A Mixed-Effects Model to Predict COVID-19 Hospitalizations Using Wastewater Surveillance

Maria L. Daza–Torres[1, +], J. Cricelio Montesinos-López[1, +], Angel N. Desai[2], Colleen C. Naughton[3], Heather N. Bischel[4], and Miriam Nuño[1,*]

[1] *Department of Public Health Sciences, University of California Davis, California 95616, United States*

[2] *Department of Internal Medicine, University of California Davis, Sacramento, California 95817, United States*

[3] *Department of Civil and Environmental Engineering, University of California Merced, Merced, California 95343, United States*

[4] *Department of Civil and Environmental Engineering, University of California Davis, Davis, California 95616, United States*

[+] *co-first author*

[*] *corresponding author: mnuno@ucdavis.edu*

Abstract

During the COVID-19 pandemic, many countries have investigated the potential usefulness of wastewater-based epidemiology as an early warning system. Initially, methods were created to detect the presence of SARS-CoV-2 RNA in wastewater. Researchers have since conducted extensive studies to examine the link between viral concentration in wastewater and COVID-19 cases in areas served by sewage treatment plants over time. However, only a few reports have attempted to create predictive models for hospitalizations based on SARS-CoV-2 RNA concentrations in wastewater. This study implemented a linear mixed-effects model that connects the levels of virus in wastewater (WW) to hospitalizations. The model was then utilized to predict short-term trends in 21 counties in California, based on data from March 14, 2022, to May 21, 2023. The modeling framework proposed here permits repeated measurements as well as fixed and random effects. The mathematical model that utilized wastewater data showed strong performance and successfully identified trends in hospitalizations. Additionally, the proposed model allows for predicting SARS-CoV-2 hospitalizations two weeks ahead, which is critical for optimizing hospital resource allocation.

Keywords: COVID–19; SARS-CoV-2; Wastewater-based epidemiology (WBE); Hospitalizations; Generalized mixed models; Model performance.

1 Introduction

Wastewater-based disease surveillance (WDS) is a valuable tool for identifying and monitoring the spread of infectious diseases within a community. By detecting the presence of SARS-CoV-2 RNA in municipal sewer systems, WDS systems can provide early detection of outbreaks and track changes in the concentration of the virus over time. They can also help identify new variants as they emerge.

Wastewater surveillance has been used effectively prior to the COVID-19 (Coronavirus Disease 2019) pandemic, most notably as a tool to monitor and respond to poliovirus and the presence of pharmaceutical and illicit drugs (Paul et al., 1940; Trask et al., 1938). The pandemic led to the widespread implementation of wastewater surveillance in communities across the United States. This approach, known as wastewater-based epidemiology (WBE), has proven to be a cost-effective method for tracking COVID-19 and other infectious diseases(To et al., 2010; Hughes et al., 2022; Boehm et al., 2023a; Kumblathan et al., 2021). To bolster the nation's capacity to monitor SARS-CoV-2 effectively, the US Centers for Disease Control and Prevention (CDC) launched the National Wastewater Surveillance System (NWSS) in September 2020, establishing over 1250 sampling sites across the US, encompassing more than 133 million people by October 2022. An essential factor contributing to the success of WDS is the extensive coverage of municipal wastewater collection systems, which connects approximately eighty percent of households in the United States. WDS offers a cost-effective strategy to monitor trends in SARS-CoV-2 community infection rates as well as other infectious diseases.

In contrast, conducting widespread testing in a community requires substantial testing capacity, including access to testing kits, trained personnel, and testing facilities. Implementing community-wide testing can be financially burdensome, especially in resource-constrained settings. Even with adequate testing capacity, achieving high participation rates can be challenging. When individuals get tested for a disease, their behaviors can lead to biases and inaccuracies in estimating the prevalence of the disease. This can make the tested population less representative and affect the accuracy of disease prevalence estimates. Therefore, relying solely on confirmed cases to determine disease prevalence in a community can lead to biased estimates (Daza-Torres et al., 2023).

In order to reduce the biases introduced by testing practices, the test positivity rate (TPR) emerges as a crucial alternate measure. The TPR is the proportion of individuals who have tested positive for COVID-19 relative to the total number of tests administered on a given day. TPR has been used as a valuable alternative metric to track the spread of COVID-19 and to assess the prevalence cases of the disease compared to the number of tests being conducted (Dallal et al., 2021; Dowdy and D'Souza, 2020). Previous research has demonstrated that the TPR is a reliable indicator for monitoring the growth of the COVID-19 pandemic, exhibiting a strong correlation with the number of patients admitted in hospitals and intensive care units Fenga and Gaspari (2021); Montesinos-López et al. (2021). However, the interpretation and use of the TPR can be challenging in certain contexts, as it relies on accurate and comprehensive testing data, which may be hindered by factors such as limited testing capacity, variable testing practices, and variations in reporting standards.

One solution to the challenges faced in estimating disease prevalence in a community is through the use of WDS. This approach is relatively simple and accurate in gathering data from both symptomatic and asymptomatic individuals across the entire community. Wastewater-based epidemiology (WBE) provides a more accurate estimation of disease prevalence as it represents a

sample of the population regardless of individual testing behaviors or access to healthcare. It has been proven to be a useful tool for COVID-19 surveillance and early warning. Numerous studies have observed a strong correlation between SARS-CoV-2 RNA concentrations in wastewater with the number of daily new cases in corresponding wastewater treatment plants catchment areas and, more recently, with the TPR. Montesinos-López et al. (2023).

However, there is limited available data relating wastewater viral concentration with hospital admission rates Galani et al. (2022). A correlation between wastewater data and hospitalizations, with a time lag typically ranging from 8 to 14 days, was described in Kadonsky et al. (2023); Galani et al. (2022). Galani et al. (2022) implemented an artificial neural network model that combines clinical test results (COVID-19-positive cases) and wastewater concentrations to model new hospital and ICU admissions. These findings emphasize the potential of wastewater surveillance to improve preparedness and response to the spread of infectious diseases.

Building upon this existing knowledge, we propose using a linear mixed-effects (LMM) model to predict COVID-19 hospitalizations by examining viral concentrations found in county-level wastewater. Our analysis covers various counties in California to offer a comprehensive view of the correlation between wastewater data and hospitalization rates. The LMM model takes into account random and fixed effects at the county level and combines wastewater viral concentrations and hospital admission rates to capture the overall regional trends of COVID-19 hospitalizations.

We are able to examine the correlation between SARS-CoV-2 wastewater indicators and hospitalization rates, and how this relationship may differ across counties and over time. Our theory is that the link between wastewater and hospitalizations will fluctuate within and between counties as a result of different factors such as population demographics, medical resources, and preventive measures. These variables can greatly influence disease transmission patterns and affect the accuracy of hospitalization estimates based on wastewater data at the county level.

We have assessed and compared the predictive performance of the LMM model in predicting COVID-19 hospitalizations by using three different factors: reported cases, TPR, and wastewater data. This thorough evaluation of various models provides valuable information about the advantages and potential drawbacks of utilizing wastewater-based modeling as a means of surveillance to enhance mitigation strategies and decrease the impact of outbreaks.

2 Materials and Methods

The analytical framework was developed using data from twenty-one counties across California, USA, actively collecting wastewater data. The analysis includes clinical and wastewater data from March 21, 2022, to May 21, 2023.

2.1 Data Sources

Clinical Data

We used publicly available data for COVID-19 daily cases and hospitalization at the county level from the California Department of Public Health (CDPH) website (CDPH). According to the CDPH, hospitalization counts represent the number of patients hospitalized in an inpatient bed with a laboratory-confirmed COVID-19 diagnosis. However, it's worth noting that some hospitalizations may not be solely due to SARS-CoV-2 infection. The count includes all inpatients, including those in ICU and Medical/Surgical units, but excludes patients waiting for an inpatient bed in affiliated clinics, outpatient departments, emergency departments, and overflow locations.

