Wastewater-based epidemiology for COVID-19 using dynamic artificial neural networks

Jesús M. Zamarreño [a,c,*], Andrés F. Torres-Franco [a,b,*], José Gonçalves [a,b], Raúl Muñoz [a,b], Elisa Rodríguez [a,b], José María Eiros [d], Pedro García-Encina [a,b]

[a] *Institute of Sustainable Processes, Dr. Mergelina, s/n, 47011 Valladolid, Spain*
[b] *Department of Chemical Engineering and Environmental Technology, School of Industrial Engineering, Universidad de Valladolid, C/ Dr. Mergelina, s/n, 47011 Valladolid, Spain*
[c] *Department of System Engineering and Automatic Control, School of Industrial Engineering, Universidad de Valladolid, C/ Dr. Mergelina s/n, 47011 Valladolid, Spain*
[d] *Microbiology Service, Hospital Universitario Río Hortega, Gerencia Regional de Salud, Paseo de Zorrilla 1, 47007 Valladolid, Spain*

HIGHLIGHTS

- SARS-CoV-2 RNA in wastewater showed a dynamic correlation with hospitalizations.
- A dynamic ANN model was designed and trained using N1 and 1st-dose vaccinations.
- The ANN model provided accurate predictions of hospitalizations and risk levels.
- ANN modeling is a powerful tool for water-based epidemiology assessments.

GRAPHICAL ABSTRACT
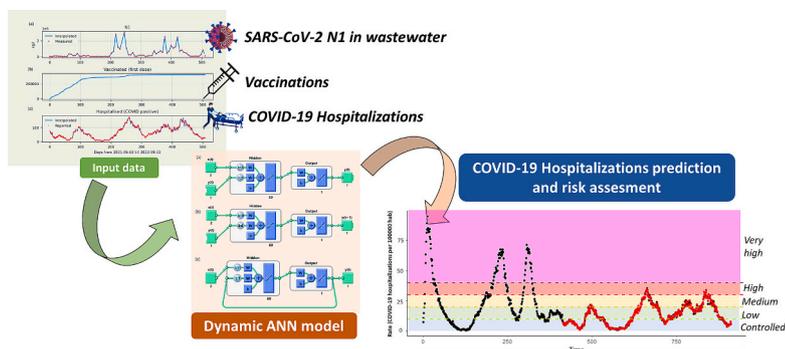
ARTICLE INFO

Editor: Warish Ahmed

ABSTRACT

Global efforts in vaccination have led to a decrease in COVID-19 mortality but a high circulation of SARS-CoV-2 is still observed in several countries, resulting in some cases of severe lockdowns. In this sense, wastewater-based epidemiology remains a powerful tool for supporting regional health administrations in assessing risk levels and acting accordingly. In this work, a dynamic artificial neural network (DANN) has been developed for predicting the number of COVID-19 hospitalized patients in hospitals in Valladolid (Spain). This model takes as inputs a wastewater epidemiology indicator for COVID-19 (concentration of RNA from SARS-CoV-2 N1 gene reported from Valladolid Wastewater Treatment Plant), vaccination coverage, and past data of hospitalizations. The model considered both the instantaneous values of these variables and their historical evolution. Two study periods were selected (from May 2021 until September 2022 and from September 2022 to July 2023). During the first period, accurate predictions of hospitalizations (with an overall range between 6 and 171) were favored by the correlation of this indicator with N1 concentrations in wastewater ($r = 0.43$, $p < 0.05$), showing accurate forecasting for 1 day ahead and 5 days ahead. The second period's retraining strategy maintained the overall accuracy of the model despite lower hospitalizations. Furthermore, risk levels were assigned to each 1 day ahead

---

* Corresponding authors at: Institute of Sustainable Processes, Dr. Mergelina, s/n, 47011 Valladolid, Spain.
 E-mail addresses: jmzamarreno@uva.es (J.M. Zamarreño), andresfelipe.torres@uva.es (A.F. Torres-Franco).

prediction during the first and second periods, showing agreement with the level measured and reported by regional health authorities in 95 % and 93 % of cases, respectively. These results evidenced the potential of this novel DANN model for predicting COVID-19 hospitalizations based on SARS-CoV-2 wastewater concentrations at a regional scale. The model architecture herein developed can support regional health authorities in COVID-19 risk management based on wastewater-based epidemiology.

1. Introduction

Wastewater-based epidemiology (WBE) is based on detecting and quantifying chemical and biological biomarkers that provide valuable information on a community's health and lifestyle habits (Sims and Kasprzyk-Hordern, 2020). With the recent COVID-19 pandemic, WBE emerged as a useful tool for COVID-19 surveillance and prediction at the local scale, constituting an indirect population-level diagnosis tool (Polo et al., 2020). COVID-19 WBE is based on SARS-CoV-2 RNA detection and quantification in wastewater, as the virus is excreted into the sewer system via the feces of infected individuals (Jones et al., 2020). In many studies, SARS-CoV-2 concentrations in wastewater were significantly correlated with COVID-19 cases, particularly in hospitals (de Araújo et al., 2023) or elderly residences (Pico-Tomàs et al., 2023). Furthermore, mathematical models have been applied to associate RNA excretion with COVID-19 cases (Vallejo et al., 2020), showing an estimated early prediction capacity of cases ranging from 2 to 24 days, depending on variables such as clinical evidence, wastewater sampling frequency, duration of viral excretion, and environmental conditions (Jiang et al., 2022).

Despite this growing body of work assessing SARS-CoV-2 RNA concentrations in wastewater, few studies have created epidemiological models or environmental surveillance methods that successfully correlate SARS-CoV-2 RNA and the number of COVID-19 cases (either hospitalized or deceased individuals) quantified by health administration services (Galani et al., 2022), indicating the complexity of estimating associations between RNA shedding and cases. Besides parametric algorithms, autoregressions, moving average methods, non-parametric approaches such as SEIR (e.g., IHME, 2021; Rodríguez et al., 2021; Patón et al., 2022), and especially, artificial neural networks (ANN) (Haykin, 1999) have been used as modeling tools for case number predictions (Jiang et al., 2022; Rauch et al., 2022). These last models imitate the human brain by finding the best parametrization of nodes, layers and functions through trial and error methods and are widely used as a universal functional approximator (Ghiassi et al., 2005). During the COVID-19 pandemic, several ANN models were applied worldwide for case forecasting or sewer epidemiology. For instance, ANN-based methods were applied in Utah, USA, achieving high precision in estimating the number of cases and effective reproduction rate using viral load, vaccination rate, clinical testing positive rate, and weather data as model inputs (Jiang et al., 2022). ANN was also applied in forecasting COVID-19 cases in Brazil (Braga et al., 2021), ICU hospitalization and mortality in Greece (Asteris et al., 2022), as well as daily patients and deaths using historical data from various countries (Kuvvetli et al., 2021). The latter surveyed different approaches for predicting COVID-19 indicators based on ANN. Their methodology considered 10 days forecast based on 3 ANN models that resulted in 10 outputs (one for each forecast).

Overall, these previous works in which ANN investigated the relationship of COVID-19 cases with the concentrations of SARS-CoV-2 RNA in wastewater were based on a static approach in terms of ANN architecture. However, the intrinsically dynamic nature of COVID-19 cases (or hospitalizations or deaths) and RNA concentration in wastewater suggest that they are more likely to follow an intrinsically dynamic behavior. In this sense, Dynamic ANN (DANN) models, which can rapidly adapt to these variations, may result in better estimations since they can learn and gather knowledge at each layer, propagating and modifying this knowledge to the following layer and repeating these

actions until the desired network performance criteria are met (Ghiassi et al., 2005).

In this context, a novel approach based on a DANN model of the NARX (Nonlinear Auto-Regressive with eXogenous inputs) type was developed and validated to predict the number of hospitalizations based on the evolution of recent historical data, considering WBE indicators and vaccination rates as inputs in order to evaluate the association between the model's outcomes and the risk levels reported by regional health authorities.

2. Material and methods

2.1. Area of study

The area of study was located in the province of Valladolid (Castilla y Leon, Spain), with nearly 520,000 inhabitants. About 90 % of inhabitants had all COVID-19 hospitalizations assigned to Valladolid University Clinical Hospital and the Río Hortega University Hospital, located in Valladolid's capital. The other 10 % of the province's population was assigned to the Medina del Campo Hospital (not included in this study). About 80 % of the population assigned to the University Clinical Hospital and the Río Hortega University Hospital had their municipal wastewater treated in the Valladolid wastewater treatment plant (WWTP), including the municipalities of Valladolid (capital), Arroyo de la Encomienda, La Cistérniga, Laguna de Duero, Simancas, and Zaratán. The study area is presented in Fig. 1, showing the province population assigned to the Hospitals in Valladolid (capital), and whose wastewater is treated in Valladolid WWTP.

