

COVID-19 Forecasting from U.S. Wastewater Surveillance Data: A Retrospective Multi-Model Study (2022–2024)*

Faharudeen Alhassan^{a,b,*,1}, Hamed Karami^{a,b}, Amanda Bleichrodt^b, James M. Hyman^e, Isaac C. H. Fung^c, Ruiyan Luo^b and Gerardo Chowell^{b,d,*}

^aDepartment of Mathematics and Statistics, Georgia State University, Atlanta, GA, 30303, USA

^cDepartment of Biostatistics, Epidemiology and Environmental Health Sciences, Jiann-Ping Hsu College of Public Health, Georgia Southern University, Statesboro, GA, 30460, USA

^eDepartment of Mathematics, Tulane University, New Orleans, LA, 70118, USA

^bDepartment of Population Health Sciences, Georgia State University, Atlanta, GA, 30303, USA

^dDepartment of Applied Mathematics, Kyung Hee University, Yongin, GA, 17104, Korea

ARTICLE INFO

Keywords:

COVID-19

Wastewater surveillance

Ensemble modeling

n-sub-epidemic framework

Time series forecasting

ABSTRACT

Accurate and reliable forecasting models are critical for guiding public health responses and policy decisions during pandemics such as COVID-19. Retrospective evaluation of model performance is essential for improving epidemic forecasting capabilities. In this study, we used COVID-19 wastewater data from CDC's National Wastewater Surveillance System to generate sequential weekly retrospective forecasts for the United States from March 2022 through September 2024, both at the national level and for four major regions (Northeast, Midwest, South, and West). We produced 133 weekly forecasts using 11 models, including ARIMA, generalized additive models (GAM), simple linear regression (SLR), Prophet, and the *n*-sub-epidemic framework (top-ranked, weighted-ensemble, and unweighted-ensemble variants). Forecast performance was assessed using mean absolute error (MAE), mean squared error (MSE), weighted interval score (WIS), and 95% prediction interval coverage. The *n*-sub-epidemic unweighted ensembles outperformed all other models at 3–4-week horizons, particularly at the national level and in the Midwest and West. ARIMA and GAM performed best at 1–2-week horizons in most regions, whereas Prophet and SLR consistently underperformed across regions and horizons. These findings highlight the value of region-specific modeling strategies and demonstrate the utility of the *n*-sub-epidemic framework for real-time outbreak forecasting using wastewater surveillance data.

1. Introduction

The COVID-19 pandemic has highlighted the urgent need for efficient, scalable, and timely disease surveillance methods. Traditional epidemiological approaches rely primarily on clinical testing and hospital records, which, while valuable, are often constrained by delays in reporting, limited testing availability, and the under-detection of asymptomatic infections [Nixon, Jindal, Parker, Marshall, Reich, Ghobadi, Lee, Truelove, and Gardner, 2022, Alvarez, Bielska, Hopkins, Belal, Goldstein, Slick, Pavalagantharajah, Wynfield, Dakey, Gedeon, et al., 2023]. These limitations have underscored the importance of alternative surveillance methods, especially wastewater-based epidemiology (WBE), which has emerged as a cost-effective, non-invasive, and real-time tool to monitor SARS-CoV-2 transmission at the community level. Viral particles are shed in human waste even before individuals exhibit symptoms or seek medical attention [Medema, Heijnen, Elsinga, Italiaander, and Brouwer, 2020, Kitajima, Ahmed, Bibby, Carducci, Gerba, Hamilton, Haramoto, and Rose, 2020]. Unlike traditional testing, which depends on individual participation and healthcare access, WBE provides a comprehensive snapshot of community-wide infection levels. As a result, it has been widely implemented across multiple countries to track and forecast COVID-19 trends [Peccia, Zulli, Brackney, Grubaugh, Kaplan, Casanovas-Massana, Ko, Malik, Wang, Wang, et al., 2020]. Its potential for early detection makes WBE especially valuable for improving calibration and lead-time accuracy in epidemic forecasting.

*Corresponding author: gchowell@gsu.edu

✉ falhassan@gsu.edu (F. Alhassan); hkarami@gsu.edu (H. Karami); ableichrodt@gsu.edu (A. Bleichrodt); mhyman@tulane.edu (J.M. Hyman); cfung@georgiasouthern.edu (I.C.H. Fung); rluo@gsu.edu (R. Luo); gchowell@gsu.edu (G. Chowell)

ORCID(s):

Recent research has demonstrated a strong and consistent correlation between SARS-CoV-2 RNA concentrations in wastewater and the number of reported COVID-19 cases [Li, Zhang, Sherchan, Orive, Lertxundi, Haramoto, Honda, Kumar, Arora, Kitajima, et al., 2023b]. This relationship establishes WBE as an effective early warning system, allowing public health officials to detect emerging outbreaks and monitor infection trends [Gonzalez, Curtis, Bivins, Bibby, Weir, Yetka, Thompson, Keeling, Mitchell, and Gonzalez, 2020]. WBE has been successfully applied to epidemic forecasting on different geographical scales, from localized wastewater treatment plants to national surveillance networks [Joseph-Duran, Serra-Compte, Sàrrias, Gonzalez, López, Prats, Català, Alvarez-Lacalle, Alonso, and Arnaldos, 2022, Daza-Torres, Montesinos-López, Kim, Olson, Bess, Rueda, Susa, Tucker, García, Schmidt, et al., 2023]. For instance, a study in Catalonia (Spain) demonstrated the potential of WBE to predict SARS-CoV-2 incidence across multiple regions, reinforcing its role in epidemiological forecasting [Joseph-Duran et al., 2022]. Similarly, research in the United States (U.S.) showed that wastewater data could predict weekly COVID-19 hospital admissions up to four weeks in advance, emphasizing its value in public health preparedness [Li, Liu, Gao, Sherchan, Zhou, Khan, Van Loosdrecht, and Wang, 2023a]. In rural communities where clinical testing is often limited, WBE has proven particularly beneficial. In a recent multi-state analysis of 189 wastewater treatment plants during 2023-2024, it was demonstrated that SARS-CoV-2 RNA concentrations precede hospitalization admissions by 2 to 12 days, with a lead time varying by state-level WBE coverage and facility characteristics [Schenk, Rauch, Zulli, and Boehm, 2024]. A study in Idaho successfully integrated wastewater data with SEIR modeling to forecast outbreaks with lead times of up to 11 days, showcasing its ability to enhance surveillance in underserved areas [Meadows, Coats, Narum, Top, Ridenhour, and Stalder, 2025]. However, most of these studies have focused on short-term forecasts or single geographic areas, leaving a gap in understanding long-term WBE forecasting performance across diverse U.S. regions. With the gradual de-escalation of widespread clinical testing and case-based reporting in the U.S., wastewater surveillance is becoming an increasingly critical data source for public health forecasting. This shift further underscores the need to evaluate the forecasting performance of models applied to wastewater data across different geographic and temporal contexts.

Retrospective forecasting plays a crucial role in evaluating the performance of COVID-19 prediction models across diverse settings. By systematically comparing past forecasts with actual observed outcomes, researchers can refine methodologies, optimize model parameters, and improve predictive accuracy [Drews, Kumar, Singh, De La Sen, Singh, and Pandey, 2022]. Such evaluations are essential for identifying sources of uncertainty, enhancing forecast calibration, and determining the effectiveness of various statistical and mechanistic modeling approaches [Colonna, Nane, Choma, Cooke, and Evans, 2022]. Despite these advancements, many forecasting models struggle to account for abrupt shifts in outbreak dynamics, such as sudden spikes in infections or unexpected declines in hospitalizations, critical changes that directly influence public health decision-making [Lopez, Cramer, Pagano, Drake, O'Dea, Adey, Ayer, Chhatwal, Dalgic, Ladd, et al., 2024]. Improving the ability of models to detect these turning points is crucial for enhancing forecasting systems and ensuring timely interventions. Integrating dynamic modeling approaches with wastewater surveillance data has been proposed as a promising strategy to enhance predictive accuracy and real-time outbreak monitoring [Phan, Brozak, Pell, Oghuan, Gitter, Hu, Ribeiro, Ke, Mena, Perelson, et al., 2023, Proverbio, Kemp, Magni, Ogorzaly, Cauchie, Gonçalves, Skupin, and Aalto, 2022].

In this study, we conducted a comprehensive retrospective evaluation of 1- to 4-week forecast performance across 11 different models. Our analysis includes n -sub-epidemic models alongside statistical approaches such as autoregressive integrated moving average (ARIMA), generalized additive models (GAMs), Simple Linear Regression (SLR), and Meta (Facebook's) Prophet model. We assess forecasting performance at both national and regional levels, covering four U.S. regions: the Midwest, Northeast, South, and West. The evaluation period spans from March 5, 2022, to September 14, 2024. To ensure a rigorous and standardized assessment, we utilize key epidemic forecasting performance metrics, including mean absolute error (MAE), mean squared error (MSE), 95% prediction interval (PI) coverage, and weighted interval score (WIS). Integrating dynamic modeling approaches with wastewater surveillance data has been proposed as a promising strategy to enhance predictive accuracy and real-time outbreak monitoring. Our findings provide insights into the optimal selection of forecasting models for wastewater-based epidemic surveillance across different geographic contexts and time horizons.

2. Methods

2.1. Data Source and Preparation

The wastewater data used for forecasting COVID-19 trends was obtained from the National Wastewater Surveillance System (NWSS), which is hosted by the Centers for Disease Control and Prevention (CDC) [CDC, 2024a]. Wastewater data provides an early warning system, indicating whether infection levels in a particular region may be increasing or decreasing.

The dataset covers national-level data and data from four regions: Midwest, Northeast, South, and West. The regions are defined as follows:

- **West:** Alaska, Arizona, California, Colorado, Guam, Hawaii, Idaho, Montana, Nevada, New Mexico, Oregon, Utah, Washington, Wyoming (N=14).
- **Midwest:** Illinois, Indiana, Iowa, Kansas, Michigan, Minnesota, Missouri, Nebraska, North Dakota, Ohio, South Dakota, Wisconsin (N=12).
- **Northeast:** Connecticut, Maine, Massachusetts, New Hampshire, New Jersey, New York, Pennsylvania, Puerto Rico, Rhode Island, Vermont (N=10).
- **South:** Arkansas, Alabama, Delaware, District of Columbia, Florida, Georgia, Kentucky, Louisiana, Maryland, Mississippi, North Carolina, Oklahoma, South Carolina, Tennessee, Texas, Virginia, and West Virginia (N=17).

Wastewater viral activity levels are used to assess the concentration of the virus in wastewater, which may indicate the infection risk in a given area. These levels are categorized as follows:

- **Minimal:** Up to 1.5
- **Low:** 1.5–3
- **Moderate:** 3–4.5
- **High:** 4.5–8
- **Very High:** Greater than 8

The viral activity level is calculated based on the number of standard deviations above the baseline and transformed to a linear scale using the formula:

$$\text{Wastewater Viral Activity Level (WVAL)} = e^x,$$

where x is the number of standard deviations relative to baseline [CDC, 2024b].

It is important to note that WVAL represents a normalized metric rather than absolute viral concentration, allowing for comparison across different wastewater treatment facilities with varying population sizes and sampling protocols.

Wastewater samples were collected by state, tribal, local, and territorial partners as part of CDC's National Wastewater Surveillance System (NWSS), following CDC guidance on site selection, sampling, and laboratory methods [Kirby, Walters, Jennings, Fugitt, LaCross, Mattioli, Marsh, Roberts, Mercante, Yoder, and Hill, 2021]. The network expanded from 209 sampling sites in September 2020 to more than 1,500 by December 2022, covering roughly 47% of the U.S. population [Adams, Bias, Welsh, Webb, Reese, Delgado, Person, West, Shin, and Kirby, 2024]. Catchment sizes range from systems serving thousands of residents to large utilities serving over one million people, and participating laboratories implemented standardized methods and quality controls recommended by CDC and NWSS partners [Adams et al., 2024, Kirby et al., 2021]

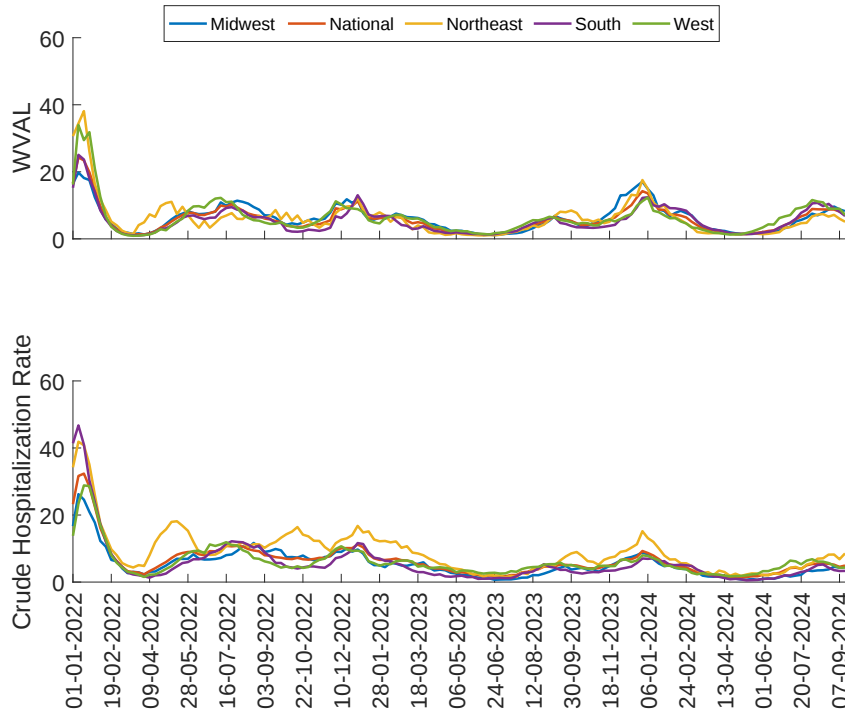


Figure 1: Time-series comparison of wastewater viral activity level (WVAL) and crude COVID-19 hospitalization rates from January 2022 to September 2024. The top panel illustrates the WVAL detected in wastewater surveillance data, and the bottom panel represents the crude hospitalization rate over the same period. Both datasets exhibit periodic fluctuations, with notable peaks aligning at several time points, suggesting a potential correlation between wastewater viral activity and hospital admissions. The observed trends indicate that increases in wastewater viral concentration often precede corresponding surges in hospitalization rates, highlighting the potential utility of wastewater surveillance as an early warning system for monitoring COVID-19 outbreaks. This relationship emphasizes the importance of integrating wastewater-based epidemiology into public health decision-making. The lead time between WVAL peaks and hospitalization surges appears to vary by region and outbreak intensity, typically ranging from 2-12 days based on recent U.S. surveillance data [Schenk et al., 2024]

2.2. Model calibration and specification

We used a total of 11 models in our analysis, including the top-ranked models, ensemble models (weighted and unweighted) from the n -sub-epidemic framework, along with ARIMA, GAM, GLM, and SLR (each estimated under a normality assumption), and the Prophet model. A detailed description of the methodologies for these models is provided below.

The CDC publishes wastewater data weekly. For each model and location in our analysis, we used all data between January 1, 2022, and September 14, 2024, released by the CDC on October 26, 2024, to construct rolling calibration windows. The forecasting period began on March 5, 2022 (10 weeks after the first considered data point), and advanced in one-week increments, with models refit each week using the most recent 10 weeks of data to generate 1- to 4-week-ahead forecasts; the 10-week calibration window balances sufficient data for capturing recent dynamics while minimizing the influence of changing variability and structural breaks, consistent with short-term epidemic forecasting practice [Bleichrodt, Luo, Kirpich, and Chowell, 2024a, Bleichrodt, Dahal, Maloney, Casanova, Luo, and Chowell, 2023].

We assumed a normal error structure in the observed data. This decision was based on two main considerations: (i) the short 10-week calibration window helped minimize the influence of changing variability in the data over time, and (ii) the models demonstrated consistently high calibration coverage, suggesting that the normality assumption provided an adequate approximation of the residual distribution over the fitting period.

2.3. The n -sub-epidemic modeling framework

This framework is a flexible approach used to capture complex epidemic patterns by combining multiple overlapping sub-epidemics [Chowell, Dahal, Bleichrodt, Tariq, Hyman, and Luo, 2024, Chowell, Dahal, Tariq, Roosa, Hyman, and Luo, 2022]. Rather than viewing an outbreak as a single continuous event, this framework breaks it

into smaller, more manageable components, each representing a distinct sub-epidemic. These sub-epidemics may correspond to different geographic regions, demographic groups, or periods within the course of an epidemic.

Each sub-epidemic is modeled using a generalized logistic growth function [Chowell et al., 2024] that captures its growth rate, final size, and scaling behavior. This allows the model to represent a wide range of early outbreak dynamics, from slow, sub-exponential growth to rapid, exponential expansion. The mathematical structure is consistent across all sub-epidemics, but each sub-epidemic is characterized by a unique set of parameters.

An n -sub-epidemic trajectory consists of n overlapping sub-epidemics and is described by the system of differential equations

$$\frac{dC_i(t)}{dt} = A_i(t) r_i C_i^{p_i}(t) \left(1 - \frac{C_i(t)}{K_{0i}}\right), \quad i = 1, \dots, n. \quad (1)$$

In the original n -sub-epidemic formulation, $C_i(t)$ denotes the cumulative number of cases associated with the i -th sub-epidemic. In our wastewater setting, we instead treat $C_i(t)$ as the wastewater viral activity level (WVAL) contributed by the i -th sub-epidemic, and we use the same dynamical form as a flexible phenomenological model for the rise and fall of the WVAL signal. The term $A_i(t)$ is a binary activation indicator (0 or 1) that determines when the i -th sub-epidemic is active. Under this interpretation, the i -th sub-epidemic evolves according to Eq. (1).

where:

- $C_i(t)$ is the wastewater viral activity level (WVAL) contributed by the i -th sub-epidemic at time t ,
- $\frac{dC_i(t)}{dt}$ is the rate of change in the modeled WVAL signal for the i -th sub-epidemic at time t ,
- $r_i C_i^{p_i}(t)$ is the growth term governing how rapidly the wastewater signal increases, with p_i modulating the growth regime,
- $\left(1 - \frac{C_i(t)}{K_{0i}}\right)$ is a saturation term that slows growth as $C_i(t)$ approaches the asymptotic level K_{0i} .

The model parameters have the following interpretations:

- r_i : intrinsic growth rate of the WVAL signal for the i -th sub-epidemic,
- $p_i \in [0, 1]$: scaling parameter controlling the effective growth regime (with $p_i = 0$ corresponding to approximately constant growth, $p_i = 1$ to exponential growth, and $0 < p_i < 1$ to sub-exponential growth),
- K_{0i} : asymptotic (plateau) level of the WVAL signal for the i -th sub-epidemic.