The number of positive COVID-19 cases is determined by the total count of molecular tests that were positive using polymerase chain reaction (PCR) techniques. We took into account both "Total Tests" and "Positive Tests", which reflect the cumulative totals based on the estimated date when the specimen was collected for testing. However, it's essential to keep in mind that there could be a delay between the date of specimen collection and when the test results are reported.

Hospital admissions and COVID-19 cases were aggregated weekly. To calculate population-level hospitalization rates for a county, we take the number of hospital admissions and new cases in the past 7 days, divide it by the population in that same county, and then multiply the result by 100K. To calculate the TPR for a specific county, the number of positive COVID-19 tests from the past 7 days is divided by the total number of tests conducted in that county during that same time frame.

Wastewater Data

Wastewater data used in this study are available publicly through the Verily Wastewater Testing Dashboard (Verily). Since the pre-analytical and analytical methods used for measuring the SARS-CoV-2 RNA N gene and PMMoV are registered and previously described in peer-reviewed publications Boehm et al. (2023b); Montesinos-López et al. (2023); Daza-Torres et al. (2023), they are not presented herein. WastewaterSCAN (WastewaterSCAN) provides testing, sampling supplies, and free shipping to participating treatment plants. Sewage treatment plants receive kits to sample wastewater solids three times or more weekly and shipping for lab analysis at no cost. Verily makes results available through its dashboard within 48 hours of receiving samples (Verily).

Each location reports viral copies per gram of dry solid waste in wastewater samples, including SARS-CoV-2 and Pepper Mild Mottle Virus (PMMoV), a virus ingested with pepper-containing food, which has been widely measured and frequently used for normalizing SARS-CoV-2 concentration data. The SARS-CoV-2 viral load is normalized by the PMMoV concentration, resulting in the dimensionless metric N/PMMoV. The N gene is present in all variants of the virus. PMMoV is a highly abundant RNA virus detected broadly in wastewater (Symonds et al., 2018). PMMoV serves as a process control for virus extraction efficiency and helps account for variations in population size, rainfall, and water usage. If the target is not detected in the sample, the concentration is

reported as 0 c/g. Results are reported as viral copies per gram (dry weight) of wastewater solids, indicating the number of virus copies found in a gram of solid waste from a location. Figure 1 shows the normalized N gene concentration and hospitalization rate per 100K population.

The study's analyses solely depend on openly available data and do not contain any private health information. Moreover, we did not receive authorization to access codes or links that could reveal the identities of individuals. As a result, our study is exempt from ethical review and the requirement for informed consent, according to the Common Rule 45 CFR46.102 guidelines.

2.2 Model Outcome

To determine the average hospitalization rates per week, we combined the number of hospital admissions with confirmed COVID-19 test results and divided it by seven. We then adjusted this metric to reflect every 100,000 people in each county, making it easier to compare data across different counties (refer to Figure 1). For the sake of clarity, we will refer to this standardized measurement as "hospitalization rate" moving forward.

Similarly, we utilized the normalized concentration aggregated weekly and divided by the number of samples collected that week. We used data from the wastewater treatment plant serving the highest population within each county. We applied a logarithmic transformation to the weekly average wastewater data and hospitalization rates to facilitate the analysis. These two metrics were the basis for exploring the relationship between hospitalization rates and SAS-CoV-2 normalized wastewater concentrations.

This study analyzed data from twenty-one counties in California. These counties were chosen based on specific criteria to ensure that each one had at least one treatment plant reporting wastewater data and had started collecting that data before January 2023. San Benito County was excluded from the analysis due to the low quality of hospital data. The study encompassed two datasets: 1) SARS-CoV 2 RNA concentrations in wastewater and 2) COVID-19 hospitalization rates from March 21, 2022 to May 21, 2023.

The analysis in this study encompassed three distinct COVID-19 waves, each corresponding to a specific Omicron variant. These variants were BA.2 + BA.4 + BA.5, which dominated from March 21, 2022, to June 19, 2022; BQ Omicron, which dominated from June 20, 2022, to February 6, 2023; and XBB Omicron, which dominated from February 7, 2023, until the end of the study period (May 21, 2023). These periods were defined based on the reported dominance of each variant by the California Department of Public Health (SCAN, 2023). Consequently, we will utilize these variant dominance periods to define the three COVID-19 waves within our analysis.

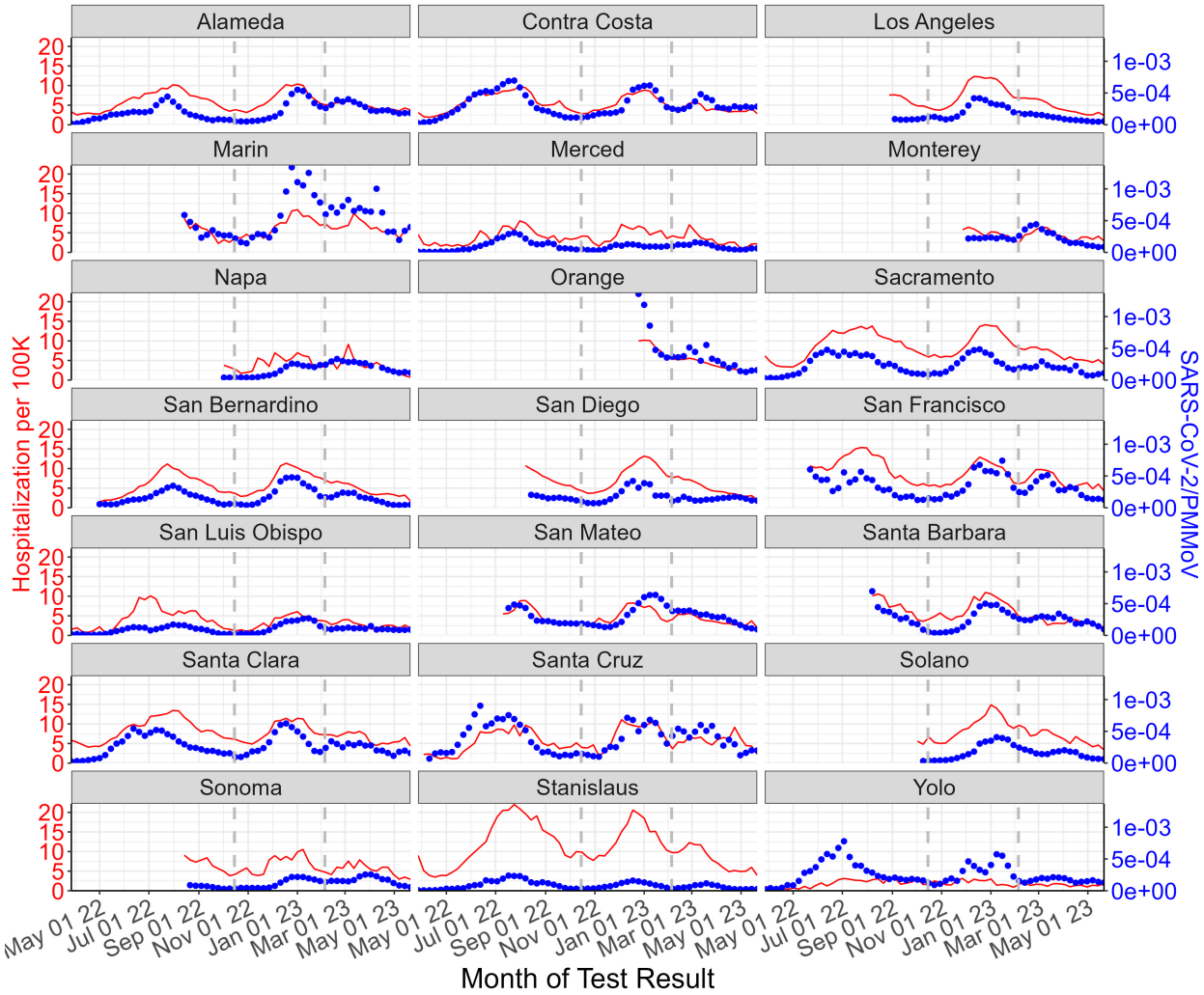

Figure 1: COVID-19 hospitalizations per 100K population (red line) and normalized SARS-CoV-2 WW concentrations (blue dots) from March 21, 2022, and May 21, 2023, by county. Vertical gray lines represent the end of each wave.