2.2. SARS-CoV-2 N1 RNA concentrations

SARS-CoV-2 N1 RNA concentrations in the influent wastewater to the Valladolid WWTP were retrieved from https://aquavall.es/sistema-de-alerta-temprana-para-la-deteccion-de-covid-19/. Aquavall has published N1 data after performing sampling collection, viral concentration, and N1 quantification by RT-qPCR, as indicated in Randazzo et al. (2020). The first study period spans from 5/2/2021 to 9/22/2022, with approximately a weekly reporting frequency. This first period was used for developing the model. A second study period was included to validate the model performance, comprising from 9/23/2022 to 7/7/2023.

2.3. Epidemiological data collection for the development of dynamic ANN

Since September 2022, the Spanish Public Health Service has only reported daily COVID-19-positive cases in individuals over 60 years old. Moreover, due to the easy accessibility to antigen auto tests in pharmacies and widespread vaccination coverage, a large number of cases are currently not reported to the Spanish Public Health Authorities. Therefore, in the current COVID-19 surveillance scenario in Spain, the reported number of infected individuals is no longer a reliable indicator as it may result in substantial underestimations of the ongoing cases. On the other hand, given that the primary concern of current surges in COVID-19 infections is the potential saturation of hospitals, the number of hospitalized cases has been considered a more reliable measure of the virus' incidence in the population, as hospitalization becomes unavoidable if a person's symptoms worsen. In addition, the regional health administration has recently established alert levels based on the number of daily new hospitalizations, facilitating a rapid qualitative

assessment of regional epidemiological situations (MSCBS, 2022). The vaccination status of the population is also expected to be an important attenuation parameter that influences the number of hospitalized cases as the vaccinated population acquires defenses against the most severe symptoms of the disease. In this context, the number of COVID-19 hospitalizations at the Valladolid University Clinical Hospital and the Río Hortega University Hospital (both located in the capital city) was retrieved from the regional health administration website (https://anali sis.datosabiertos.jcyl.es/pages/coronavirus/?seccion=descargas) using the provided API for data filtering and acquisition. As explained above, these hospitals were selected because they are located in the area served by Valladolid WWTP. Data on vaccinated individuals were retrieved from the same website; it was only accessible at the provincial level, and there are no reports at the municipal level.

Nevertheless, since 90 % of the province's population was assigned

to the studied hospitals, the total number of vaccinated individuals at the province level was regarded as a good estimation for the vaccination in the population in the studied area. Only the first dose was included in the model for vaccination status. The subsequent doses were ignored, assuming that the first doses were the main parameter affecting the increase in immunity during the study period, also simplifying the model by avoiding the consideration of heterogeneous time lags. The warning levels established by the regional health administration in Castilla y Leon (Spain), which are based on COVID-19 bed occupation in hospitals, were also retrieved (https://analisis.datosabiertos.jcyl.es/pa ges/coronavirus/?seccion=descargas) and compared with the corresponding level assigned to the prediction after applying the same criteria.
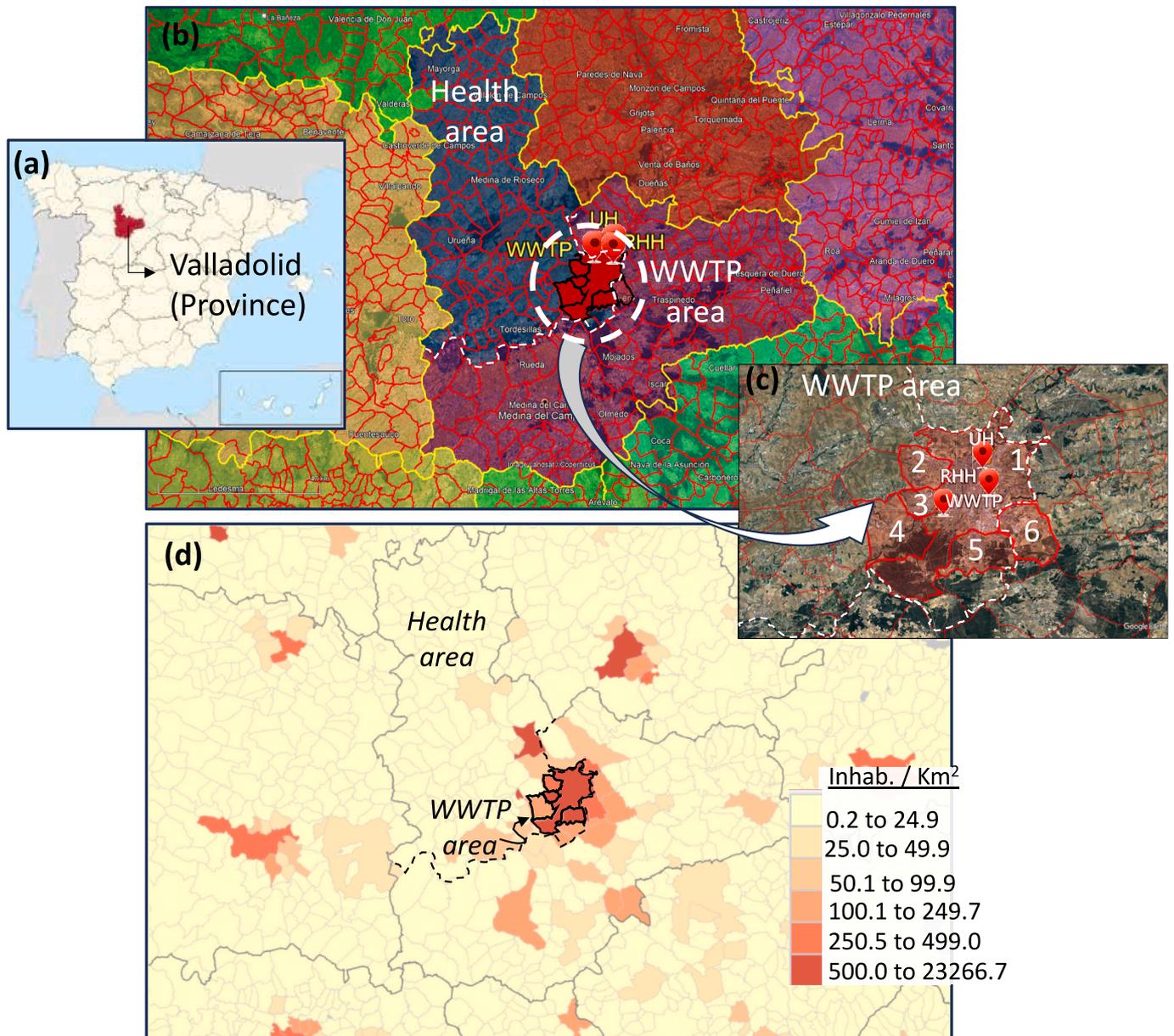

Fig. 1. Area of study. (a) Valladolid (Province). (b) The health administrative area assigned to the University Clinical Hospital (HP), Rio Hortega Hospital (RHH) and the area served by Valladolid WWTP, which includes (c) the municipalities of (1) Valladolid (capital), (2) Zaratán, (3) Arroyo de la encomienda; (4) Simancas and (5) Laguna de Duero. (d) Population densities indicate that the vast majority of inhabitants served by Valladolid WWTP were attended by Clinical Hospital (HP) and Rio Hortega Hospital (RHH). Data retrieved from https://idecyl.jcyl.es/geonetwork/static/api/records/SPAGOBCYLCITDTSHHARS and https://atlasau.mitma.gob. es/#bbox=-770943,5233384,457132,281228&c=indicator&i=pobevo.densidad&s=2021&view=map4

2.4. Associations between SARS-CoV-2 N1 concentrations and number of hospitalizations

Before the application of the ANN model, the level of association between SARS-CoV-2 N1 concentrations in wastewater and the number of hospitalizations in the raw data was assessed in terms of the statistical distribution of the data series and the correlation intensity at 0, 7 and 14 days in lead time. For the modeling step, the mismatch in the frequency of data series (i.e., daily records of hospitalizations but genetic footprint in wastewater reported twice a week) was addressed by applying a linear interpolation algorithm to estimate SARS-CoV-2 N1 concentrations on the days with no data reported (Fig. 2a). This interpolation was made using the class *scipy.interpolate.interp1d* in *Python* for each day between two measurements.

2.5. Dynamic ANN (DANN) model

To associate SARS-CoV-2 N1 concentrations in Valladolid WWTP influent wastewater with the number of COVID-19 hospitalizations, a DANN model of the NARX type was retrospectively applied for the period between May-2021 and Sept-2022, which corresponded to a period with high data availability and several variations in the risk levels associated to COVID-19 hospitalizations. Data on SARS-CoV-2 N1 concentrations, first-dose vaccination and the number of hospitalizations used for DANN modeling is presented in Fig. 2.