To determine when a new sub-epidemic begins, the model employs an indicator $A_i(t)$ that activates the $(i + 1)$ -th sub-epidemic once the modeled wastewater viral activity level $C_i(t)$ for the i -th sub-epidemic exceeds a specified threshold C_{thr} .

$$A_i(t) = \begin{cases} 1, & \text{if } C_{i-1}(t) > C_{\text{thr}}, \\ 0, & \text{otherwise,} \end{cases} \quad \text{for } i = 2, \dots, n, \quad (2)$$

with $A_1(t) = 1$ to activate the first sub-epidemic by default.

The initial value of observed viral loads is given by:

$$C_1(0) = I_0, \quad (3)$$

where I_0 is the initial viral load at the start of the outbreak.

The cumulative epidemic curve is defined as the sum of all individual sub-epidemic trajectories:

$$C_{\text{tot}}(t) = \sum_{i=1}^n C_i(t). \quad (4)$$

When modeling a single sub-epidemic ($n = 1$), the model reduces to the classic three-parameter logistic growth model. For more complex epidemic dynamics (such as multiple peaks or plateaus, as seen in Figure 1), using multiple sub-epidemics ($n > 1$) increases modeling flexibility and accuracy. A full n -sub-epidemic model requires estimation of $3n + 1$ parameters. In this paper, we set $n \leq 2$ in the n -sub-epidemic trajectory.

To enhance the accuracy of the estimation and evaluation process, we used 30 initial guesses for the parameter estimation process (`numstartpoints = 30`). We conducted 300 bootstrap realizations (B) to characterize parameter uncertainty effectively.

The n -sub-epidemic modeling framework offers a powerful and adaptable tool for analyzing and forecasting disease outbreaks. It is particularly well-suited to epidemics that display multi-wave progression. This framework has demonstrated excellent performance in modeling various infectious diseases, including Ebola, Zika, COVID-19, and mpox [Chowell et al., 2024, 2022, Bleichrodt et al., 2024a].

2.3.1. Parameter estimation and model selection

To estimate the model parameters, we used a nonlinear least-squares fitting method, aligning the model's predictions with the observed COVID-19 wastewater data [Chowell et al., 2024]. After fitting several models, each using a different threshold value C_{thr} to define when new sub-epidemics begin. The C_{thr} values are generated by uniformly partitioning the cumulative sum of the smoothed signal into several levels equal to the total data points. We evaluated their performance using the corrected Akaike Information Criterion (AIC_c). This criterion helps identify models that best balance goodness-of-fit and model complexity.

The formula for AIC_c is:

$$AIC_c = n_d \log(\text{SSE}) + 2m + \frac{2m(m+1)}{n_d - m - 1}, \quad (5)$$

where each component serves a specific purpose in model selection:

- $n_d \log(\text{SSE})$: Measures the goodness-of-fit, where SSE is the sum of squared errors between observed and predicted values
- $2m$: Penalty term for model complexity to prevent overfitting, where m is the number of parameters
- $\frac{2m(m+1)}{n_d - m - 1}$: Finite sample correction that becomes important when the sample size n_d is small relative to the number of parameters

The parameters are defined as follows:

- SSE is the sum of squared errors between the observed data and model predictions,
- m is the number of model parameters,
- n_d is the number of data points.

This method assumes that the residuals (errors) in the data follow a normal distribution. More details about the parameter estimation procedure are provided in [Chowell et al., 2022].

2.3.2. Parametric bootstrapping

To assess uncertainty in the parameters of the best-fitting model $f(t, \hat{\Theta})$, we used a parametric bootstrapping approach [Hastie, 2009]. Here $f(t, \hat{\Theta})$ denotes the model's predicted value at time t using the estimated parameters $\hat{\Theta}$. This method involves generating multiple datasets by resampling from the fitted model and then reestimating the parameters for each sample. This helps quantify variability in parameter estimates and model forecasts without relying on closed-form solutions.

In addition to estimating standard errors and confidence intervals, we used bootstrapping to produce one- to four-week-ahead forecasts with associated uncertainty bounds. In this analysis, we generated 300 bootstrap realizations to characterize the parameter uncertainty and uncertainty in forecasting.

2.3.3. Building ensemble n -sub-epidemic models

The ensemble models were constructed by combining several of the top-performing sub-epidemic models. Some ensembles were unweighted, meaning each selected model contributed equally. Others were weighted according to their AIC_c rankings, for example, by combining the top 2 models (Ensemble 2) or the top 3 models (Ensemble 3), based on the ordering $AIC_{c_1} \leq \dots \leq AIC_{c_r}$, where $i = 1, \dots, r$.

The construction of these ensemble models follows the methodology detailed in [Chowell et al., 2022]. The 95% Prediction intervals (PIs) for the ensemble forecasts were computed using the same bootstrapping method described above, allowing us to quantify uncertainty in the ensemble outputs.

2.4. Auto-regressive integrated moving average model

The ARIMA model is a well-established statistical technique commonly used for forecasting time series data. It has been successfully applied by [Mondal, Shit, and Goswami, 2014] in finance, by [Tektaş, 2010, Dimri, Ahmad, and Sharif, 2020] in weather prediction, and in infectious disease modeling by [Chowell et al., 2022, Long, Tan, and Newman, 2023, Iftikhar, Khan, Khan, and Khan, 2023, Benvenuto, Giovanetti, Vassallo, Angeletti, and Ciccozzi, 2020]. Due to its flexibility and simplicity, the ARIMA model has become a standard benchmark in epidemic forecasting [Fattah, Ezzine, Aman, El Moussami, and Lachhab, 2018, Bleichrodt et al., 2024a].

An ARIMA model is defined by three components: the autoregressive (AR) part, which captures dependencies between current and past observations; the integrated (I) part, which applies differencing to stabilize the series; and the moving average (MA) part, which models the relationship between an observation and past forecast errors.

The general form of an ARIMA(p, d, q) model is given by:

$$\phi(B)(1 - B)^d y_t = c + \theta(B)\epsilon_t, \quad (6)$$

where y_t is the observed time series at time t , B is the backward shift operator such that $By_t = y_{t-1}$, and ϵ_t is a white noise error term. The polynomial $\phi(B) = 1 - \phi_1 B - \dots - \phi_p B^p$ represents the autoregressive component, while $\theta(B) = 1 + \theta_1 B + \dots + \theta_q B^q$ represents the moving average component. The differencing operator $(1 - B)^d$ is employed to transform the time series into a stationary form by removing trends. In the special case where $d = 1$, the model operates on the first-differenced series $y'_t = y_t - y_{t-1}$. Higher-order differencing is used when required to remove more complex trends.

For this study, we fitted ARIMA models to weekly COVID-19 wastewater data. Model selection was performed automatically using the Hyndman-Khandakar algorithm [Hyndman and Khandakar, 2008], which selects optimal values of p , d , and q based on information criteria. Forecasts were generated using an interactive R Shiny dashboard [Bleichrodt, Phan, Luo, Kirpich, and Chowell-Puente, 2024b], and any negative predictions are set to zero.

2.5. Generalized additive model

The generalized additive model (GAM) extends the generalized linear model by allowing smooth, non-linear functions of the predictors, enabling flexibility in capturing complex relationships between covariates and the response variable [Wood, 2017, Baayen and Linke, 2020]. This flexibility makes GAMs particularly useful for modeling time series data where trends may be complex and may not follow simple parametric forms.

In this study, we used GAMs to model weekly COVID-19 viral load with time as the sole predictor. Assuming a Gaussian error, the model is specified as follows:

$$y_t = \beta_0 + s(t) + \epsilon_t, \quad (7)$$

where β_0 is an intercept, $s(t)$ is an unspecified smooth function of time, and $\epsilon_t \sim N(0, \sigma^2)$ represents the modeling error at time t [Bleichrodt et al., 2024b, A., 2021]. Although we adopt the Gaussian family here, the GAM framework readily accommodates alternative distributions if the normality assumption is violated.

The smooth function $s(t)$ is represented as a linear combination of basis functions:

$$s(t) = \sum_{k=1}^K \beta_k b_k(t), \quad (8)$$

where $\{b_k(t)\}$ are predefined basis functions and $\{\beta_k\}$ are the coefficients estimated from the data. In our implementation, we used the `gam()` function from the `mgcv` package in R [Wood, n.d.], which by default employs penalized

regression splines (e.g., thin plate splines or cubic splines). All the forecasts for this model were generated on the R-shiny dashboard [Bleichrodt et al., 2024b].

The number of basis functions K was selected based on the length of the calibration window for each forecast period. To avoid overfitting and control the level of smoothness, a penalty was applied to the coefficients $\{\beta_k\}$, and the optimal smoothness parameter was selected via generalized cross-validation [Wood, 2017, Bleichrodt et al., 2024b].

GAMs offer several advantages for epidemic forecasting, including flexibility to model non-linear dynamics, interpretability through additive components, and the capacity to incorporate domain knowledge via the choice of smooth terms [A., 2021, Wood, 2017]. Although GAMs capture short-term patterns effectively, their utility diminishes when extrapolating beyond the calibration window. Because the spline basis functions are fitted to the calibration data, the model quickly loses predictive validity once the series extends past that window, causing forecast accuracy to decline after only a few time steps (in our case, one to two weeks). Consequently, we use the GAM solely as a near-term benchmark whose forecasts complement, rather than replace, those from mechanistic or ensemble models.

2.6. Simple Linear Regression (SLR)

Simple Linear Regression is a foundational statistical method that models the relationship between a response variable and a single predictor using a linear equation. For our time series forecasting application, we model the WVAL as a linear function of time:

$$y_t = \beta_0 + \beta_1 t + \epsilon_t, \quad (9)$$

where y_t is the WVAL at time t , β_0 is the intercept, β_1 is the slope coefficient representing the linear trend, and $\epsilon_t \sim N(0, \sigma^2)$ is the error term. The parameters β_0 and β_1 are estimated using ordinary least squares, minimizing the sum of squared residuals. While SLR provides a simple baseline for comparison and can capture overall trends, its assumption of a constant linear relationship limits its ability to model the complex, non-linear dynamics often observed in epidemic time series. Any negative predictions from the SLR model were set to zero. Despite its limitations, including SLR in our analysis provides a useful benchmark to assess the relative performance gains achieved by more sophisticated modeling approaches.

2.7. Meta's (Facebook's) Prophet model

Originally introduced for business analytics, Facebook's Prophet model [Taylor and Letham, 2018] has become increasingly popular across different fields, including epidemiology. It has been applied to model and forecast disease trajectories such as those observed in COVID-19 [Satrio, Darmawan, Nadia, and Hanafiah, 2021, Battineni, Chintalapudi, and Amenta, 2020, Kirpich, Shishkin, Weppelmann, Tchernov, Skums, and Gankin, 2022] and mpox outbreaks [Long et al., 2023]. Prophet is built on the assumption that the underlying time series $y(t)$ can be decomposed into interpretable components representing different temporal patterns:

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t, \quad (10)$$

where each component captures different aspects of the time series:

- $g(t)$: The trend component that captures long-term, non-seasonal changes in the time series
- $s(t)$: The seasonal component that models recurring patterns (e.g., weekly, monthly, or yearly cycles)
- $h(t)$: The holiday/event component that accounts for the impact of special events or anomalies
- ϵ_t : The error term representing random fluctuations not captured by the other components

For our analysis, we used the default settings of the R prophet function from the 'prophet' package, as described by Taylor [Taylor, n.d.]. Forecasting was conducted using the `predict` function from the same package [RDocumentation, n.d.]. Any negative forecasted values were set to zero. Additional details on the model fitting process are available in [Bleichrodt et al., 2024b, Taylor, n.d.].

2.8. Model Nomenclature and Forecasting Framework

The n -sub-epidemic framework generates multiple candidate models at each forecast date by fitting models with $n = 1$ and $n = 2$ sub-epidemics to the calibration data (we set $n_{\max} = 2$). We rank these candidates using corrected Akaike Information Criterion (AIC_c) and select the top three performers. Model names reference this ranking:

- **Rank 1:** Best-performing n -sub-epidemic model (lowest AIC_c)
- **Rank 2:** Second-best n -sub-epidemic model
- **Rank 3:** Third-best n -sub-epidemic model

Ensemble models combine these ranked models using two weighting approaches:

- **EM2 W / EM3 W:** Weighted ensembles of the top 2 or 3 models, with weights derived from AIC_c differences using Akaike weights (Burnham & Anderson, 2002). Specifically, model i receives weight $w_i = \exp(-\Delta_i/2) / \sum_j \exp(-\Delta_j/2)$, where $\Delta_i = AIC_{c,i} - \min(AIC_c)$.
- **EM2 UW / EM3 UW:** Unweighted ensembles where each of the top 2 or 3 models contributes equally ($w_i = 1/k$). This approach provides robustness when model rankings are unstable or AIC_c differences are negligible.

Comparison models (ARIMA, GAM, SLR, Prophet) maintain their conventional names and are fitted using the procedures described in preceding subsections. All forecasts use epidemiological weeks (epiweeks) as defined by the CDC's MMWR calendar, where each week runs Saturday through Friday. This aligns with NWSS reporting structure and ensures temporal consistency.

For a forecast generated at the end of epiweek t , the targets are:

- 1-week ahead: epiweek $t + 1$ (7 days)
- 2-week ahead: epiweek $t + 2$ (14 days)
- 3-week ahead: epiweek $t + 3$ (21 days)
- 4-week ahead: epiweek $t + 4$ (28 days)

We generated retrospective out-of-sample forecasts using a rolling, fixed-length calibration window. At each forecast origin t (March 2022–September 2024), we:

1. Fit all models using only the **most recent 10 weeks of data**, i.e., $[t - 9, t]$;
2. Generate 1–4 week-ahead forecasts;
3. Use the bootstrap method to construct the 95% PIs and get the medians;
4. Advance one week and repeat.

This protocol ensures no future information leaks into training data. Model parameters are re-estimated at each forecast time, allowing the models to adapt to evolving dynamics. The resulting forecast-observation pairs reflect operational deployment conditions where public health agencies would use only past data to predict future trends.

3. Model Evaluation

In this study, we assessed the forecasting performance of each model by computing the average values of four key metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), Weighted Interval Score (WIS), and 95% Prediction Interval (PI) coverage. Here, the term "average" denotes the mean of each metric, calculated across all forecasting periods for every combination of model, location, and forecast horizon.

3.1. Mean Absolute Error (MAE)

The mean absolute error represents the average magnitude of errors in the model predictions [Willmott and Matsuura, 2005, Kaundal, Kapoor, and Raghava, 2006]. It is calculated as the average difference between the predicted values and the actual values in the test dataset, with equal weight assigned to all individual differences. The range of MAE is from 0 to infinity, where smaller values indicate better model performance. Due to this property, it is sometimes referred to as a negatively oriented score [Baran and Nemoda, 2016].

$$\text{MAE} = \frac{1}{N} \sum_{j=1}^N |y_j - \hat{y}_j|. \quad (11)$$

3.2. Mean Square Error (MSE)

The mean square error is another widely used metric for evaluating the performance of regression models [Willmott and Matsuura, 2005]. It computes the squared differences between the predicted and actual values, summing them up and averaging over all data points. Squaring the differences removes negative signs and gives greater emphasis to larger errors. A lower MSE indicates a better fit to the data. The formula for MSE is given by:

$$\text{MSE} = \frac{1}{N} \sum_{j=1}^N (y_j - \hat{y}_j)^2. \quad (12)$$

3.3. Weighted Interval Score (WIS)

The Weighted Interval Score quantifies how closely a probabilistic forecast distribution aligns with the observed outcome, using the scale of the data. It combines the absolute error of the median forecast with a weighted sum of interval scores at different prediction levels, providing a comprehensive measure of forecast accuracy [Cramer, Ray, Lopez, Bracher, Brennen, Castro Rivadeneira, Gerding, Gneiting, House, Huang, et al., 2022].

WIS was introduced by [Bracher, Ray, Gneiting, and Reich, 2021] as an effective scoring rule for probabilistic forecasting of epidemics. For a forecast with prediction intervals at nominal coverage levels

$$\alpha \in \{0.02, 0.05, 0.10, \dots, 0.90, 0.95, 0.98\},$$

the Interval Score (IS) at level α is defined as:

$$\text{IS}_\alpha(F, y) = (u_\alpha - l_\alpha) + \frac{2}{\alpha}(l_\alpha - y)\mathbb{1}\{y < l_\alpha\} + \frac{2}{\alpha}(y - u_\alpha)\mathbb{1}\{y > u_\alpha\} \quad (13)$$

where l_α and u_α are the lower and upper bounds of the $(1-\alpha) \times 100\%$ prediction interval, and y is the observed value. The first term rewards narrow intervals (sharpness), while the penalty terms penalize under-coverage when observations fall outside the interval.

The Weighted Interval Score aggregates these interval scores:

$$\text{WIS}_{\alpha_0:K}(F, y) = \frac{1}{K + \frac{1}{2}} \left(w_0 |y - m| + \sum_{k=1}^K w_k \text{IS}_{\alpha_k}(F, y) \right), \quad (14)$$

where $w_k = \frac{\alpha_k}{2}$ for $k = 1, 2, \dots, K$, m is the median (50th percentile) of the forecast distribution F , α_k (for $k = 1, \dots, K$) is the nominal coverage level of the central $(1-\alpha_k) \times 100\%$ prediction interval and $w_0 = \frac{1}{2}$. Lower WIS indicates better forecast performance. Because WIS inherits the units of the forecast target ($\log_{10}$ WVAL in our case), differences in WIS values can be interpreted directly in terms of forecast error magnitude.

3.4. Coverage rate of 95% Prediction Interval (95% PI)

The 95% PI is a range within which future observations are expected to fall with a 95% probability, based on the predictive distribution [Montgomery, Jennings, and Kulahci, 2015]. It provides a measure of the uncertainty in a forecast and is crucial for assessing the reliability of predictive models. A well-calibrated model should produce prediction intervals that closely match the observed proportion of actual values falling within the interval.

The coverage rate of the 95% PI can be expressed as:

$$\frac{1}{N} \sum_{j=1}^N \mathbf{1}\{L_j < Y_j < U_j\}, \quad (15)$$

where L_j and U_j represent the lower and upper bounds of the 95% prediction interval, respectively. The variable Y_j denotes the observed value, and $\mathbf{1}\{\cdot\}$ is an indicator function that equals 1 if Y_j falls within the prediction interval, and 0 otherwise [Chowell et al., 2024].

3.5. Baseline Model and Forecast Skill Scores

To contextualize model performance improvements, we use Simple Linear Regression (SLR, described in Section 2.7) as our baseline reference model. SLR provides a minimal-skill benchmark representing simple trend extrapolation without capturing complex epidemic dynamics.