2.3 Data Analysis

This study focused on associations between hospitalization rates and longitudinal changes in the SARS-CoV-2 wastewater concentration over time. COVID-19 hospitalization data have been collected repeatedly from counties over time; the traditional linear regression model may not be appropriate due to the presence of correlated measurements within each county. Longitudinal studies often involve correlated data points, and the assumption of independence in linear regression is violated. LMMs allow us to explicitly model the dependence between observations within counties by incorporating random effects.

Our study hypothesized that SARS-CoV-2 wastewater concentration's association (slope) with

hospitalization rate varies by county (Figure 2). Additionally, it is important to consider the influence of virus variants, human behavior, viral evolution, vaccination status, prior immunity, and other time-dependent factors associated with each wave on the wastewater-to-hospitalization ratio, as depicted in Figure 3. To account for these factors, our model incorporates a fixed effect that captures their potential influences on the relationship between wastewater and hospitalization rates.

Previous research by Kadonsky et al. (2023) observed that the wastewater-to-hospitalization ratio remained relatively stable during two COVID-19 waves in two of three Central Valley communities between October 2021 and September 2022. It is possible for the slope of the wastewater-to-hospitalization ratio to remain constant, but the intercept may differ depending on the county.

Moreover, we explore the predictive capabilities of different COVID-19 transmission indicators, including reported cases, TPR, and wastewater data, to determine the most effective real-time indicator to monitor hospitalizations within a region of study. We assess how well various predictive models perform, including one that solely uses wastewater data, and compare them to others that use clinical data. By evaluating multiple models that rely on different data sources, we gain valuable insight into the strengths and limitations of wastewater-based modeling as a surveillance tool. This can help improve mitigation strategies and reduce the impact of disease outbreaks.

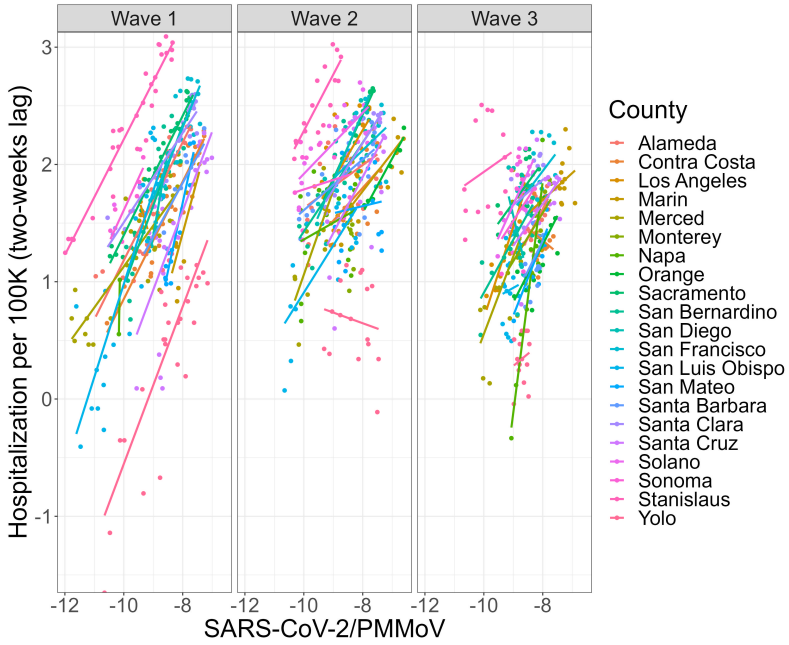

Figure 2: Log transformation of COVID-19 hospitalization rate two-week lagged versus log transformation of the normalized wastewater concentration separated by wave. Regression lines are fitted to the data and depicted; model intercepts and slopes vary by county.

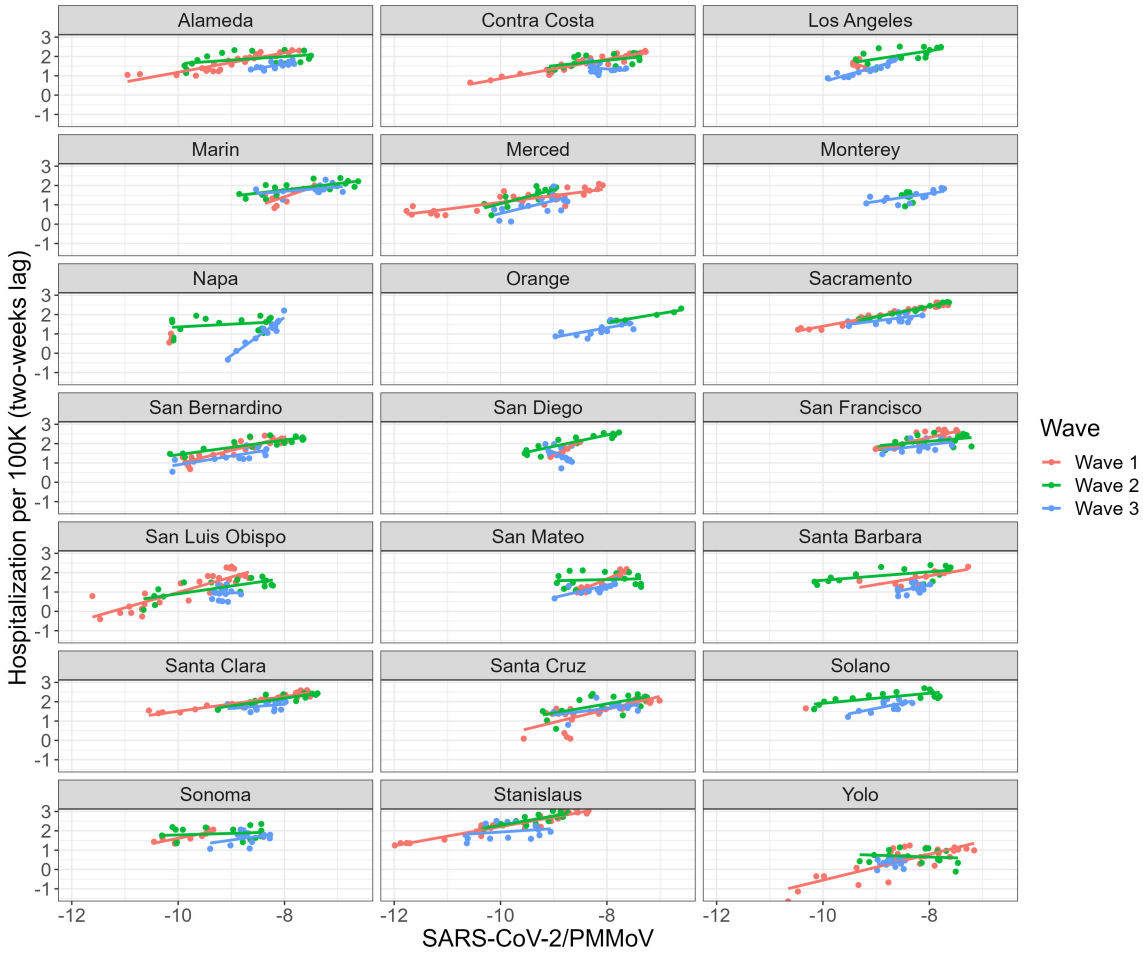

Figure 3: Log transformation of COVID-19 hospitalization rate two-week lagged and log transformation of the normalized wastewater concentration by county. Regression lines are fitted for each county and wave.

