FORECASTING INFLUENZA DYNAMICS IN THE US: A COMPARATIVE ANALYSIS USING THE N-SUB-EPIDEMIC MODEL AND THE STATMODPREDICT, A USER-FRIENDLY R-SHINY INTERFACE WITH STATISTICAL MODELS.

by

Clement Ampong

Under the Direction of Alexandra Smirnova, Ph.D.

and co-adviser Gerardo Chowell, Ph.D.

A Non-Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of

Master of Science in Mathematics

in the College of Arts and Sciences

Georgia State University

April 18, 2025

ACKNOWLEDGEMENT

I give all glory to the Almighty God for how far He has brought me. His grace and guidance have been my strength throughout this journey; praise be to His name.

I am extremely thankful to my supervisors, Professor Alexandra Smirnova, and Professor Gerardo Chowell, for their patience and insightful guidance that enabled the creation of this report.

I extend my profound gratitude to all the lecturers in my department for the impact they have had on my academic and personal life.

ABSTRACT

Forecasting Influenza Dynamics in the U.S is a project that I collected case data from the Centers for Disease Control and Prevention (CDC) between May and October 2024. I used the `n-sub epidemic` model [8] and the `STAT MOD Predict` toolbox [7] to evaluate the predictive performance of various forecasting models across the U.S.

The analysis focused on the five regions in the U.S including: National, Midwest, Northeast, South, and West. The models that were evaluated include Autoregressive Integrated Moving Average (ARIMA), Generalized Linear Model (GLM), Generalized Additive Model (GAM), Prophet, as well as top-ranked and weighted ensemble models. Performance metrics of these models were measured using the Mean Absolute Error (MAE), Mean Squared Error (MSE), 95% Prediction Interval (PI) coverage, and the Weighted Interval Score (WIS). Among all the models tested, ARIMA regularly outperformed others in both accuracy and reliability, while the weighted ensemble models performed nearly as well. On the contrary, the models that showed higher error rates and less accurate prediction intervals are the GLM and Prophet models.

These results show how important reliable forecasting tools are for guiding public health initiatives and improving epidemic prevention.

INDEX WORDS: Influenza Forecasting, n-sub-epidemic Model, Time Series, Forecast Accuracy.

1 Introduction

Influenza is a respiratory infection caused by various strains of the influenza virus, with outcomes that can range from mild illness to severe complications requiring hospitalization or resulting in death [1]. According to the Centers for Disease Control and Prevention (CDC), an average of 8% of Americans contracts influenza every season [2]. However, during the COVID-19 epidemic, this usual trend was significantly changed. Public health measures like mask-wearing, social distancing, lockdowns, and decreased foreign travel were widely used during the 2020–2021 flu season, which caused influenza activity to fall to historically low levels [3]. Influenza remains a significant public health challenge in the United States, with its seasonal and regional variability necessitating robust forecasting systems. Accurate predictions of influenza trends are critical for enabling timely healthcare interventions, optimizing medical resources, and reducing the disease's societal impact [4, 5]. Reliable forecasts empower public health officials to respond proactively, minimizing morbidity and mortality associated with influenza outbreaks. To encourage researchers to predict regional and national influenza activity, the CDC has been holding the "Predict the Influenza Season Challenge" (also known as the CDC FluSight challenge) since 2013. [6].

This study utilizes data collected through reported influenza cases and wastewater surveillance to evaluate predictive models' performance in forecasting influenza dynamics across the U.S. The dataset, spanning 26 weeks (May 4, 2024, to October 26, 2024), includes decimal-normalized case counts organized by five regions: National, Midwest, Northeast, South, and West. Wastewater surveillance provides an innovative early-warning tool, detecting viral activity within communities even before symptomatic cases are reported [5]. The viral activity levels from wastewater data are categorized based on infection risk, which offers a detailed understanding of community transmission dynamics. These levels of viral activity are divided into the following groups:

- **Minimal**: Up to 1.6
- **Low**: Greater than 1.6 and up to 4.5
- **Moderate**: Greater than 4.5 and up to 12.2
- **High**: Greater than 12.2 and up to 20.1
- **Very High**: Greater than 20.1

These levels help assess the risk of infection in an area.

To evaluate forecasting accuracy, two complementary frameworks were employed: the *n-sub epidemic model* and the *STAT MOD PREDICT toolbox*. These frameworks assessed a range of models, including *ARIMA* (AutoRegressive Integrated Moving Average), *GLM* (Generalized Linear Models), *GAM* (Generalized Additive Models), *Prophet*, and various ranked and weighted ensemble approaches [7, 8]. Each model was analyzed for its ability to predict influenza case trajectories and infection risks accurately.

1.1 Literature Review

Influenza forecasting is critical for managing public health responses and minimizing the disease's impact. In the case of newly emerging infectious diseases that spread rapidly through the population, producing short-term projections of outbreak trajectories across various geographic regions becomes especially important. These Forecast play a key role in informing timely public health policies and guiding targeted interventions strategies [9]. Various statistical models, such as ARIMA, GLM, and GAM, have been widely employed for predicting influenza trends.

ARIMA models are particularly effective for short-term forecasts due to their ability to capture temporal patterns. GLM and GAM models enhance flexibility by accommodating non-linear trends and additional covariates, such as vaccination rates and environmental factors [7].

Ensemble modeling approaches, such as Chowell's n-subepidemic framework, further improve forecasting accuracy by combining predictions from multiple models [8]. These methods are particularly effective in addressing the variability in regional epidemic patterns. For example, the SubEpiPredict toolbox has been specifically designed for this framework, allowing for the fitting and forecasting of epidemic growth trajectories using an ensemble of sub-epidemic models [12].

Frameworks like the STAT MOD PREDICT toolbox [7] systematically evaluate the performance of multiple models, enabling the selection of the most accurate and reliable forecasting tools. Additionally, the QuantDiffForecast toolbox [10] has been used for parameter estimation and forecasting based on systems of ordinary differential equations, offering quantified uncertainty in its predictions. The GrowthPredict toolbox also provides robust tools for forecasting growth trajectories in epidemics using phenomenological models, contributing significantly to forecasting capabilities [11].

1.2 Research Questions

To investigate the effectiveness and reliability of various forecasting models, the study addresses the following research questions:

- What are the most accurate and reliable predictive models for forecasting influenza dynamics across different regions in the United States?
- How can predictive models support timely public health interventions and resource allocation to mitigate the impact of influenza outbreaks?

2 Methodology

This study employs a systematic approach to forecast influenza dynamics in the United States, leveraging historical data and advanced modeling techniques. The methodology is structured as follows:

2.1 Data Source

Historical influenza case data were derived from wastewater surveillance datasets provided by CDC [5]. The dataset includes region-specific, decimal-normalized case counts spanning 26 weeks, from May 4, 2024, to October 26, 2024. The regions analyzed include National, Midwest, Northeast, South, East, and West, representing collections of states with varying demographic and environmental characteristics influencing influenza trends.

2.2 Modeling Influenza Outbreaks

The influenza data were modeled using two primary approaches: The `StatModPredict` toolbox, which offers a suite of statistical techniques for predicting time-series data [7], and the *n-sub epidemic model*, which is specifically designed to capture the temporal progression of influenza outbreaks across multiple subpopulations [8]. These models were applied to the CDC influenza data to generate forecasts of case counts and assess their performance in capturing outbreak dynamics. Detailed analysis of each approach is given next.

2.3 The StatModPredict toolbox

The `StatModPredict` toolbox includes four built-in models that is used for epidemic modeling and time series analysis. These models are GLM, GAM, ARIMA, and Facebook's prophet model. Detailed analysis of each model is given next.

2.3.1 GLM Model

The Generalized-Logistic Growth Model (GLM) is a mathematical tool used to predict the progression of an epidemic by calculating the cumulative number of cases over time. The model is defined by differential equation [11], given below:

$$\frac{dC(t)}{dt} = C'(t) = rC^p(t)\left(1 - \frac{C(t)}{K_0}\right),$$

where $c(t)$ is the total number of cases at time t and $c'(t)$ is the rate of change that corresponds to new cases reported. The parameter r is the growth rate of the epidemic, while k_0 represent the final epidemic size (that is, the maximum number of cumulative cases the outbreak is expected to reach). $p \in [0, 1]$ is the scaling parameter that controls the early dynamics of the outbreak. A value of $p = 0$ implies that the number of new cases remains constant, while a value of $p = 1$ indicates that the outbreak begins with exponential growth. Lastly, for $0 < p < 1$, the model captures sub-exponential growth behavior.

2.3.2 GAM Model

The The Generalized Additive Model (GAM) is an extension of the Generalized Linear Model [13], that uses smooth functions to explain nonlinear relationships between predictors and a response variable. The model [15] is defined as:

$$y_t = \beta_0 + s(t) + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, \sigma^2),$$

where y_t is the response variable at time t, β_0 is the intercept term, ε_t is the error term that is assumed to be normally distributed with mean of 0, and variance σ^2.

$s(t)$ is an unknown smooth function of time, that captures nonlinear patterns in the data over time. It allows the model to adjusts itself to fit the shape of the data instead of assuming a linear relationship. It is express as a linear combination of basis function (splines), given below:

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

Here, $b_k(t)$ are the basis functions, β_k are the coefficients associated with each basis function, and K is the number of basis functions that controls the flexibility of $s(t)$.

2.3.3 ARIMA Model

The Auto-Regressive Integrated Moving Average (ARIMA) model is a widely used statistical approach for modeling and forecasting time series data [14] [16].. The model consists of three parts which are : Auto-regression (AR), Integration (I), and Moving Average (MA).

The (AR) part models the relationship between an observation and its past value. The (MA) part models the error of the series as a linear combination of past forecast errors. The (I) component refers to differencing the data to remove trends and achieve stationarity, which ensures constant mean and variance over time. The general mathematical form of an ARIMA model is given by:

$$\phi(B)(1-B)^d y_t = c + \theta(B)\varepsilon_t$$

Here, y_t represent the value of the series at time t, and ε_t is the error. $(1-B)^d$ means conducting the differencing d times. B is the backhsif operator, where $By_t = y_{t-1}$, $B^2y_t = y_{t-2}$, and so on [17].

The function $\phi(B)$ defines the (AR) part of the model and is given as

$$\phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p,$$

where p is the order of the autoregressive component.
The (MA) part of the model is also given as

$$\theta(B) = 1 - \theta_1 B - \cdots - \theta_q B^q,$$

where q is the order of the moving average component, and d represents the degree of differencing, which helps to stabilize the mean of the time series.

2.3.4 Facebook's Prophet Model

The Prophet model [18] is a forecasting algorithm that was developed by Facebook. It was originally designed for business applications but now widely used in epidemiological modeling, including forecasting for COVID-19 [19]. The model breaks down a time series into multiple multiple additive components: a trend component, a seasonality component, a holiday or event component, and an error term. The model is given below:

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

where $y(t)$ is the observed time series value at time t, $g(t)$ is the non-periodic trend, $s(t)$ captures the seasonal effects like weekly or yearly cycles, $h(t)$ accounts for the holiday, and ϵ_t is the error term, assumed to be normally distributed. The default settings of the `prophet` function from the R `prophet` was applied, following the approach by Taylor [20].

2.4 The n-sub Epidemic Model

The `n-sub epidemic` model is a compartmental modeling technique that combines several overlapping sub-epidemics that represent the overall epidemic trajectory. Each sub-epidemic is modeled using a modified logistic growth formulation. The method allows for flexibility in modeling epidemic curves that may not follow a single peak structure due to multiple waves. The model [11] is given as:

$$\frac{dC_i(t)}{dt} = C_i'(t) = A_i(t)\, r_i\, C_i^{p_i}(t) \left(1 - \frac{C_i(t)}{K_{0i}}\right),$$

where, $C_i(t)$ is the cumulative number of cases for the $i-$th sub-epidemic at time t, and $C_i'(t)$ is the rate of change(thus, the number of new cases per unit time). r_i denotes the growth rate parameter for the i-th sub epidemic, p_i is the scaling parameter that controls the shape of the early epidemic growth. K_{0i} is the final epidemic size for the i-th sub-epidemic.
$A_i(t)$ is the activation function of each sub-epidemic, given as

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

This function controls whether the i-th sub-epidemic becomes active based on whether the cumulative count of the previous sub-epidemic exceeds a predefined threshold C_{thr}.
The total number of parameters in an n-sub epidemic is $3n + 1$, which account for the growth rate, scaling parameter, and the final size for each of the n$-$sub epidemic.
The cumulative counts for each sub-epidemic are added up to determine the total cumulative

case count for the epidemic:

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

This formulation enables the model to simulate complex epidemic behaviors by layering simple epidemic curves.

2.4.1 Top-Ranked Model Selection

The corrected Akaike Information Criterion (AICc) is used to evaluate and compare various sub-epidemic models. This criterion helps identify the best-fitting model parameter by using a nonlinear least-square optimization method to fit the model solution to the observed data . The AICc is giuven as:

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

where $SSE = \sum_{j=1}^{n_d} \left(f(t_j, \hat{\Theta}) - y_j \right)^2$ is the sum of square errors, m is the number of model parameters, and n_d is the number of data points used during the model fitting. The model with the lowest AICc value is considered to be the best-fitting model, which balance the goodness of fit and model complexity.

2.4.2 Quantifying Uncertainty with Parametric Bootstrapping

To assess uncertainty in parametric estimates and forecast, the model uses a parametric bootstrap approach [11]. This method involves resampling the model residuals and refitting the model repeatedly. For this work, a 300 bootstrap realization was used. The process allows for the construction of prediction intervals (PIs) and estimation of forecast uncertainty without requiring analytical solutions. The same method was used to generate a short-term forecast with quantified uncertainty.

