant, and neither is the decay of RNA viruses as they travel down the sewers (Ahmed et al., 2020a). Therefore, WBE brings with it many challenges which are addressed here using ML techniques.

Objectives

- Provide a short-term (1 –7 day) forecast of COVID-19 case counts using a deep ML model.
- Determine whether the performance of this model can be enhanced using wastewater data.
- Determine what wastewater characteristics are most helpful to produce accurate predictions.

Data collection

The epidemiological data used for training are released by the Québec province public health agency. The wastewater data originate from sampling campaigns undertaken at the two main wastewater treatment plants of Québec City ("East" and "West"). SARS-CoV-2 RNA was measured using a RT-qPCR assay targeting the N2 gene of the virus (Ahmed et al., 2020b). A viral faecal indicator, the Pepper Mild Mottle Virus (PMMV), was measured as internal control. Wastewater was characterized daily at the treatment plants.

Model Structure

The model structure used for this study relied on multiple convolutional layers (Figure 1a). At a time instant t_i, the model takes in wastewater data from the preceding 14 days as input and projects the data into a multi-dimensional manifold space. Then, from the shared projection, three task-specific heads extract a vector of predictions for days 1-7 after t_i. The multiple heads encourage generalization across the common layers. Each head predicts a metric produced by public health authorities (daily case counts, total active cases, percentage of positive test results).

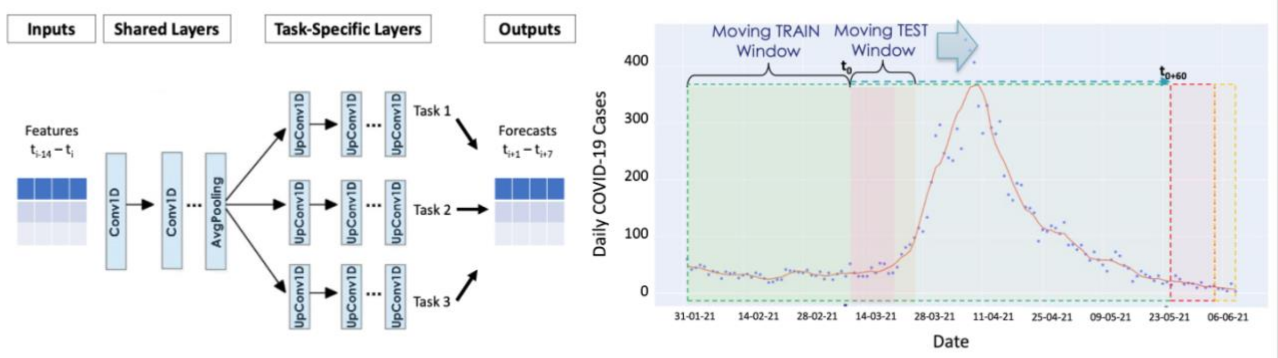

Figure 1. a) Model Structure of the proposed deep machine-learning model. b) Evolution of the training and testing windows throughout the training loop.

Selection of features

The proposed model structure makes it simple to create multiple candidate models that differ only in their input features. This makes it possible to compare the added value provided by different inputs by comparing model versions that contain the feature to others that don't. The different candidate input features used to analyse the added value of wastewater characteristics are shown in Table 1.

Table 1. Versions of the model compared in this study

Recipes	Public health features	Wastewater features
No WW	Reported cases, Active cases, Test positivity ratio	None
N2 All	Idem	N2; N2 x flow; N2 / PMMV
Flow Load	Idem	N2 x flow; PMMV x flow; COD x flow; NH_4^+-N x flow; BOD_{5C} x flow
Flow Load Pre-trained	Same as "Flow Load" Model: trained on "Québec East" and tested on "Québec West"	

Training and testing

One of the major difficulties encountered in this study was the scarcity of good quality data (roughly 120 days, which include a single pandemic wave, across 2 sampling sites, Figure 1b). To maximize learning, the model was first trained on data from February 1 to March 1. The model was then tested on data it had never seen by trying to predict cases after March 14. The training window was then expanded by a day, the model was retrained, and the model was tested again for March 15, and so on until May 31 (see Figure 1b). Training was performed on "Québec East". To test whether the model would perform well on a new case wave, the model was tested on "Québec West", which has the same case data but different wastewater data.

Results and discussion

The results, shown in Figure 2 and 3, show that all models perform similarly on the short-term forecast. However, for longer-term forecasts, the models using wastewater data, and more particularly the "Flow Load" model, perform much better. More precisely, the "Flow Load" model reduces the error of the "No WW" model by 42% on its 6-day forecast. Additionally, our comparative Mean Absolute Error (MAE) results demonstrate that the pre-trained model performs much better and provides large reductions in MAE on 4-7-day forecasts. It can also be seen in Figure 3 that the Flow Load models perform well at the beginning and crest of the wave, which are critical moments for pandemic management.