2.4 Mathematical Model

The log-transformed COVID-19 hospitalization rate (henceforth hospitalization rate) at the county $i = 1, \ldots, 21$ and week $t = 1, \ldots, T$ is modelled as

$$y_{it} = \mu + \mathbf{x}_i \boldsymbol{\beta} + v_t + g_i + Z_{it} u_i + \varepsilon_{it} \tag{1}$$

where μ is the average hospitalization rate across the study area, $\mathbf{x}_i = \{x_{1it}, x_{2it}, \ldots x_{pit}\}$ is the vector of p covariates (e.g., wastewater concentration, reported cases, and TPR). The v_t term is a temporal-level fixed effect indicating the pandemic's three waves (grouped time component) as specified above. The g_i term is a county-level random effect corresponding to the County where the hospital is located. County-level random effects are modeled as $g_k \sim N(0, \sigma_g^2)$, with $k = 1, \ldots, 21$.

Finally, we allow for additional flexibility by including a random slope, u_i, for a fitted covariate Z_{it} (e.g., wastewater concentration). The ε_{it} term represent the residuals distributed as $N(0, \sigma^2)$ with σ^2 is the measurement error variance. The random effects g_i and u_i are assumed to be independent of the residuals ε_{it}.

2.5 Model Implementation

A two-week lag was incorporated to address the temporal dynamics between hospitalization and COVID-19 transmission indicators. This lag duration was chosen based on previous research showing a nearly two-week delay between hospitalizations and measurements of wastewater, confirmed cases, and test positivity rate (Kadonsky et al., 2023; Montesinos-López et al., 2021). Specifically, we used the hospitalization data two weeks ahead (denoted as Y_i) as the response variable for concentration, cases, and TPR at time i.

Our goal was to accurately capture the relationship between predictor variables and hospitalization outcomes by incorporating a two-week lag. This enabled us to assess the model's predictive capabilities with greater accuracy, looking ahead to two weeks. It's important to note that the duration of this lag can be customized for each county. Nevertheless, our primary objective is to evaluate the model's ability to predict hospitalizations two weeks in advance, using data from wastewater.

In order to account for the interdependence in hospitalization data, all models presented below will consider the variable 'county' (i.e., the county where the hospital is situated) as a random effect in the predictor. The Linear Mixed Models (LMMs) were executed using the "lme" function from the "nlme" package in R.

2.6 Model Evaluation

We tested 9 statistical prediction models to determine their accuracy in forecasting, using the mean absolute scaled error (MASE) as our benchmark. To create these models, we used different predictors in Equation 1, including the number of confirmed cases (cases), TPR, N gene concentration (N), normalized N gene concentration (N/PMMoV), a wave factor, and interaction effects between normalized concentration and a wave factor (N/PMMoV×wave). The wave factor was added to account for several factors related to the emergence of a new variant, which cannot be measured directly due to the lack of available data.

The models M1, M2, M3, and M4 assume fixed effects predictor variables such as cases, TPR, N gene concentration, and normalized N gene concentration, respectively. All four models also consider a county-level random effect. On the other hand, models M5, M6, M7, and M8 include a wave effect factor to models M1, M2, M3, and M4, respectively. Model M9 is similar to model M8 but includes an N/PMMoV×wave interaction. A summary of these models and their comparative metrics are described in Table 1.

To evaluate the performance of the models, we used ten-fold cross-validation. The data was split into training and testing sets at random. The training set received 80% of the data, while the testing set was allocated 20%. The model was created using the training set, and its prediction performance was evaluated using the test set. This process was repeated 10 times, creating 10 random partitions. The overall prediction performance was reported as the average of the 10 partitions.

The MASE was used to compare the observed values with the forecasted values to measure the accuracy of predictions. The MASE was computed across 10 partitions, $\text{MASE} = \frac{1}{10}\sum_{k=1}^{10} \text{MASE}_k$, where MASE_k of the k-th fold is estimated by

$$\text{MASE}_k = \frac{\frac{1}{J}\sum_{j}^{J} |y_j - \hat{y}_j|}{\frac{1}{T-1}\sum_{t=2}^{T} |y_t - y_{t-1}|}$$

where y_t and $\hat{y}_t$ represent the observed and predicted values at time t, respectively. In addition to the MASE, the Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC) were also used to evaluate model fitness, with lower values of AIC and BIC indicating superior model fit.

3 Results

We present the findings for the nine models used to predict COVID-19 hospitalizations for 21 counties in California. Using the mean absolute scaled error (MASE) as a measure of the accuracy of the model forecasts, we calculated differences in accuracy percent between models in order to specify the reduction percentage of the mean absolute error relative to comparative models. We found that M2 outperformed models M1, M3, and M4 by 27.7%, 29.3%, and 32.6%, respectively. Models M1, M3, and M4 exhibited comparable performance.

By adding the variable of wave, models M1-M4 were updated to M5-M8, respectively. This update significantly enhanced the forecasting accuracy of the models which previously did not include wave as a variable. Out of all the models, M5 had the highest prediction capacity (25% better than M1), while M6 had the lowest (6.4% compared to model M2). Adding a wave effect to M3 and M4 also improved their performance, with a performance increment of 12.1% and 14.6%, respectively. Model M9, which included an interaction term for normalized concentrations and wave interaction along with a fixed effect of wave variable, did not result in any improvement compared to Model 8. Therefore, while it is crucial to include the wave effect, the concentration×wave interaction does not enhance prediction accuracy. Furthermore, adding interaction terms of Cases×wave, TPR×wave, and N×wave to models M5, M6, and M7 did not enhance their performance. It should be noted that the modifications made to models M5-M7 were not thoroughly evaluated in terms of their impact on the models' outcomes.

Our study demonstrates the feasibility of WW data to provide accurate predictions for hospitalizations at a level that is similar to the accuracy observed in predicting confirmed cases and TPR. The implications of these findings are significant in enhancing the real-time tracking of infectious

diseases.

The hospitalization rate predictions for California counties, based on Model M8, are presented in Figure 4. The predicted hospitalization rates align well with the observed data, indicating the model's effectiveness in forecasting hospitalization trends.

Table 1: Model prediction performance based on the mean absolute scaled error (MASE), across 10 random partitions, assuming 80% of the total data set for training and the remaining 20% for testing.

Model	**MASE****	**AIC**	**BIC**
M1 = Cases + County*	0.976 (0.075)	872	891
M2 = TPR + County*	0.764 (0.063)	417	437
M3 = N + County*	0.988 (0.071)	845	864
M4 = N/PMMoV + County*	1.013 (0.071)	804	824
M5 = M1 + Wave	0.781 (0.069)	550	579
M6 = M2 + Wave	0.718 (0.067)	327	356
M7 = M3 + Wave	0.881 (0.078)	641	670
M8 = M4 + Wave	0.884 (0.08)	634	664
M9 = M8 + N/PMMoV×Wave	0.863 (0.077)	619	658

∗ Included as a random effect.
∗∗ Mean (standard deviation).