We compute forecast skill scores to quantify improvement over this baseline by comparing the *average* error across all forecast dates. For example, for MAE, we define

$$\text{Skill Score} = 1 - \frac{\overline{\text{MAE}}_{\text{model}}}{\overline{\text{MAE}}_{\text{SLR}}}, \quad (16)$$

where $\overline{\text{MAE}}_{\text{model}}$ and $\overline{\text{MAE}}_{\text{SLR}}$ denote the mean MAE across all forecast dates for a given model and for the SLR baseline, respectively. Positive values indicate that the model outperforms SLR (e.g., $\text{SS} = 0.40$ implies a 40% reduction in error compared to linear trend extrapolation). Skill scores provide a scale-independent performance metric across regions with different baseline error magnitudes. Skill scores for MAE, MSE, and WIS across all models, regions, and horizons are presented in Supplementary Figures S6–S10 and referenced in the Results where appropriate.

4. Results

Our retrospective forecasting study generated 133 weekly forecast sets spanning March 2022 through September 2024, producing over 10,000 individual forecast-observation pairs across all regions, models, and horizons. Figures 4 through 8 present aggregated performance metrics (mean values computed across all 133 forecast weeks) to identify systematic patterns in model performance. To complement these averages, we also present the full distributions of MAE, MSE, and WIS across forecast weeks in Supplementary Figures S1–S5 (boxplots for each region), which highlight variability and consistency of model performance over time. This aggregation approach balances comprehensive evaluation with interpretability. Throughout the Results, we reference models using the nomenclature established in Section 2.8: Rank 1-3 (top n -sub-epidemic models by AIC_c), EM2 W/UW and EM3 W/UW (2- and 3-model ensembles, weighted and unweighted), and comparison models ARIMA, GAM, SLR, and Prophet.

We start the Results section by highlighting forecasts from September 7, 2024, one of the final evaluation dates in our retrospective study, to show how different models behave when predicting beyond the calibration period. Figure 2 displays national-level forecasts from all models, while Figure 3 shows the same for the Midwest region. In both cases, sub-epidemic models closely follow the observed decline in WVAL and maintain prediction intervals that are wide enough to include future observations. In contrast, statistical models like SLR and Prophet tend to produce much narrower intervals that quickly diverge from the actual data. Together, these plots preview two key patterns we explore in the following sections: the strong performance of ensemble models at 3–4 week horizons and their ability to better anticipate regional turning points.

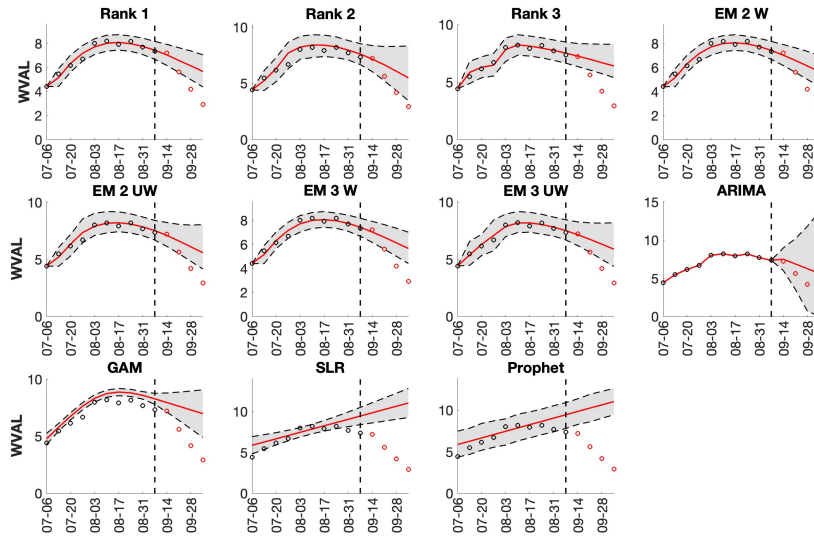


Figure 2: Forecasts of WVAL at the national level for an example forecast date (originating on September 7, 2024) across all models. Each panel shows the median prediction (red solid line) and the corresponding 95% prediction interval (gray shaded region bounded by black dashed lines). Black circles denote calibration data, while red circles indicate observations used for evaluation. The vertical dashed line marks the end of the calibration period. This example illustrates the varying widths of prediction intervals across models, with ensemble approaches generally providing wider but more reliable uncertainty bounds than single models such as SLR or Prophet.

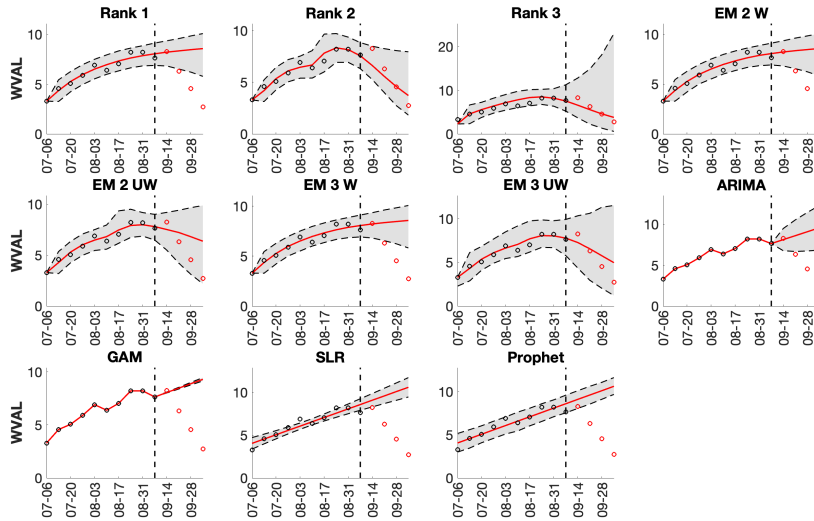


Figure 3: Forecasts of WVAL for the Midwest region for an example forecast date (originating on September 7, 2024) across all models. Each panel shows the median prediction (red solid line) and the corresponding 95% prediction interval (gray shaded region bounded by black dashed lines). Black circles denote calibration data, while red circles indicate observations used for evaluation; the vertical dashed line marks the end of the calibration period. The GAM model exhibits particularly tight prediction intervals during calibration, making its gray region difficult to distinguish from the median curve. The sub-epidemic models more accurately track the observed declining trend, whereas several statistical models tend to overestimate future WVAL values.

4.1. National

Results for the National region are summarized in Fig. 4, which pools all metrics (MAE, MSE, WIS, Coverage95) across 1–4-week horizons into a single panel. Errors increase with forecast horizon, but relative model rankings remain stable.

At the 1-week horizon, ARIMA is among the best-performing models, achieving the lowest MSE (0.65) and WIS (0.35). GAM also performs strongly, with MAE 0.47, MSE 0.65, and WIS 0.37. Several n -sub-epidemic models (EM2 UW and EM3 UW) provide similarly competitive results, with WIS values of 0.42 and 0.38, respectively. For uncertainty quantification, EM3 UW attains the highest 95% PI coverage (90.98%), followed closely by EM2 UW (86.47%).

At the 2-week horizon, GAM again is part of the set of models that provided the lowest error metrics, achieving MAE 0.59, MSE 1.03, and WIS 0.47. ARIMA performs similarly, with MAE 0.60, MSE 0.98, and WIS 0.45. EM3 UW and EM2 UW produce MAE values of 0.63 and 0.66 and WIS values of 0.47 and 0.50, remaining competitive. EM3 UW again yields the highest 95% PI coverage at this horizon (88.72%).

At the 3-week horizon, EM3 UW emerges as the strongest overall model, achieving the lowest MAE (0.74), MSE (1.25), and WIS (0.56). Rank3 and EM2 UW also perform well, with MAE 0.78 & 0.77 respectively and the same WIS value of 0.60. ARIMA also provided competitive values at this horizon (MAE 0.78, MSE 1.45, WIS 0.61). EM3 UW again attains the highest 95% PI coverage (85.96%).

At the 4-week horizon, EM3 UW continued to be among the strongest models, achieving the lowest WIS (0.65) and one of the lowest MAE values (0.84). EM2 UW also performs well, with MAE 0.87, MSE 1.62, and WIS 0.70. For uncertainty quantification, EM3 UW again provides the highest 95% PI coverage (84.40%), indicating consistently well-calibrated forecast intervals at the national scale.

Overall, these results indicate that GAM and ARIMA offer highly accurate 1–2-week-ahead forecasts for the National region, while EM2 UW and EM3 UW provide the most reliable 3–4-week-ahead forecasts. Some models achieve very similar accuracy at certain horizons—for example, at the 1-week horizon, GAM and ARIMA produced comparable MAE values (0.47 and 0.48).

To help evaluate the magnitude and variability of these differences, we present the distributions of all metrics across models, forecast periods, and horizons at the national level in Supplementary Fig. S1 (with corresponding regional distributions in Supplementary Figs. S2–S5). Performance metric distributions across all 133 forecast periods (Supplementary Figs. S1–S5) reveal the consistency of model rankings. Focusing on the national level, Supplementary Fig. S1 shows that GAM and EM3 UW consistently achieve the lowest MAE values among all models, with their distributions clearly separated from SLR and Prophet. The boxplots show partial overlap among the top-performing models (GAM, EM3 UW, ARIMA, Rank 3), indicating that while ensemble approaches offer reliable advantages, performance differences among leading models vary across forecast periods. This pattern suggests that the ensemble approach provides dependable, though not universal, improvements across changing epidemic dynamics.

Skill scores relative to the SLR baseline (Section 3.5) quantify forecast improvement beyond simple trend extrapolation. At the national level, EM3 UW achieves MAE skill scores of 46% at 1-week ahead and 40–41% at 2–4 week horizons, representing substantial error reduction compared to linear trend forecasting (Supplementary Figure S6). For MSE, skill scores reach 55–70%, while WIS skill scores range from 49–54% across all horizons. These consistently strong values demonstrate that ensemble forecasting provides meaningful improvements over naive baseline methods throughout the forecast window. Regional skill score patterns (Supplementary Figures S7–S10) show similar results, with ensemble models consistently achieving positive skill scores across all regions and horizons.

EM3 UW	0.53	0.63	0.74	0.84	MAE
EM2 UW	0.57	0.66	0.77	0.87	
ARIMA	0.48	0.60	0.79	0.93	
Rank 3	0.56	0.68	0.78	0.89	
GAM	0.47	0.59	0.78	0.95	
EM2 W	0.66	0.77	0.87	0.98	
EM3 W	0.66	0.77	0.87	0.98	
Rank 1	0.66	0.77	0.87	0.98	
Rank 2	0.71	0.79	0.90	1.01	
Prophet	0.83	0.91	1.05	1.15	
SLR	0.83	0.91	1.05	1.15	
EM3 UW	0.63	0.93	1.25	1.56	MSE
EM2 UW	0.72	1.01	1.32	1.62	
ARIMA	0.62	0.98	1.45	1.83	
Rank 3	0.74	1.11	1.42	1.73	
GAM	0.65	1.03	1.59	2.02	
EM2 W	0.92	1.23	1.54	1.84	
EM3 W	0.92	1.23	1.54	1.84	
Rank 1	0.92	1.23	1.54	1.84	
Rank 2	1.45	1.72	2.04	2.35	
Prophet	1.36	1.57	1.95	2.23	
SLR	1.36	1.57	1.95	2.23	
EM3 UW	0.38	0.47	0.56	0.65	WIS
EM2 UW	0.42	0.50	0.60	0.70	
ARIMA	0.35	0.45	0.61	0.73	
Rank 3	0.41	0.51	0.60	0.70	
GAM	0.37	0.47	0.63	0.78	
EM2 W	0.50	0.60	0.71	0.82	
EM3 W	0.50	0.60	0.71	0.82	
Rank 1	0.50	0.60	0.71	0.82	
Rank 2	0.52	0.60	0.70	0.82	
Prophet	0.67	0.76	0.92	1.03	
SLR	0.70	0.78	0.93	1.04	
EM3 UW	90.98	88.72	85.96	84.40	95% PI
EM2 UW	86.47	82.71	79.20	75.94	
ARIMA	82.71	80.08	75.94	74.06	
Rank 3	81.95	80.83	78.20	75.75	
GAM	66.17	65.41	65.91	64.29	
EM2 W	63.16	56.02	52.13	48.68	
EM3 W	63.16	56.02	52.13	48.68	
Rank 1	64.66	57.14	52.88	49.06	
Rank 2	81.20	76.69	74.94	71.99	
Prophet	43.61	34.21	31.33	28.76	
SLR	33.83	29.70	27.57	26.88	
	1-wk	2-wk	3-wk	4-wk	

Figure 4: Log-transformed averages of MAE, MSE, and WIS, and the (untransformed) average 95% PI coverage across all models for 1–4 week horizons in the National region, spanning the period from March 5, 2022, to September 14, 2024 (133 forecasts). Lighter shades indicate smaller (better) values for MAE, MSE, and WIS, whereas darker shades indicate larger errors. For coverage, values closer to 95% indicate better performance. Black cells highlight the best-performing model(s) for each forecast horizon and metric.

4.2. Midwest

Results for the Midwest region are summarized in Fig. 5. Forecast errors increase with horizon length, but overall model rankings remain relatively stable. At the 1-week horizon, GAM and ARIMA are among the best-performing models, with GAM attaining the lowest MAE (0.63) and the lowest MSE (0.96), while ARIMA also performs competitively, particularly for MAE (0.64) and WIS (0.47). The EM3 UW ensemble provides the highest 95% PI coverage (87.97%), indicating strong uncertainty quantification even for near-term forecasts.

At the 2-week horizon, ARIMA and GAM are again among the best-performing models. ARIMA attained the lowest MAE (0.71), MSE (1.15), and WIS (0.54), while GAM provided a very competitive MAE (0.73), MSE (1.27) and WIS (0.57) values. EM3 UW also performed with providing very close MAE, MSE, and WIS values. EM3 UW again attains the highest 95% PI coverage (85.71%), indicating a well-calibrated ensemble option at shorter horizons.

Performance patterns shift at horizons 3 and 4. At the 3-week horizon, EM3 UW is part of the best-performing set of models, with the lowest MAE (0.91), the lowest MSE (1.65), and the lowest WIS (0.70), while also maintaining high 95% PI coverage (83.46%). At the 4-week horizon, EM3 UW continues to be among the top-performing models in terms of MAE (1.03), MSE (1.98), WIS (0.81), and PI coverage (81.58%). The EM2 UW also provided very similar values compared to EM3 UW at horizons 3 and 4.

Overall, these patterns suggest that GAM and ARIMA are well suited for 1–2-week forecasting in the Midwest, whereas the unweighted ensembles, specifically EM3 UW and EM2 UW, are among the most reliable options for 3–4 weeks, balancing accuracy and uncertainty quantification. Several models exhibit very similar performance in this region; for example, at the 3-week horizon, EM2 UW had an average WIS of 0.71 compared with 0.70 for EM3 UW.

To help assess the magnitude of these small differences, we present the distributions of the metrics across all models, forecast periods, and horizons in Supplementary Fig. S2, and the corresponding skill scores relative to the SLR baseline in Supplementary Fig. S7, rather than relying solely on point estimates.

ARIMA	0.64	0.71	0.93	1.06	MAE
EM3 UW	0.70	0.79	0.91	1.03	
EM2 UW	0.73	0.82	0.93	1.04	
GAM	0.63	0.73	0.95	1.08	
EM2 W	0.82	0.92	1.02	1.12	
EM3 W	0.82	0.92	1.02	1.12	
Rank 1	0.82	0.92	1.02	1.12	
Prophet	0.95	1.00	1.16	1.28	
SLR	0.95	1.00	1.16	1.28	
Rank 3	0.75	0.93	1.08	1.22	
Rank 2	0.75	0.80	0.92	1.04	
ARIMA	0.96	1.15	1.79	2.16	MSE
EM3 UW	1.04	1.29	1.65	2.00	
EM2 UW	1.11	1.37	1.68	1.98	
GAM	0.96	1.27	1.95	2.33	
EM2 W	1.34	1.61	1.90	2.19	
EM3 W	1.34	1.61	1.90	2.19	
Rank 1	1.34	1.61	1.90	2.19	
Prophet	1.69	1.81	2.30	2.63	
SLR	1.69	1.81	2.30	2.63	
Rank 3	1.38	2.25	2.71	3.19	
Rank 2	1.60	1.30	1.69	2.16	
ARIMA	0.47	0.54	0.73	0.85	WIS
EM3 UW	0.50	0.59	0.70	0.81	
EM2 UW	0.53	0.61	0.71	0.82	
GAM	0.50	0.57	0.79	0.89	
EM2 W	0.63	0.73	0.84	0.95	
EM3 W	0.63	0.73	0.84	0.95	
Rank 1	0.63	0.74	0.84	0.95	
Prophet	0.76	0.84	1.02	1.15	
SLR	0.80	0.86	1.03	1.15	
Rank 3	0.55	0.70	0.83	0.97	
Rank 2	0.57	1.04	5.15	6.17	
ARIMA	78.95	75.56	72.18	72.18	95% PI
EM3 UW	87.97	85.71	83.46	81.58	
EM2 UW	84.21	80.45	77.44	75.38	
GAM	72.18	70.30	65.41	67.67	
EM2 W	58.65	54.51	50.88	48.31	
EM3 W	58.65	54.51	50.88	48.31	
Rank 1	59.40	54.14	50.88	48.50	
Prophet	48.12	39.85	37.34	32.33	
SLR	35.34	30.83	31.08	29.14	
Rank 3	83.46	81.95	79.20	75.00	
Rank 2	81.20	77.44	74.94	72.93	
	1-wk	2-wk	3-wk	4-wk	

Figure 5: Log-transformed averages of MAE, MSE, and WIS, and the (untransformed) average 95% PI coverage across all models for 1–4 week horizons in the Midwest region, spanning the period from March 5, 2022, to September 14, 2024 (133 forecasts). Lighter shades indicate smaller (better) values for MAE, MSE, and WIS, whereas darker shades indicate larger errors. For coverage, values closer to 95% indicate better performance. Black cells highlight the best-performing model(s) for each forecast horizon and metric.

4.3. Northeast

Fig. 6 summarizes the results for the Northeast region. Across all horizons, this region shows some of the highest MAE, MSE, and WIS values, indicating that the Northeast was among the most challenging settings for forecasting in our study. We discuss potential reasons for this in detail in Section 5.1.

At the 1-week horizon, ARIMA and EM3 UW are the leading models. EM3 UW attains the lowest MAE (0.80), MSE (1.33), and WIS (0.59), with ARIMA, Rank2, and EM2 UW yielding very similar WIS values (0.61, 0.62, and 0.61, respectively). Both EM3 UW and the Rank3 of the n -sub-epidemic model attain the highest 95% PI coverage (86.47%), indicating well-calibrated uncertainty at this horizon.