2.4.3 Ensemble Model

The study combines the forecast of the top-ranked sup-epidemic models to create ensemble models. The ensemble forecast were generated using the two highest performing individual models, which improves the prediction robustness. The ensemble approach helps to reduce the variance of forecast and improve the reliability of prediction intervals [12].

2.5 Performance Metrics Evaluation

The performance metrics to assess the accuracy and model calibration was evaluated using four metrics: Mean Absolute Error (MAE), Mean Square Error (MAE), 95% Prediction Interval (PI) coverage, and the Weighted Interval Score (WIS).

MAE measures the average magnitude of absolute differences between predicted and observed values, given as

$$\text{MAE} = \frac{1}{N}\sum_{h=1}^{N}\left|f(t_h,\hat{\Theta}) - y_{t_h}\right|,$$

where N is the total number of time point used during the forecast evaluation, $f(t_h,\hat{\Theta})$ is the predicted value at time t_h based on the estimated model parameter $\hat{\Theta}$, and y_{t_h} is the actual observed value at that time.

MSE measures the average of squared difference between the predicted and observed values:

$$\text{MAE} = \frac{1}{N}\sum_{h=1}^{N}\left(f(t_h,\hat{\Theta}) - y_{t_h}\right)^2$$

95% PI coverage measures the reliability of the model's forecast uncertainty. It evaluate the proportion of observed data points that fall with the 95% prediction interval generated by the model. The model is given as:

$$\text{95\% PI coverage} = \frac{1}{N}\sum_{t=1}^{N} \mathbf{1}\{L_t < Y_t < U_t\},$$

where Y_t is the observed value at time t, L_t and U_t represent the the lower and upper bounds of the of the 95% prediction interval respectively. The indicator function $1\{.\}$ returns 1 if the observed value lies within the interval and 0 otherwise.

WIS provides a comprehensive evaluation of forecast quality by integrating multiple aspects of prediction uncertainty and accuracy. It combine several interval scores, each corresponding to different prediction intervals, and incorporate the absolute error of the predictive median. WIS is given as:

$$\text{WIS}\alpha 0 : K(F, y) = \frac{1}{K+1}\left(\frac{1}{2}\left|y - \tilde{y}\right| + \sum_{k=1}^{K} w_k \cdot \text{IS}_{\alpha_k}(F,\ y)\right),$$

where K is the number of central prediction intervals used, $\tilde{y}$ is the predictive median, $w_0 = 0.5$ is the weight assigned to the absolute error of the predictive median, $w_k = \alpha_k/2$ is the weight for each interval score, and $\text{IS}_{\alpha_k}(F,\ y)$ is the interval score at level α_k that accounts for both the width of the interval and whether the observed value falls inside it.

For a central $(1-\alpha)\%$ prediction interval, the interval score is defined as:

$$IS_\alpha(F, y) = (u - l) + \frac{2}{\alpha}(l - y) \cdot \mathbf{1}(y < l) + \frac{2}{\alpha}(y - u) \cdot \mathbf{1}(y\ >\ u)$$

Here, α is the significance level, l and u are the lower and upper bounds of the $(1-\alpha)\%$ prediction interval, and $1(\cdot)$ is the indicator function, which equal to 1 if the condition is true, and 0 otherwise.

3 Forecasting Results

We generated a 5-week forecast for the weekly influenza cases in the USA from September 28, 2024, to October 26, 2024. The performance of the top-ranking models, ensemble models, ARIMA, GLM, GAM, and Prophet models were evaluated individually to assess their forecasting accuracy, reliability, and overall effectiveness.

Note that the tables in this section provide a brief overview of the model performance, with a more comprehensive analysis discussed in the **discussion section**.

The colors indicated in the tables represent the performance of models based on key metrics:

- **Blue:** Indicates the best-performing model with the highest accuracy and reliability.
- **Green:** Represents models that are close contenders, performing well but slightly below the best model.
- **Red:** Highlights the worst-performing models, characterized by high errors or low reliability.

The results are given below:

3.1 Results for National

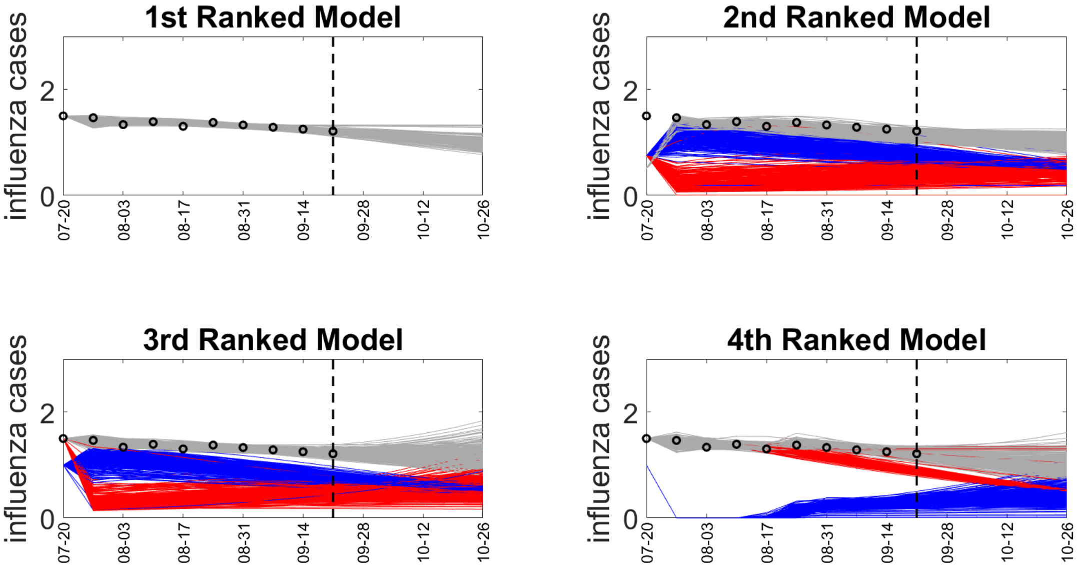


Figure 1: 5-week forecast sub-epidemic profiles derived from the top-ranking sub-epidemic models of the National weekly influenza cases. The blue and red curves represent different sub-epidemic waves generated with 300 bootstrap realizations, while the gray curves show the overall epidemic trajectory. The vertical line marks the start of the forecast, separating the calibration and forecast periods.

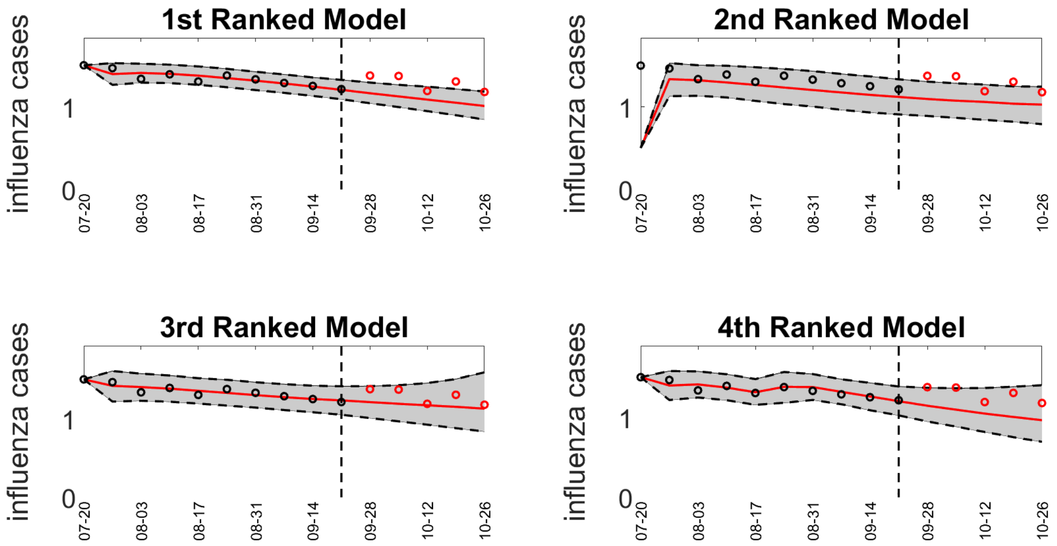


Figure 2: Shows 5-week forecasts from the top-ranking sub-epidemic models of the National weekly influenza cases. The 95% prediction interval is represented by a shaded area around the solid line, with a vertical line marking the start of the forecast to separate the calibration and forecast periods. Circles indicate the observed data points.

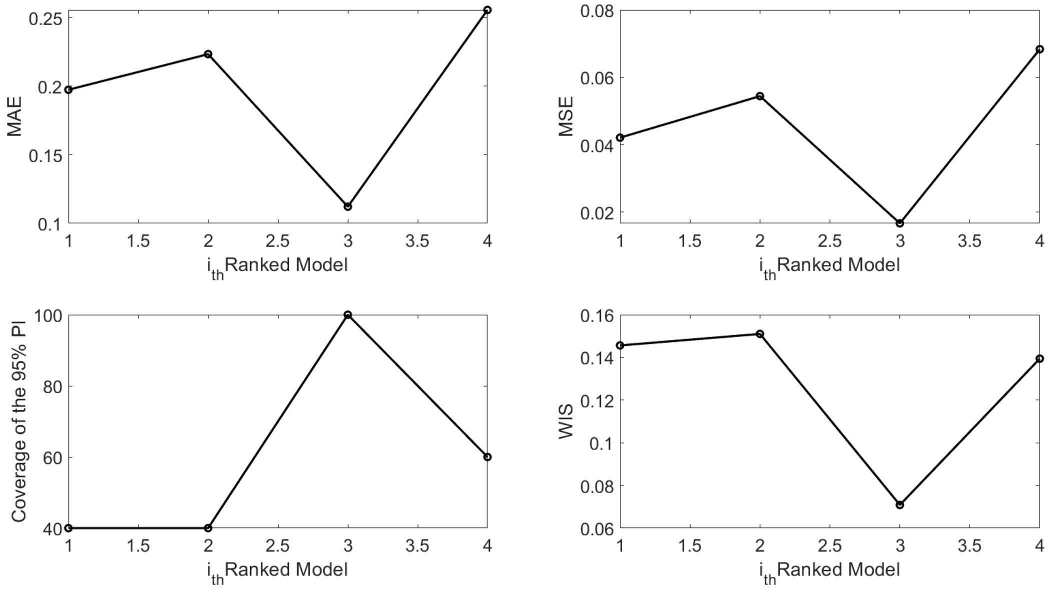


Figure 3: Performance metrics for 5-week forecasts based on the top-ranking sub-epidemic models for the National weekly influenza cases.

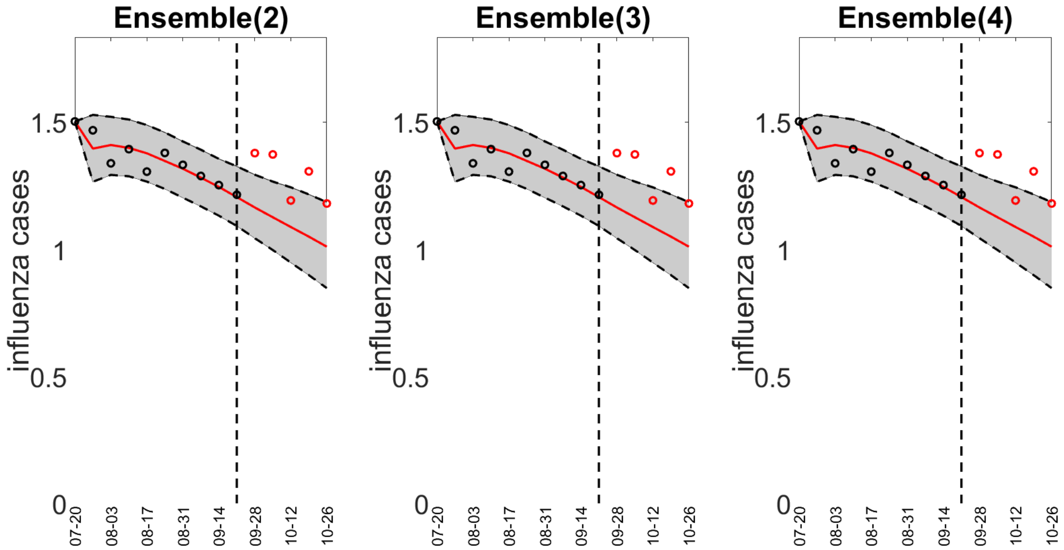


Figure 4: Displays 5-week sub-epidemic ensemble model forecasts for the National weekly influenza cases. The vertical line marks the start time of the forecast. The circles represent the data points, with model fits shown as solid lines and 95% prediction intervals indicated by shaded areas.

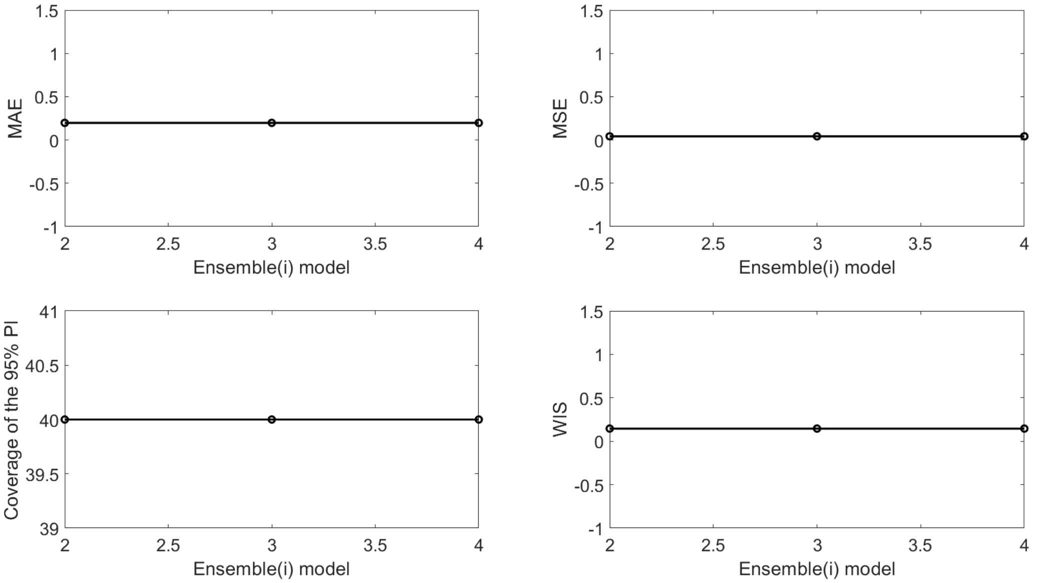


Figure 5: Performance metrics for 5-week forecasts based on the ensemble sub-epidemic models for the National weekly influenza cases.