4 Discussion

Analyzing wastewater has been shown to be an effective way to estimate SAR-CoV-2 incidence, detect new variants, and correlate with hospitalizations. Therefore, it can play a crucial role in epidemiological surveillance. This study evaluated the effectiveness of using wastewater-based surveillance in 21 California counties as a predictive tool for hospitalizations. The study population corresponds to the wastewater measurement of 21 wastewater treatment plants and covers a population of 13,503,427 inhabitants.

The WW sampling periodicity was weekly at all the studied counties but in situations with low disease surveillance, less frequent monitoring may be sufficient to keep close monitoring. As the COVID-19 pandemic is no longer classified as a public health emergency, the measures taken to control the spread of the disease have significantly reduced. The recent change in emergency status has led to a significant reduction in SARS-CoV-2 testing. As a result, there is a shortage of data that could help us better understand how the virus is spreading. Due to the difficulties in conducting regular SARS-CoV-2 testing, utilizing wastewater-based surveillance presents a distinctive and unprecedented opportunity to enhance disease monitoring not just for SARS-CoV-2 but also for other respiratory illnesses like influenza and respiratory syncytial virus (RSV).

We employed mixed effects models to evaluate associations between hospital admission rates for COVID-19 and the concentration of SARS-CoV-2 RNA in wastewater across 21 counties in

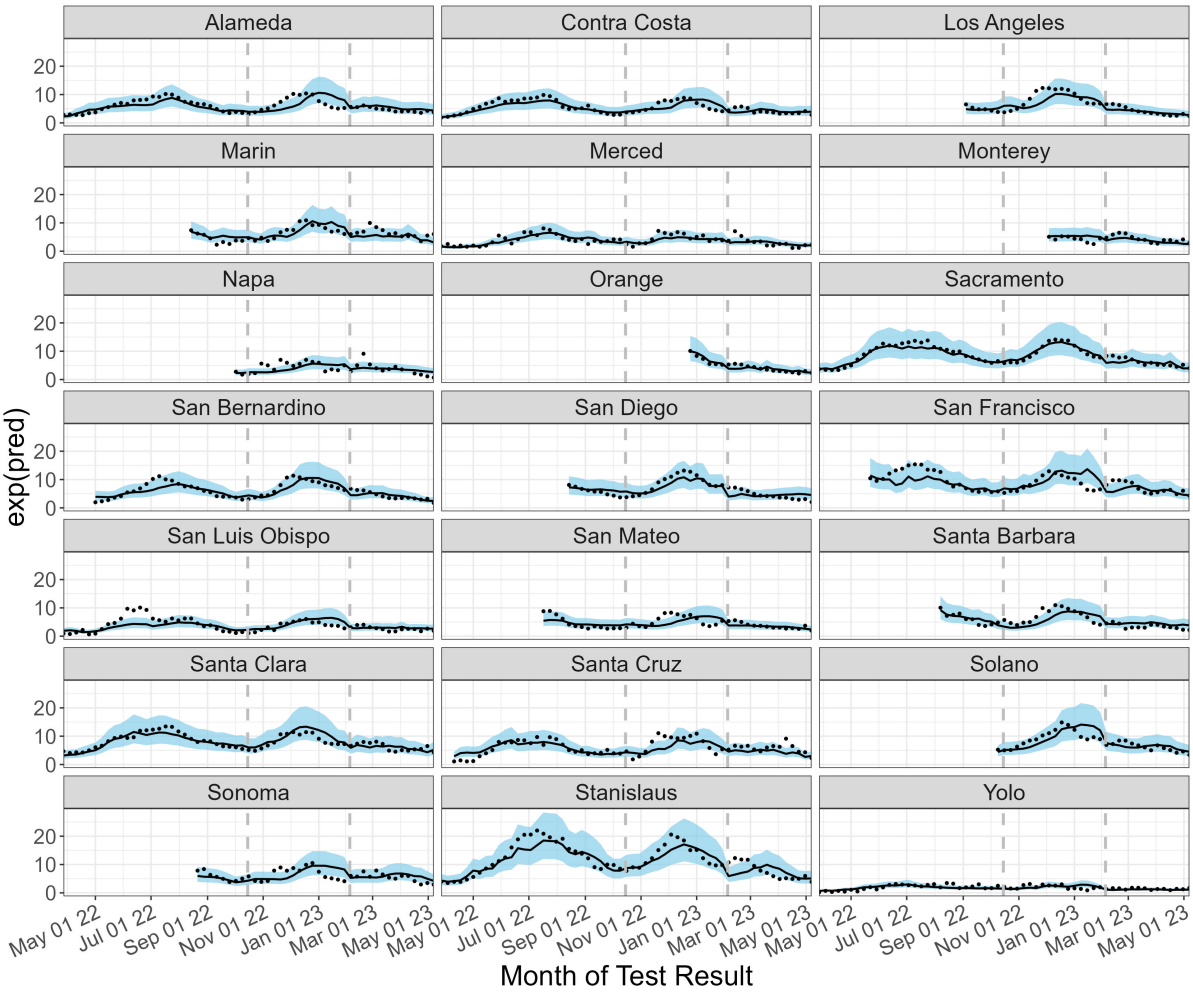

Figure 4: Estimated COVID-19 hospitalization rate assuming Model 8. Black-solid lines and blue-shadow area describe the mean and 80% prediction intervals, respectively. Black dots represent the observed data.

California. Based on the results, using various predictors and accounting for the wave effect can greatly affect the accuracy of model predictions. Specifically, Model M6, which includes TPR, performs better than Model M5, which only considers the number of confirmed cases. This finding is in agreement with previous research Dallal et al. (2021), indicating that TPR is a more reliable indicator of disease spread than confirmed cases.

Furthermore, including the wave effect enhances the accuracy of their predictions compared to earlier models that lacked it. Model M6, which utilizes both TPR and the wave effect, displays the most precise prediction performance, with lower mean absolute scaled error (MASE) values than the other models. It is worth noting that the wave effect is introduced to account for the non-autonomous nature of the extended COVID-19 pandemic outbreak. Epidemic outbreaks are intricate systems influenced by a variety of factors, including virus variants, human behavior, vaccination status, viral

evolution, prior immunity, and other time-dependent variables.

It is worth noting that Model M9, which examines the interaction between the normalized concentration and the wave, does not significantly improve the performance of Model M8. However, it is important to keep in mind that this analysis only included three COVID-19 waves, and only a few counties in the study had wastewater data for all three waves. Additionally, some counties only began collecting wastewater data after the emergence of the Omicron BA.2+BA.4+BA.5 (wave 1). Therefore, there is not enough information to determine if the relationship between hospitalization and viral load is consistent across all waves. Future studies should comprehensively explore the time-dependent factors associated with each wave.

Our study has shown that wastewater data can accurately predict hospitalizations two weeks in advance, without losing much predictive ability compared to traditional testing methods. This is especially important when testing capacity is limited. These results have implications for improving real-time monitoring of infectious diseases. Firstly, the ability to predict hospitalizations using wastewater data two weeks in advance allows decision-makers to take timely action by implementing targeted interventions and allocating resources. Secondly, relying on wastewater data bypasses the limitations and challenges associated with testing capacity, offering a comprehensive and alternative surveillance strategy.

A limitation of our study is that we didn't consider the correlation between different communities or sewersheds in our model. Adding a spatial correlation structure to the models could have helped to show how communities are interdependent when needed. However, we did use the classification of regions provided by California, like Northern California, Greater Sacramento, San Joaquin, Bay Area, and Southern California, to include a regional effect in our model. Even though we included this information, it didn't lead to a significant improvement in our models' performance.

One further limitation of our study pertains to the hospitalization data utilized. The data accounted for hospital admission numbers of patients diagnosed with COVID-19 during their hospital stay. Nevertheless, it is crucial to acknowledge that these numbers may not precisely reflect the seriousness of COVID-19 complications. The dataset may involve individuals admitted to the hospital for reasons unrelated to COVID-19 but who tested positive during their hospital stay. Therefore, relying solely on hospitalization numbers may not accurately gauge the impact and severity of COVID-19 on the population.