At the 2-week horizon, ARIMA is clearly the best-performing model on error-based metrics, achieving the lowest MAE (0.77), MSE (1.28), and WIS (0.58). EM3 UW remains competitive, with slightly larger error values (MAE 0.85, MSE 1.46, WIS 0.63), and again attained high 95% PI coverage (86.09%), comparable to the Rank3 model (86.47%).

Performance patterns changed at horizons 3 and 4. At the 3-week horizon, EM3 UW provided the lowest MAE (0.92), MSE (1.66), and WIS (0.69) among all models. It also delivered high 95% PI coverage (85.96%), only slightly below the Rank3 model (86.47%). ARIMA remains reasonably competitive but with higher errors (MAE 1.41, MSE 2.83, WIS 1.00).

At the 4-week horizon, EM3 UW is the clear top performer, achieving the lowest MAE (1.00), MSE (1.95), and WIS (0.77), while also providing the highest 95% PI coverage (85.15%). The Rank 2 and EM2–UW ensembles are the next-best models on the error metrics at this horizon, with MAE values of 1.03, MSE values of 2.05, and WIS values of 0.82 and 0.80, respectively, illustrating the strength of unweighted ensemble methods for 3- and 4-week-ahead forecasts in this region.

Overall, these results indicate that ARIMA and the unweighted ensemble models offer the strongest 1–2-week-ahead performance in the Northeast, whereas unweighted n -sub-epidemic ensembles, particularly EM3 UW, provide the most reliable 3–4-week-ahead performance. In this region, some models yield very similar error values; for example, at the 1-week horizon, the EM3 UW model achieves an MAE of 0.80 compared with 0.82 for ARIMA. To help assess the magnitude of these differences, we present the distributions of all metrics across models, forecast periods, and horizons in Supplementary Fig. S3, and the corresponding skill scores relative to the SLR baseline in Supplementary Fig. S8, rather than relying solely on point estimates.

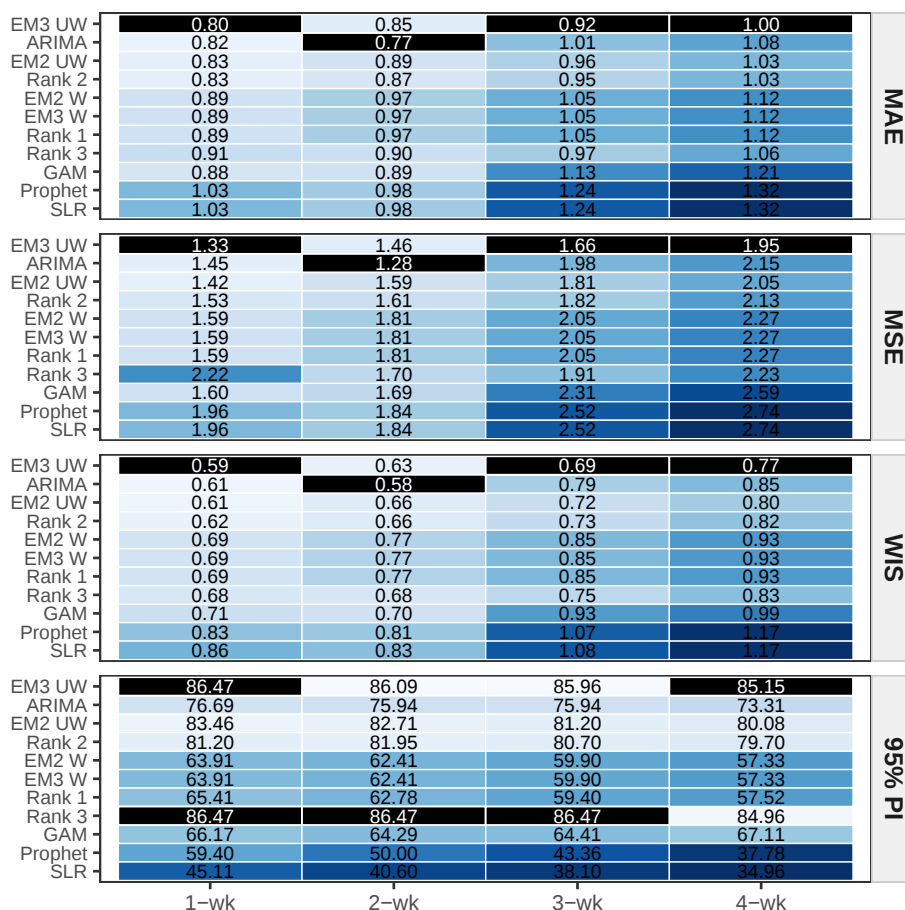


Figure 6: Log-transformed averages of MAE, MSE, and WIS, and the (untransformed) average 95% PI coverage across all models for 1–4 week horizons in the Northeast region, spanning the period from March 5, 2022, to September 14, 2024 (133 forecasts). Lighter shades indicate smaller (better) values for MAE, MSE, and WIS, whereas darker shades indicate larger errors. For coverage, values closer to 95% indicate better performance. Black cells highlight the best-performing model(s) for each forecast horizon and metric.

4.4. South

Fig. 7 summarizes the results for the South region. This region exhibits moderate error levels relative to the Northeast and Midwest, with clearer separation among models and more stable performance patterns across horizons. Overall, ARIMA and EM3 UW consistently rank among the best-performing models across all forecast horizons, with GAM also performing well at 1–2 weeks.

At the 1-week horizon, ARIMA is the strongest model on error-based metrics, achieving the lowest MAE (0.56), MSE (0.80), and WIS (0.42). GAM and several n -sub-epidemic models also perform well, including EM3 UW and Rank3, with MAE values of 0.58 and 0.63 and WIS values of 0.43 and 0.48, respectively. GAM attains MAE 0.59, MSE 0.88, and WIS 0.46. In terms of uncertainty quantification, EM3 UW attains the highest 95% PI coverage (90.23%), closely followed by EM2 UW (88.72%).

At the 2-week horizon, ARIMA again provided the lowest error values, with MAE 0.74, MSE 1.34, and a very competitive WIS of 0.58 compared with EM3 UW (0.57). GAM and the unweighted ensembles remain competitive: GAM attains MAE 0.78, MSE 1.54, and WIS 0.64, while EM3 UW and EM2 UW produce similar MAE values (0.75 and 0.78) and strong predictive interval coverage (87.22% and 82.33%, respectively), highlighting their consistency at this horizon.

At the 3- and 4-week horizons, ARIMA continues to provide the lowest error values. At the 3-week horizon, ARIMA attains MAE 0.86, MSE 1.68, and WIS 0.68. At the same horizon, EM3 UW and EM2 UW yield very competitive results, with MAE 0.89 and 0.90, MSE 1.88 and 1.84, and WIS 0.73 and 0.81, respectively. At the 4-week horizon, ARIMA remains the top performer, providing the lowest MAE (0.98), MSE (1.99), and WIS (0.79), while EM3 UW remains a strong competitor with similar values.

Across all horizons, EM3 UW attains the highest 95% PI coverage, indicating well-calibrated forecast uncertainty in this region.

Overall, these results indicate that ARIMA, GAM, EM2 UW, and EM3 UW provide the strongest 1–2-week-ahead performance for the South region, while ARIMA and EM3 UW remain the best-performing models for 3–4-week-ahead forecasts. In this region, some models produce very similar metric values; for example, at the 2-week horizon, ARIMA yields a WIS of 0.58 compared with 0.57 for EM3 UW. To help assess the magnitude and variability of these differences, we present the distributions of all metrics across models, forecast periods, and horizons in Supplementary Fig. S4, and the corresponding skill scores relative to the SLR baseline in Supplementary Fig. S9, rather than relying solely on point estimates.

ARIMA	0.56	0.74	0.86	0.98	MAE
EM3 UW	0.58	0.75	0.89	1.01	
EM2 UW	0.63	0.78	0.90	1.02	
GAM	0.59	0.78	0.91	1.04	
EM2 W	0.71	0.85	0.97	1.08	
EM3 W	0.71	0.85	0.97	1.08	
Rank 1	0.71	0.85	0.97	1.08	
Rank 3	0.63	0.78	0.95	1.12	
Prophet	0.91	1.04	1.14	1.23	
SLR	0.91	1.04	1.14	1.23	
Rank 2	0.71	1.18	1.65	2.34	
ARIMA	0.80	1.34	1.68	1.99	MSE
EM3 UW	0.85	1.41	1.88	2.27	
EM2 UW	0.92	1.41	1.84	2.20	
GAM	0.88	1.54	1.89	2.21	
EM2 W	1.07	1.50	1.87	2.20	
EM3 W	1.07	1.50	1.87	2.20	
Rank 1	1.07	1.50	1.87	2.20	
Rank 3	1.15	1.53	2.24	3.11	
Prophet	1.49	1.84	2.09	2.35	
SLR	1.49	1.84	2.09	2.35	
Rank 2	1.91	5.25	7.43	9.73	
ARIMA	0.42	0.58	0.68	0.79	WIS
EM3 UW	0.43	0.57	0.73	0.93	
EM2 UW	0.46	0.63	0.81	1.08	
GAM	0.46	0.64	0.74	0.85	
EM2 W	0.53	0.67	0.79	0.91	
EM3 W	0.53	0.67	0.79	0.91	
Rank 1	0.53	0.67	0.79	0.91	
Rank 3	0.48	0.59	0.75	0.89	
Prophet	0.72	0.87	0.98	1.09	
SLR	0.76	0.89	0.99	1.09	
Rank 2	0.51	0.82	1.16	1.69	
ARIMA	81.20	79.32	76.94	74.25	95% PI
EM3 UW	90.23	87.22	84.46	81.02	
EM2 UW	88.72	82.33	78.95	75.19	
GAM	67.67	63.91	63.16	61.47	
EM2 W	68.42	61.28	55.14	50.94	
EM3 W	68.42	61.28	55.14	50.94	
Rank 1	68.42	60.90	54.64	51.50	
Rank 3	84.21	81.95	77.94	75.00	
Prophet	48.12	39.85	33.33	28.57	
SLR	33.83	31.58	27.32	25.00	
Rank 2	83.46	75.19	71.93	69.36	
	1-wk	2-wk	3-wk	4-wk	

Figure 7: Log-transformed averages of MAE, MSE, and WIS, and the (untransformed) average 95% PI coverage across all models for 1–4 week horizons in the South region, spanning the period from March 5, 2022, to September 14, 2024 (133 forecasts). Lighter shades indicate smaller (better) values for MAE, MSE, and WIS, whereas darker shades indicate larger errors. For coverage, values closer to 95% indicate better performance. Black cells highlight the best-performing model(s) for each forecast horizon and metric.

4.5. West

Fig. 8 presents the results for the West region. There is clear separation among models and strong performance from both ARIMA and the unweighted n -sub-epidemic ensembles, particularly EM3–UW. Overall, these models consistently rank among the top-performing approaches across all forecast horizons.

At the 1-week horizon, EM3 UW and EM2 UW are the leading models based on metrics. EM3 UW achieved the lowest MAE (0.57), the lowest MSE (0.76), and the lowest WIS (0.42). EM2 UW performed similarly, with MAE 0.60, MSE 0.83, and WIS 0.43. GAM also performed well, providing MAE 0.59, MSE 0.86, and WIS 0.47. For uncertainty quantification, EM3 UW yields exceptionally high 95% PI coverage (96.24%), the highest of any model in this region and significantly above the other regions.

At the 2-week horizon, EM3 UW remains the best-performing model on all three error metrics, achieving the lowest MAE (0.70), MSE (1.17), and WIS (0.52). EM2 UW again performs competitively, with MAE 0.72, MSE 1.18, and WIS 0.55. ARIMA provided MAE 0.79, MSE 1.37, and WIS 0.62, remaining competitive but with slightly larger errors. EM3 UW continued to provide the highest 95% PI coverage (94.36%) at this horizon.

At the 3-week horizon, EM3 UW again is part of the best-performing models, with one of the lowest MAEs (0.84), a competitive MSE (1.62) compared to EM2 UW (1.51) and the lowest WIS (0.67). Most of the sub-epidemic models provided similar values across metrics. EM3 UW further provides the highest 95% PI coverage (90.48%) at this horizon.

At the 4-week horizon, EM3 UW continues to be among the top performers, achieving competitive MAE (0.95) and MSE (1.97), compared to EM2 UW MAE (0.94) and MSE (1.84). The Rank 1 model also performs competitively with

MAE 0.99, MSE 1.90, and WIS 0.80. For this horizon, EM3 UW also attains the highest 95% PI coverage (87.78%), highlighting its consistently strong calibration in this region.

Overall, these results show that the West region benefits from strong performance by both ARIMA, GAM, and the unweighted n -sub-epidemic ensembles across a 1–2 week forecast. The n -sub-epidemic models provide the most reliable performance across all horizons, achieving the lowest error metrics and the highest PI coverage values. Clearly several models produce similar results at some horizons; for example, at the 3-week horizon, GAM and ARIMA yielded MAE values of 0.91 and 0.92, respectively. To help assess the magnitude and variability of these differences, we present the distributions of all metrics across models, forecast periods, and horizons in Supplementary Fig. S5, and the corresponding skill scores relative to the SLR baseline in Supplementary Fig. S10, rather than relying solely on point estimates.

EM3 UW	0.57	0.70	0.84	0.95	MAE
EM2 UW	0.60	0.72	0.82	0.94	
EM2 W	0.68	0.79	0.89	0.99	
EM3 W	0.68	0.79	0.89	0.99	
Rank 1	0.68	0.79	0.89	0.99	
ARIMA	0.62	0.79	0.92	1.05	
GAM	0.59	0.85	0.91	1.03	
Rank 3	0.62	0.78	0.93	1.08	
Prophet	0.91	1.05	1.14	1.25	
SLR	0.91	1.05	1.14	1.25	
Rank 2	0.69	0.88	1.02	1.16	
EM3 UW	0.76	1.17	1.62	1.97	MSE
EM2 UW	0.83	1.18	1.51	1.84	
EM2 W	1.00	1.32	1.62	1.90	
EM3 W	1.00	1.32	1.62	1.90	
Rank 1	1.00	1.32	1.62	1.90	
ARIMA	0.94	1.37	1.72	2.03	
GAM	0.86	1.64	1.80	2.15	
Rank 3	0.84	1.45	2.04	2.58	
Prophet	1.49	1.86	2.10	2.38	
SLR	1.49	1.86	2.11	2.38	
Rank 2	1.40	2.19	2.82	3.39	
EM3 UW	0.42	0.52	0.67	1.63	WIS
EM2 UW	0.43	0.55	1.13	1.48	
EM2 W	0.49	0.60	0.70	0.80	
EM3 W	0.49	0.60	0.70	0.80	
Rank 1	0.49	0.60	0.70	0.80	
ARIMA	0.47	0.62	0.74	0.86	
GAM	0.47	0.70	0.74	0.86	
Rank 3	0.45	0.58	0.76	1.60	
Prophet	0.73	0.89	1.00	1.11	
SLR	0.77	0.91	1.01	1.12	
Rank 2	0.66	2.57	6.32	7.49	
EM3 UW	96.24	94.36	90.48	87.78	95% PI
EM2 UW	91.73	87.97	84.46	82.33	
EM2 W	74.44	68.05	62.91	60.71	
EM3 W	74.44	68.05	62.91	60.71	
Rank 1	74.44	68.80	63.16	60.15	
ARIMA	79.70	72.93	70.18	67.11	
GAM	67.67	66.17	64.91	63.72	
Rank 3	91.73	90.23	86.72	83.27	
Prophet	42.11	35.34	31.58	27.26	
SLR	31.58	30.08	26.57	24.44	
Rank 2	87.97	84.59	81.45	77.82	
	1-wk	2-wk	3-wk	4-wk	

Figure 8: Log-transformed averages of MAE, MSE, and WIS, and the (untransformed) average 95% PI coverage across all models for 1–4 week horizons in the West region, spanning the period from March 5, 2022, to September 14, 2024 (133 forecasts). Lighter shades indicate smaller (better) values for MAE, MSE, and WIS, whereas darker shades indicate larger errors. For coverage, values closer to 95% indicate better performance. Black cells highlight the best-performing model(s) for each forecast horizon and metric.

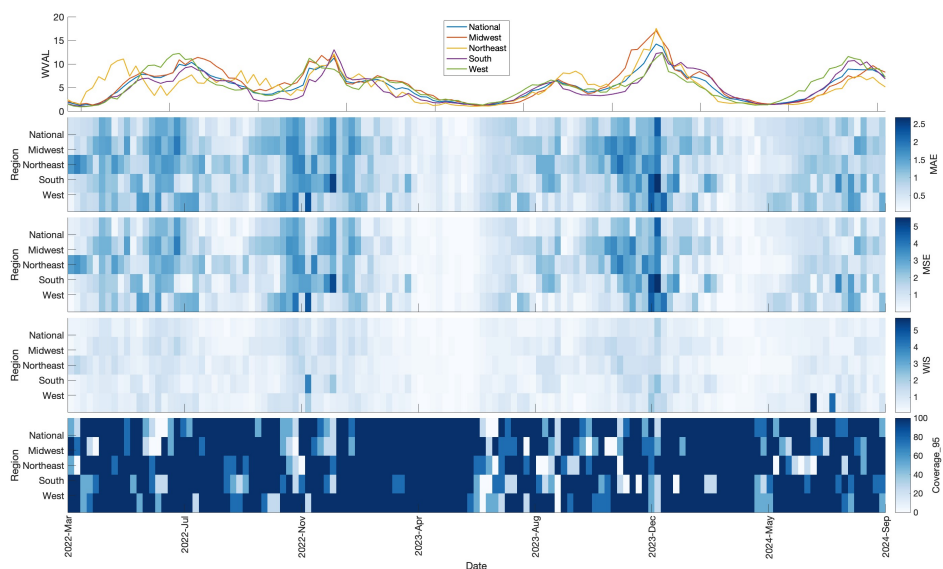


Figure 9: Temporal dynamics of forecast performance for EM3 UW model. Top panel: Weekly viral activity levels (WVAL, $\log_{10}$ scale) across five U.S. regions from March 2022 to September 2024. Lower panels: Heatmaps showing temporal variation in four performance metrics (MAE, MSE, WIS, and 95% PI coverage) for 4-week ahead forecasts. Darker blue in error metric panels indicates higher errors. The coverage panel uses inverted coloring (darker = higher coverage = better performance). Error magnitudes increase during high WVAL periods, but this is partly expected because both WVAL and errors are on the same log scale; higher baseline values naturally yield larger absolute errors. The key scientific question is whether errors increase *disproportionately* during surges, indicating true model degradation during critical periods. Future analysis should examine skill scores (errors relative to baseline) across different WVAL regimes to isolate genuine performance degradation from scale-dependent error increases.