For developing the DANN model, it was assumed that the number of hospitalized patients depends on their own previous evolution, the past history of N1 concentrations and vaccination coverage. The justification for this election is as follows: Nowadays, everybody is familiar with the concept of infection waves, where the evolution of cases, hospitalized patients, and deaths follow a characteristic curve where any of these three indicators start rising exponentially, reaching a maximum and afterwards following a decreasing trend until a new wave comes. This means that the present value of any of these indicators depends on its previous trend. Similarly, the SARS-CoV-2 concentration at the given time will affect any of the three indicators in the following days, as it takes a few days from the date an individual becomes infected until it is detectable in the feces and a few more days until hospitalization is required. A similar reasoning applies to vaccination coverage and its influence on hospitalized patients.

For DANN models, an estimation of the number of required past inputs and past outputs for a more accurate estimation of the current output was essential. The number of past daily samples required as inputs to train the model was estimated following the method developed by He and Asada (1993), selecting the best value based on an order index. In this method, the appropriate model orders can be determined easily and reliably by evaluating the change of an index defined as a Lipschitz number with the successive change of model orders. This index improves (lower values) as long as more past values are considered until reaching a plateau, with no further decreases in the order index at increasing numbers of past values. Hence, there is a tradeoff between selecting a low value of the temporal horizon (simpler model) and enough temporal data for better model accuracy. By implementing this index by He and Asada (1993) in Matlab (Norgaard, 2023) through the *lipschit* function, it was inferred that considering more than 11 past inputs of N1 concentration and hospitalized patients resulted in no further decrease in the index (Fig. 3a). This result was confirmed by the 2D representation (Fig. 3b) of the order index, which did not experience a significant change after 11 past inputs. This number aligns with estimations reported in the literature (Jiang et al., 2022). A similar analysis was performed for the number of vaccinated persons (first dose). In this particular case, considering more than 7 past outputs did not significantly decrease the index (Fig. 3c). Although 3 past inputs could be enough in this context, 11 past values were considered due to the restrictions of the Deep Learning toolbox used for training (all the inputs must consider the same number of past values). Nevertheless, using more past inputs than required was not problematic, except for certain increase in the computational burden (the neural network will learn not to consider the inputs with less influence by decreasing their weights).

Similar to other works in the field (e.g., Zahmatkesh et al., 2023), a feedforward neural network was designed to identify the relationship between inputs and outputs, but instead of considering a static relationship, the design of a dynamic neural network where the past history of inputs and outputs determines the present output was preferred, as justified previously. This dynamic model is called NARX model and is mathematically represented as (Eqs. (1) and (2)):

$$\varphi(t) = [y(t-1)\cdots y(t-n)\; u(t-k)\cdots u(t-m-k+1)]^T \tag{1}$$

$$\widehat{y}(t|\theta) = \widehat{y}(t|t-1,\theta) = g(\varphi(t),\theta) \tag{2}$$

where $\varphi(t)$ is the regressor vector at time t, composed of n past outputs (y) and m past inputs (u) considering a possible delay k. $\widehat{y}(t|t-1,\theta)$ is the prediction of the output at time *t*, considering measurements known at time $t-1$ (in this case, this means 1 day earlier) and given some values of the model parameters θ. $g(\varphi(t),\theta)$ is the nonlinear model
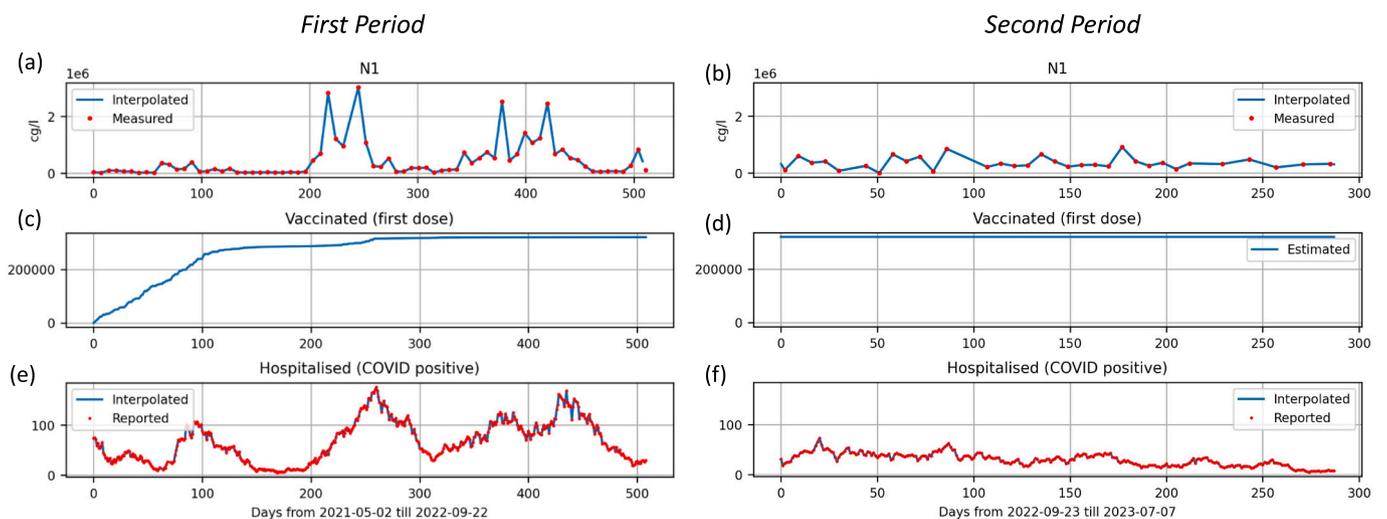

Fig. 2. Time course of (a-b) SARS-CoV-2 N1 RNA concentrations in influent wastewater of the Valladolid WWTP (Aquavall), (c-d) the number of vaccinated people (first dose), (e-f) number of COVID-19 hospitalizations. All data covers the period from May-2021 to Sept-2022 (first period) and from sept-2022 to July-2023 (second period).
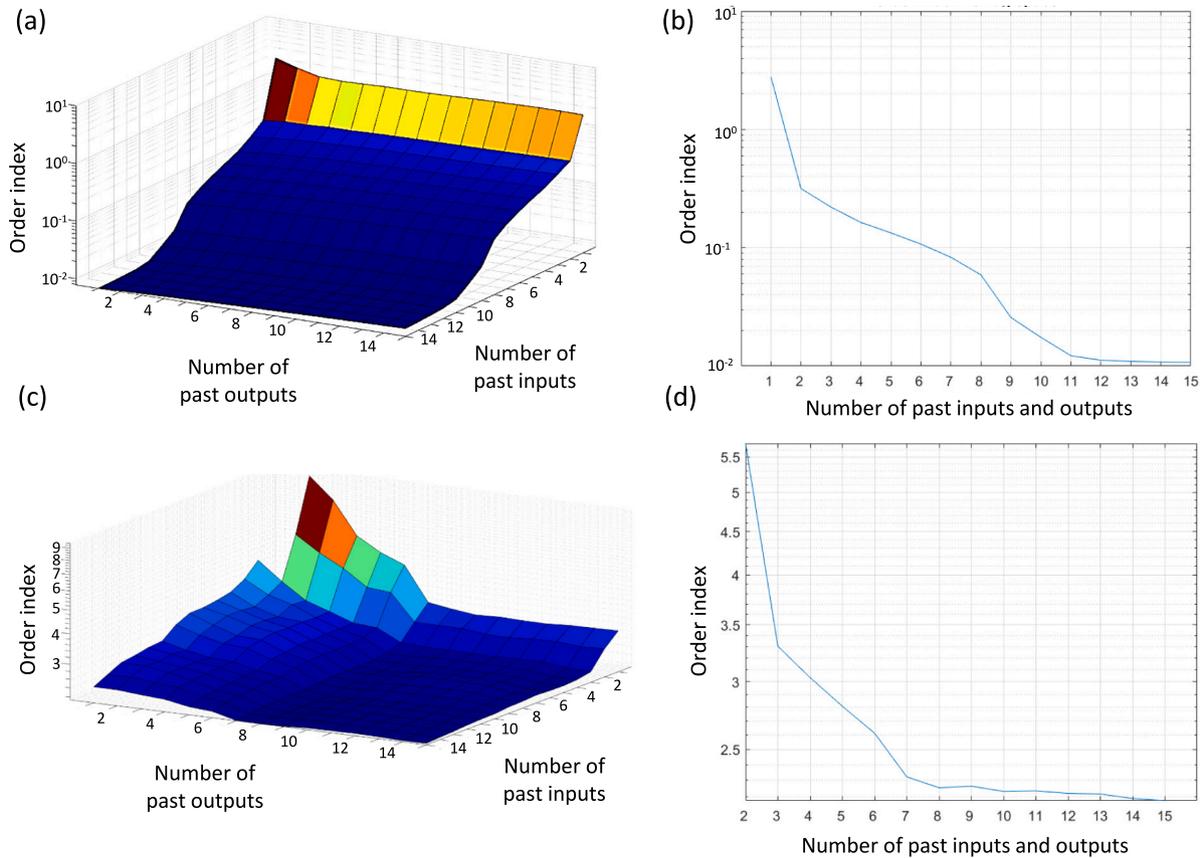
Fig. 3. Estimation of the number of past inputs (N1 concentration) and past outputs (hospitalized patients) (a) 3-D and (b) 2-D; Estimation of the number of past inputs (vaccinated persons) and past outputs (hospitalized patients). (c) 3-D and (d) 2-D.

function (in this study, an artificial neural network) that depends on the regressor vector and the model parameters. Hornik et al. (1989) demonstrate that an ANN with one hidden layer of sigmoid neurons can be considered a universal approximator. The neural network structure is represented in Supplementary Fig. A1, where each circle represents a processing element (neuron), and each connection is associated with a weight (parameter to be adjusted). The model included two inputs (SARS-CoV-2 N1 concentrations and number of vaccinated persons) and one output (number of hospitalized patients).