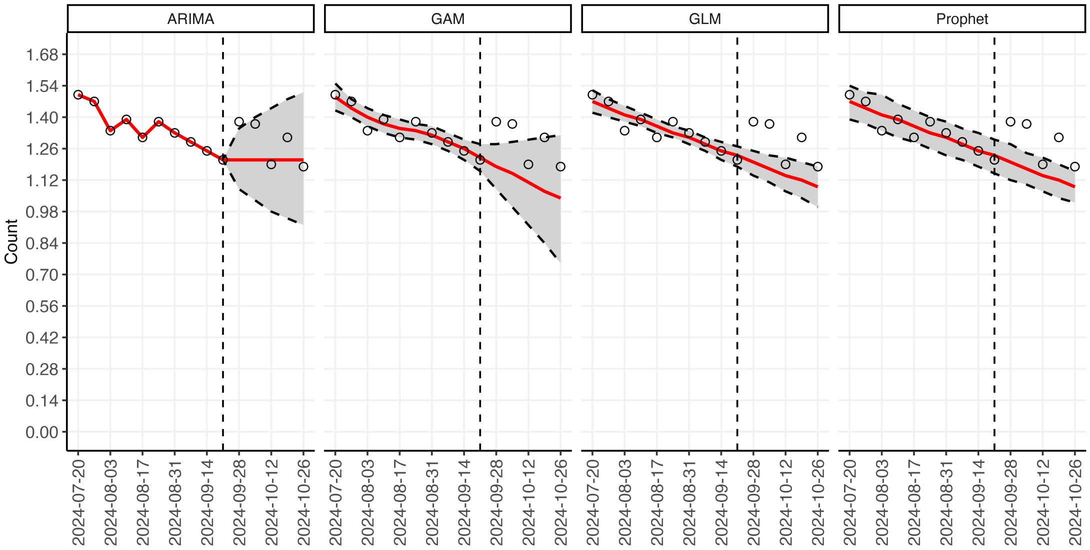
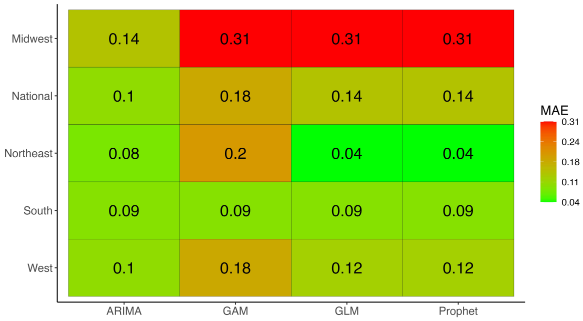
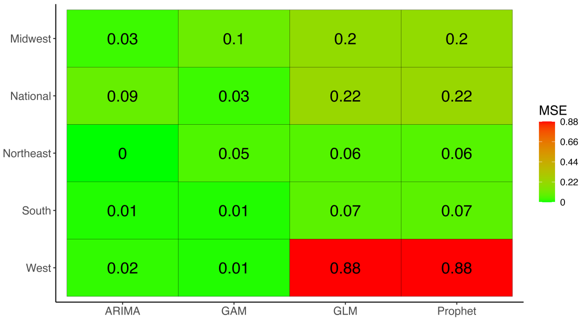


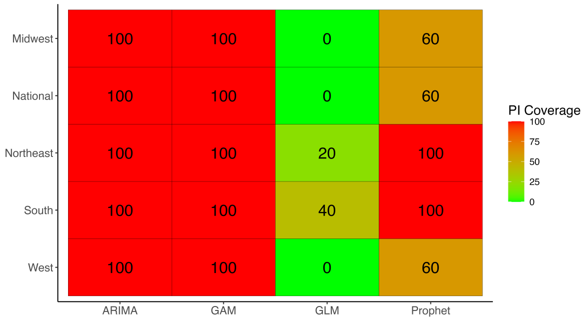
Figure 6: Shows 5-week forecasts from the ARIMA,GLM,GAM AND THE PROPHET models of the National weekly influenza cases. The 95% prediction interval is represented by a shaded area around the solid line, with a vertical line marking the start of the forecast to separate the calibration and forecast periods. The circles indicate the observed data points.



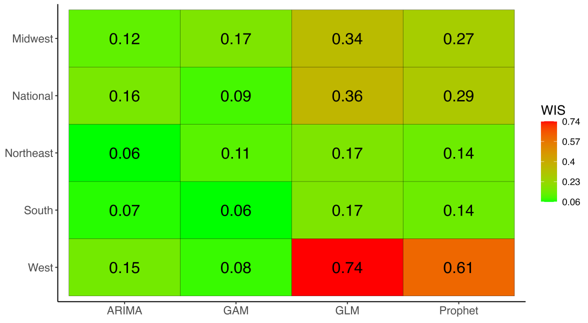
MAE - Calibration



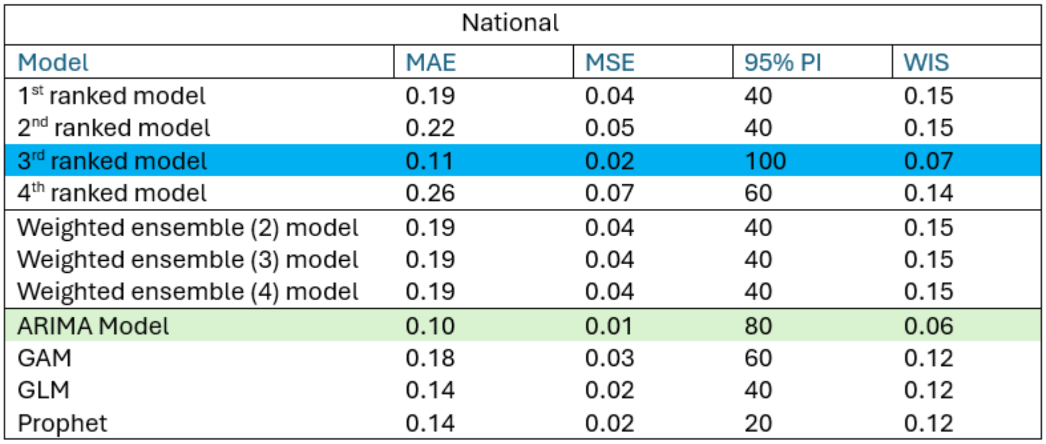
MSE - Calibration



PI - Calibration



WIS - Calibration



National				
Model	MAE	MSE	95% PI	WIS
1st ranked model	0.19	0.04	40	0.15
2nd ranked model	0.22	0.05	40	0.15
3rd ranked model	0.11	0.02	100	0.07
4th ranked model	0.26	0.07	60	0.14
Weighted ensemble (2) model	0.19	0.04	40	0.15
Weighted ensemble (3) model	0.19	0.04	40	0.15
Weighted ensemble (4) model	0.19	0.04	40	0.15
ARIMA Model	0.10	0.01	80	0.06
GAM	0.18	0.03	60	0.12
GLM	0.14	0.02	40	0.12
Prophet	0.14	0.02	20	0.12

Table 1: 5-weeks forecasting performance metrics for National.

The figures in **Table 1** compare the performance metrics of various models for forecasting weekly influenza cases. The MAE is lowest for the ARIMA model (0.10), indicating the best accuracy, followed closely by the 3rd ranked model (0.11). Similarly, the ARIMA model also achieves the lowest MSE (0.01), reflecting superior precision. While the 3rd ranked model achieves 100% coverage of the 95% prediction interval, the ARIMA model shows a narrower but reliable 80% coverage. Lastly, the WIS score highlights the ARIMA model's superior calibration (0.06), slightly outperforming the 3rd ranked model (0.07).

3.2 Results for Midwest

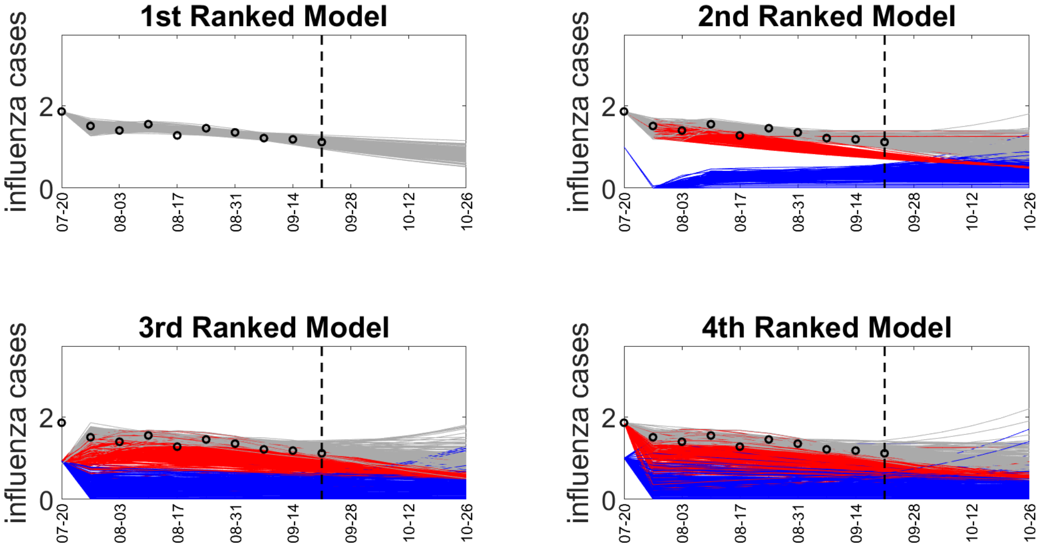


Figure 7: 5-week forecast sub-epidemic profiles derived from the top-ranking sub-epidemic models of the Midwest weekly influenza cases. The blue and red curves represent different sub-epidemic waves generated with 300 bootstrap realizations, while the gray curves show the overall epidemic trajectory. The vertical line marks the start of the forecast, separating the calibration and forecast periods.

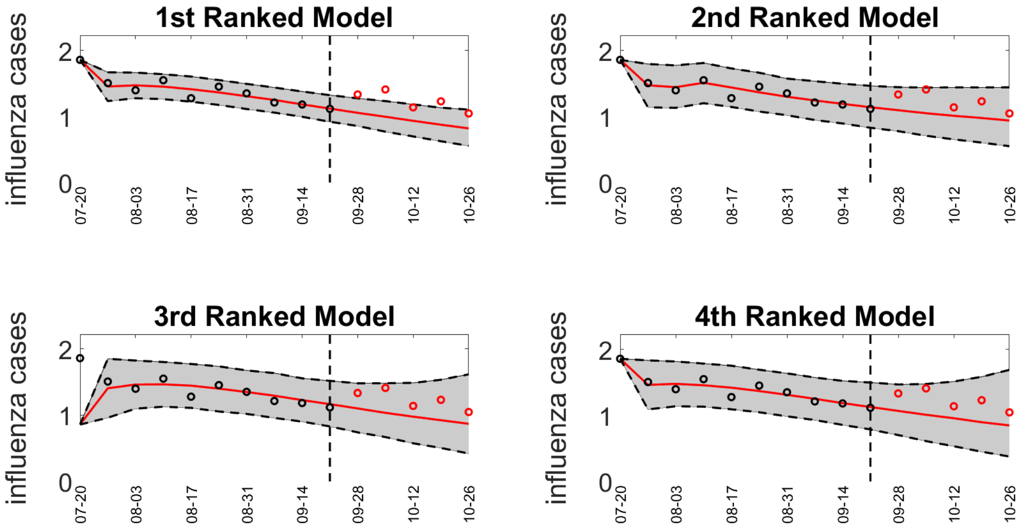


Figure 8: Shows 5-week forecasts from the top-ranking sub-epidemic models of the Midwest weekly influenza cases. The 95% prediction interval is represented by a shaded area around the solid line, with a vertical line marking the start of the forecast to separate the calibration and forecast periods. Circles indicate the observed data points.

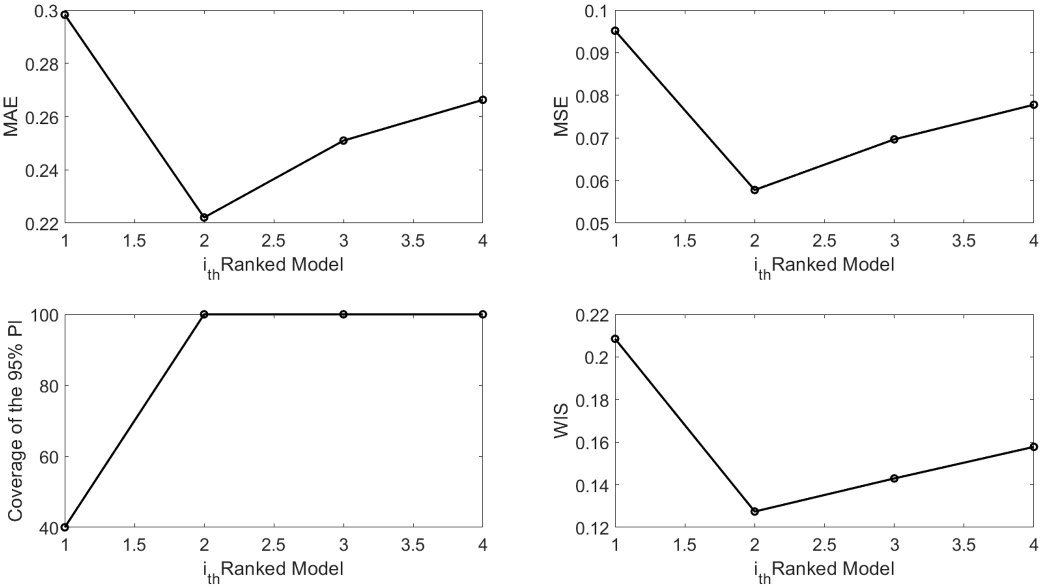


Figure 9: Performance metrics for 5-week forecasts based on the top-ranking sub-epidemic models for the Midwest weekly influenza cases.