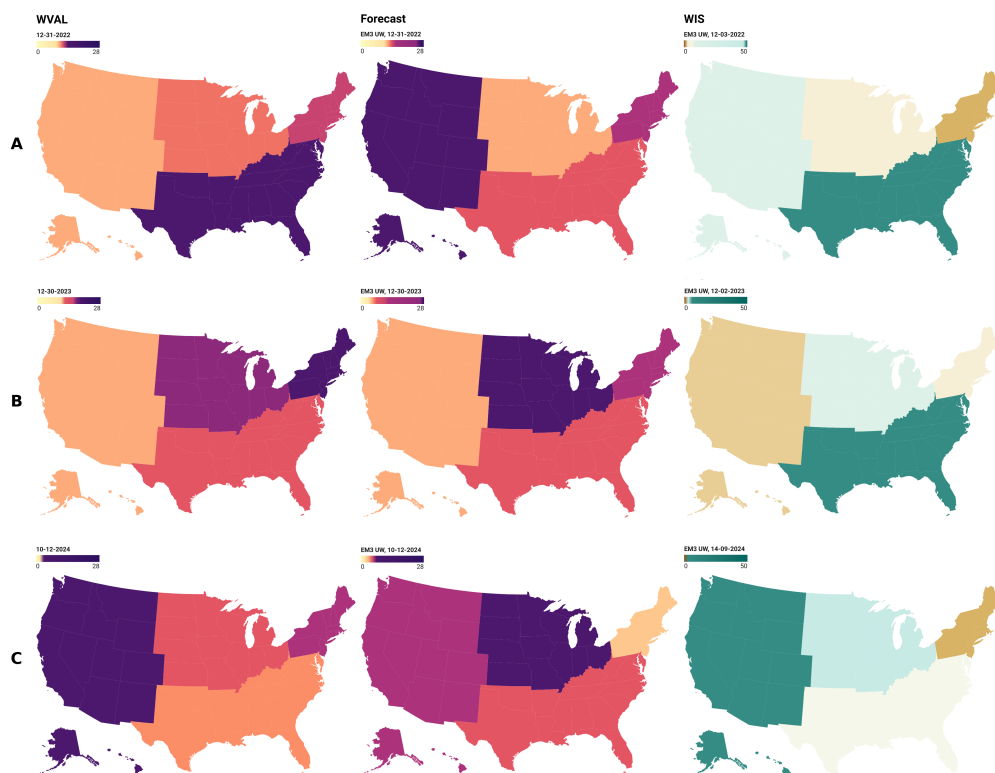


Figure 10: The figure displays the WVAL values and model performance for three different forecasting periods, each with a 4-week forecast horizon: (A) a forecast conducted on 12-03-2022 targeting the week of 12-31-2022; (B) a forecast conducted on 12-02-2023 targeting the week of 12-30-2023; and (C) a forecast conducted on 09-14-2024 targeting the week of 10-12-2024. Each row corresponds to one of these forecast periods. The first column shows the actual WVAL values observed during the target week. The second column presents the 4-week-ahead predictions from the EM3 UW model, and the third column displays the associated WIS. This figure highlights how the model performs across different periods and regions. The spatial heterogeneity evident in these maps underscores the importance of region-specific modeling approaches, as viral activity patterns and forecast accuracy vary substantially across geographic areas.

5. Discussion

This study highlights the effectiveness of retrospective forecasting using wastewater surveillance data to track COVID-19 trends across U.S. regions. By evaluating multiple models across various forecast horizons, we identified key differences in predictive accuracy and uncertainty quantification. The unweighted ensemble models within the n -sub-epidemic framework, especially EM3 UW and EM2 UW, stood out for their reliability in forecasting 3- to 4-week ahead. On the other hand, statistical models like ARIMA and GAM were better suited for 1- to 2-week forecasts. These findings reinforce the advantage of combining overlapping sub-epidemics to capture complex transmission dynamics, especially in settings characterized by multiple peaks and extended plateaus in wastewater signals [Chowell et al., 2022].

Our results align with previous research showing that ensemble models improve epidemic forecast performance by balancing the strengths and weaknesses of individual models [Cramer et al., 2022, Yamana, Kandula, and Shaman, 2017]. Specifically, EM3 UW consistently outperformed other models across metrics such as MAE, WIS, and 95% prediction interval (PI) coverage, while EM2 UW in some cases minimized large prediction errors (MSE). This supports the use of ensemble modeling to enhance both accuracy and robustness in epidemic forecasting [Cramer et al., 2022]. The robustness of ensemble approaches is further supported by large-scale international validation: analysis of 48 forecasting models across 32 European countries over 52 weeks found that ensembles outperformed 83% of individual models for cases and 91% for deaths, with median ensemble aggregation proving more robust than mean aggregation [Sherratt, Gruson, Grah, Johnson, Niehus, Prasse, Sandmann, et al., 2023]. This cross-national consistency validates the generalizability of ensemble forecasting strategies. We also observed trade-offs between accuracy and uncertainty quantification. EM3 UW delivered accurate forecasts and maintained the best 95% PI coverage, highlighting its ability

to provide sharp, reliable predictions. These results align with earlier findings from [Gneiting and Raftery, 2007, Funk et al., 2019], which emphasized the importance of well-calibrated forecasts that balance precision with uncertainty.

Model performance varied across regions and forecast horizons. For example, EM3 UW dominated in the Midwest, especially at 3- to 4-week forecast horizons, while GAM and ARIMA excelled at 1- to 2-week-ahead forecasts. In the South and Northeast regions, ARIMA consistently performed well in the short term (1- to 2-week), but ensemble models like EM3 UW and EM2 UW showed more substantial 3- to 4-week reliability. In the West, EM2 UW and EM3 UW performed very well at 3- to 4-week forecasts. These results emphasize the need for context-specific modeling strategies, particularly when geographic, demographic, and reporting differences can impact model performance [Pei, Kandula, and Shaman, 2020].

5.1. Regional Heterogeneity in Forecast Performance

Our results reveal substantial regional variation in forecast accuracy, with the Northeast consistently exhibiting higher prediction errors across models and horizons. Understanding why certain regions present greater forecasting challenges offers insights into the relationship between epidemic dynamics and predictability.

Several intersecting factors likely contribute to the Northeast's reduced forecast accuracy. First, the region encompasses both dense urban centers (New York City, Boston, Philadelphia) and lower-density suburban areas within the same wastewater catchment aggregations. Epidemic dynamics in high-density settings exhibit more volatile week-to-week fluctuations due to rapid transmission when susceptible populations are depleted and greater sensitivity to behavioral changes. Second, Northeastern cities often operate older combined sewer systems in which stormwater mixes with sewage. During precipitation events, dilution effects can substantially reduce measured WVAL without corresponding reductions in actual infection levels, introducing measurement noise independent of true epidemic dynamics. Third, pronounced seasonality with harsh winters drives indoor congregation and affects both viral shedding rates and RNA degradation in wastewater. This environmental variability compounds forecasting challenges. Finally, substantial non-resident populations (daily commuters, seasonal tourists) contribute to wastewater but may not reflect local residential infection patterns, creating a spatial mismatch between catchments and epidemic dynamics.

These findings suggest region-specific modeling adaptations. In high-density regions like the Northeast, models that account for rapid behavioral responses (GAM with adaptive smoothing) or capture multiple epidemic phases (n -sub-epidemic framework) may outperform simple autoregressive approaches. In contrast, the South and Midwest's more stable dynamics allow simpler models (ARIMA, SLR) to achieve competitive performance with lower computational cost. The West's geographic diversity (urban coast versus rural interior) suggests ensemble approaches combining models with different structural assumptions may provide robustness against this heterogeneity.

The regional performance differences have practical implications for public health surveillance. Decision-makers in the Northeast should anticipate wider forecast uncertainty bounds, and decision protocols should account for this irreducible uncertainty. Regions with volatile dynamics may benefit from increased sampling frequency or more sophisticated protocols (24-hour composite samples, flow-weighted sampling) to reduce measurement noise. Rather than deploying a single national model, our results support a portfolio approach where regional models are tailored to local epidemic characteristics. The consistent superiority of certain models in specific regions suggests local agencies could adopt region-optimized approaches.

5.2. Reconciling Ensemble and Region-Specific Forecasting

Our study advocates both ensemble forecasting and region-specific model selection, approaches that might appear contradictory but operate at different levels of the forecasting problem. Ensemble methods address *temporal uncertainty*: at any given forecast time, we cannot know which model will perform best in the upcoming weeks. ARIMA excels during stable periods with predictable autocorrelation; n -sub-epidemic models succeed during growth and decline phases when sub-epidemic structure is evident; GAM adapts well to changing smoothness requirements. By combining models, ensembles achieve robustness across varying conditions. Our results demonstrate this: The unweighted ensemble EM3 UW frequently outperforms individual models in most regions, particularly at the 3- and 4-week horizons, by hedging against temporal uncertainty.

Regional adaptation addresses *spatial heterogeneity*: different regions exhibit systematically different epidemic characteristics. The Northeast's volatile dynamics favor models that adapt quickly (GAM), while the South's more stable patterns allow simpler approaches (ARIMA) to perform competitively. Region-specific modeling means selecting which models to include in regional ensembles (perhaps excluding Prophet if it consistently underperforms),

adjusting model hyperparameters for regional characteristics (e.g., shorter smoothing windows for volatile regions), and interpreting forecast uncertainty differently across regions based on regional predictability.

We recommend a hierarchical strategy for public health deployment. At the national level, maintain an ensemble of diverse models (ARIMA, GAM, n -sub-epidemic, Prophet) to ensure robustness across all regions and time periods. At the regional level, optimize which models to include in regional ensembles, adjust ensemble weighting schemes (equal weights versus performance-based), and tune model-specific hyperparameters. At the temporal monitoring level, continuously assess which models currently perform best and adjust ensemble compositions accordingly. This hierarchical approach leverages both strengths: ensemble robustness protects against model failures, while regional adaptation improves baseline performance.

The temporal performance variation we observe (9) reveals an important pattern: model performance deteriorates significantly during periods of rapid change (surges) in viral activity, suggesting potential for adaptive weighting strategies. Recent work demonstrated that adaptive weighted ensembles with exponential time decay (assigning the highest weight to the most recent performance) achieved 52% error reduction compared to static weighting approaches in contexts with irregular epidemic patterns [Tsang, Du, Cowling, and Viboud, 2024]. Such approaches could address the performance degradation during surge periods that is evident in our analysis.

Our study period spans a critical surveillance transition: clinical COVID-19 case reporting was largely discontinued in March 2024, with hospitalization reporting ending in May 2024 [Schenk et al., 2024]. This shift elevates wastewater surveillance from complementary to essential status. Our findings validate WBE-based forecasting across this transition, demonstrating stable performance despite changes in clinical reporting infrastructure.

Our findings also suggest a degree of model stability over time, with the top-performing models maintaining consistent ranks across multiple forecasting periods, suggesting that the ensemble models do not exhibit drastic fluctuations in performance. This consistency is critical for real-time applications, as it suggests these models can remain reliable without frequent recalibration. This is an important consideration for public health use [Shaman et al., 2013, Brooks, Farrow, Hyun, Tibshirani, and Rosenfeld, 2018].

Wastewater surveillance remains a promising tool for early detection of COVID-19 and other infectious diseases. The ability of ensemble models to extract meaningful trends from this data source enhances its value for public health decision-making [Peccia et al., 2020, Wu, Xiao, Zhang, Moniz, Endo, Armas, Bushman, Chai, Duvallet, Erickson, et al., 2021]. Short-term forecasts (1- to 2-week) can benefit from flexible statistical models like ARIMA and GAM, while long-term (3- to 4-week) projections are best captured by the n -sub-epidemic ensemble framework. This strategic allocation of models can support more responsive and informed intervention planning.

An important practical consideration is the computational cost of different modeling approaches. While the n -sub-epidemic framework requires more computational resources than simple statistical models, the improved forecast accuracy and uncertainty quantification justify this additional cost, especially for public health applications where reliable predictions are critical. Future work could explore more computationally efficient implementations of ensemble methods without sacrificing performance.

Nonetheless, several limitations must be acknowledged. This is a retrospective study, meaning forecasts were tested against past data rather than generated in real time. While helpful in evaluating model accuracy, this approach does not fully reflect the challenges of real-time forecasting, such as data delays and changing conditions, which can affect operational performance. Wastewater data quality varies across jurisdictions due to differences in WWTP size, sampling protocols, and laboratory methods. Plant-level variability in viral recovery efficiency, flow normalization, and PMMoV adjustment could contribute to regional performance differences [Schenk et al., 2024]. Our focus on phenomenological and statistical models excluded mechanistic SEIR-type models incorporating transmission dynamics and intervention effects. Hybrid approaches combining mechanistic structure with ensemble phenomenological forecasting could improve performance during periods of rapidly changing transmission rates. Our study does not explicitly stratify model performance by epidemic “phase” (rising, peak/plateau, declining). Our rolling, 10-week calibration framework is designed to mimic operational use and naturally spans multiple phases, but it does not allow formal phase-specific comparisons without additional assumptions and loss of sample size. Future work could incorporate explicit phase definitions to evaluate performance across distinct stages of epidemic dynamics.

Additionally, while unweighted ensembles performed well, further research is needed to refine ensemble strategies, such as developing adaptive or data-driven weighting schemes. Exploring hybrid approaches that integrate statistical, machine learning, and mechanistic elements could also improve adaptability and accuracy across forecast horizons. Finally, limitations in the wastewater dataset, such as uneven geographic coverage, reporting inconsistencies, and noise in signal detection, may have impacted model performance in ways not fully captured in this study.

6. Conclusion

Overall, our findings reinforce the benefits of ensemble forecasting approaches for WVAL, particularly unweighted ensembles, which performed well across multiple metrics. These results emphasize the importance of balancing accuracy and uncertainty quantification when designing epidemic forecasting models. Future research should explore alternative weighting schemes, investigate the impact of model diversity within ensembles, and extend this analysis to additional regions and time periods. Additionally, integrating domain knowledge with machine learning techniques could further enhance the interpretability and robustness of epidemic forecasts [Viboud, Sun, Gaffey, Ajelli, Fumanelli, Merler, Zhang, Chowell, Simonsen, Vespignani, et al., 2018].

Our study provides actionable insights for public health practitioners: use ARIMA or GAM for immediate (1- to 2-week) operational planning, and rely on n -sub-epidemic ensemble models for medium-term (3-4 week) strategic planning. This tiered approach can optimize resource allocation and intervention timing based on forecast reliability at different horizons.

These insights contribute to the ongoing development of effective and reliable forecasting tools for epidemic response and public health decision-making. Future work should focus on optimizing ensemble model structures to maximize predictive performance while ensuring adaptability to different epidemiological contexts.

Author Contributions

F.A. and G.C. conceived the study. F.A., H.K., A.B., and G.C. contributed to the conceptualization and methodology of the study, participated in validation of results, curated the data, wrote the original manuscript draft, and prepared the visualizations (figures and tables). All contributed to the writing and revising of earlier versions of the manuscript.

Funding

G.C. was supported by NSF Awards 2125246 and 2026797. F.A. is supported through a 2CI Fellowship from Georgia State University.

References

- Shafi A. Medium. what are generalized additive models? <https://towardsdatascience.com/generalised-additive-models-6dfbedf1350a>, 2021.
- Carly Adams, Megan Bias, Rory M Welsh, Jenna Webb, Heather Reese, Stephen Delgado, John Person, Rachel West, Soo Shin, and Amy Kirby. The national wastewater surveillance system (nwss): from inception to widespread coverage, 2020–2022, united states. *Science of The Total Environment*, 924:171566, 2024.
- Elizabeth Alvarez, Iwona A Bielska, Stephanie Hopkins, Ahmed A Belal, Donna M Goldstein, Jean Slick, Sureka Pavalagantharajah, Anna Wynfield, Shruthi Dakey, Marie-Carmel Gedeon, et al. Limitations of covid-19 testing and case data for evidence-informed health policy and practice. *Health Research Policy and Systems*, 21(1):11, 2023.
- R Harald Baayen and Maja Linke. An introduction to the generalized additive model. *A practical handbook of corpus linguistics*, pages 563–591, 2020.
- Sándor Baran and Dóra Nemoda. Censored and shifted gamma distribution based emos model for probabilistic quantitative precipitation forecasting. *Environmetrics*, 27(5):280–292, 2016.
- Gopi Battineni, Nalini Chintalapudi, and Francesco Amenta. Forecasting of covid-19 epidemic size in four high hitting nations (usa, brazil, india and russia) by fb-prophet machine learning model. *Applied Computing and Informatics*, 2020.
- Domenico Benvenuto, Marta Giovanetti, Lazzaro Vassallo, Silvia Angeletti, and Massimo Ciccozzi. Application of the arima model on the covid-2019 epidemic dataset. *Data in brief*, 29:105340, 2020.
- Amanda Bleichrodt, Sushma Dahal, Kevin Maloney, Lisa Casanova, Ruiyan Luo, and Gerardo Chowell. Real-time forecasting the trajectory of monkeypox outbreaks at the national and global levels, july–october 2022. *BMC medicine*, 21(1):19, 2023.
- Amanda Bleichrodt, Ruiyan Luo, Alexander Kirpich, and Gerardo Chowell. Evaluating the forecasting performance of ensemble sub-epidemic frameworks and other time series models for the 2022–2023 mpox epidemic. *Royal Society Open Science*, 11(7):240248, 2024a.
- Amanda Bleichrodt, Amelia Phan, Ruiyan Luo, Alexander Kirpich, and Gerardo Chowell-Puente. Statmodpredict: A user-friendly r-shiny interface for fitting and forecasting with statistical models. *Available at SSRN 4849702*, 2024b.
- Johannes Bracher, Evan L Ray, Tilmann Gneiting, and Nicholas G Reich. Evaluating epidemic forecasts in an interval format. *PLOS Computational Biology*, 17(2):e1008618, 2021. doi: 10.1371/journal.pcbi.1008618. URL <https://doi.org/10.1371/journal.pcbi.1008618>.
- Logan C. Brooks, David C. Farrow, Sangwon Hyun, Robert J. Tibshirani, and Roni Rosenfeld. A flexible modeling framework for seasonal and pandemic influenza forecasting. *PLoS Computational Biology*, 14(10):e1006352, 2018. doi: 10.1371/journal.pcbi.1006352.
- CDC. National wastewater surveillance system (nwss). <https://www.cdc.gov/nwss/rv/COVID19-nationaltrend.html>, 2024a.
- CDC. National wastewater surveillance system (nwss). <https://www.cdc.gov/nwss/about-data.html#data-method>, 2024b.
- Gerardo Chowell, Sushma Dahal, Amna Tariq, Kimberlyn Roosa, James M Hyman, and Ruiyan Luo. An ensemble n-sub-epidemic modeling framework for short-term forecasting epidemic trajectories: Application to the covid-19 pandemic in the usa. *PLoS Computational Biology*, 18(10):e1010602, 2022.
- Gerardo Chowell, Sushma Dahal, Amanda Bleichrodt, Amna Tariq, James M Hyman, and Ruiyan Luo. Subepipredict: A tutorial-based primer and toolbox for fitting and forecasting growth trajectories using the ensemble n-sub-epidemic modeling framework. *Infectious Disease Modelling*, 2024.
- Kyle J Colonna, Gabriela F Nane, Ernani F Choma, Roger M Cooke, and John S Evans. A retrospective assessment of covid-19 model performance in the usa. *Royal Society open science*, 9(10):220021, 2022.
- Estee Y Cramer, Evan L Ray, Velma K Lopez, Johannes Bracher, Andrea Brennen, Alvaro J Castro Rivadeneira, Aaron Gerding, Tilmann Gneiting, Katie H House, Yuxin Huang, et al. Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the united states. *Proceedings of the National Academy of Sciences*, 119(15):e2113561119, 2022. doi: 10.1073/pnas.2113561119.
- Maria L Daza-Torres, J Criselio Montesinos-López, Minji Kim, Rachel Olson, C Winston Bess, Lezlie Rueda, Mirjana Susa, Linnea Tucker, Yury E Garcia, Alec J Schmidt, et al. Model training periods impact estimation of covid-19 incidence from wastewater viral loads. *Science of The Total Environment*, 858:159680, 2023.
- Tripti Dimri, Shamshad Ahmad, and Mohammad Sharif. Time series analysis of climate variables using seasonal arima approach. *Journal of Earth System Science*, 129:1–16, 2020.
- Max Drews, Pankaj Kumar, Rahul K. Singh, Manuel De La Sen, Shikha S. Singh, and Arun K. Pandey. Model-based ensembles: Lessons learned from retrospective analysis of covid-19 infection forecasts across 10 countries. *Science of The Total Environment*, 806:150639, 2022. doi: 10.1016/j.scitotenv.2021.150639.
- Jamal Fattah, Latifa Ezzine, Zineb Aman, Haj El Moussami, and Abdeslam Lachhab. Forecasting of demand using arima model. *International Journal of Engineering Business Management*, 10:1847979018808673, 2018.
- Sebastian Funk et al. Assessing the performance of real-time epidemic forecasts. *PLoS Comput Biol*, 2019.
- Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *J Am Stat Assoc*, 2007.
- Raul Gonzalez, Kyle Curtis, Aaron Bivins, Kyle Bibby, Mark H Weir, Kathleen Yetka, Hannah Thompson, David Keeling, Jamie Mitchell, and Dana Gonzalez. Covid-19 surveillance in southeastern virginia using wastewater-based epidemiology. *Water research*, 186:116296, 2020.
- Trevor Hastie. The elements of statistical learning: data mining, inference, and prediction, 2009.
- Rob J Hyndman and Yeasmin Khandakar. Automatic time series forecasting: the forecast package for r. *Journal of statistical software*, 27:1–22, 2008.
- Hasnain Iftikhar, Murad Khan, Mohammed Saad Khan, and Mehak Khan. Short-term forecasting of monkeypox cases using a novel filtering and combining technique. *Diagnostics*, 13(11):1923, 2023.
- Bernat Josep-Duran, Albert Serra-Compte, Miquel Sàrrias, Susana Gonzalez, Daniel López, Clara Prats, Martí Català, Enric Alvarez-Lacalle, Sergio Alonso, and Marina Arnaldos. Assessing wastewater-based epidemiology for the prediction of sars-cov-2 incidence in catalonia. *Scientific reports*, 12(1):15073, 2022.