2.6. NARX architecture for prediction

For achieving a prediction capability using the NARX model as starting point (Fig. A1), some adjustments were made to the original model's architecture, as shown in Fig. 4a, in which $x(t)$ is the time series for the 2 inputs, $y(t)$ is the time series for the output, W and b are parameters of the NN (weights and bias, respectively) and 1:11 or 1:7 stand for the range of past values used in the calculations (in days). However, for forecasting and decision-making, as is our case, it is desirable to predict the output at time $t + 1$ (i.e., tomorrow) once that output at time t (today) is available. Thus, the architecture of the network was modified to return its output a timestep early by removing one delay so that its minimal tap delay is now 0 instead of 1. The new network returned the same outputs as the original network. However, results are shifted one timestep to the left (Fig. 4b), behaving like a model predicting the number of patients hospitalized one step (one day) in advance.

Although the ANN in Fig. 4b was trained to make one-step-ahead predictions, it can also be used to make long-term predictions, but in this case, more imprecision is expected. To be able to give more predictions in advance, the way to proceed is to feedback the predictions to the ANN input, i.e., the one-step-ahead prediction is fed back to the

input and a two-step-ahead prediction is obtained, which is repeated consecutively for more predictions. This is called multistep prediction. The previous architecture is used when past data are available and a closed loop (Fig. 4c) is used when it is not available, simulating a network in open-loop form as long as there are known output data and then switching to closed-loop form to perform multistep prediction while providing only the external inputs. To assess this multistep prediction capability, all but 5 timesteps of the input and target series were taken to simulate the network in open-loop form, taking advantage of the higher accuracy that the target series produces. Next, the network and its final states were converted to closed-loop form to make five predictions with only the five inputs provided.

Once the final model architecture was achieved, the number of hospitalized patients for each day was determined based on their own value for the preceding 7 days and the input values (vaccination coverage and N1 concentration) for the preceding 11 days, as previously determined with the order index.

2.7. Model training and back-estimation

After achieving the final model architecture, a training procedure was applied to adjust the neural network parameters. This procedure was applied by splitting the dataset into three subsets: training subset, validation subset, and test subset using interleaved indices through the *divideint* Matlab function (Deep Learning Toolbox). The training subset was used to improve the model through an optimization algorithm (usually called training procedure). Then, the validation subset was used to evaluate the model's performance on a different subset and set an early stop of the training algorithm to avoid overfitting. Finally, the test subset is reserved so, at the end, the model can be evaluated on fresh data that has not participated in the training procedure. The neural
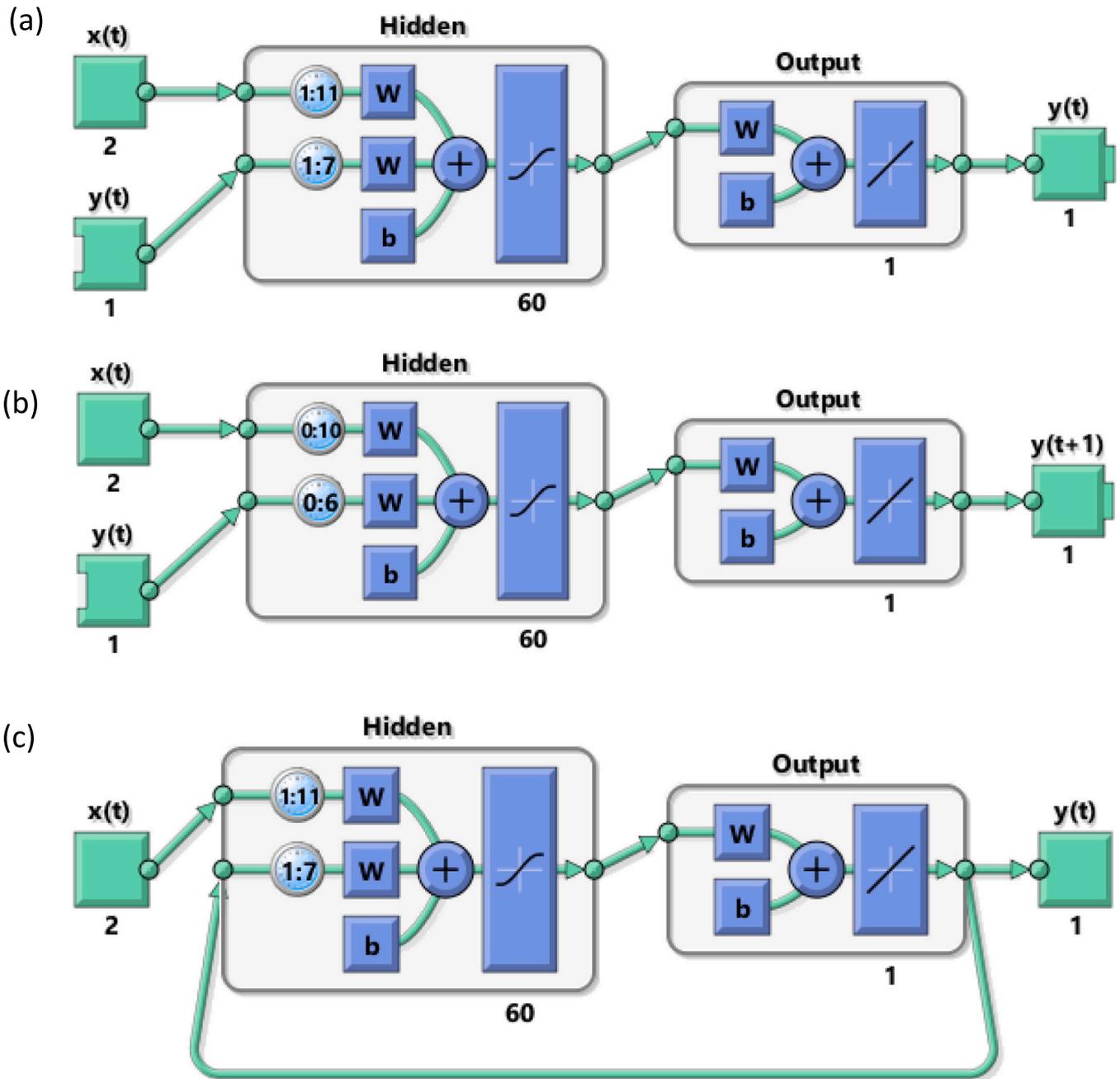
Fig. 4. (a) ANN model; (b) One step-ahead prediction ANN, and (c) Closed loop predictions. All three images were taken from the Deep Learning Toolbox, Matlab.

network's performance was assessed through the Mean Squared Error (MSE). The training procedure was carried out by the Levenberg-Marquardt algorithm, which is one of the most widely used due to its good performance and low memory footprint. As this type of algorithm is a local search optimization method, repeating the training procedure from different initial conditions is a good strategy for getting an optimal solution. The number of neurons in the hidden layer was selected by trial and error, using as selection criteria to minimize the MSE values achieved in the training and validation subsets.

After obtaining all parameters in the final architecture of the DANN model, it was applied to forecasting the number of hospitalized patients over the whole dataset, including the test subset. Furthermore, the model's capacity to predict COVID-19 cases under different virus circulation patterns was assessed in a second study period, allowing for testing the model in a real-life scenario. During this second period,

incremental retraining was performed every time new data was available, as it could include new information for the model, such as new phases with different epidemiological dynamics. A two-stage procedure was employed to test the model on this second period of data. Firstly, the model predicted the number of patients who would be hospitalized the following day. Secondly, once available, the current value of hospitalized patients for that day was incorporated into the training set to improve the model. This procedure was repeated daily, as if it were being used in real life, allowing the model's performance to be assessed using new data that was not included in the training phase. In this sense, all samples from this second data period were considered a new test subset.

2.8. Modeling software, error calculation, risk assessment and data analysis

The Neural Net Series app, from the Deep Learning Toolbox v14, from Matlab (Matlab, 2020) was used to implement the designed ANN.