For future research, it would be beneficial to include a spatial correlation structure in the models to acknowledge the interdependence among various communities or sewersheds. This would provide a better understanding of how geographical factors affect the COVID-19 and SARS-CoV-2 RNA concentration in wastewater, thereby increasing the accuracy of predictions.

Additionally, it's essential to consider an adaptive modeling framework to tackle the ongoing pandemic's non-autonomous nature. This framework would enable us to adjust model parameters when a new variant becomes dominant or when significant changes occur in the disease dynamics. With this approach, we can generate more reliable and up-to-date forecastings that can help

decision-makers implement timely interventions and strategies. It is important to note that the current study did not specifically address this aspect, leaving it open for future research. In order to gain a deeper understanding of the dynamics of COVID-19 and increase the accuracy of predictions in wastewater-based surveillance systems, it would be highly beneficial to incorporate a spatial correlation structure and utilize an adaptive modeling framework.

Availability Statement

All code and underlying data are publicly available through the GitHub repository (link). Analyses were carried out using R version 4.2.2.

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.

Acknowledgments

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.

Credit authorship contribution statement

M.L.D.T and **J.C.M.L.**: Conceptualization, Methodology, Software, Formal analysis, Writing - Original draft preparation, Writing - Reviewing & Editing. **A.N.D**, **C.C.N.**, and **H.N.B.**: Writing - methods, Writing - Reviewing & Editing. **M.N.**: Methodology, research oversight/supervision, Writing - Reviewing & Editing.

References

A. B. Boehm, B. Hughes, D. Duong, V. Chan-Herur, A. Buchman, M. K. Wolfe, and B. J. White. Wastewater concentrations of human influenza, metapneumovirus, parainfluenza, respiratory syncytial virus, rhinovirus, and seasonal coronavirus nucleic-acids during the covid-19 pandemic: a surveillance study. *The Lancet Microbe*, 4(5):e340–e348, 2023a. doi: 10.1016/S2666-5247(22)00386-X.

A. B. Boehm, M. K. Wolfe, K. R. Wigginton, A. Bidwell, B. J. White, B. Hughes, D. Duong, V. Chan-Herur, H. N. Bischel, and C. C. Naughton. Human viral nucleic acids concentrations in

wastewater solids from central and coastal california usa. *Scientific Data*, 10(1):396, 2023b. doi: 10.1038/s41597-023-02297-7.

CDPH. California department of public health (cdph). Available online: https://data.chhs.ca.gov/organization/california-department-of-public-health. (Accessed on 05/23/2023).

A. A. Dallal, U. AlDallal, and J. A. Dallal. Positivity rate: an indicator for the spread of covid-19. *Current Medical Research and Opinion*, 37(12):2067–2076, 2021. doi: 10.1080/03007995.2021.1980868. URL https://doi.org/10.1080/03007995.2021.1980868. PMID: 34517740.

M. L. Daza-Torres, J. C. Montesinos-López, M. Kim, R. Olson, C. W. Bess, L. Rueda, M. Susa, L. Tucker, Y. E. García, A. J. Schmidt, C. C. Naughton, B. H. Pollock, K. Shapiro, M. Nuño, and H. N. Bischel. Model training periods impact estimation of covid-19 incidence from wastewater viral loads. *Science of The Total Environment*, 858:159680, 2023. ISSN 0048-9697. doi: 10.1016/j.scitotenv.2022.159680.

D. Dowdy and G. D'Souza. Covid-19 testing: understanding the "percent positive". *Johns Hopkins Bloomberg School of Public Health*, 2020. URL https://publichealth.jhu.edu/2020/covid-19-testing-understanding-the-percent-positive.

L. Fenga and M. Gaspari. Predictive capacity of covid-19 test positivity rate. *Sensors*, 21(7), 2021. ISSN 1424-8220. doi: 10.3390/s21072435. URL https://www.mdpi.com/1424-8220/21/7/2435.

A. Galani, R. Aalizadeh, M. Kostakis, A. Markou, N. Alygizakis, T. Lytras, P. G. Adamopoulos, J. Peccia, D. C. Thompson, A. Kontou, A. Karagiannidis, E. S. Lianidou, M. Avgeris, D. Paraskevis, S. Tsiodras, A. Scorilas, V. Vasiliou, M.-A. Dimopoulos, and N. S. Thomaidis. Sars-cov-2 wastewater surveillance data can predict hospitalizations and icu admissions. *Science of The Total Environment*, 804:150151, 2022. ISSN 0048-9697. doi: https://doi.org/10.1016/j.scitotenv.2021.150151. URL https://www.sciencedirect.com/science/article/pii/S0048969721052281.

B. Hughes, D. Duong, B. J. White, K. R. Wigginton, E. M. G. Chan, M. K. Wolfe, and A. B. Boehm. Respiratory syncytial virus (rsv) rna in wastewater settled solids reflects rsv clinical positivity rates. *Environmental Science & Technology Letters*, 9(2):173–178, 2022. doi: 10.1021/acs.estlett.1c00963. URL https://doi.org/10.1021/acs.estlett.1c00963.

K. F. Kadonsky, C. Naughton, R. Olson, M. Susa, S. G.L., M. Nuno, J. Daza-Torres, M.L. andf Montesinos-López, et al. Expansion of wastewater-based disease surveillance to improve health equity in california's central valley: sequential shifts in case-to-wastewater and hospitalization-to-wastewater ratios. *Frontiers in Public Health*, 11, 2023. doi: 10.3389/fpubh.2023.1141097.

T. Kumblathan, Y. Liu, G. K. Uppal, S. E. Hrudey, and X.-F. Li. Wastewater-based epidemiology for community monitoring of sars-cov-2: Progress and challenges. *ACS Environmental Au*, 1(1):18–31, 2021. doi: 10.1021/acsenvironau.1c00015. URL https://doi.org/10.1021/acsenvironau.1c00015.

J. C. Montesinos-López, M. L. Daza-Torres, Y. E. García, L. A. Barboza, F. Sanchez, A. J. Schmidt, B. H. Pollock, and M. Nuño. The role of sars-cov-2 testing on hospitalizations in california. *Life*, 11:1336, 2021. doi: 10.3390/life11121336.

J. C. Montesinos-López, M. L. Daza–Torres, Y. E. García, C. Herrera, C. W. Bess, H. N. Bischel, and M. Nuño. Bayesian sequential approach to monitor covid-19 variants through positivity rate from wastewater. *medRxiv*, 2023. doi: 10.1101/2023.01.10.23284365. URL https://www.medrxiv.org/content/early/2023/01/13/2023.01.10.23284365.

J. R. Paul, J. D. Trask, and S. Gard. Ii. poliomyelitic virus in urban sewage. *The Journal of Experimental Medicine*, 71(6):765–77, 1940. doi: 10.1084/jem.71.6.765.

SCAN. Sewer coronavirus alert network. https://soe-wbe-pilot.wl.r.appspot.com/charts#page=locationcompare-variants, 2023. (Accessed on 05/23/2023).

E. Symonds, K. H. Nguyen, V. Harwood, and M. Breitbart. Pepper mild mottle virus: a plant pathogen with a greater purpose in (waste) water treatment development and public health management. *Water research*, 144:1–12, 2018. doi: https://doi.org/10.1016/j.watres.2018.06.066.