- Rakesh Kaundal, Amar S Kapoor, and Gajendra PS Raghava. Machine learning techniques in disease forecasting: a case study on rice blast prediction. *BMC bioinformatics*, 7:1–16, 2006.
- Amy E Kirby, Maroya Spalding Walters, William C Jennings, Ronald Fugitt, Nathan LaCross, Mia Mattioli, Zachary A Marsh, Virginia A Roberts, Jeffrey W Mercante, Jonathan Yoder, and Vincent R Hill. Using wastewater surveillance data to support the covid-19 response—united states, 2020–2021. *Morbidity and Mortality Weekly Report*, 70(36):1242–1244, 2021. doi: 10.15585/mmwr.mm7036a2.
- Alexander Kirpich, Aleksandr Shishkin, Thomas A Weppelmann, Alexander Perez Tchernov, Pavel Skums, and Yuriy Gankin. Excess mortality in belarus during the covid-19 pandemic as the case study of a country with limited non-pharmaceutical interventions and limited reporting. *Scientific reports*, 12(1):5475, 2022.
- Masaaki Kitajima, Warish Ahmed, Kyle Bibby, Annalaura Carducci, Charles P Gerba, Kerry A Hamilton, Eiji Haramoto, and Joan B Rose. Sars-cov-2 in wastewater: State of the knowledge and research needs. *Science of the total environment*, 739:139076, 2020.
- Xuan Li, Huan Liu, Li Gao, Samendra P Sherchan, Ting Zhou, Stuart J Khan, Mark CM Van Loosdrecht, and Qilin Wang. Wastewater-based epidemiology predicts covid-19-induced weekly new hospital admissions in over 150 usa counties. *Nature communications*, 14(1):4548, 2023a.
- Xuan Li, Shuxin Zhang, Samendra Sherchan, Gorka Orive, Unax Lertxundi, Eiji Haramoto, Ryo Honda, Manish Kumar, Sudipti Arora, Masaaki Kitajima, et al. Correlation between sars-cov-2 rna concentration in wastewater and covid-19 cases in community: a systematic review and meta-analysis. *Journal of Hazardous Materials*, 441:129848, 2023b.
- Bowen Long, Fangya Tan, and Mark Newman. Forecasting the monkeypox outbreak using arima, prophet, neuralprophet, and lstm models in the united states. *Forecasting*, 5(1):127–137, 2023.
- Velma K Lopez, Estee Y Cramer, Robert Pagano, John M Drake, Eamon B O’Dea, Madeline Adee, Turgay Ayer, Jagpreet Chhatwal, Ozden O Dalgic, Mary A Ladd, et al. Challenges of covid-19 case forecasting in the us, 2020–2021. *PLoS computational biology*, 20(5):e1011200, 2024.
- Tyler Meadows, Erik R Coats, Solana Narum, Eva M Top, Benjamin J Ridenhour, and Thibault Stalder. Epidemiological model can forecast covid-19 outbreaks from wastewater-based surveillance in rural communities. *Water Research*, 268:122671, 2025.
- Gertjan Medema, Leo Heijnen, Goffe Elsinga, Ronald Italiaander, and Anke Brouwer. Presence of sars-coronavirus-2 rna in sewage and correlation with reported covid-19 prevalence in the early stage of the epidemic in the netherlands. *Environmental science & technology letters*, 7(7): 511–516, 2020.
- Prapanna Mondal, Labani Shit, and Saptarsi Goswami. Study of effectiveness of time series modeling (arima) in forecasting stock prices. *International Journal of Computer Science, Engineering and Applications*, 4(2):13, 2014.
- Douglas C. Montgomery, Cheryl L. Jennings, and Murat Kulahci. *Introduction to Time Series Analysis and Forecasting*. John Wiley & Sons, Hoboken, NJ, 2nd edition edition, 2015. ISBN 9781118745113. URL <https://www.wiley.com/en-us/Introduction+to+Time+Series+Analysis+and+Forecasting,+2nd+Edition-p-9781118745113>.
- Kristen Nixon, Sonia Jindal, Felix Parker, Maximilian Marshall, Nicholas G Reich, Kimia Ghobadi, Elizabeth C Lee, Shaun Truelove, and Lauren Gardner. Real-time covid-19 forecasting: challenges and opportunities of model performance and translation. *The Lancet Digital Health*, 4(10): e699–e701, 2022.
- Jordan Peccia, Alessandro Zulli, Doug E Brackney, Nathan D Grubaugh, Edward H Kaplan, Arnau Casanovas-Massana, Albert I Ko, Aryn A Malik, Dennis Wang, Mike Wang, et al. Measurement of sars-cov-2 rna in wastewater tracks community infection dynamics. *Nature biotechnology*, 38 (10):1164–1167, 2020.
- Sen Pei, Sasikiran Kandula, and Jeffrey Shaman. Differential effects of intervention timing on covid-19 spread in the united states. *Science advances*, 6(49):eabd6370, 2020.
- Tin Phan, Samantha Brozak, Bruce Pell, Jeremiah Oghuan, Anna Gitter, Tao Hu, Ruy M Ribeiro, Ruian Ke, Kristina D Mena, Alan S Perelson, et al. Making waves: integrating wastewater surveillance with dynamic modeling to track and predict viral outbreaks. *Water research*, 243:120372, 2023.
- Daniele Proverbio, Françoise Kemp, Stefano Magni, Leslie Ogorzaly, Henry-Michel Cauchie, Jorge Gonçalves, Alexander Skupin, and Atte Aalto. Model-based assessment of covid-19 epidemic dynamics by wastewater analysis. *Science of the Total Environment*, 827:154235, 2022.
- RDocumentation. *predict: Model predictions*, n.d. URL <https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/predict>. Accessed November 28, 2024.
- Christophorus Benedetto Aditya Satrio, William Darmawan, Bellatasya Unrica Nadia, and Novita Hanafiah. Time series analysis and forecasting of coronavirus disease in indonesia using arima model and prophet. *Procedia Computer Science*, 179:524–532, 2021.
- H Schenk, W Rauch, A Zulli, and AB Boehm. SARS-CoV-2 surveillance in US wastewater: Leading indicators and data variability analysis in 2023–2024. *PLOS ONE*, 19(11):e0313927, 2024. doi: 10.1371/journal.pone.0313927.
- Jeffrey Shaman et al. Real-time influenza forecasts during the 2012–2013 season. *Nat Commun*, 2013.
- K Sherratt, H Gruson, R Grah, H Johnson, R Niehus, B Prasse, F Sandmann, et al. Predictive performance of multi-model ensemble forecasts of COVID-19 across European nations. *eLife*, 12:e81916, 2023. doi: 10.7554/eLife.81916.
- Sean Taylor. *prophet: Prophet forecaster*, n.d. URL <https://www.rdocumentation.org/packages/prophet/versions/1.0/topics/prophet>. Accessed November 28, 2024.
- Sean J Taylor and Benjamin Letham. Forecasting at scale. *The American Statistician*, 72(1):37–45, 2018.
- Mehmet Tektaş. Weather forecasting using anfis and arima models. *Environmental Research, Engineering and Management*, 51(1):5–10, 2010.
- Tim K Tsang, Qiurui Du, Benjamin J Cowling, and Cécile Viboud. An adaptive weight ensemble approach to forecast influenza activity in an irregular seasonality context. *Nature Communications*, 15(1):8625, 2024.
- Cécile Viboud, Kaiyuan Sun, Robert Gaffey, Marco Ajelli, Laura Fumanelli, Stefano Merler, Qian Zhang, Gerardo Chowell, Lone Simonsen, Alessandro Vespignani, et al. The rapid ebola forecasting challenge: Synthesis and lessons learnt. *Epidemics*, 22:13–21, 2018.
- Cort J Willmott and Kenji Matsuura. Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance. *Climate research*, 30(1):79–82, 2005.
- Simon N Wood. *Generalized additive models: an introduction with R*. chapman and hall/CRC, 2017.

- Simon N. Wood. *gam: Generalized Additive Models with Integrated Smoothness Estimation*, n.d. URL <https://www.rdocumentation.org/packages/mgcv/versions/1.9-1/topics/gam>. Accessed November 28, 2024.
- Fuqing Wu, Amy Xiao, Jianbo Zhang, Katya Moniz, Noriko Endo, Federica Armas, Mary Bushman, Peter R Chai, Claire Duvallet, Timothy B Erickson, et al. Wastewater surveillance of sars-cov-2 across 40 us states from february to june 2020. *Water research*, 202:117400, 2021.
- Teresa K Yamana, Sasikiran Kandula, and Jeffrey Shaman. Individual versus superensemble forecasts of seasonal influenza outbreaks in the united states. *PLoS computational biology*, 13(11):e1005801, 2017.

7. Supplementary

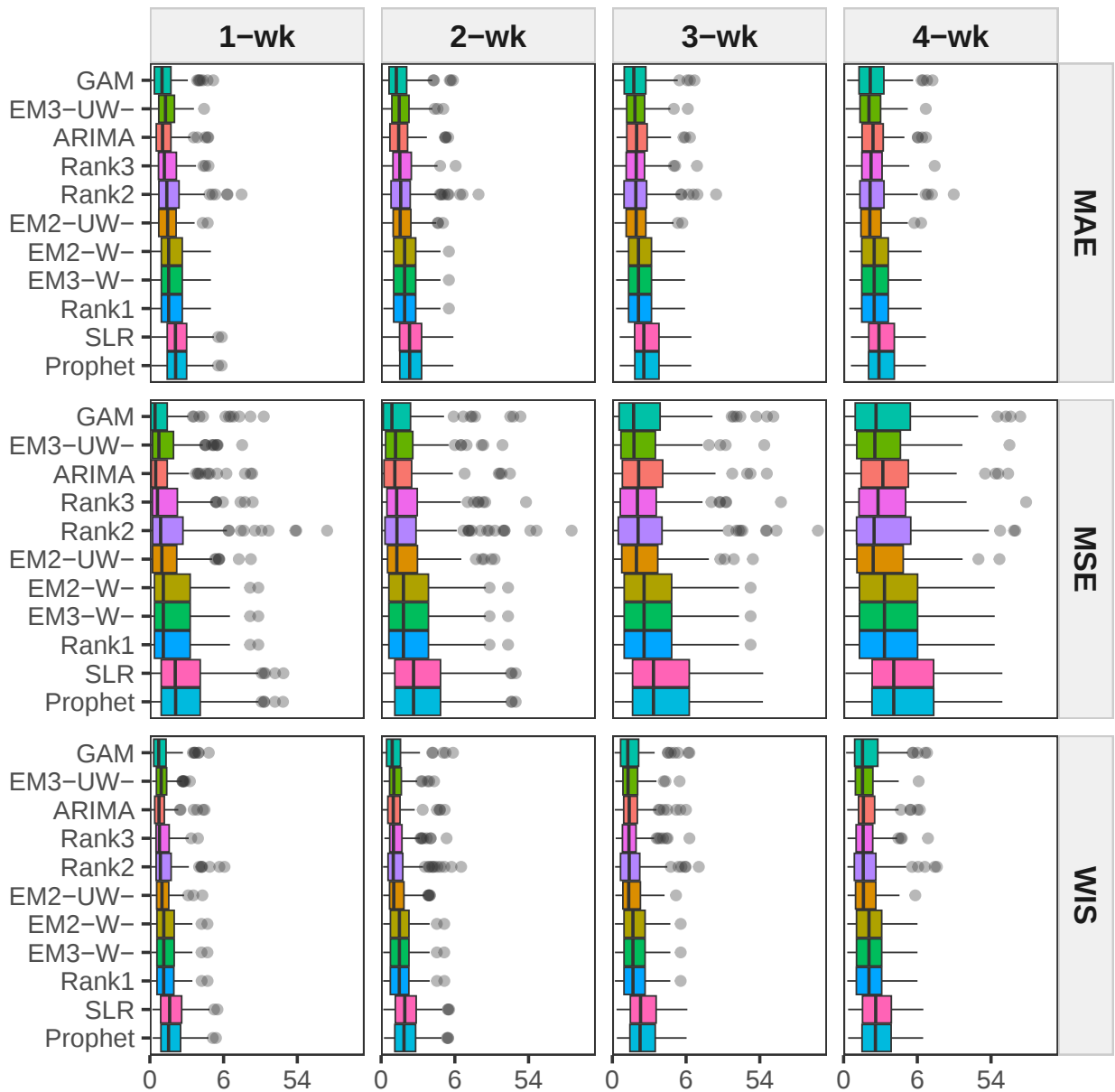


Figure S1: National-level distributions of forecast performance across models and horizons. Boxplots show the distributions of MAE, MSE, and WIS (rows; plotted on a log scale) for 1–4 week-ahead forecasts (columns) over all national forecast dates. For each model, the vertical line inside the box marks the median, and the box width spans the interquartile range. Whiskers extend to $1.5 \times \text{IQR}$, and gray points indicate outliers.

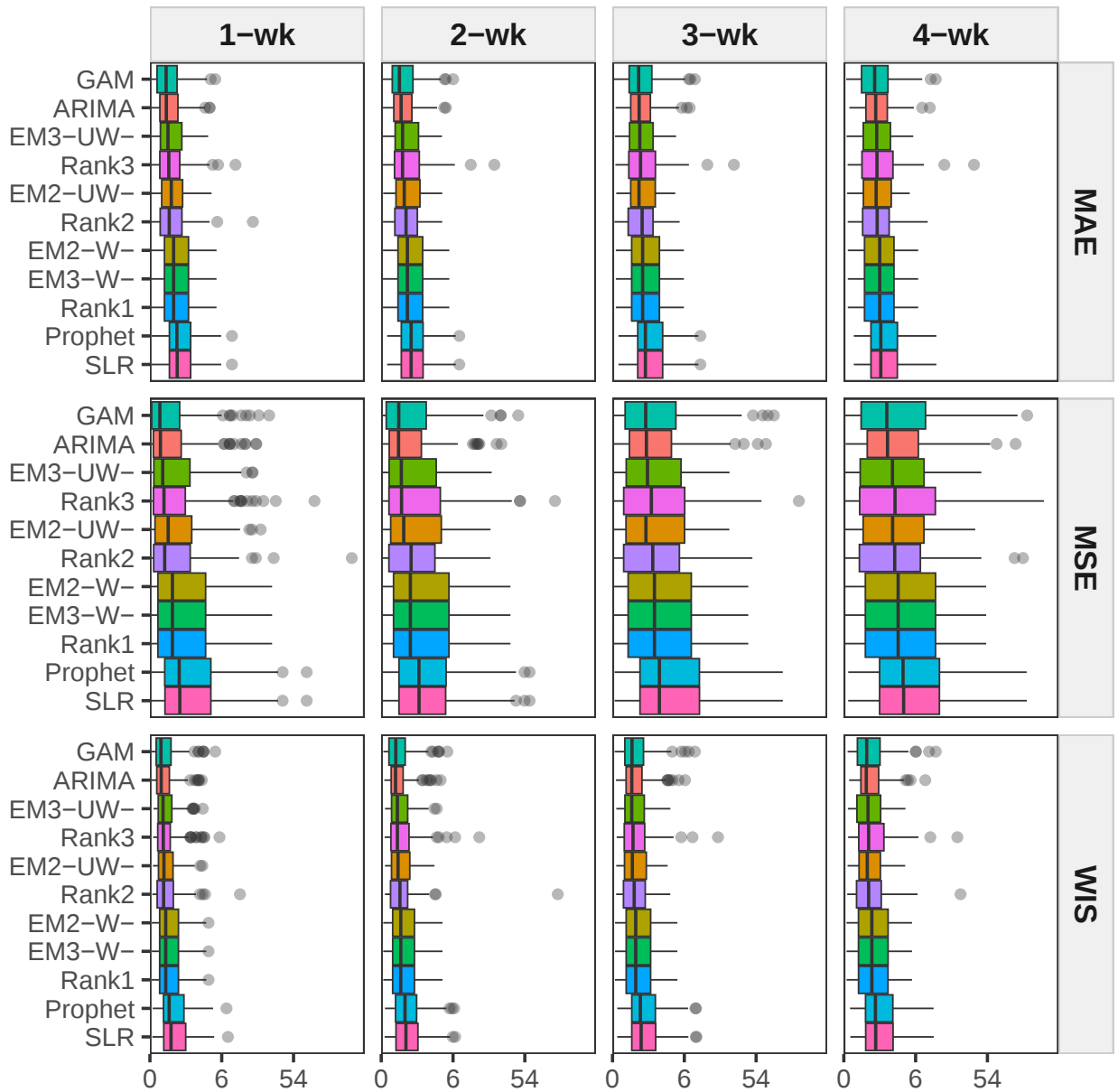


Figure S2: Midwest region distributions of forecast performance across models and horizons. Boxplots show the distributions of MAE, MSE, and WIS (rows; plotted on a log scale) for 1–4 week-ahead forecasts (columns) over all Midwest region forecast dates. For each model, the vertical line inside the box marks the median, and the box width spans the interquartile range. Whiskers extend to $1.5 \times \text{IQR}$, and gray points indicate outliers.