The error (E_t) of each prediction (P_t) was calculated according to Eqs. (3) and (4):

$$E_t = N_t - P_t, \tag{3}$$

$$E_t(\%) = \frac{N_t - P_t}{N_t} \times 100\%, \tag{4}$$

in which N_t is the number of COVID-19 hospitalizations reported by regional authorities. In addition, each prediction was associated with a hospitalization rate (R_t') estimated for the population assigned to Hospitals in Valladolid's capital ($I = 495,929$ inhab), following Eq. (5). Each rate value was associated with a risk classification, following the same criteria adopted by the regional health administration, defined as controlled (≤ 10), low (>10 to ≤ 20), medium (>20 to ≤ 30), high (>30 to ≤ 40) or very high (>40).

$$R_t' = 100,000 \times P_t/I, \tag{5}$$

Risk estimations for each case at time t were classified as matches when the rate estimated from the prediction matched the risk level of the actual COVID-19 hospitalization rate at time t (R_t). Risk predictions were classified as underestimations or overestimations when the predicted rate showed a lower or higher risk level, respectively, than the actual hospitalization rate at time t. The error associated with each predicted rate (E_{R_t}) was calculated according to Eq. (6):

$$E_{R_t} = \frac{R_t - R_t'}{R_t} \times 100\% \tag{6}$$

Finally, basic statistics analyses, including Shapiro-wilk test for normality, skewness assessment, average estimation, and correlations, were performed using Rstudio using the *stats* and *ggplot2* packages. In addition, the Epiforge checklist (Pollett et al., 2021) was followed as a guide for reporting methods and results of COVID-19 hospitalization predictions based on the DANN model developed (Supplementary Table A1).

3. Results and discussion

3.1. SARS-CoV-2 N1 concentrations and epidemiological situation in Valladolid

Since the beginning of the COVID-19 pandemic (and until 10th March 2023), the number of deaths in Valladolid was 1751, accumulating 195,473 new positive cases. Fig. 2 shows N1 concentrations in the influent wastewater of Valladolid's WWTP (Fig. 2a-b), the number of vaccinated persons (with the first dose, Fig. 2c-d), and the number of hospitalized patients (Fig. 2e-f). The data is shown for the first study period from May 2021 until September 2022, and the second period from September 2022 to July 2023. During the first period, the number of first-dose vaccinated people increased rapidly, reaching a plateau around day 250. The number of daily hospitalizations ranged between 6 and 171, showing four waves with peaks observed around days 95, 260, 390, and 440, which matched with peaks in N1 concentrations (overall values ranging from 4.0 to 6.5 $\log_{10}$ cg L^{-1}) at days 90, 250, 380, and 430. During the second period, hospitalizations decreased to a range of 5 to 73 cases, with first-dose vaccinations stabilized at maximum values and N1 concentrations from 3.3 to 6.5 $\log_{10}$ cg L^{-1}.

Considering the overall dataset, the numbers of hospitalizations, first-dose vaccinations and N1 concentrations were not normally distributed (Supplementary Fig. A.2), showing moderately to highly skewed distributions (1.0, -2.3 and -045, respectively) associated with

asymmetry towards lower and higher values, respectively. Only during the first period and for hospitalizations below 200 cases, N1 concentrations in wastewater showed significant correlations with the number of hospitalizations (Spearman rs = 0.58, p = 5.0E-08; Pearson r = 0.57, p = 1.0E-07), whereas the linear fit between these parameters (Fig. A3a) showed a relatively low R^2 value of 0.33. However, by increasing to 14 d the lag time between hospitalizations or new hospitalizations and N1 concentrations, the R^2 was increased to 0.58 (Fig. A3b) at the expense of a decrease in the correlation strength. These results suggest that a shorter delay time (7 days) produces the best combination of results (R^2 = 0.56, Spearman rs = 0.43, p = 0.0001; Pearson r = 0.41, p = 0.0002). Previous studies, such as Galani et al. (2022), identified a significant correlation between hospitalizations and SARS-CoV-2 in municipal wastewater. In contrast, other works, such as Gonçalves et al. (2021), de Araújo et al. (2023) or Pico-Tomàs et al. (2023) demonstrated correlations between hospitalizations and SARS-CoV-2 presence in hospital and elderly residences wastewaters. No correlations were observed between N1 concentrations and hospitalizations during the second period. This lack of correlation was probably linked to a virus circulation among the population but with a low number of hospitalizations due to immunity acquisition after an effective vaccination program.

In the case of the first period, similar to the aforementioned studies in which the detection of significant correlations was associated with a relatively high amount of data for N1 concentrations in wastewater and case detections, the detected correlations were supported by enough data available of N1 concentrations and hospitalizations and new hospitalizations for the case study area. Furthermore, the improvement in the significance of the correlation by adjusting the delay time between SARS-CoV-2 footprint detection in wastewater and hospitalizations has also been demonstrated, variating mainly between 2 and 8 days (Medema et al., 2020; Peccia et al., 2020; Galani et al., 2022; Olesen et al., 2021), but up to 22 days (Sangsanont et al., 2022). This lead time is probably a consequence of the effect of high SARS-CoV-2 excretion from asymptomatic patients and the time between when an individual begins shedding detectable virus in their stool and when they seek diagnostic testing and indicates the potential of N1 concentrations to be used to estimate community-level disease prevalence and trends (Olesen et al., 2021). Other relevant variable is the relatively rapid decay of viruses in wastewater (De Oliveira et al., 2021; Torres-Franco et al., 2024). However, no improvements in the correlations were achieved when considering a temperature-based decay in the sewer system (data not shown), indicating that the lag time was more critical than N1 decay for adjusting correlations with hospitalizations or new hospitalizations. Previous studies have reported similar results, showing that the strength of the relationship between wastewater and case data strongly depends on sampling frequency and population size. Furthermore, no significant improvement was achieved by normalizing wastewater data to flow rate (Schill et al., 2023), and considering the decay of SARS-CoV-2 in the sewerage system had minimal effects on WBE assessments (Mota et al., 2021; Pardo-Figueroa et al., 2022).

3.2. COVID-19 hospitalization forecasting using the DANN model

The DANN model addressed the limitations of directly correlating SARS-CoV-2 with hospitalizations or new hospitalizations, in the context of decreased cases reported to public health authorities, predicting the number of hospitalizations as a function of N1 concentrations in wastewater. After achieving the final model architecture, the DANN model obtained the number of hospitalized patients for each date as a function of past hospitalized patients for the previous 7 days and the value of the inputs (N1 concentration and vaccination coverage) for the previous 11 days (Fig. 5a).

The concentrations of SARS-CoV-2 N1 in wastewater, number of first dose vaccinations, and COVID-19 hospitalizations were used for training the neural network architecture described in Section 2.5. For the first period, the data available were divided into training (70 %), validation
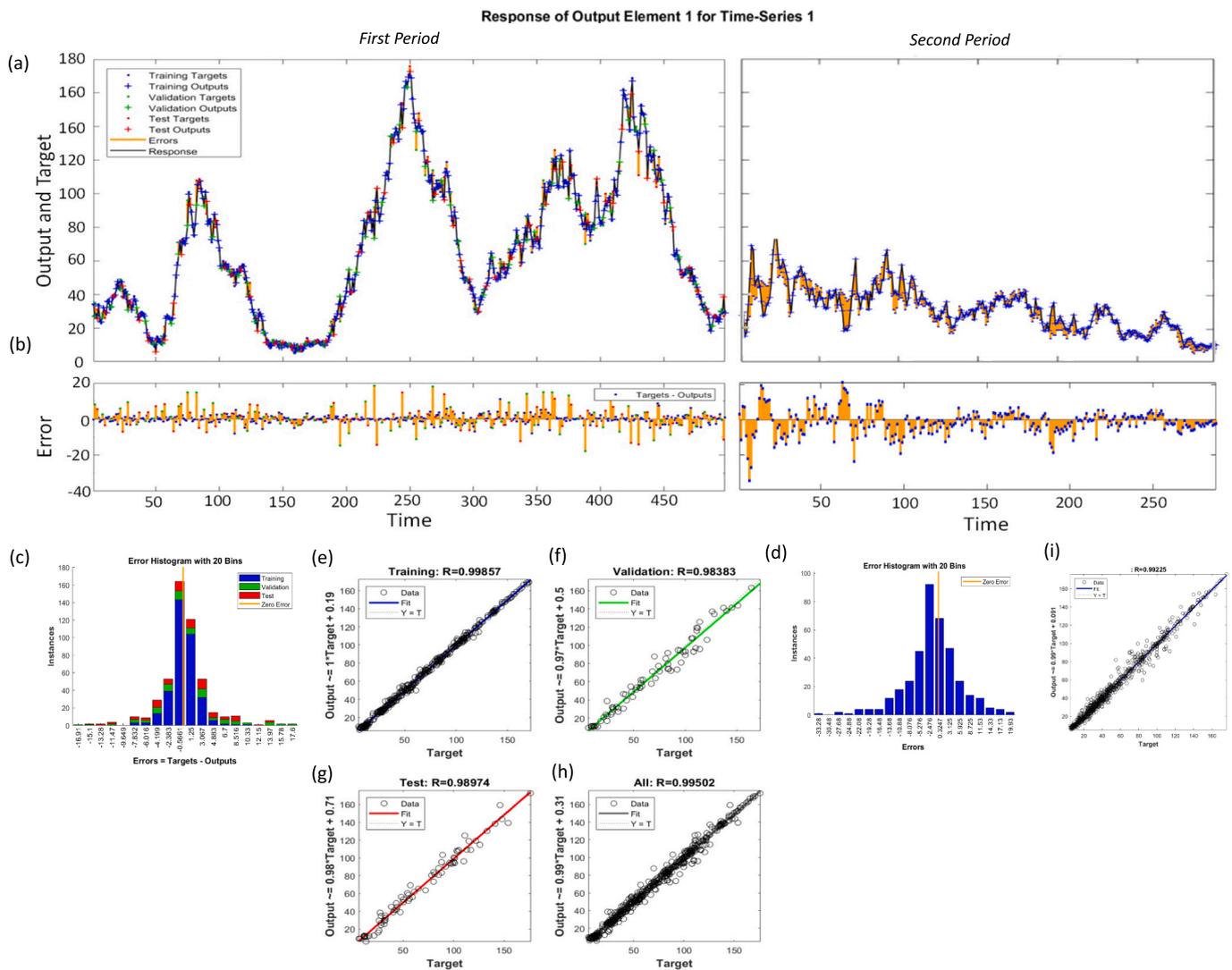
Fig. 5. (a) Number of hospitalized patients (Targets) and Outputs of the ANN for the three subsets; (b) Errors between Targets and Outputs (c, d) Distribution of errors between targets and outputs for the three subsets in the first and second study periods. (e to i) Linear regression of targets relative to outputs for each subset and the overall series.