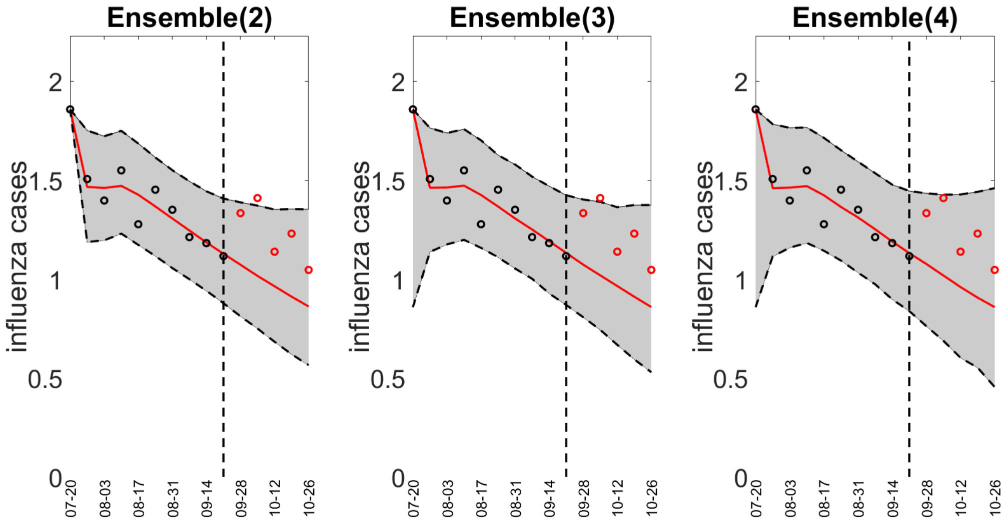


Figure 10: Displays 5-week sub-epidemic ensemble model forecasts for the Midwest weekly influenza cases. The vertical line marks the start time of the forecast. The circles represent the data points, with model fits shown as solid lines and 95% prediction intervals indicated by shaded areas.

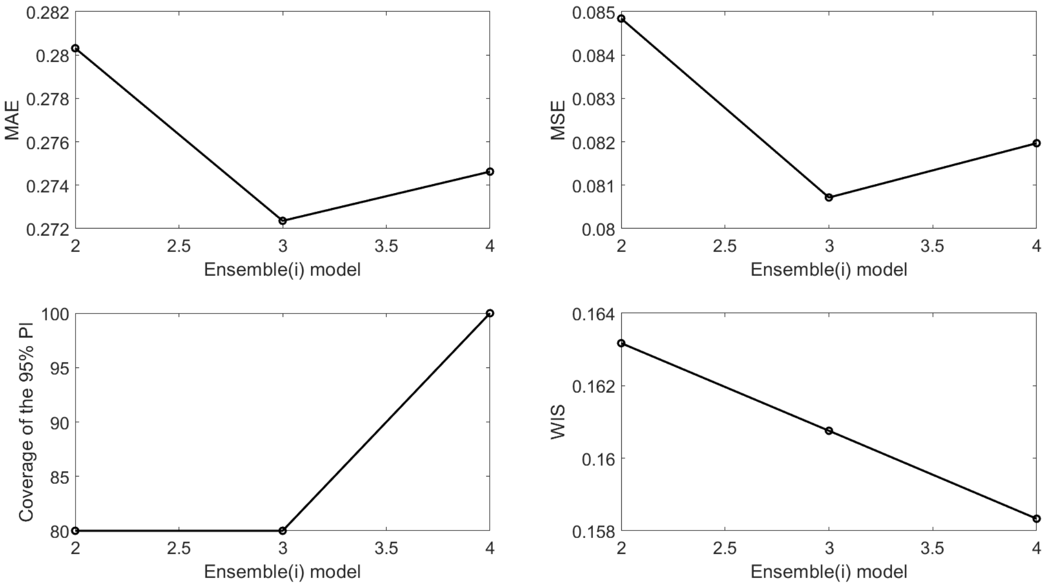


Figure 11: Performance metrics for 5-week forecasts based on the ensemble sub-epidemic models for the Midwest weekly influenza cases.

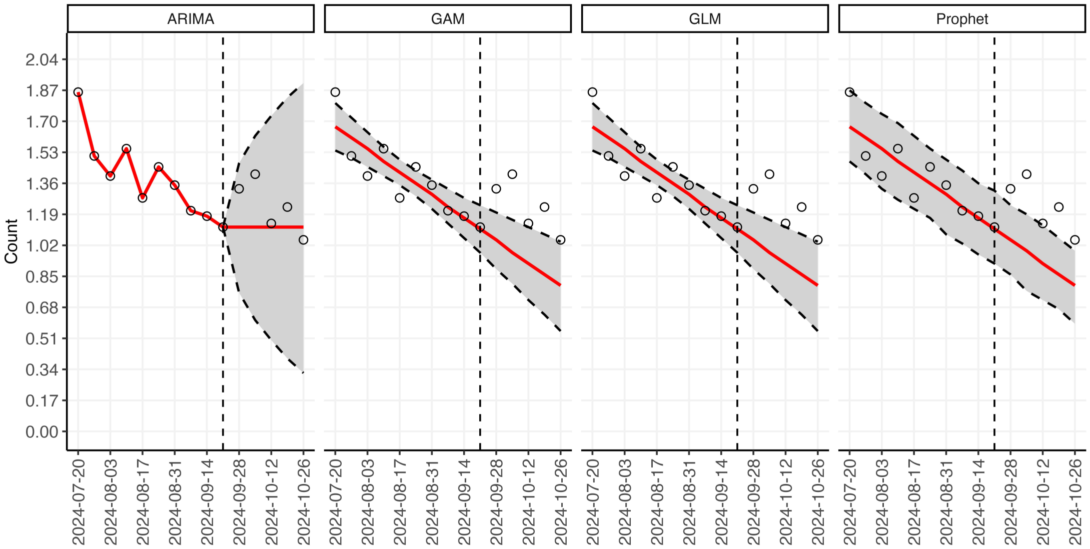
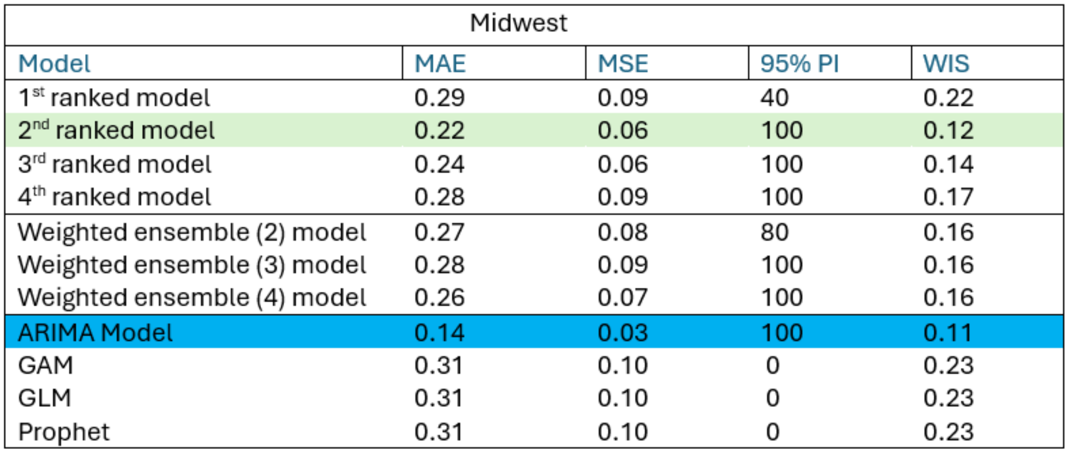


Figure 12: Shows 5-week forecasts from the ARIMA,GLM,GAM AND THE PROPHET models of the Midwest weekly influenza cases. The 95% prediction interval is represented by a shaded area around the solid line, with a vertical line marking the start of the forecast to separate the calibration and forecast periods. The circles indicate the observed data points.



Midwest				
Model	MAE	MSE	95% PI	WIS
1st ranked model	0.29	0.09	40	0.22
2nd ranked model	0.22	0.06	100	0.12
3rd ranked model	0.24	0.06	100	0.14
4th ranked model	0.28	0.09	100	0.17
Weighted ensemble (2) model	0.27	0.08	80	0.16
Weighted ensemble (3) model	0.28	0.09	100	0.16
Weighted ensemble (4) model	0.26	0.07	100	0.16
ARIMA Model	0.14	0.03	100	0.11
GAM	0.31	0.10	0	0.23
GLM	0.31	0.10	0	0.23
Prophet	0.31	0.10	0	0.23

Table 2: 5-weeks forecasting performance metrics for Midwest

The figures in **Table 2** compare the performance metrics of various models for forecasting

weekly influenza cases in the Midwest. The ARIMA model achieves the best accuracy with the lowest MAE (0.14) and MSE (0.03), reflecting superior precision. Both the ARIMA model and the 2nd ranked model achieve 100% coverage of the 95% prediction interval, demonstrating strong reliability, while other models show lower or no coverage. Finally, the ARIMA model has the lowest WIS (0.11), indicating superior calibration, closely followed by the 2nd ranked model (0.12).

3.3 Results for Northeast

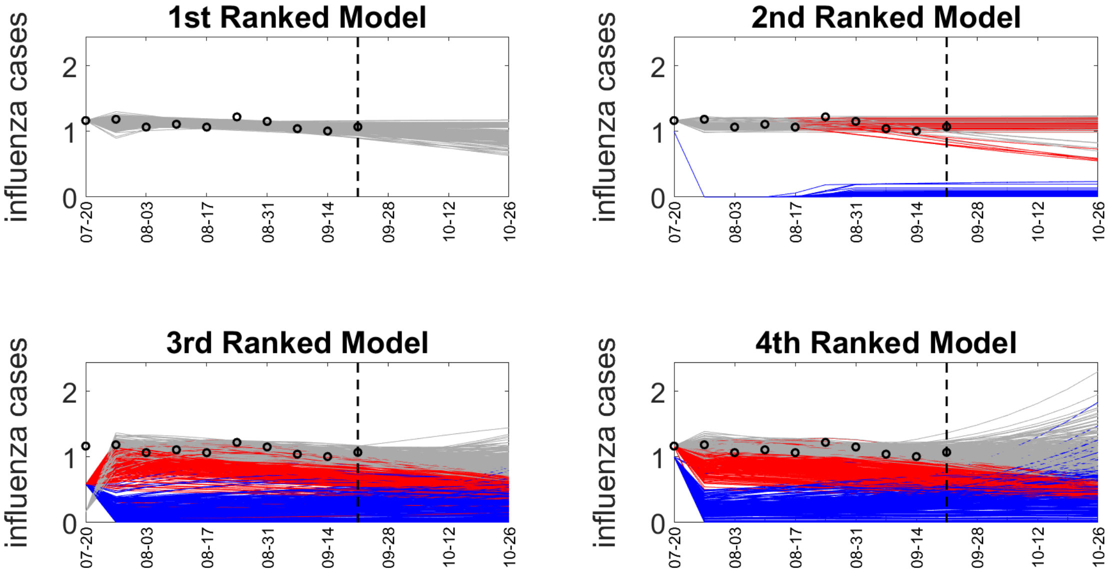


Figure 13: 5-week forecast sub-epidemic profiles derived from the top-ranking sub-epidemic models of the Northeast weekly influenza cases. The blue and red curves represent different sub-epidemic waves generated with 300 bootstrap realizations, while the gray curves show the overall epidemic trajectory. The vertical line marks the start of the forecast, separating the calibration and forecast periods.

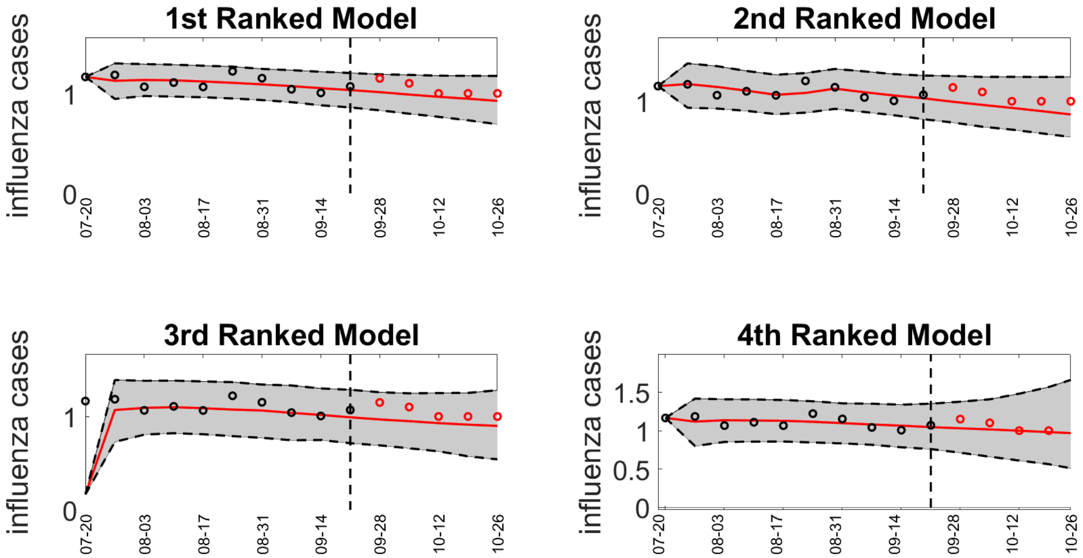


Figure 14: Shows 5-week forecasts from the top-ranking sub-epidemic models of the Northeast weekly influenza cases. The 95% prediction interval is represented by a shaded area around the solid line, with a vertical line marking the start of the forecast to separate the calibration and forecast periods. Circles indicate the observed data points.