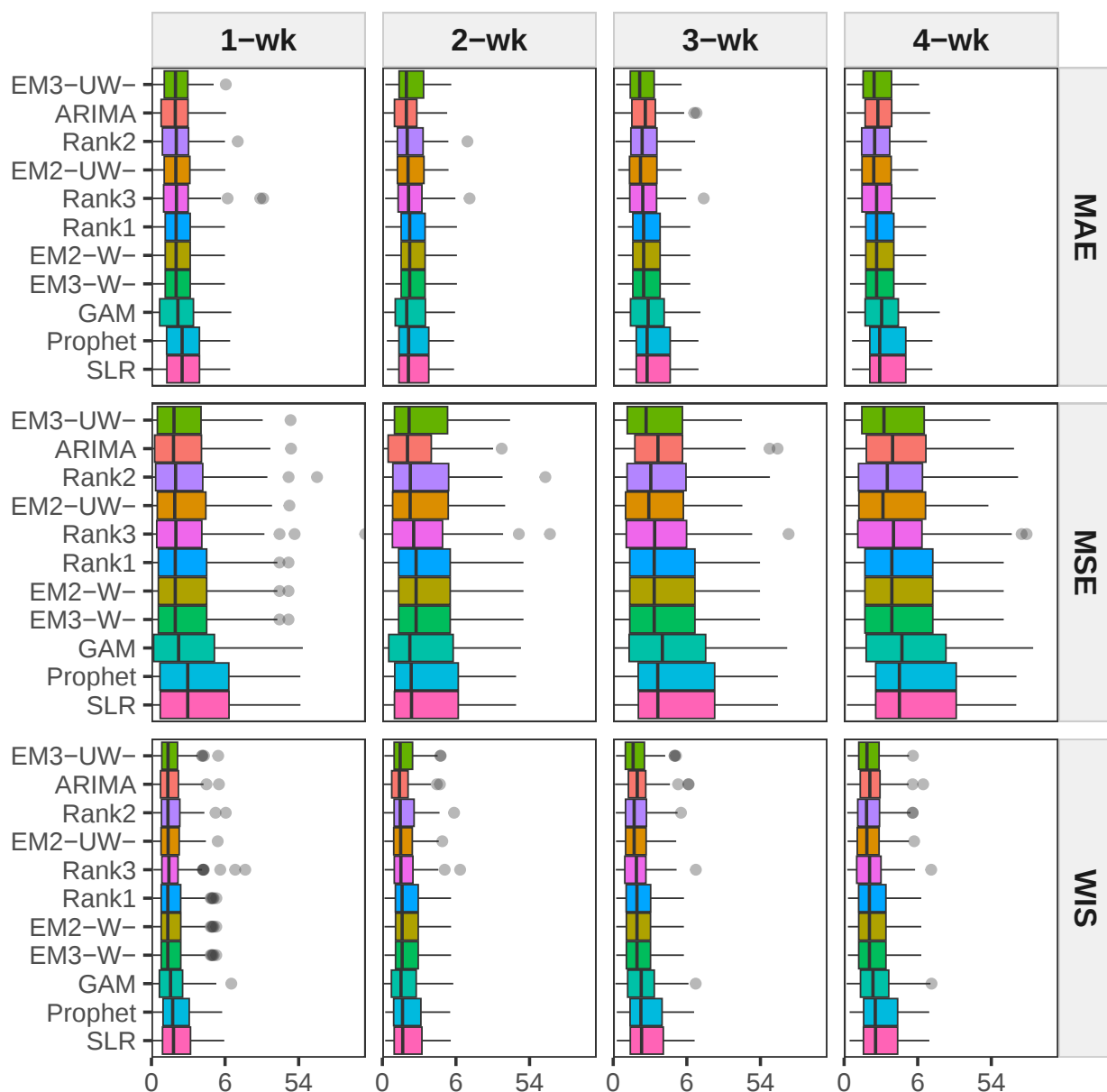


Figure S3: Northeast region distributions of forecast performance across models and horizons. Boxplots show the distributions of MAE, MSE, and WIS (rows; plotted on a log scale) for 1–4 week-ahead forecasts (columns) over all Northeast region forecast dates. For each model, the vertical line inside the box marks the median, and the box width spans the interquartile range. Whiskers extend to $1.5 \times \text{IQR}$, and gray points indicate outliers.

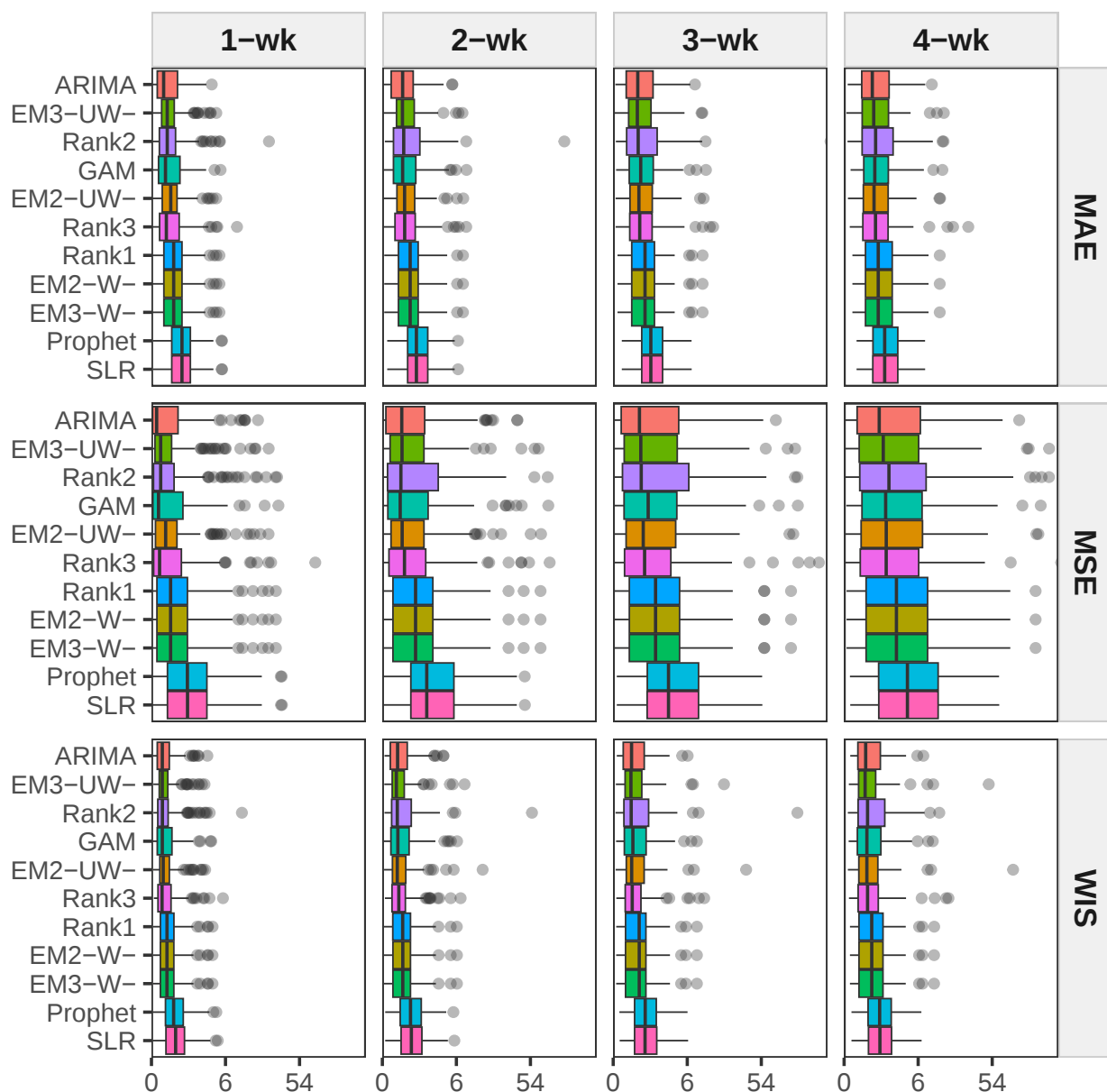


Figure S4: South region distributions of forecast performance across models and horizons. Boxplots show the distributions of MAE, MSE, and WIS (rows; plotted on a log scale) for 1–4 week-ahead forecasts (columns) over all South region forecast dates. For each model, the vertical line inside the box marks the median, and the box width spans the interquartile range. Whiskers extend to $1.5 \times \text{IQR}$, and gray points indicate outliers.

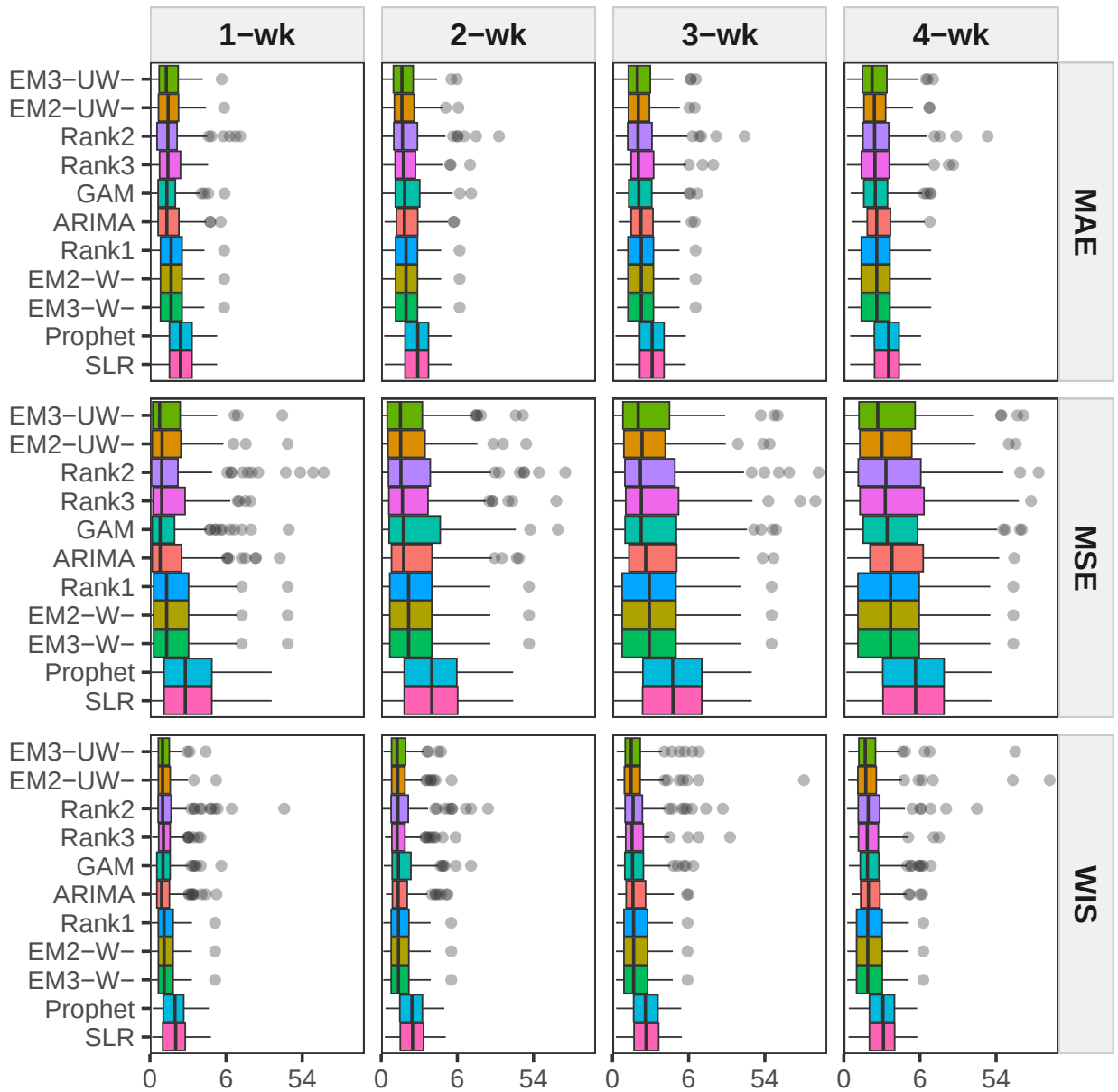


Figure S5: West region distributions of forecast performance across models and horizons. Boxplots show the distributions of MAE, MSE, and WIS (rows; plotted on a log scale) for 1–4 week-ahead forecasts (columns) over all West region forecast dates. For each model, the vertical line inside the box marks the median, and the box width spans the interquartile range. Whiskers extend to $1.5 \times \text{IQR}$, and gray points indicate outliers.

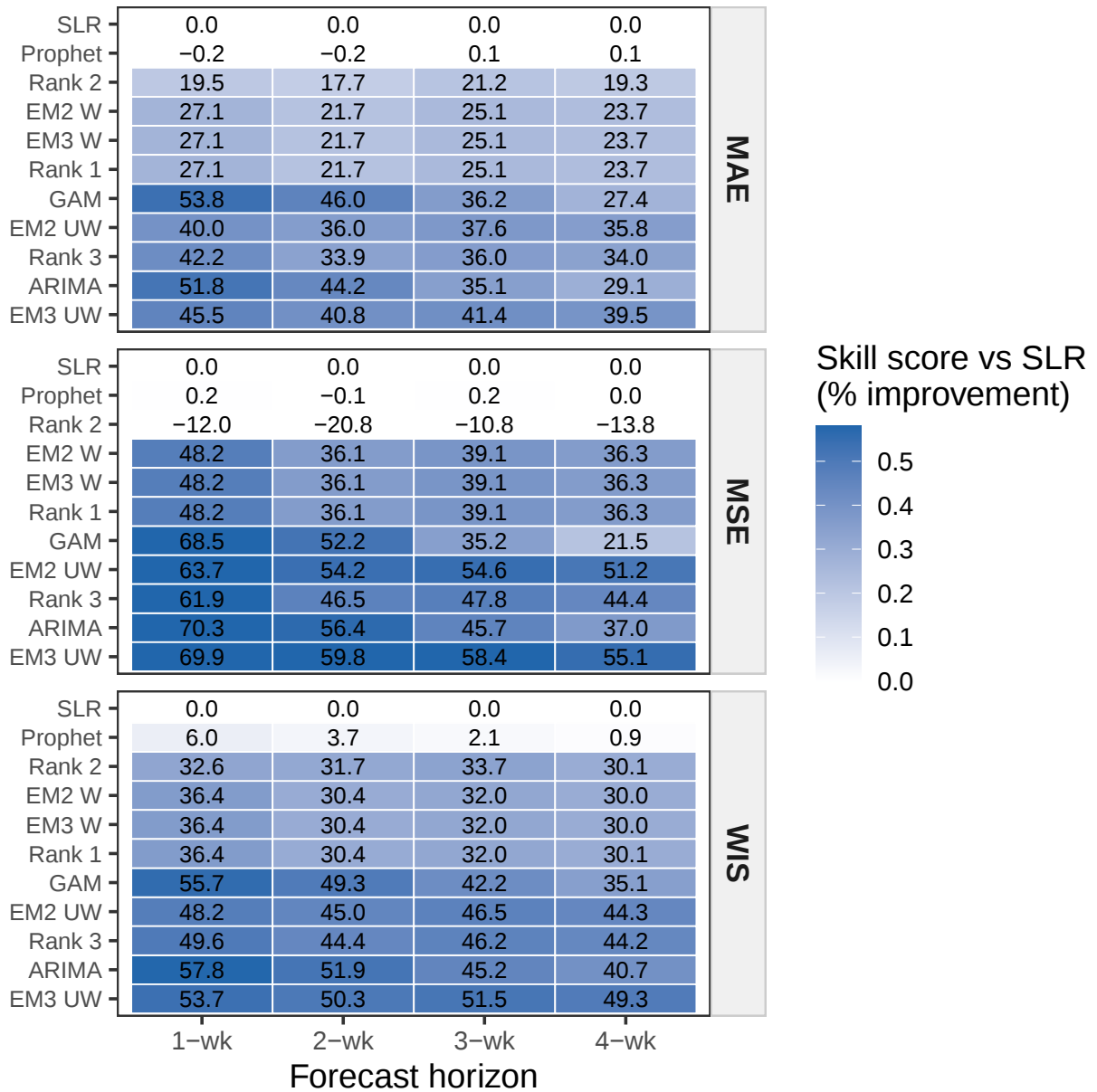


Figure S6: Skill scores relative to the SLR baseline at the national level across models, horizons, and metrics. Heatmaps display percent skill scores ($100 \times [1 - \bar{m}/\bar{m}_{\text{SLR}}]$), where $\bar{m}$ and $\bar{m}_{\text{SLR}}$ denote the average MAE, MSE, or WIS across all forecast dates for a given model and for the SLR baseline, respectively. Rows correspond to MAE, MSE, and WIS, and columns to 1–4 week-ahead horizons. Positive values (blue) indicate lower error than SLR (better performance), while values near zero (white) indicate performance similar to the baseline; SLR is therefore zero by construction. The fill scale was bounded using the 5th and 95th percentiles of the skill-score distribution, with values outside this range clipped to the endpoint colors, while cell annotations report the original percent skill values.