(15 %), and test (15 %) subsets using interleaved indices, so every independent subset was representative of the whole range of variation of the data. The assessment was measured through the Mean Squared Error (MSE) and the Levenberg-Marquardt backpropagation method was used as the training algorithm. The number of neurons in the hidden layer was optimized by trial and error in order to obtain a good performance, not only on the training subset, but also on the test subset. For this purpose, several trainings were run starting with 10 neurons and increasing the number of neurons in 10 units each time. In addition, the training procedure for each neural network was repeated 10 times in order to minimize the influence of the initial random conditions of the algorithm. After all training experiments, 60 neurons at the hidden layer were considered the best value in terms of overall performance (MSE = 17.2), showing a satisfactory performance for the test subset (MSE = 35.7). By taking the square root of the MSE, a RMSE index with a value of 4 (overall) and 6 (test) was obtained, which means that, on average, the proposed neural network model provided an error of 4 (overall) and 6 (test) hospitalized patients for the real value. For the second period, when all data was regarded as a test subset, a RMSE index of 7 was achieved for this period, indicating an acceptable network performance.

The absolute error of each prediction in Fig. 5a is presented in Fig. 5b. During the first period, all errors were less than 20 hospitalized

patients, resulting in higher error percentages (from −100 % up to 55 %), mainly during periods of particularly low cases. Nevertheless, the overall good performance of the model was demonstrated by an error of 15 % ±11 in the prediction regarding the number of hospitalized (Fig. A4). The distribution of these errors also performed well, as seen from the histogram in Fig. 5c. The regression fit (Figs. 5e to 5h) confirmed the good results obtained. During the second period, the absolute error in the predictions ranged from −35 to 21, and − 117 % to 51 %, compared to reported hospitalizations. The distribution of errors is shown in the histogram of Fig. 5d. Most of the predictions that exhibited higher errors occurred during the first days of this second period when COVID-19 showed a trend of comparatively fewer cases of hospitalizations but relatively high N1 concentrations and a constant maximum of first-dose vaccination. Nevertheless, the retraining procedure decreased the predictions' error by incorporating more past data with these characteristics, achieving relatively low errors (−10 to 10) from about day 110 onwards.

Similar to previous studies that demonstrated high accuracy in predicting hospitalizations based on SARS-CoV-2 using ANN models (e.g., Galani et al., 2022), DANN modeling overcame the influence of factors that can limit strong correlations, such as the delay time associated with a high abundance of asymptomatic individuals or variations in the

patterns of SARS-CoV-2 excretion in feces. These factors likely constituted the main constraints for achieving more simple accurate approximations, for instance, by linear fits between SARS-CoV-2 N1 concentrations in wastewater and hospitalizations, as explored above. Similarly, the DANN modeling approach minimized some uncertainties, such as the reliability of SARS-CoV-2 RNA detection and quantification in wastewater, which can be influenced by sample acquisition. On the other hand, first-dose vaccination probably acted as a relevant covariate, mainly during the first study period, mediating relatively similar peak values in hospitalizations and N1 concentrations in wastewater at increasing numbers of positive cases (Fig. A5). This effect was likely compared to immunity waning, as evidenced by the relatively low error in the predictions, probably due to high effectiveness of first-dose vaccinations in reducing COVID-19 symptoms severity and hospitalizations (Mohammed et al., 2022), as demonstrated mainly during peak in hospitalizations at day 251 (Fig. 2e, first period), when a high number of first dose vaccinations was already achieved, accounting for 128 cases while the number of new COVID-19 positive cases was 2242. These values were in contrast with those, for instance, at day 5, when there was still low coverage in first-dose vaccination in Valladolid, with 76 hospitalizations but about a hundred new positive cases.

In this sense, the proposed DANN model demonstrated its capability to achieve high-accuracy predictions based on a few variables. Furthermore, the decrease in the prediction error over time reinforced the fact that the training process can effectively develop the model's capacity to learn the combined effect of a large number of variables that were not included in the model but were tested in other studies using static ANN models, such as climate and clinical testing coverage (Jiang et al., 2022) or other typical variables included in the WBE assessment whose influence was somewhat included in the past outputs of the model.

In addition to the accurate results achieved for 1 day-ahead predictions, it was identified that the 5 step-ahead predictions were still rather good, with an MSE of 3.5, corresponding to an RMSE of 2 hospitalized patients, for the last 5 predicted days of the historical series (first period). These predictions, together with the last 7 data from the same series and the last 11 data from N1 concentration, are represented in Fig. 6a and b, respectively. As seen in Fig. 6, the predictions given by the ANN are logical: the number of hospitalized patients for the last 7 days in the dataset (days 492 to 498) shows a decreasing trend as a result of the decrease in N1 concentrations, except for the last 2 days, but these increasing concentrations will only affect the number of hospitalized patients in the future. Hence, the developed DANN predicted that the

number of hospitalized patients would decrease for the following 5 days.

3.3. Early COVID-19 warning system based on N1 concentrations

The regional health administration in Castilla y León has established risk levels on several indicators, such as the incidence in people above 60 years, the rate of new COVID-19 hospitalization, and/or occupation of beds per 100,000 inhabitants. Fig. 7 shows the risk levels based on COVID-19 hospitalizations and calculated for the population of the area of influence of the hospitals Clinico Universitario and Rio Hortega in Valladolid, throughout the pandemic (from March 18th, 2020, to March 10th, 2023,). It can be noted from Fig. 7 that the period selected for the ANN model was characterized by risk levels ranging from "controlled" to "high" during the first period and from "controlled" to "low" during the second period. The proposed dynamic ANN also allowed the assessment of the risk levels associated with each 1 day-ahead prediction by transforming the number of hospitalizations caused by COVID-19 into a rate per 100,000 inhabitants (Fig. 7a) and assigning a risk level to each prediction. For the first period, it was estimated that in 194, 167, 98, and 15 cases in which respectively the "controlled", "low", "medium" and "high" levels assigned to each prediction coincided with the actual risk level (Table 1), leading to an overall coincidence of 95.2 % of cases in which the prediction matched the risk level determined by the regional health authority.