K. K. To, K.-H. Chan, I. W. Li, T.-Y. Tsang, H. Tse, J. F. Chan, I. F. Hung, S.-T. Lai, C.-W. Leung, Y.-W. Kwan, Y.-L. Lau, T.-K. Ng, V. C. Cheng, J. S. Peiris, and K.-Y. Yuen. Viral load in patients infected with pandemic h1n1 2009 influenza a virus. *Journal of Medical Virology*, 82 (1):1–7, 2010. doi: https://doi.org/10.1002/jmv.21664. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/jmv.21664.

J. D. Trask, A. Vignec, and J. R. Paul. Poliomyelitis virus in human stools. *Journal of the American Medical Association*, 111(1):6–11, 1938. doi: 10.1001/jama.1938.02790270008002.

Verily. Wastewater testing dashboard. URL http://publichealth.verily.com/?v=SC2_N&l=Coeur+d%27Alene%2C+IDh. (Accessed on 05/23/2023).

WastewaterSCAN. Wastewaterscan dashboard. URL https://data.wastewaterscan.org/tracker. (Accessed on 05/23/2023).

Appendix

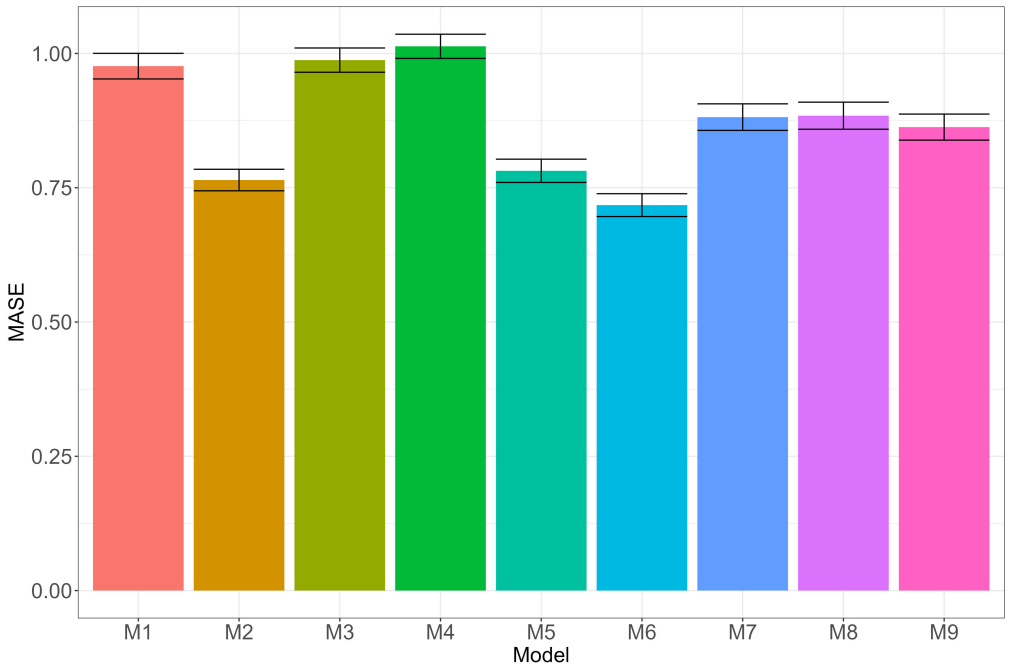

Figure 5: **Performance prediction of models described in Table 1:** mean absolute scaled error (MASE) with its standard deviation across the 10 partitions.

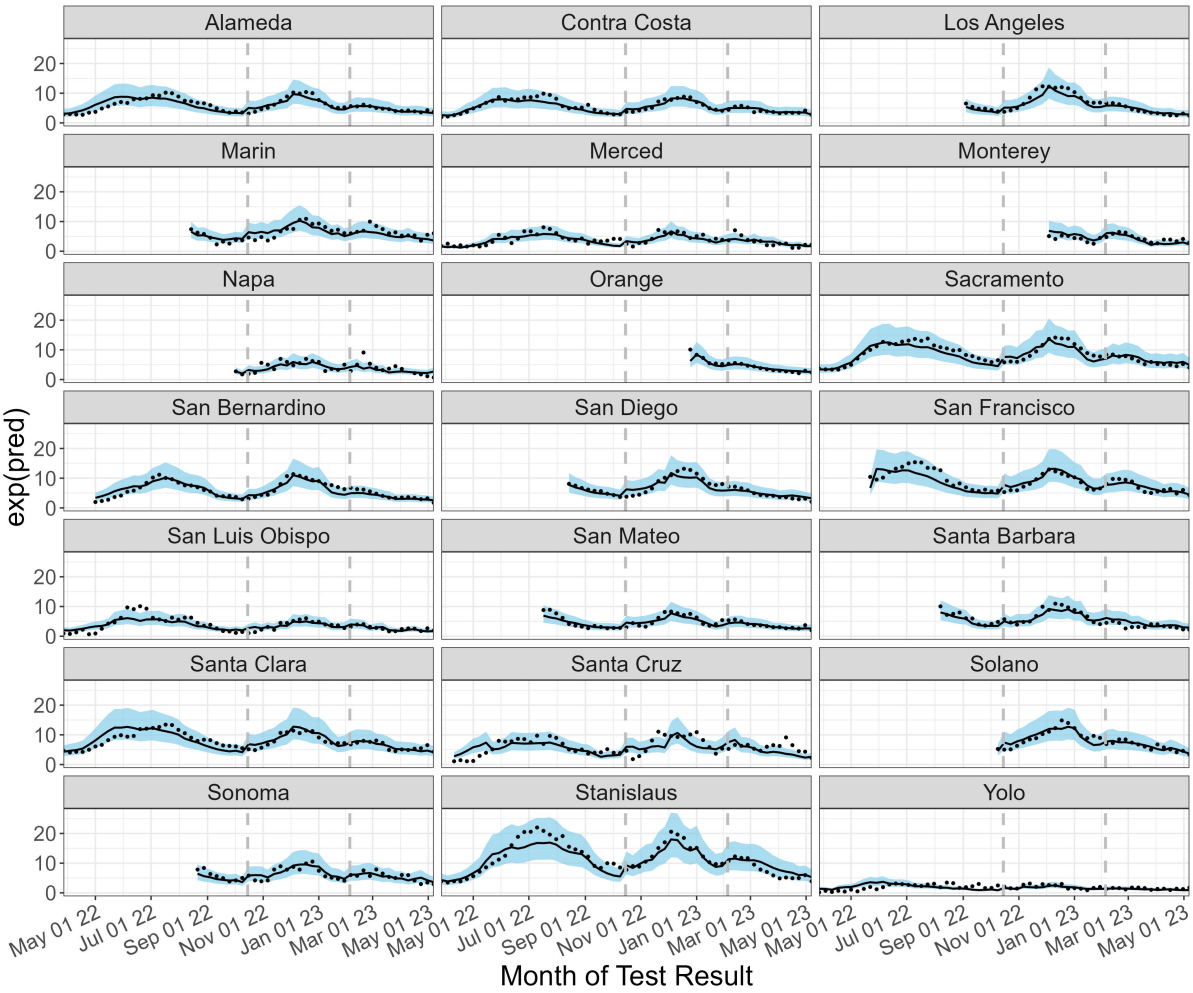

Figure 6: **Model M5 = Cases + wave + County.** Estimated COVID-19 hospitalization rate. Black-solid lines and blue-shadow area describe the mean and 80% prediction intervals, respectively. Black dots represent the observed data.