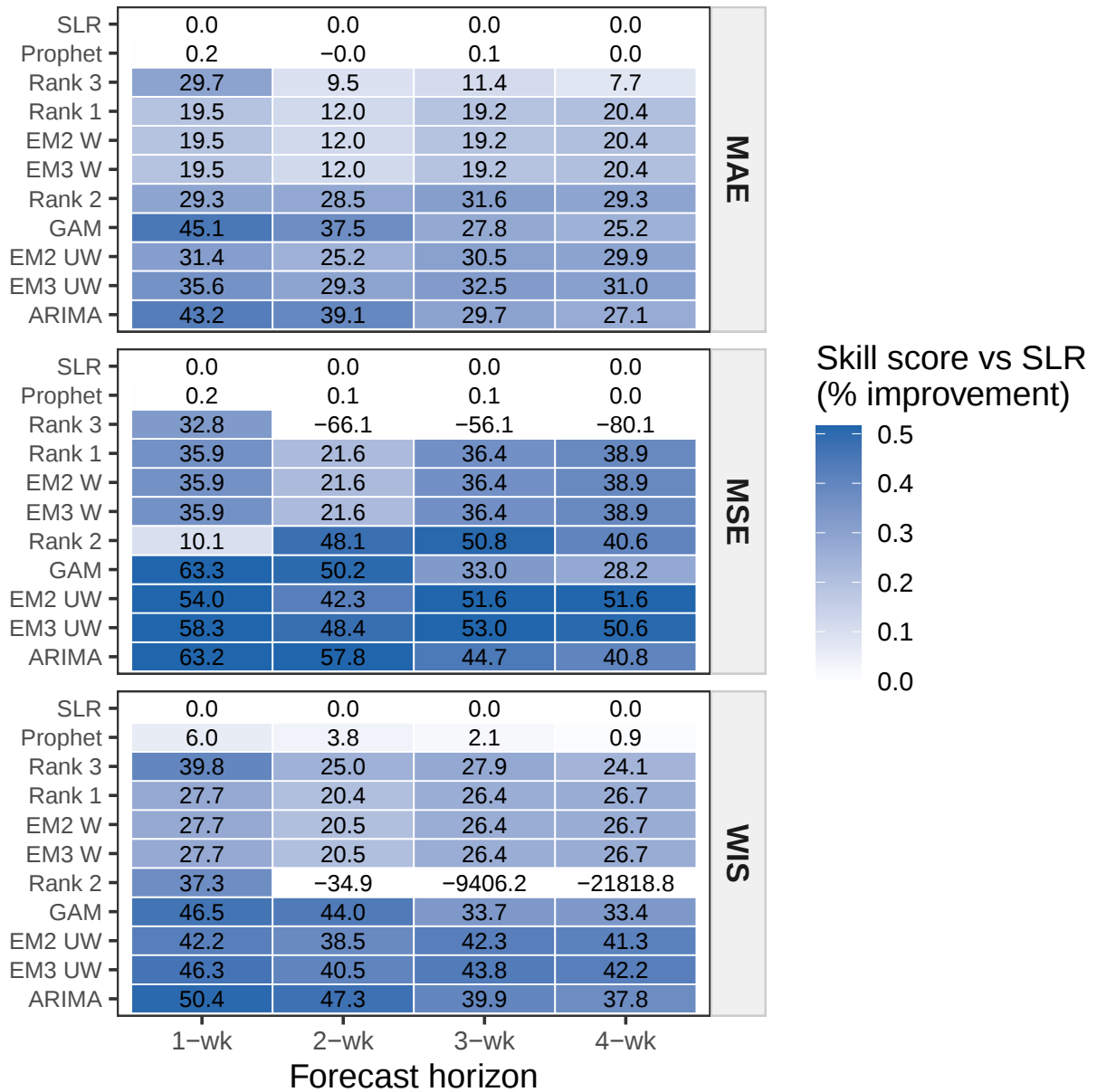


Figure S7: Midwest region skill scores relative to the SLR baseline across models, horizons, and metrics. Heatmaps display percent skill scores ($100 \times [1 - \bar{m}/\bar{m}_{\text{SLR}}]$), where $\bar{m}$ and $\bar{m}_{\text{SLR}}$ denote the average MAE, MSE, or WIS across all forecast dates for a given model and for the SLR baseline, respectively. Rows correspond to MAE, MSE, and WIS, and columns to 1–4 week-ahead horizons. Positive values (blue) indicate lower error than SLR (better performance), while values near zero (white) indicate performance similar to the baseline; SLR is therefore zero by construction. The fill scale was bounded using the 5th and 95th percentiles of the skill-score distribution, with values outside this range clipped to the endpoint colors, while cell annotations report the original percent skill values.

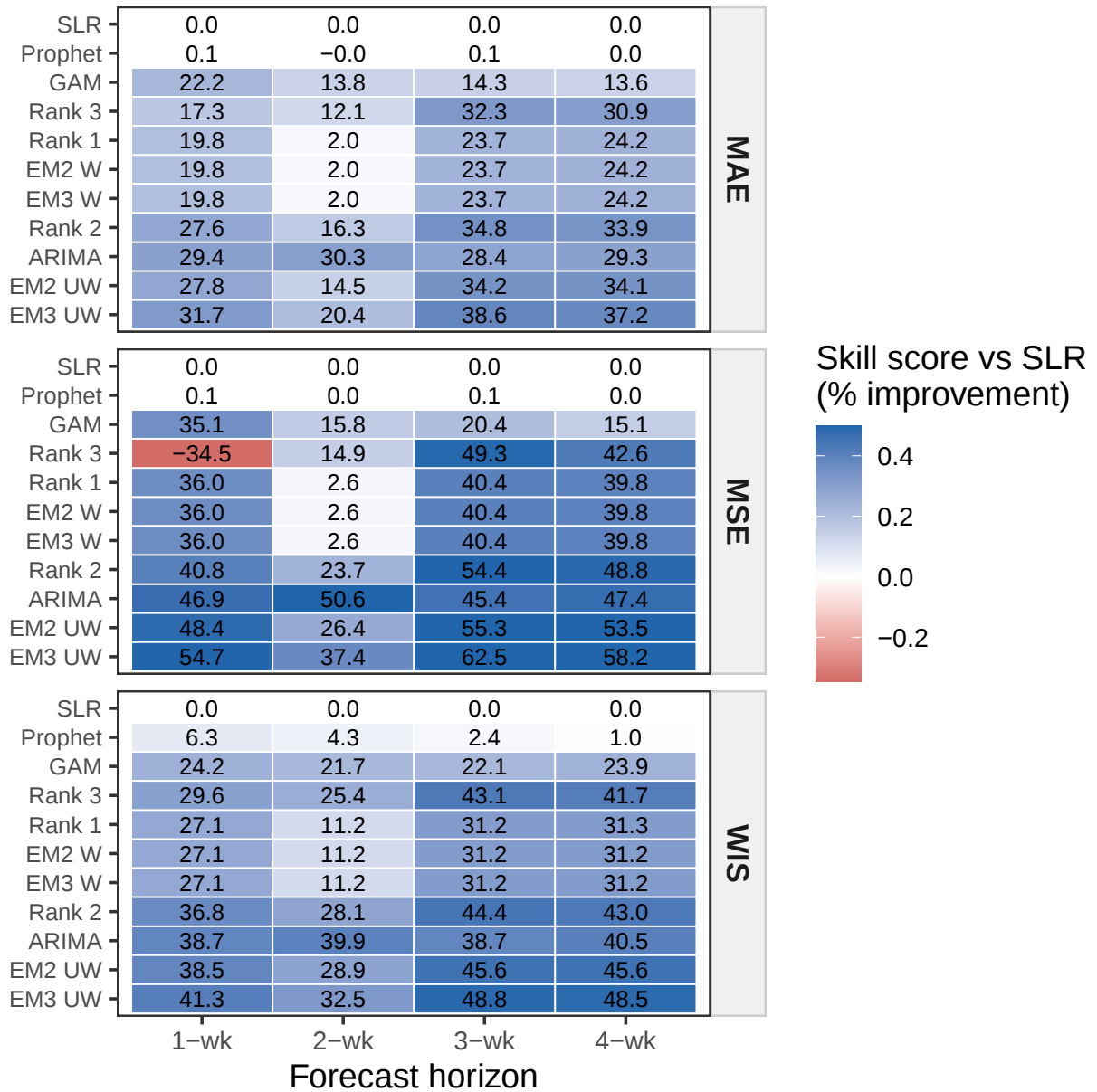


Figure S8: Northeast region skill scores relative to the SLR baseline across models, horizons, and metrics. Heatmaps display percent skill scores ($100 \times [1 - \bar{m}/\bar{m}_{\text{SLR}}]$), where $\bar{m}$ and $\bar{m}_{\text{SLR}}$ denote the average MAE, MSE, or WIS across all forecast dates for a given model and for the SLR baseline, respectively. Rows correspond to MAE, MSE, and WIS, and columns to 1–4 week-ahead horizons. Positive values (blue) indicate lower error than SLR (better performance), whereas negative values (red) indicate higher error than SLR (worse performance); by definition, SLR has a skill score of zero. To reduce the influence of extreme negative outliers on the color mapping, the fill scale was bounded using the 5th and 95th percentiles of the skill-score distribution: values below the 5th percentile and above the 95th percentile were clipped to the endpoint colors, while cell annotations report the original percent skill values.

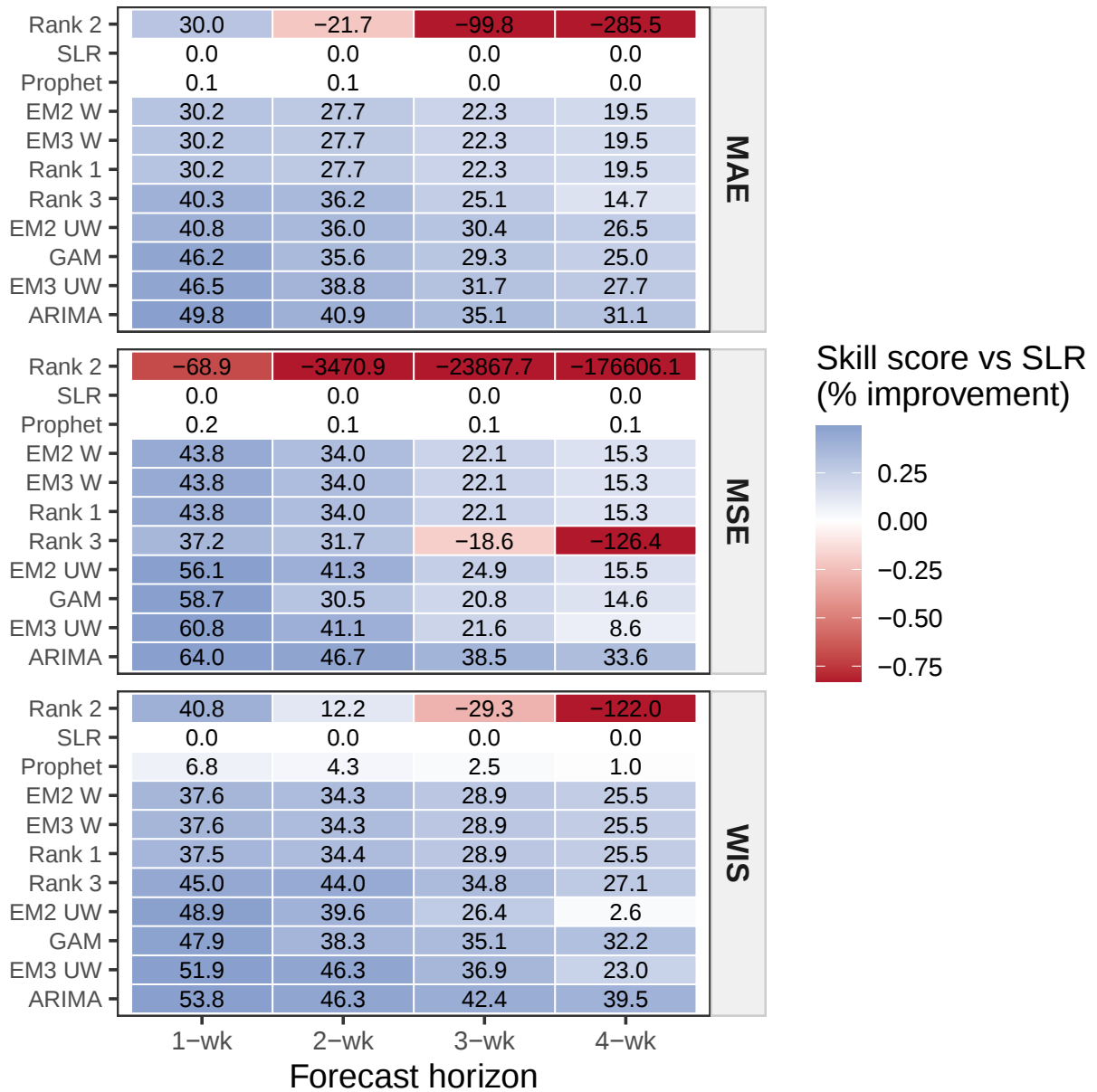


Figure S9: South region skill scores relative to the SLR baseline across models, horizons, and metrics. Heatmaps display percent skill scores ($100 \times [1 - \bar{m}/\bar{m}_{\text{SLR}}]$), where $\bar{m}$ and $\bar{m}_{\text{SLR}}$ denote the average MAE, MSE, or WIS across all forecast dates for a given model and for the SLR baseline, respectively. Rows correspond to MAE, MSE, and WIS, and columns to 1–4 week-ahead horizons. Positive values (blue) indicate lower error than SLR (better performance), whereas negative values (red) indicate higher error than SLR (worse performance); by definition, SLR has a skill score of zero. To reduce the influence of extreme negative outliers on the color mapping, the fill scale was bounded using the 5th and 95th percentiles of the skill-score distribution: values below the 5th percentile and above the 95th percentile were clipped to the endpoint colors, while cell annotations report the original percent skill values.

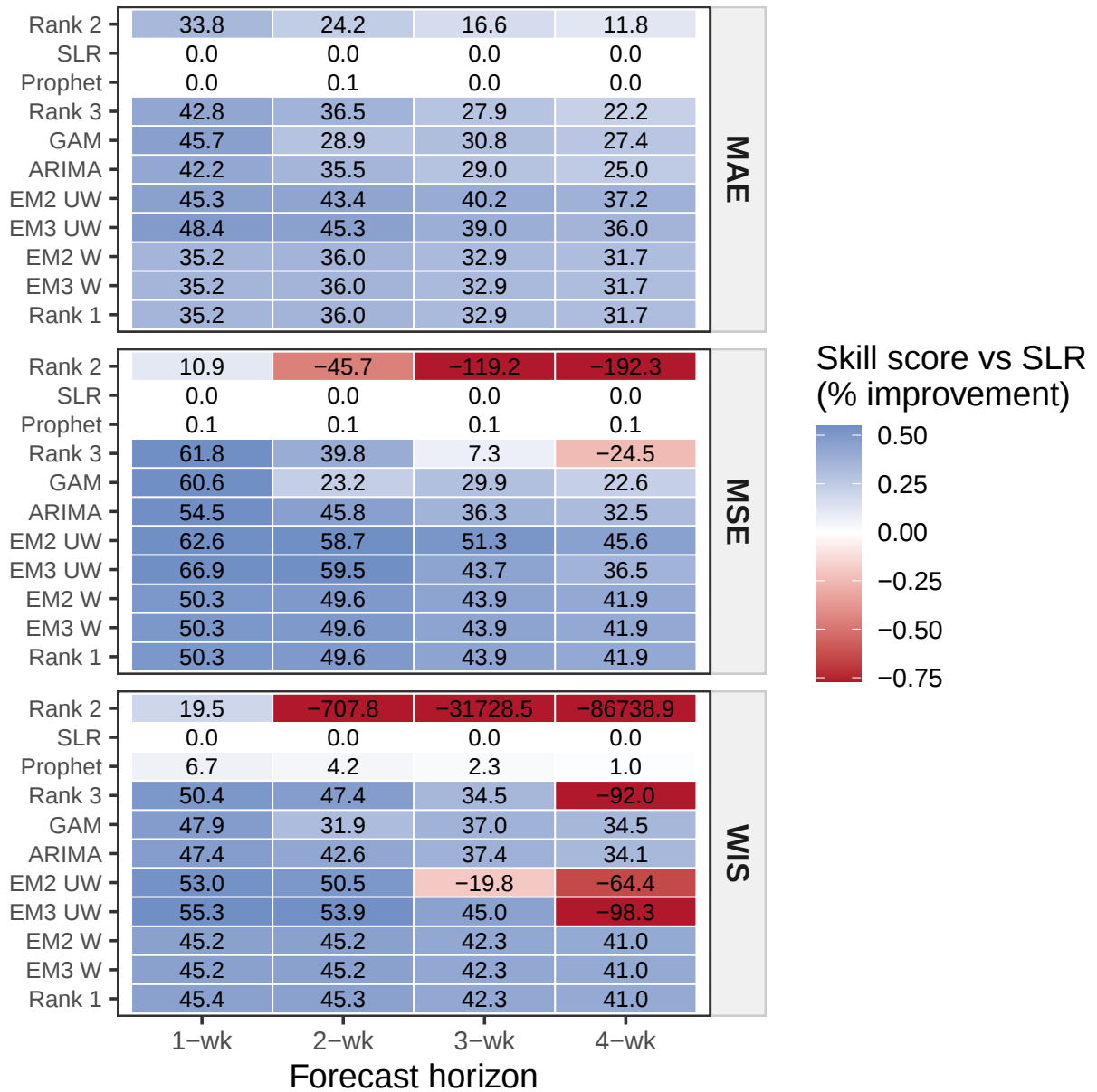


Figure S10: West region skill scores relative to the SLR baseline across models, horizons, and metrics. Heatmaps display percent skill scores ($100 \times [1 - \bar{m}/\bar{m}_{\text{SLR}}]$), where $\bar{m}$ and $\bar{m}_{\text{SLR}}$ denote the average MAE, MSE, or WIS across all forecast dates for a given model and for the SLR baseline, respectively. Rows correspond to MAE, MSE, and WIS, and columns to 1–4 week-ahead horizons. Positive values (blue) indicate lower error than SLR (better performance), whereas negative values (red) indicate higher error than SLR (worse performance); by definition, SLR has a skill score of zero. To reduce the influence of extreme negative outliers on the color mapping, the fill scale was bounded using the 5th and 95th percentiles of the skill-score distribution: values below the 5th percentile and above the 95th percentile were clipped to the endpoint colors, while cell annotations report the original percent skill values.