On the other hand, the predictions resulted in 2, 10 and 2 underestimations of "low", "medium", and "high" risk levels, respectively. In these cases, the predictions showed average errors of −8.6 %, −10 % and − 7.2 % for "low", "medium" and "high" risk levels, respectively (Fig. 7b, Table 1). Overestimations occurred in 6, 3 and 1 cases, respectively, for "controlled", "low", and "medium" risk levels, coinciding with error percentages of 12 %, 7.0 %, 9 %, and 5 % for the same categories. Overall, whereas the ANN model showed high accuracy in predicting the risk level for the "controlled" and "low" situations, it showed a slightly lower accuracy for the "medium" and "high" levels mainly due to the higher variability of the past-days values of hospitalizations in each situation of risk increase and peaks occurrence. During the second period, 17 predictions overestimated the "controlled" risk level, whereas the "low" risk level was overestimated in 1 case and underestimated in 5 cases. Overall, the DANN model matched the risk level in 92.3 % of the 287 hospitalization cases in this period. In addition, the "controlled" and "low" levels were respectively overestimated and underestimated, with error percentages of 30 % and 17 %, with no other risk categories recorded. All error percentages in risk level
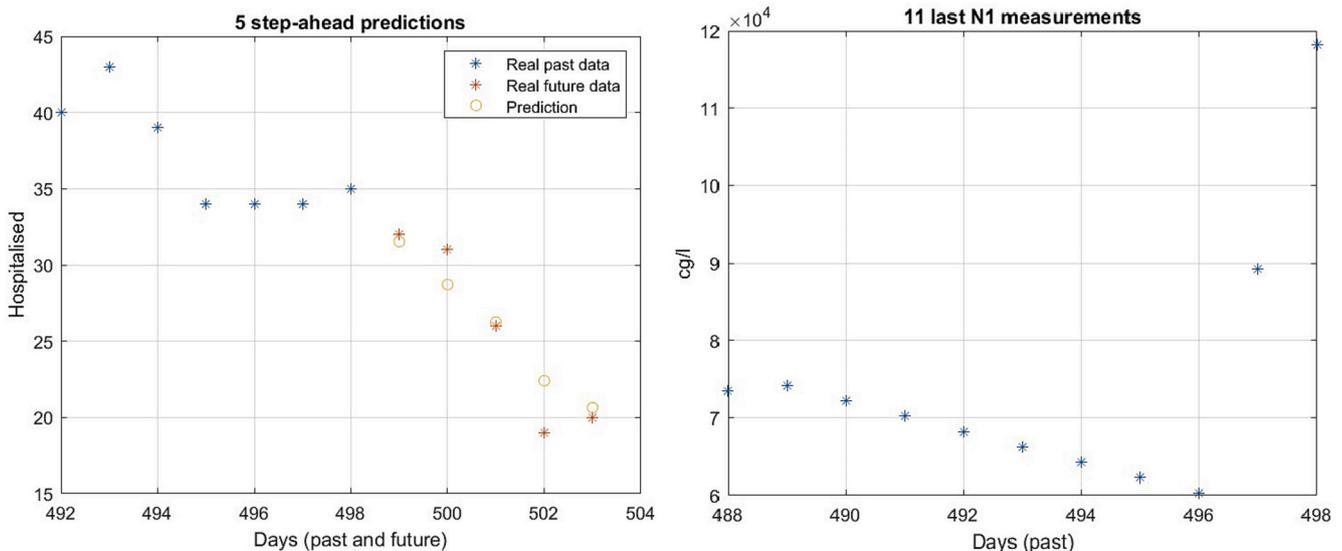

Fig. 6. (a) 5 step-ahead predictions at the end of the historical series and (b) 11 last N1 measurements at the end of the historical series.
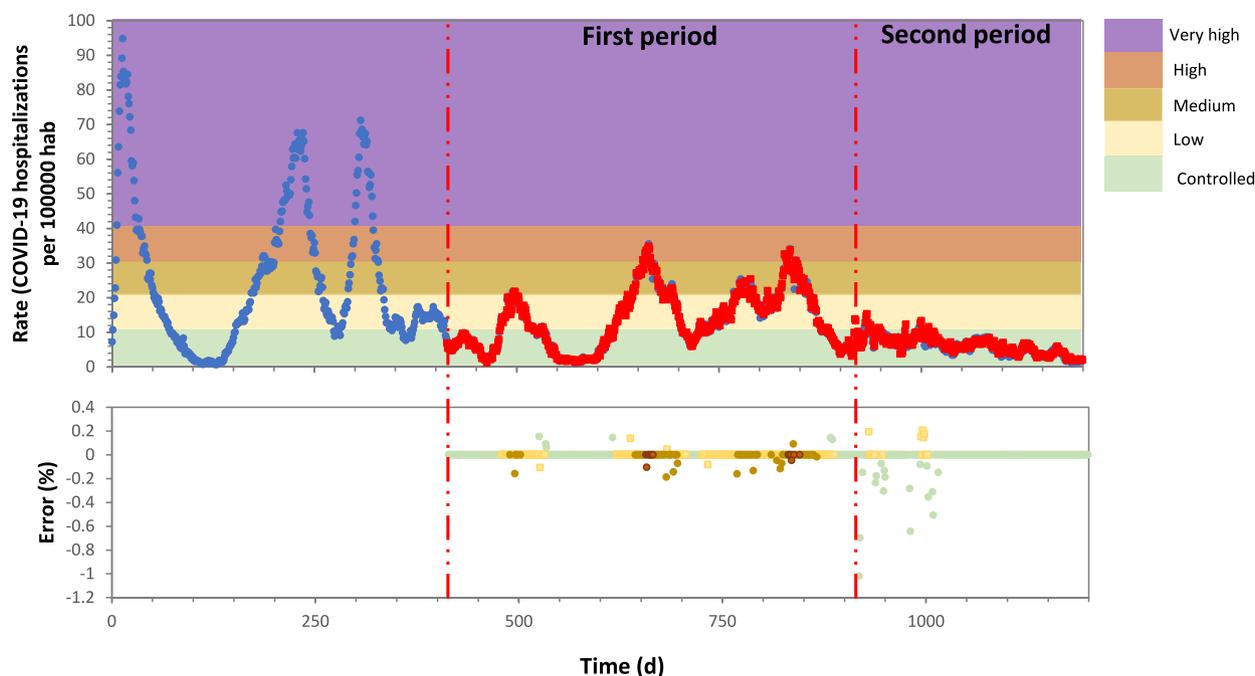
Fig. 7. (a) Warning levels established by the Regional Health Administration in Castilla y Leon (Spain) based on COVID-19 occupation of beds in hospitals (data retrieved from (https://analisis.datosabiertos.jcyl.es/pages/coronavirus/?seccion=descargas, (■). (b) Predictions achieved by dynamic ANN (■) and percentage of error between prediction and hospitalizations associated with (●) "Controlled", (●) "Low", (●) "Medium" and (●) "High" and (●) very high risk levels.

Table 1
Summary of matches, underestimations, and overestimations of risk levels based on COVID-19 hospitalization predictions. Prediction errors correspond to average values.

Risk level	Parameter	First period			Second period		
		Cases		Prediction average error (%)	Cases		Prediction average error (%)
		Counts	%		Counts	%	
Controlled	Cases	200	100	1.3	270	100	11
	Matches	194	97	0.9	253	94	10
	Underestimations	N.A	N.A	N.A	N.A	N.A	N.A
	Overestimations	6	3	12	17	6	30
Low	Cases	172	100	−0.5	17	100	1
	Matches	167	97	−0.6	12	70	9
	Underestimations	2	1	−8.6	5	30	17
	Overestimations	3	2	7.0	0	0	0
Medium	Cases	109	100	−0.7	0	0	0
	Matches	98	90	0.1	0	0	0
	Underestimations	10	9.2	−10	0	0	0
	Overestimations	1	0.9	9	0	0	0
High	Cases	17	100	−1.2	0	0	0
	Matches	15	88	−0.4	0	0	0
	Underestimations	2	12	−7.2	0	0	0
	Overestimations	0	0	N.A	0	0	0

predictions were lower than those obtained for the net number of hospitalization predictions. Indeed, by associating the predictions of hospitalizations to risk levels, the overall accuracy during the first study period increased, on average, from 85 % to 99.8 %, and from 90 % to 98.5 % during the second period, resulting in a powerful strategy to manage COVID-19 warning levels when applying DANN models for WBE. The accuracies achieved were similar to or higher than values in ANN applications reported in the literature, such as 86 % and 87 % in predictions of mortality rate and ICU admissions (Kuvvetli et al., 2021) or 89.5 % in morbidity and mortality in Greece (Asteris et al., 2022). Despite these limitations, the ANN showed high potential to provide regional early alarms for managing the COVID-19 pandemic.

4. Conclusions

The dynamic ANN model developed in this work was able to accurately predict the number of hospitalized patients with COVID-19 in Valladolid (Spain) based on historical data, the SARS-CoV-2 N1 RNA concentration in wastewater (registered from the WWBE at Valladolid) and the vaccination coverage in the last 11 days. The influence of uncertainties and unknown variables, such as SARS-CoV-2 RNA recovery efficiencies, delay times between SARS-CoV-2 RNA detection in wastewater and hospitalizations, weather conditions, wastewater flowrates, and immunity acquisition and waning among the population, seems to be somehow included in the dynamic design architecture of the ANN model. The predictions achieved through the DANN model were successfully associated with the risk levels defined by regional health authorities, showing the potential to support anticipatory actions when a

significant increase in the number of hospitalized patients is expected due to possible increases in SARS-CoV-N1 detections in wastewater. Also, the average error in the number of predicted cases of hospitalization was decreased by associating the predictions with the risk levels defined by the regional health authorities, resulting in a powerful strategy for WBE assessments. This model architecture and its association with risk estimations could be applied in cases with similar data availability, emerging as a valuable tool for the continued application of WBE for COVID-19 epidemiology, especially in the current scenario of declined cases monitoring.

CRediT authorship contribution statement

Jesús M. Zamarreño: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Andrés F. Torres-Franco:** Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **José Gonçalves:** Writing – review & editing. **Raúl Muñoz:** Writing – review & editing, Project administration, Funding acquisition, Conceptualization. **Elisa Rodríguez:** Writing – review & editing, Project administration, Funding acquisition, Conceptualization. **José María Eiros:** Writing – review & editing, Conceptualization. **Pedro García-Encina:** Writing – review & editing, Project administration, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.scitotenv.2024.170367.