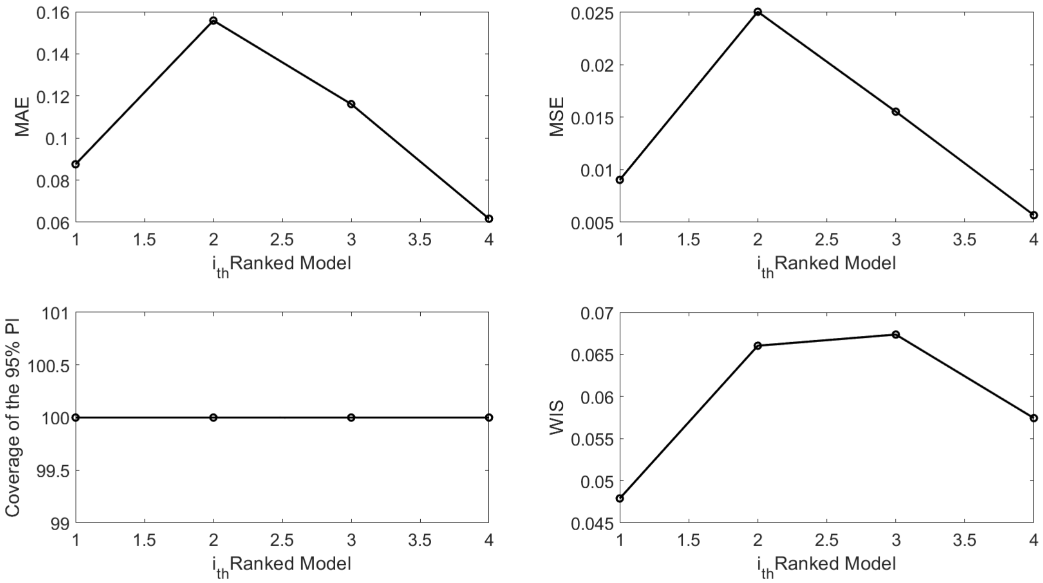


Figure 15: Performance metrics for 5-week forecasts based on the top-ranking sub-epidemic models for the Northeast weekly influenza cases.

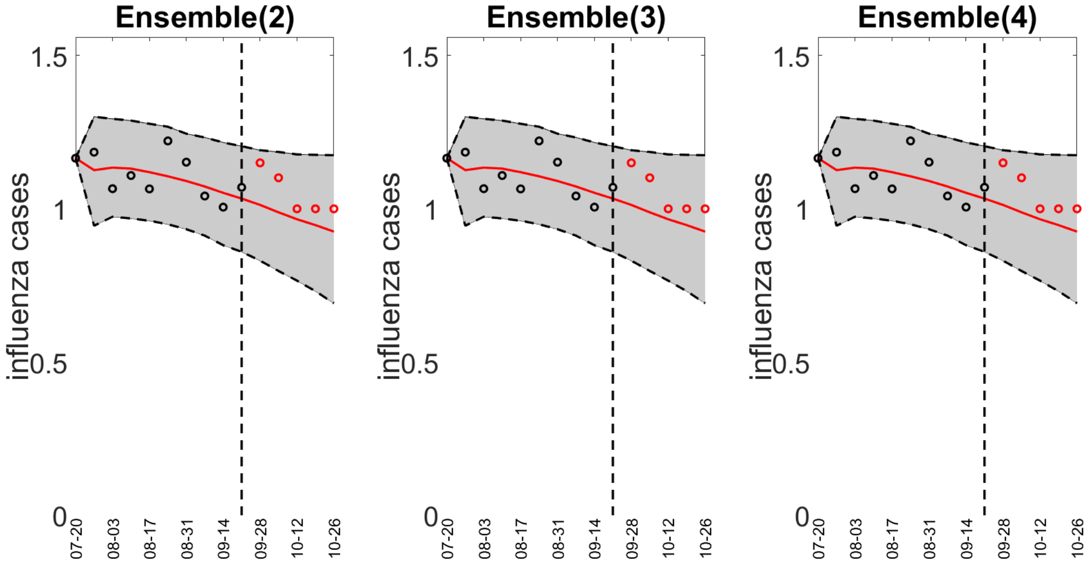


Figure 16: Displays 5-week sub-epidemic ensemble model forecasts for the Northeast weekly influenza cases. The vertical line marks the start time of the forecast. The circles represent the data points, with model fits shown as solid lines and 95% prediction intervals indicated by shaded areas.

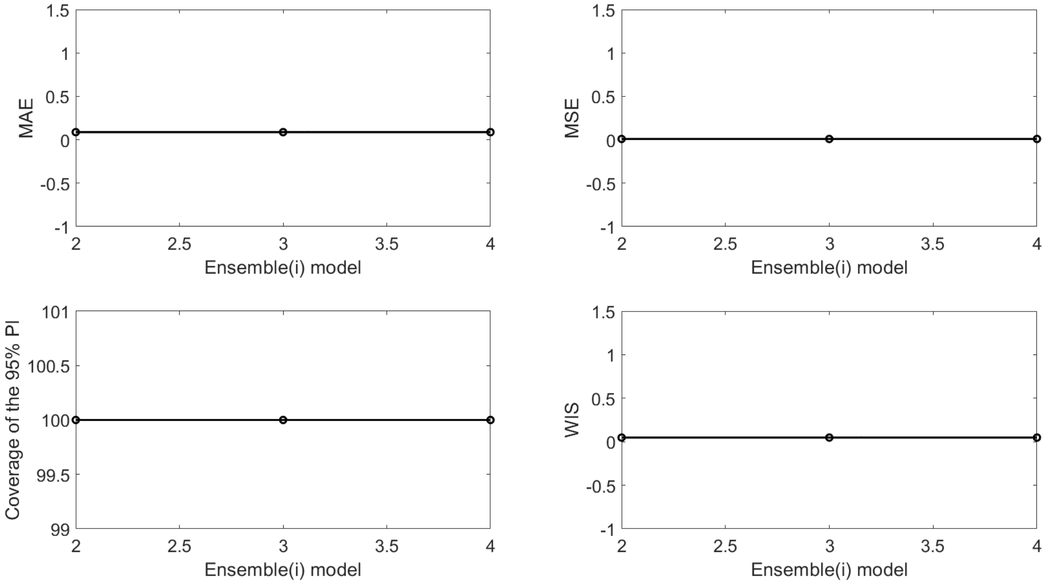


Figure 17: Performance metrics for 5-week forecasts based on the ensemble sub-epidemic models for the Northeast weekly influenza cases.

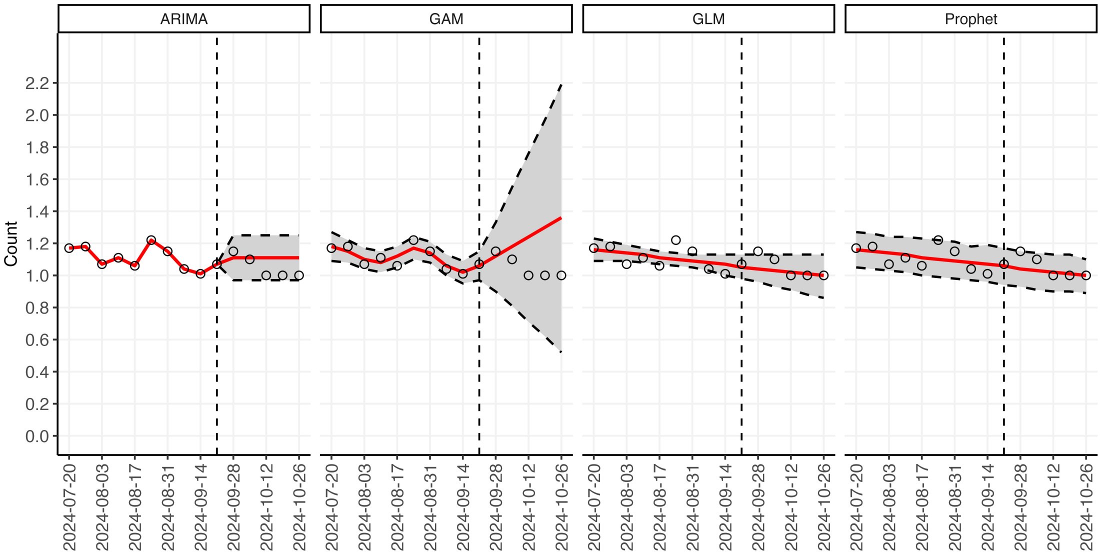
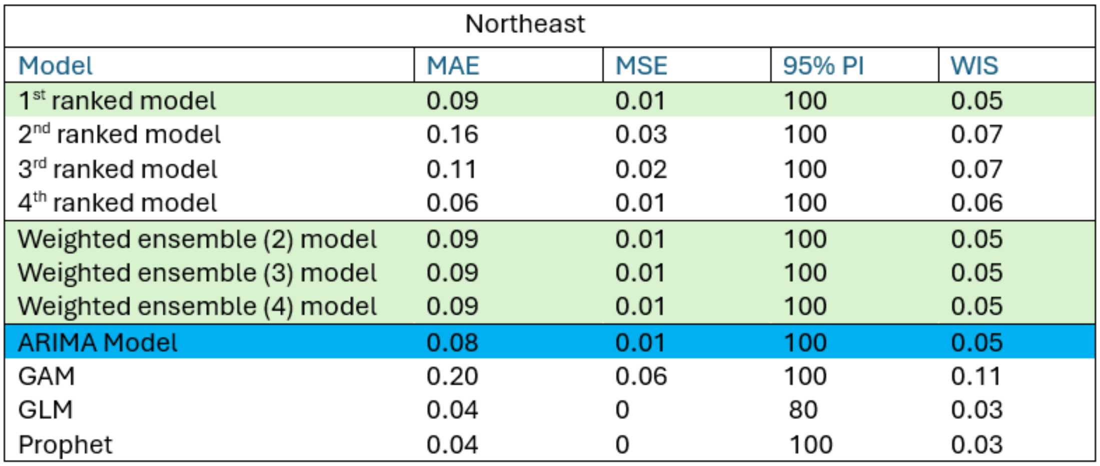


Figure 18: Shows 5-week forecasts from the ARIMA,GLM,GAM AND THE PROPHET models of the Northeast weekly influenza cases. The 95% prediction interval is represented by a shaded area around the solid line, with a vertical line marking the start of the forecast to separate the calibration and forecast periods. The circles indicate the observed data points.



Northeast				
Model	MAE	MSE	95% PI	WIS
1st ranked model	0.09	0.01	100	0.05
2nd ranked model	0.16	0.03	100	0.07
3rd ranked model	0.11	0.02	100	0.07
4th ranked model	0.06	0.01	100	0.06
Weighted ensemble (2) model	0.09	0.01	100	0.05
Weighted ensemble (3) model	0.09	0.01	100	0.05
Weighted ensemble (4) model	0.09	0.01	100	0.05
ARIMA Model	0.08	0.01	100	0.05
GAM	0.20	0.06	100	0.11
GLM	0.04	0	80	0.03
Prophet	0.04	0	100	0.03

Table 3: 5-weeks forecasting performance metrics for Northeast

The figures in **Table 3** compare the performance metrics of various models for forecasting

weekly influenza cases in the Northeast. The ARIMA model demonstrates the best overall performance with the lowest MAE (0.08) and MSE (0.01), reflecting exceptional accuracy and precision. It achieves 100% coverage of the 95% prediction interval, matching the reliability of other top-performing models like the 1st ranked model and weighted ensemble models. Additionally, the ARIMA model achieves one of the lowest WIS scores (0.05), indicating excellent calibration, comparable to the 1st ranked model and the weighted ensembles.