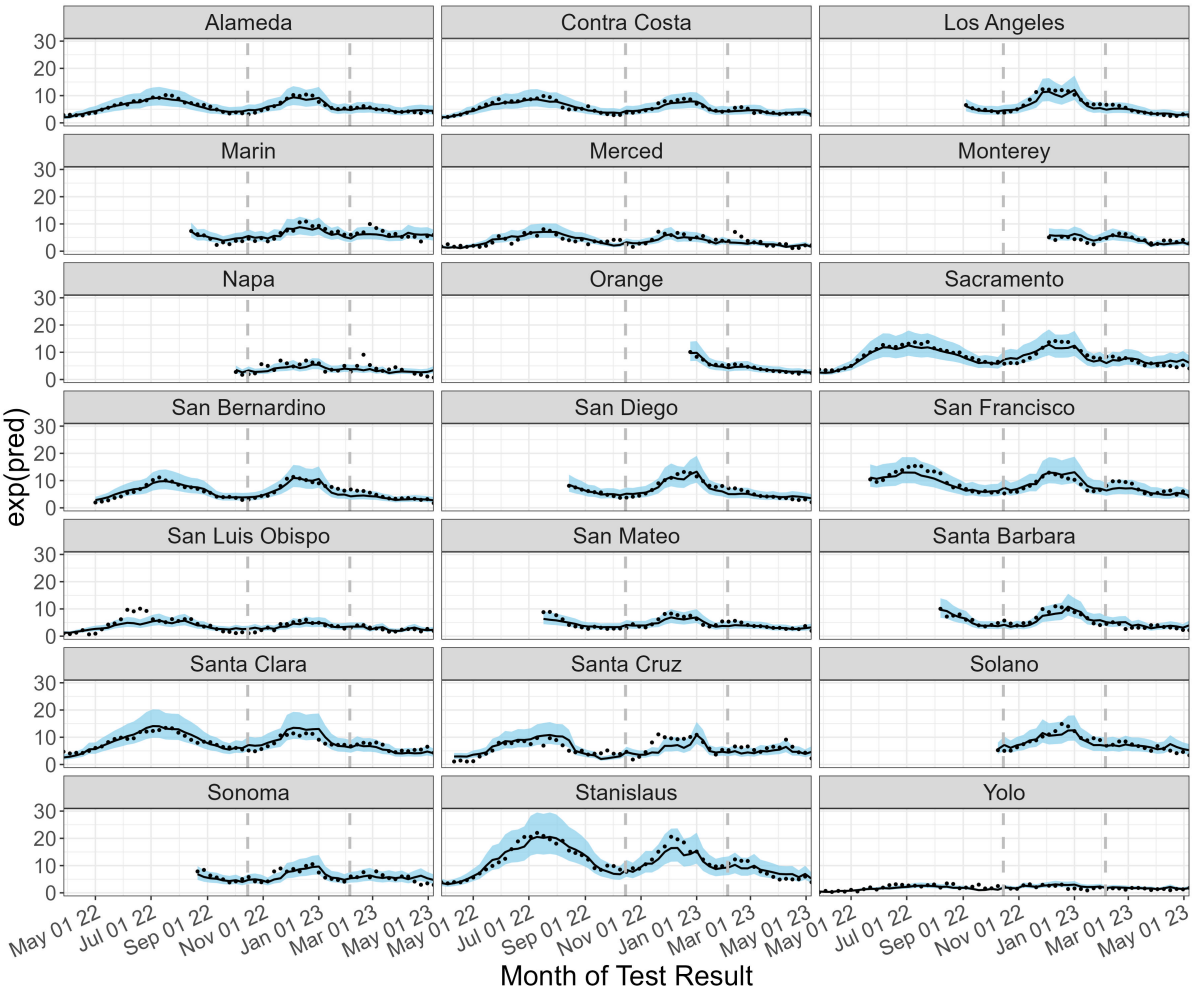

Figure 7: **Model M6 = TPR + wave + County.** Estimated COVID-19 hospitalization rate. Black-solid lines and blue-shadow area describe the mean and 80% prediction intervals, respectively. Black dots represent the observed data.

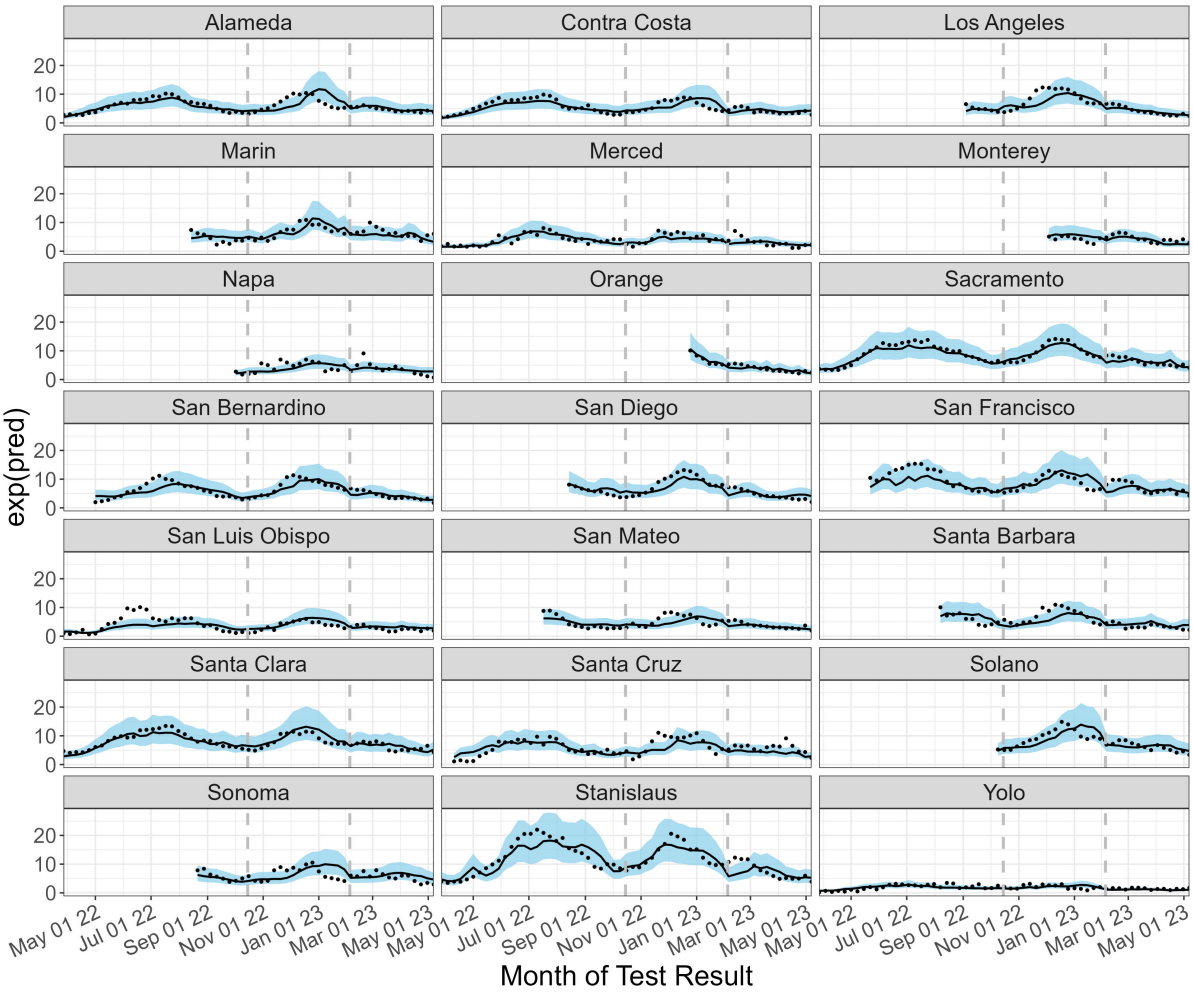

Figure 8: **Model M7 = N + wave + County.** Estimated COVID-19 hospitalization rate. Black-solid lines and blue-shadow area describe the mean and 80% prediction intervals, respectively. Black dots represent the observed data.