References

Asteris, P.G., Gavriilaki, E., Touloumenidou, T., Koravou, E.E., Koutra, M., Papayanni, P. G., Anagnostopoulos, A., 2022. Genetic prediction of icu hospitalization and mortality in covid-19 patients using artificial neural networks. J. Cell. Mol. Med. 26 (5), 1445–1455.

Braga, M.D.B., Fernandes, R.D.S., Souza Jr., G.N.D., Rocha, J.E.C.D., Dolácio, C.J.F., Tavares Jr., I.D.S., Vallinoto, A.C.R., 2021. Artificial neural networks for short-term forecasting of cases, deaths, and hospital beds occupancy in the COVID-19 pandemic at the Brazilian Amazon. PLoS One 16 (3), e0248161.

de Araújo, J.C., Madeira, C.L., Bressani, T., Leal, C., Leroy, D., Machado, E.C., Chernicharo, C.A., 2023. Quantification of SARS-CoV-2 in wastewater samples from hospitals treating COVID-19 patients during the first wave of the pandemic in Brazil. Sci. Total Environ. 860, 160498.

De Oliveira, L.C., Torres-Franco, A.F., Lopes, B.C., da Silva Santos, B.S.Á., Costa, E.A., Costa, M.S., Mota, C.R., 2021. Viability of SARS-CoV-2 in river water and wastewater at different temperatures and solids content. Water Res. 195, 117002.

Galani, A., Aalizadeh, R., Kostakis, M., Markou, A., Alygizakis, N., Lytras, T., Thomaidis, N.S., 2022. SARS-CoV-2 wastewater surveillance data can predict hospitalizations and ICU admissions. Sci. Total Environ. 804, 150151.

Ghiassi, M., Saidane, H., Zimbra, D.K., 2005. A dynamic artificial neural network model for forecasting time series events. Int. J. Forecast. 21 (2), 341–362.

Gonçalves, J., Koritnik, T., Mioč, V., Trkov, M., Bolješič, M., Berginc, N., Paragi, M., 2021. Detection of SARS-CoV-2 RNA in hospital wastewater from a low COVID-19 disease prevalence area. Sci. Total Environ. 755, 143226.

Haykin, S., 1999. Neural Networks. "A Comprehensive Foundation". Prentice-Hall.

He, X., Asada, H., 1993. A new method for identifying orders of input-output models for nonlinear dynamic systems. In: 1993 American Control Conference, San Francisco, CA, USA, pp. 2520–2523. https://doi.org/10.23919/ACC.1993.4793346.

Hornik, K., Stinchcombe, M., White, H., 1989. Multilayer feedforward networks are universal approximators. Neural Netw. 2, 359–366.

IHME COVID-19 Forecasting Team, 2021. Modeling COVID-19 scenarios for the United States. Nat. Med. 27, 94–105. https://doi.org/10.1038/s41591-020-1132-9.

Jiang, G., Wu, J., Weidhaas, J., Li, X., Chen, Y., Mueller, J., Jackson, G., 2022. Artificial neural network-based estimation of COVID-19 case numbers and effective reproduction rate using wastewater-based epidemiology. Water Res. 218, 118451.

Jones, D.L., Baluja, M.Q., Graham, D.W., Corbishley, A., McDonald, J.E., Malham, S.K., Wilcox, M.H., 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. Sci. Total Environ. 749, 141364.

Kuvvetli, Y., Deveci, M., Paksoy, T., Garg, H., 2021. A predictive analytics model for COVID-19 pandemic using artificial neural networks. Decision Anal. J. 1, 100007.

MATLAB, 2020. Version 9.8.0 (R2020a). The MathWorks Inc., Natick, Massachusetts.

Medema, G., Heijnen, L., Elsinga, G., Italiaander, R., Brouwer, A., 2020. 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. 7 (7), 511–516.

Mohammed, I., Nauman, A., Paul, P., Ganesan, S., Chen, K.H., Jalil, S.M.S., Zakaria, D., 2022. The efficacy and effectiveness of the COVID-19 vaccines in reducing infection, severity, hospitalization, and mortality: a systematic review. Hum. Vaccin. Immunother. 18 (1), 2027160.

Mota, C.R., Bressani-Ribeiro, T., Araujo, J.C., Leal, C.D., Leroy-Freitas, D., Machado, E.C., Chernicharo, C.A., 2021. Assessing spatial distribution of COVID-19 prevalence in Brazil using decentralized sewage monitoring. Water Res. 202, 117388.

MSCBS, 2022. Ministerio de Sanidad, Consumo y Bienestar Social, Instituto de Salud Carlos Tercero. Estrategia de vigilancia y control frente a Covid-19 tras la fase aguda de la pandemia. https://www.sanidad.gob.es/profesionales/saludPublica/ccayes/alertasActual/nCov/documentos/Nueva_estrategia_vigilancia_y_control.pdf.

Norgaard, M., 2023. nnsysid. https://www.mathworks.com/matlabcentral/fileexchange/87-nnsysid. MATLAB Central File Exchange. Obtained February 28, 2023.

Olesen, S.W., Imakaev, M., Duvallet, C., 2021. Making waves: defining the lead time of wastewater-based epidemiology for COVID-19. Water Res. 202, 117433.

Pardo-Figueroa, B., Mindreau-Ganoza, E., Reyes-Calderon, A., Yufra, S.P., Solorzano-Ortiz, I.M., Donayre-Torres, A.J., Santa-María, M.C., 2022. Spatiotemporal surveillance of SARS-CoV-2 in the sewage of three Major urban areas in Peru: generating valuable data where clinical testing is extremely limited. Acs Es&T Water 2 (11), 2144–2157.

Patón, M., Al-Hosani, F., Stanciole, A.E., Aden, B., Timoshkin, A., Sadani, A., Rodríguez, J., 2022. Model-based evaluation of the COVID-19 epidemiological impact on international visitors during expo 2020. Infect. Dis. Model. 7 (3), 571–579.

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

Pico-Tomàs, A., Mejías-Molina, C., Zammit, I., Rusiñol, M., Bofill-Mas, S., Borrego, C.M., Corominas, L., 2023. Surveillance of SARS-CoV-2 in sewage from buildings housing residents with different vulnerability levels. Sci. Total Environ. 872, 162116.

Pollett, S., Johansson, M.A., Reich, N.G., Brett-Major, D., Del Valle, S.Y., Venkatramanan, S., Rivers, C., 2021. Recommended reporting items for epidemic forecasting and prediction research: the EPIFORGE 2020 guidelines. PLoS Med. 18 (10), e1003793.

Polo, D., Quintela-Baluja, M., Corbishley, A., Jones, D.L., Singer, A.C., Graham, D.W., Romalde, J.L., 2020. Making waves: wastewater-based epidemiology for COVID-19–approaches and challenges for surveillance and prediction. Water Res. 186, 116404.

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

Rauch, W., Schenk, H., Insam, H., Markt, R., Kreuzinger, N., 2022. Data modelling recipes for SARS-CoV-2 wastewater-based epidemiology. Environ. Res. 214, 113809.

Rodríguez, J., Patón, M., Uratani, J.M., Acuña, J.M., 2021. Modelling the impact of interventions on the progress of the COVID-19 outbreak including age segregation. PLoS One 16 (3), e0248243.

Sangsanont, J., Rattanakul, S., Kongprajug, A., Chyerochana, N., Sresung, M., Sriporatana, N., Sirikanchana, K., 2022. 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. 809, 151169.

Schill, R., Nelson, K.L., Harris-Lovett, S., Kantor, R.S., 2023. The dynamic relationship between COVID-19 cases and SARS-CoV-2 wastewater concentrations across time and space: Considerations for model training data sets. Sci. Total Environ. 871, 162069.

Sims, N., Kasprzyk-Hordern, B., 2020. Future perspectives of wastewater-based epidemiology: monitoring infectious disease spread and resistance to the community level. Environ. Int. 139, 105689.

Torres-Franco, A.F., Leroy-Freitas, D., Martinez-Fraile, C., Rodríguez, E., García-Encina, P.A., Muñoz, R., 2024. Partitioning and inactivation of enveloped and

nonenveloped viruses in activated sludge, anaerobic and microalgae-based wastewater treatment systems. Water Res. 248, 120834.

Vallejo, J., Rumbo-Feal, S., Conde, K., López-Oriona, Á., Tarrío, J., Reif, R., Poza, M., 2020. Highly Predictive Regression Model of Active Cases of COVID-19 in a Population by Screening Wastewater Viral Load.

Zahmatkesh, S., Rezakhani, Y., Chofreh, A.G., Karimian, M., Wang, C., Ghodrati, I., Khan, R., 2023. SARS-CoV-2 removal by mix matrix membrane: a novel application of artificial neural network based simulation in MATLAB for evaluating wastewater reuse risks. Chemosphere 310, 136837.