3.4 Results for South

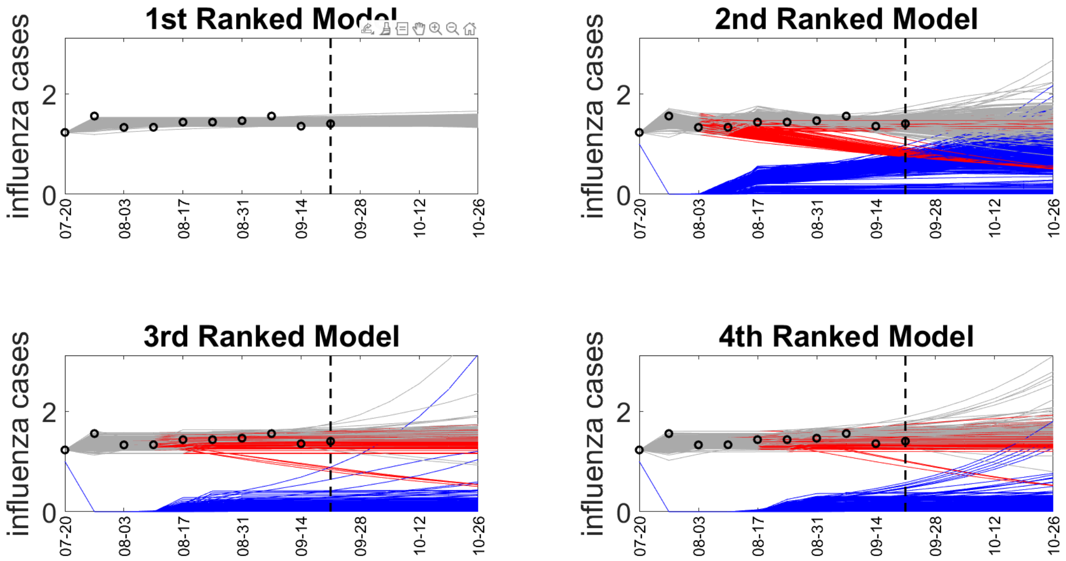


Figure 19: 5-week forecast sub-epidemic profiles derived from the top-ranking sub-epidemic models of the South weekly influenza cases. The blue and red curves represent different sub-epidemic waves generated with 300 bootstrap realizations, while the gray curves show the overall epidemic trajectory. The vertical line marks the start of the forecast, separating the calibration and forecast periods.

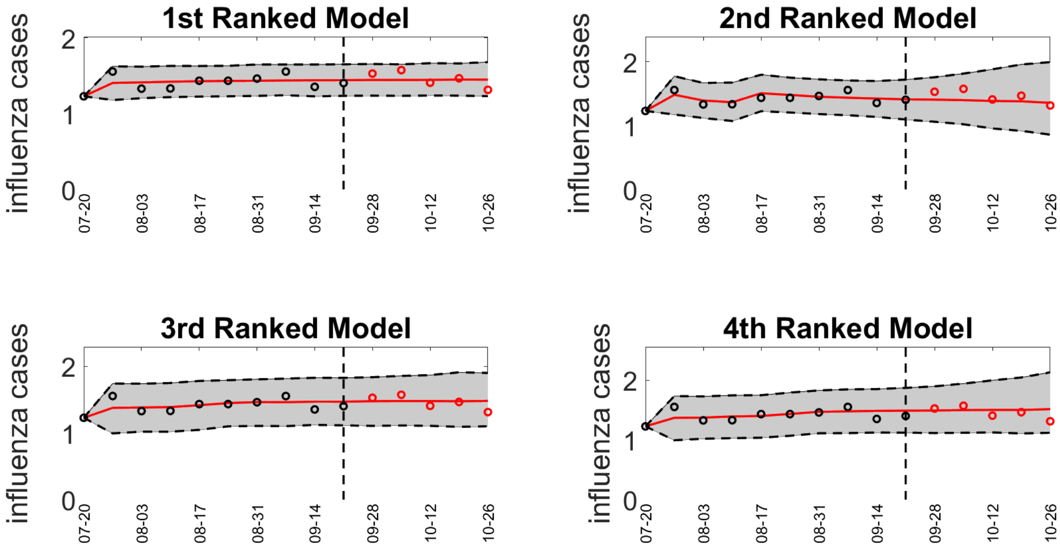


Figure 20: Shows 5-week forecasts from the top-ranking sub-epidemic models of the South weekly influenza cases. The 95% prediction interval is represented by a shaded area around the solid line, with a vertical line marking the start of the forecast to separate the calibration and forecast periods. Circles indicate the observed data points.

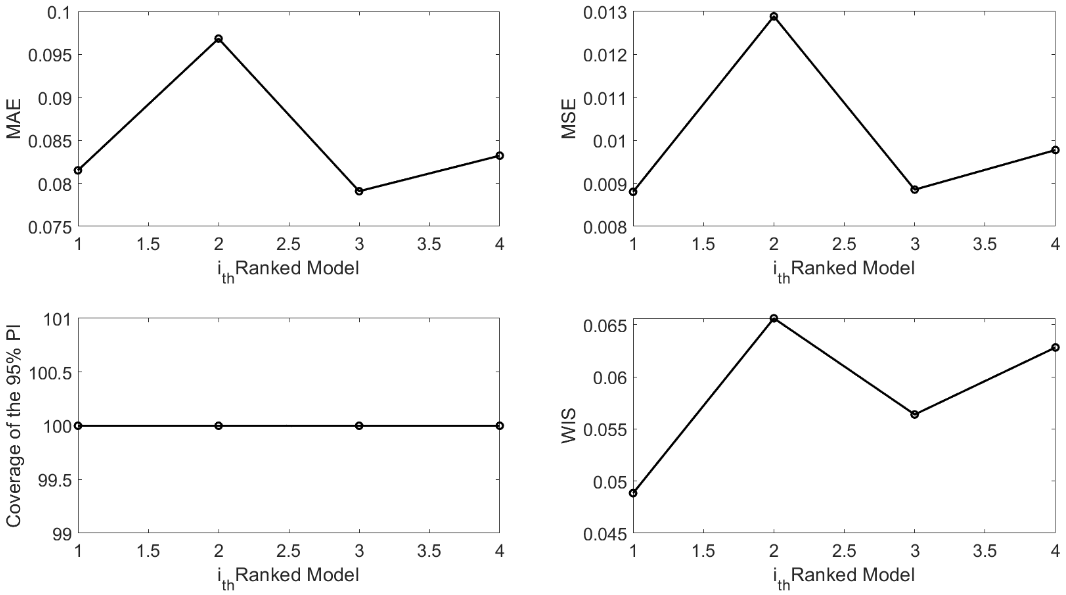


Figure 21: Performance metrics for 5-week forecasts based on the top-ranking sub-epidemic models for the South weekly influenza cases.

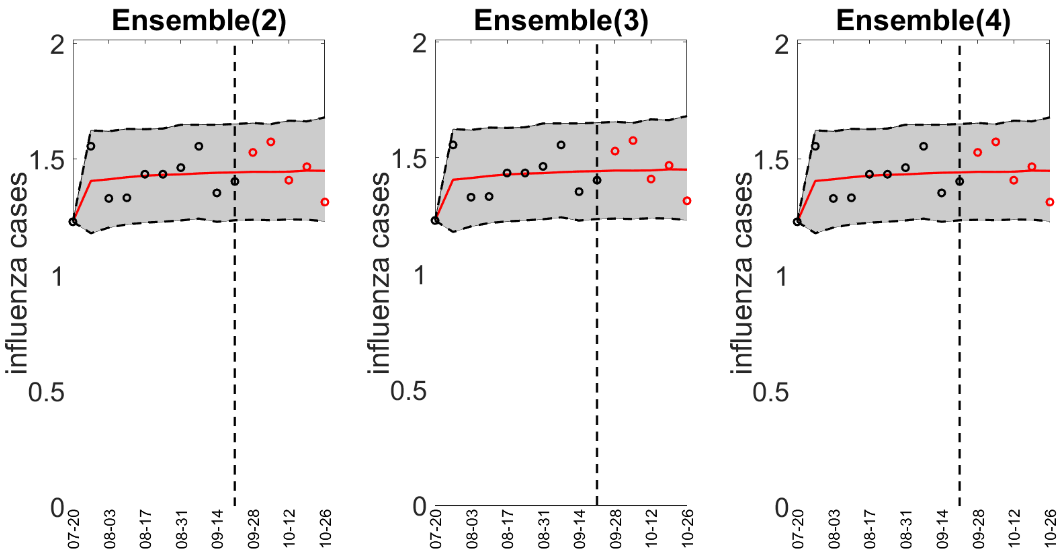


Figure 22: Displays 5-week sub-epidemic ensemble model forecasts for the South weekly influenza cases. The vertical line marks the start time of the forecast. The circles represent the data points, with model fits shown as solid lines and 95% prediction intervals indicated by shaded areas.

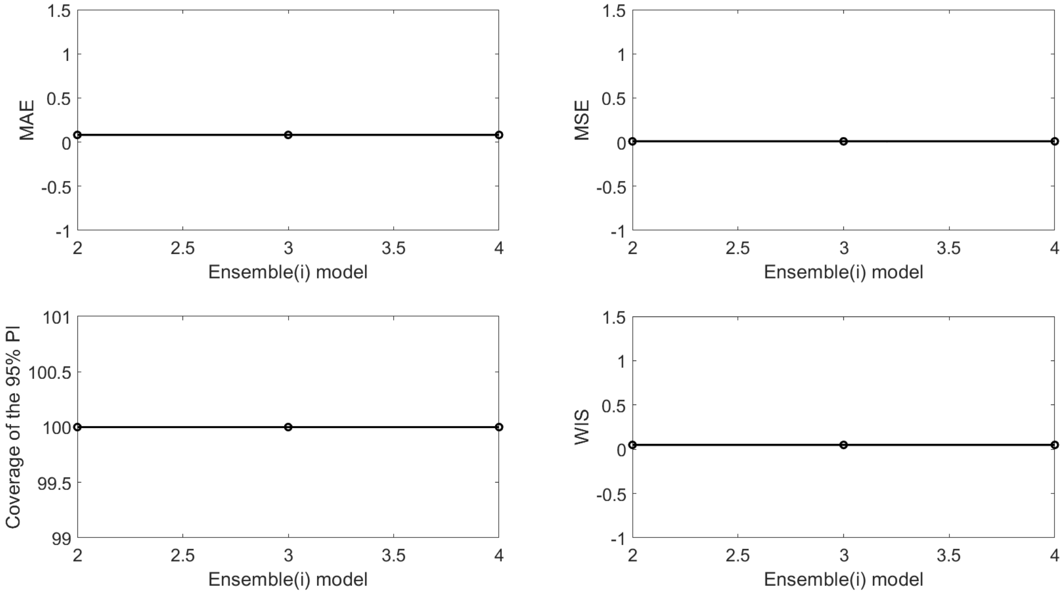


Figure 23: Performance metrics for 5-week forecasts based on the ensemble sub-epidemic models for the South weekly influenza cases.

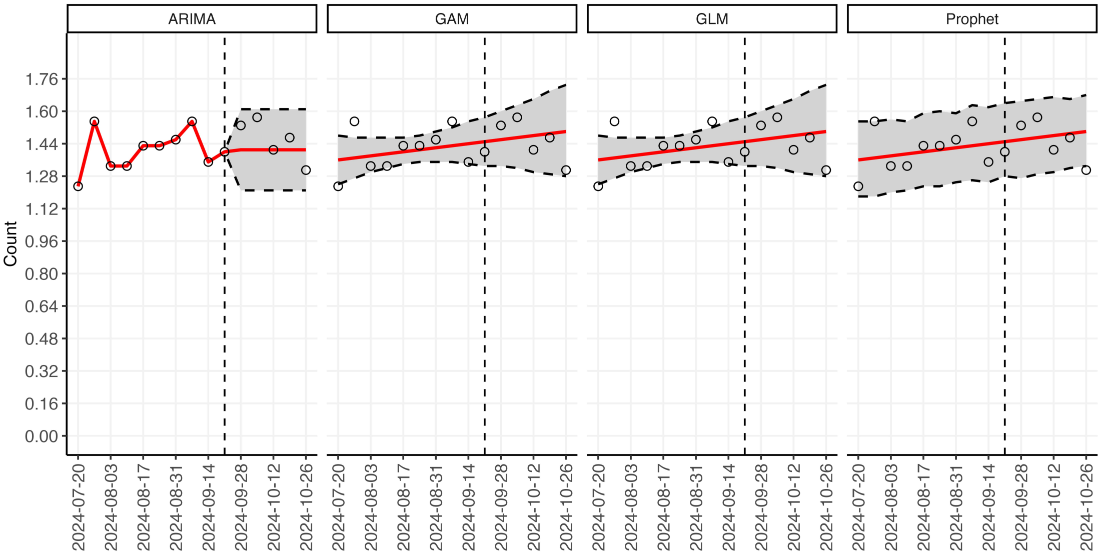
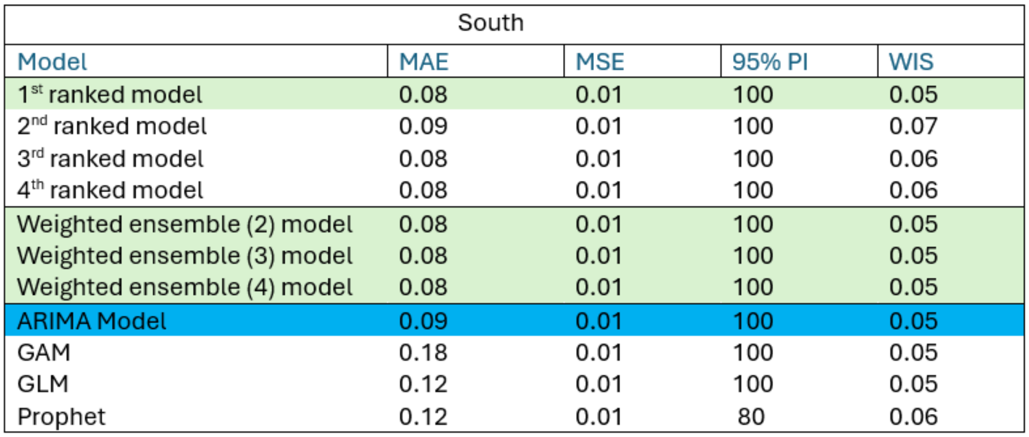


Figure 24: Shows 5-week forecasts from the ARIMA,GLM,GAM AND THE PROPHET models of the South weekly influenza cases. The 95% prediction interval is represented by a shaded area around the solid line, with a vertical line marking the start of the forecast to separate the calibration and forecast periods. The circles indicate the observed data points.