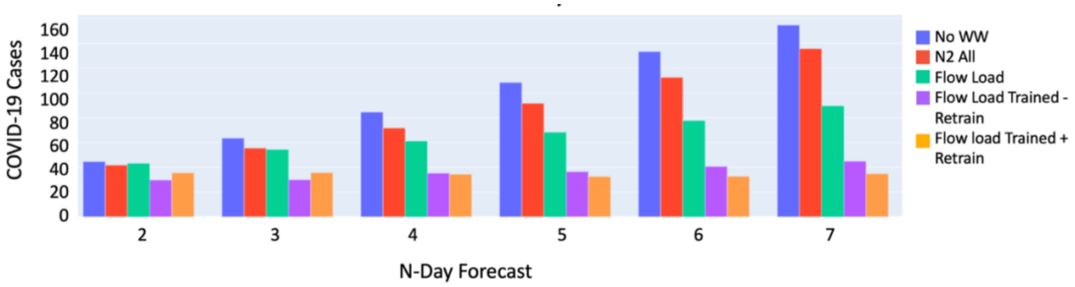

Figure 2. Mean Absolute Error of daily case forecasts for different model versions

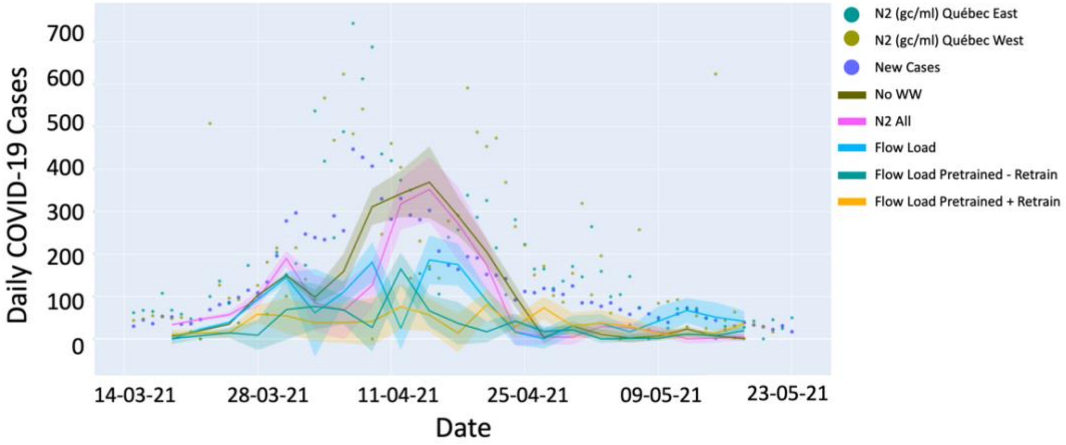

Figure 3. Absolute Error of different model versions over time

Conclusions and future work

The solution developed in this study was shown to be a useful, complementary and, on many occasions, better solution than traditional COVID-19 forecasting methods. More importantly, it was demonstrated that it can detect the start and the end of a wave, representing, for public health authorities, the most important events in a pandemic. Future work will test multiple model recipes to identify the features that most improve the forecast. Improving the model's ability to transfer what it learns from one site to another would also allow accounting for the particularities and similarities between different sewer networks and populations. A strong multi-site model would favour both the initial model's overall performance by benefiting from a large quantity of training data, as well as providing all communities with forecasting models aware of new patterns (variants and vaccinations and such) first observed on other sites.

References

Ahmed, W., Bertsch, P.M., Bibby, K., Haramoto, E., Hewitt, J., Huygens, F., Gyawali, P., Korajkic, A., Riddell, S., Sherchan, S.P., Simpson, S.L., Sirikanchana, K., Symonds, E.M., Verhagen, R., Vasan, S.S., Kitajima, M., Bivins, A., 2020a. Decay of SARS-CoV-2 and surrogate murine hepatitis virus RNA in untreated wastewater to inform application in wastewater-based epidemiology. Environmental Research 191, 110092. https://doi.org/10.1016/j.envres.2020.110092

Ahmed, W., Bertsch, P.M., Bivins, A., Bibby, K., Farkas, K., Gathercole, A., Haramoto, E., Gyawali, P., Korajkic, A., McMinn, B.R., Mueller, J.F., Simpson, S.L., Smith, W.J.M., Symonds, E.M., Thomas, K. v., Verhagen, R., Kitajima, M., 2020b. Comparison of virus concentration methods for the RT-qPCR-based recovery of murine hepatitis virus, a surrogate for SARS-CoV-2 from untreated wastewater. Science of The Total Environment 739, 139960. https://doi.org/10.1016/j.scitotenv.2020.139960

Bibby, K., Bivins, A., Wu, Z., North, D., 2021. Making waves: Plausible lead time for wastewater based epidemiology as an early warning system for COVID-19. Water Research 202, 117438. https://doi.org/10.1016/j.watres.2021.117438