South				
Model	MAE	MSE	95% PI	WIS
1st ranked model	0.08	0.01	100	0.05
2nd ranked model	0.09	0.01	100	0.07
3rd ranked model	0.08	0.01	100	0.06
4th ranked model	0.08	0.01	100	0.06
Weighted ensemble (2) model	0.08	0.01	100	0.05
Weighted ensemble (3) model	0.08	0.01	100	0.05
Weighted ensemble (4) model	0.08	0.01	100	0.05
ARIMA Model	0.09	0.01	100	0.05
GAM	0.18	0.01	100	0.05
GLM	0.12	0.01	100	0.05
Prophet	0.12	0.01	80	0.06

Table 4: 5-weeks forecasting performance metrics for South

The figures in **Table 4** compare the performance metrics of various models for forecasting weekly influenza cases in the South. The 1st ranked model and the weighted ensemble models exhibit the best overall performance with the lowest MAE (0.08), MSE (0.01), and WIS (0.05), reflecting exceptional accuracy, precision, and calibration. These models also achieve 100% coverage of the 95% prediction interval, demonstrating strong reliability. The ARIMA model closely follows, with a slightly higher MAE (0.09) but identical MSE, 95% PI coverage, and WIS, maintaining competitive performance.

3.5 Results for West

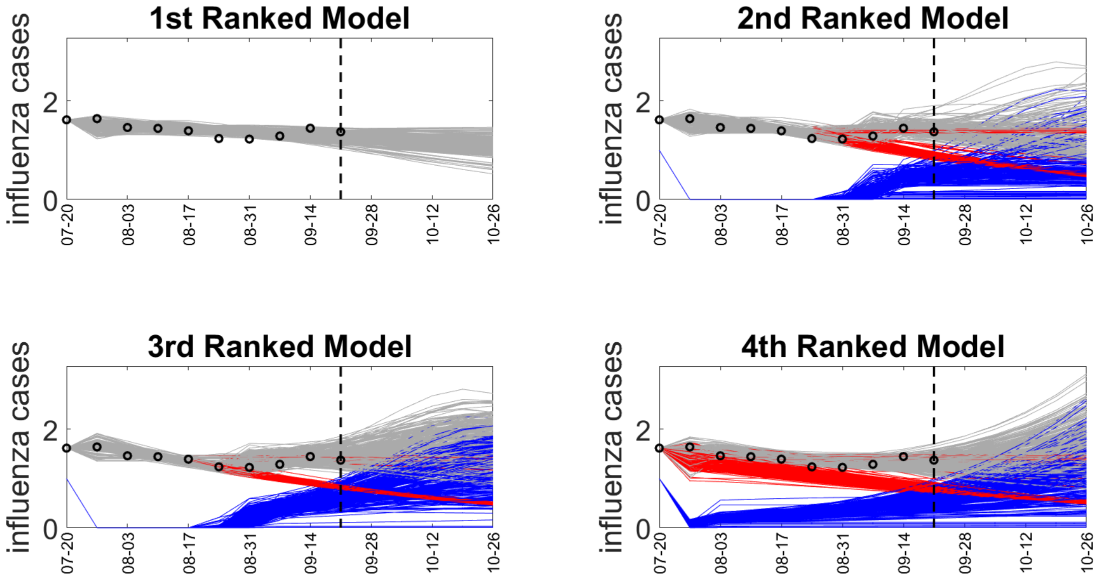


Figure 25: 5-week forecast sub-epidemic profiles derived from the top-ranking sub-epidemic models of the West weekly influenza cases. The blue and red curves represent different sub-epidemic waves generated with 300 bootstrap realizations, while the gray curves show the overall epidemic trajectory. The vertical line marks the start of the forecast, separating the calibration and forecast periods.

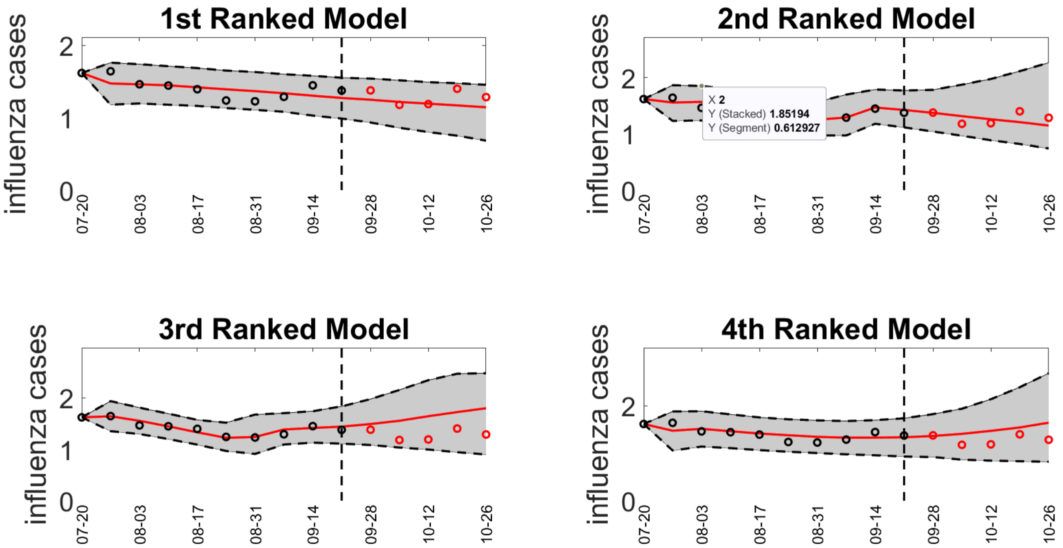


Figure 26: Shows 5-week forecasts from the top-ranking sub-epidemic models of the West weekly influenza cases. The 95% prediction interval is represented by a shaded area around the solid line, with a vertical line marking the start of the forecast to separate the calibration and forecast periods. Circles indicate the observed data points.

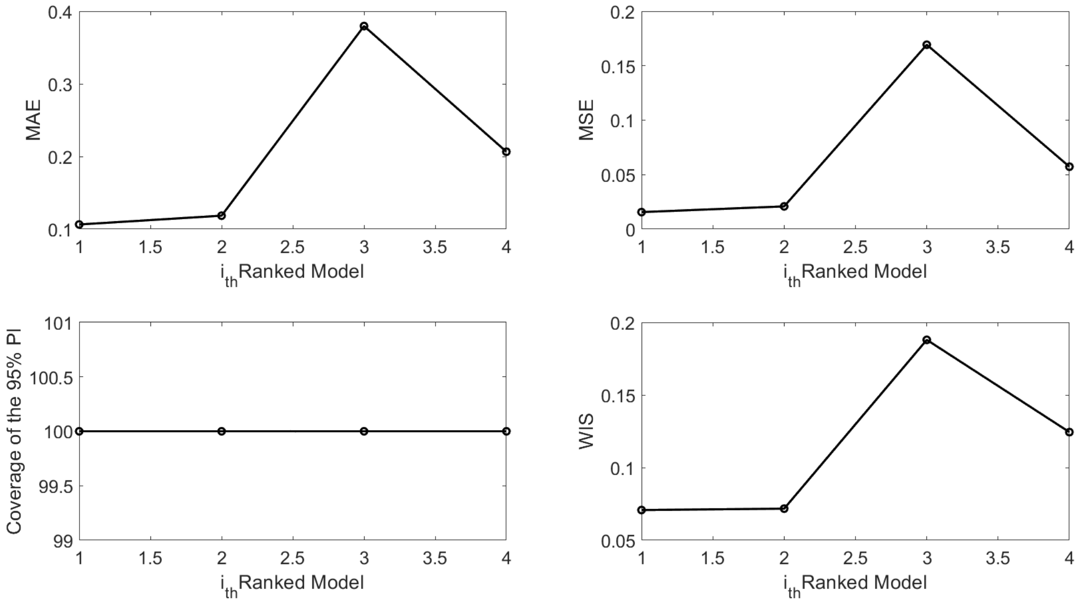


Figure 27: Performance metrics for 5-week forecasts based on the top-ranking sub-epidemic models for the West weekly influenza cases.

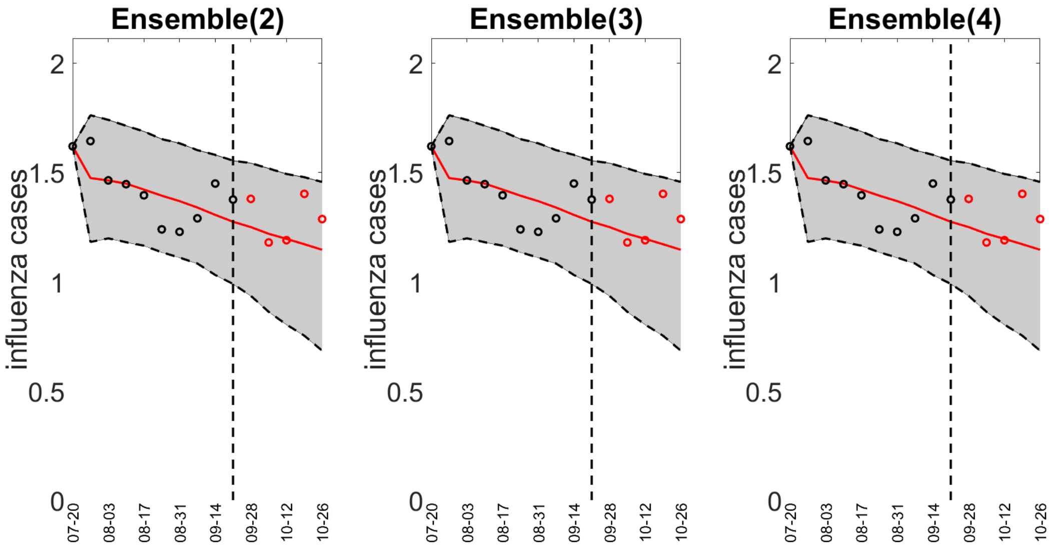


Figure 28: Displays 5-week sub-epidemic ensemble model forecasts for the West weekly influenza cases. The vertical line marks the start time of the forecast. The circles represent the data points, with model fits shown as solid lines and 95% prediction intervals indicated by shaded areas.

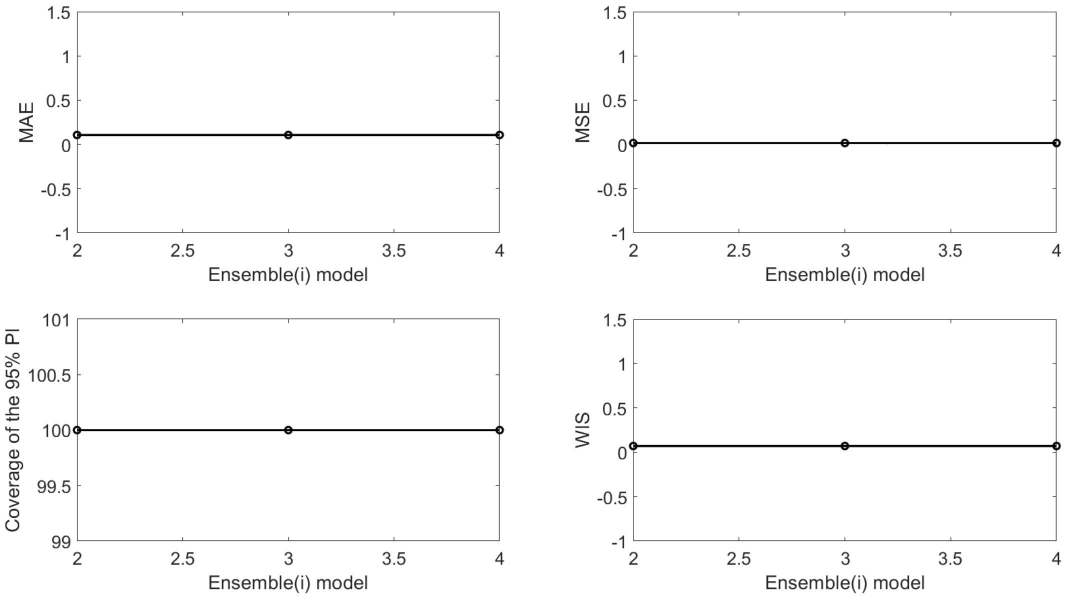


Figure 29: Performance metrics for 5-week forecasts based on the ensemble sub-epidemic models for the West weekly influenza cases.

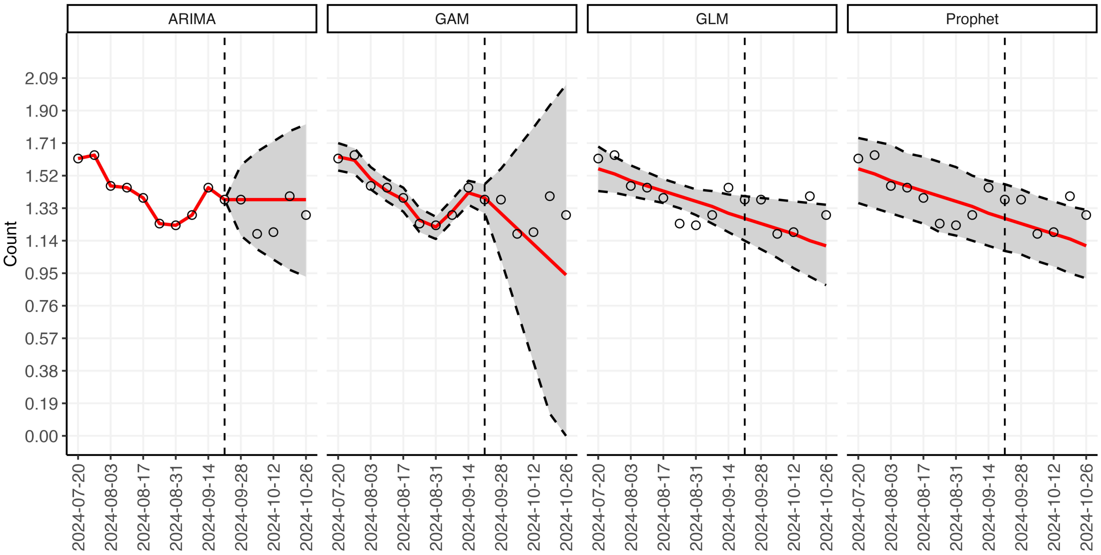
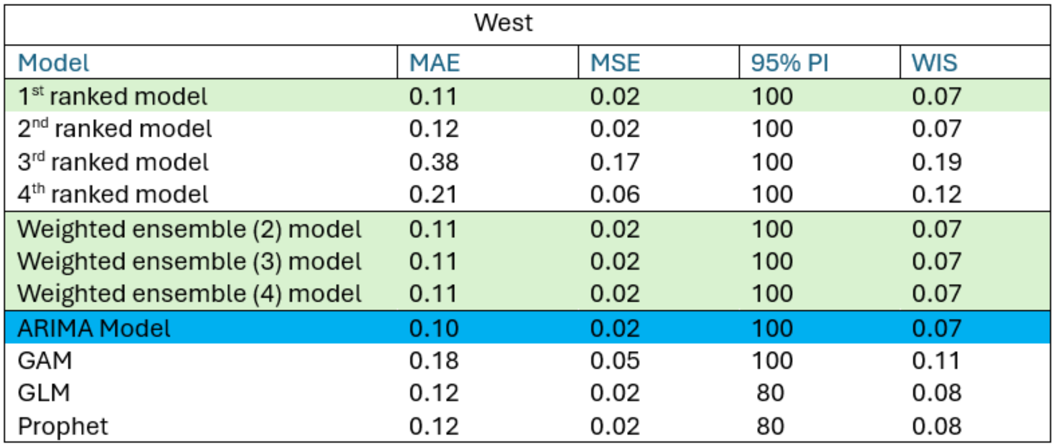


Figure 30: Shows 5-week forecasts from the ARIMA,GLM,GAM AND THE PROPHET models of the West weekly influenza cases. The 95% prediction interval is represented by a shaded area around the solid line, with a vertical line marking the start of the forecast to separate the calibration and forecast periods. The circles indicate the observed data points.



West				
Model	MAE	MSE	95% PI	WIS
1st ranked model	0.11	0.02	100	0.07
2nd ranked model	0.12	0.02	100	0.07
3rd ranked model	0.38	0.17	100	0.19
4th ranked model	0.21	0.06	100	0.12
Weighted ensemble (2) model	0.11	0.02	100	0.07
Weighted ensemble (3) model	0.11	0.02	100	0.07
Weighted ensemble (4) model	0.11	0.02	100	0.07
ARIMA Model	0.10	0.02	100	0.07
GAM	0.18	0.05	100	0.11
GLM	0.12	0.02	80	0.08
Prophet	0.12	0.02	80	0.08

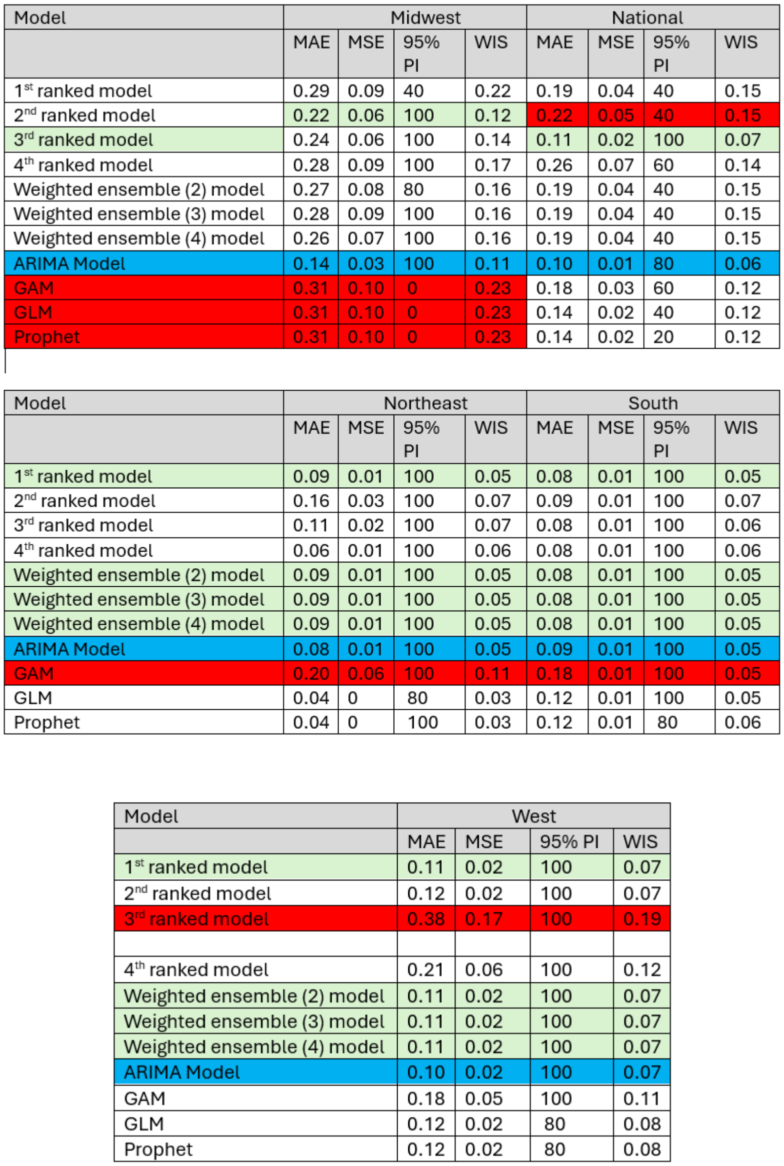
Table 5: 5-weeks forecasting performance metrics for West

The figures in **Table 5** compare the performance metrics of various models for forecasting

weekly influenza cases in the West. The ARIMA model demonstrates the best overall performance with the lowest MAE (0.10) and competitive MSE (0.02), along with 10% coverage of the 95% prediction interval, indicating strong accuracy and reliability. The 1st ranked model and weighted ensemble models also show strong performance, with identical MAE (0.11), MSE (0.02), and WIS (0.07), and full prediction interval coverage. Other models, such as GAM, GLM, and Prophet, perform slightly worse, with higher error metrics and less reliable prediction intervals.

4 Discussion

This section presents a comprehensive comparison of the model performance metrics, summarizing the evaluation of each model's forecasting accuracy and reliability based on the results presented in **Table 6** below.



Model	Midwest				National			
	MAE	MSE	95% PI	WIS	MAE	MSE	95% PI	WIS
1st ranked model	0.29	0.09	40	0.22	0.19	0.04	40	0.15
2nd ranked model	0.22	0.06	100	0.12	0.22	0.05	40	0.15
3rd ranked model	0.24	0.06	100	0.14	0.11	0.02	100	0.07
4th ranked model	0.28	0.09	100	0.17	0.26	0.07	60	0.14
Weighted ensemble (2) model	0.27	0.08	80	0.16	0.19	0.04	40	0.15
Weighted ensemble (3) model	0.28	0.09	100	0.16	0.19	0.04	40	0.15
Weighted ensemble (4) model	0.26	0.07	100	0.16	0.19	0.04	40	0.15
ARIMA Model	0.14	0.03	100	0.11	0.10	0.01	80	0.06
GAM	0.31	0.10	0	0.23	0.18	0.03	60	0.12
GLM	0.31	0.10	0	0.23	0.14	0.02	40	0.12
Prophet	0.31	0.10	0	0.23	0.14	0.02	20	0.12

Model	Northeast				South			
	MAE	MSE	95% PI	WIS	MAE	MSE	95% PI	WIS
1st ranked model	0.09	0.01	100	0.05	0.08	0.01	100	0.05
2nd ranked model	0.16	0.03	100	0.07	0.09	0.01	100	0.07
3rd ranked model	0.11	0.02	100	0.07	0.08	0.01	100	0.06
4th ranked model	0.06	0.01	100	0.06	0.08	0.01	100	0.06
Weighted ensemble (2) model	0.09	0.01	100	0.05	0.08	0.01	100	0.05
Weighted ensemble (3) model	0.09	0.01	100	0.05	0.08	0.01	100	0.05
Weighted ensemble (4) model	0.09	0.01	100	0.05	0.08	0.01	100	0.05
ARIMA Model	0.08	0.01	100	0.05	0.09	0.01	100	0.05
GAM	0.20	0.06	100	0.11	0.18	0.01	100	0.05
GLM	0.04	0	80	0.03	0.12	0.01	100	0.05
Prophet	0.04	0	100	0.03	0.12	0.01	80	0.06

Model	West			
	MAE	MSE	95% PI	WIS
1st ranked model	0.11	0.02	100	0.07
2nd ranked model	0.12	0.02	100	0.07
3rd ranked model	0.38	0.17	100	0.19
4th ranked model	0.21	0.06	100	0.12
Weighted ensemble (2) model	0.11	0.02	100	0.07
Weighted ensemble (3) model	0.11	0.02	100	0.07
Weighted ensemble (4) model	0.11	0.02	100	0.07
ARIMA Model	0.10	0.02	100	0.07
GAM	0.18	0.05	100	0.11
GLM	0.12	0.02	80	0.08
Prophet	0.12	0.02	80	0.08

Table 6: Comparison of Calibration Performance based on MAE, MSE, 95% PI, and WIS

4.1 Model Performance Discussion

Best Model Across Regions: The **ARIMA model** consistently emerges as the best performer across all regions:

- **Midwest:** It achieves the lowest MAE (0.14), MSE (0.03), and WIS (0.11) while providing 100% coverage of the 95% prediction interval, indicating excellent accuracy, reliability, and calibration.

- **National:** ARIMA excels with the lowest MAE (0.10), MSE (0.01), and WIS (0.06) and maintains 80% prediction interval coverage, slightly less than some other models but still effective.

- **Northeast:** ARIMA ties with several weighted ensemble models and the 1st ranked model in achieving the lowest WIS (0.05), 100% 95% PI coverage, and a competitive MAE (0.08), showcasing consistent reliability and precision.

- **South:** Though the ARIMA model's MAE (0.09) is slightly higher than the 1st ranked model (0.08), its MSE (0.01), WIS (0.05), and 100% prediction interval coverage keep it on par with top performers.

- **West:** ARIMA is the standout performer with the lowest MAE (0.10), excellent MSE (0.02), and WIS (0.07), tied with the 1st ranked model and weighted ensembles. Its 100% coverage of the 95% PI confirms its reliability.

Close Contenders:

- **Weighted Ensemble Models (2, 3, 4):** These models consistently perform close to ARIMA in most regions, particularly in terms of MSE and WIS.

  - In the **Northeast** and **South**, their MAE (0.08–0.09), MSE (0.01), and WIS (0.05) match the ARIMA model, showcasing their robustness.
  - In the **West**, they share the same strong performance metrics as ARIMA (MAE 0.11, MSE 0.02, WIS 0.07, and 100% PI coverage).

- **1st Ranked Model:** Performs well in specific regions:
  - In the **South**, it matches or outperforms ARIMA with the lowest MAE (0.08), MSE (0.01), and WIS (0.05), along with 100% PI coverage.
  - In the **Northeast**, its performance is comparable to ARIMA with similar metrics (MAE 0.09, MSE 0.01, WIS 0.05, and 100% PI coverage).
  - However, it underperforms slightly in the **Midwest** and **National** settings with higher MAE and WIS values compared to ARIMA.

- **2nd Ranked Model:** It is competitive in the **Midwest**, achieving low MAE (0.22) and WIS (0.12) with 100% PI coverage. In the **West**, it delivers solid performance but is outperformed by ARIMA and the weighted ensembles.

Underperforming Models:

- **3rd Ranked Model:** Performs poorly in most regions except the **Midwest** (MAE 0.24, WIS 0.14) and **National** (MAE 0.11, WIS 0.07). Particularly in the **West**, it has the highest MAE (0.38), MSE (0.17), and WIS (0.19), making it less effective model.

- **4th Ranked Model:** Performs moderately across regions, achieving reasonable 95% PI coverage in the Northeast (100%), but with relatively high MAE (0.28) and WIS (0.17) in the Midwest. While it is not the worst-performing model, its higher error metrics make it less effective compared to the top-ranked models.

- **GAM, GLM, Prophet:** Across all regions, these models consistently show higher MAE, MSE, and WIS values compared to ARIMA and other top models.
  - In the **Midwest**, GAM and GLM have a high MAE (0.31) and WIS (0.23), indicating poor accuracy and calibration.
  - In the **South**, GLM and Prophet slightly improve (MAE 0.12), but their 95% PI coverage is lower (80% for Prophet) or non-existent (GLM), reducing reliability.
  - In the **West**, GAM performs poorly with a high WIS (0.11), while GLM and Prophet offer slightly better performance but remain behind ARIMA and the weighted ensembles.

Hence from the above analysis, we observe the following:

- **Best Model Overall:** The ARIMA model consistently leads in accuracy, precision, and reliability across regions.

- **Strong Contenders:** Weighted ensemble models and the 1st ranked model often rival ARIMA, particularly in the **Northeast** and **South**.

- **Weak Performers:** The 3rd ranked model underperforms significantly in the **West**, while GAM, GLM, and Prophet are less reliable and accurate across all regions.

4.2 Insights and Public Health Implications

The detailed analysis of the *n-sub-epidemic* framework and the *STAT MOD PREDICT* toolbox models provides direct insights into how these tools can influence public health decisions. By comparing model performance across metrics such as MAE, MSE, 95% prediction interval (PI) coverage, and WIS, the analysis demonstrates the practical applications of these frameworks in improving epidemic forecasting and guiding effective public health responses.

High-performing models, such as the ARIMA model and Weighted Ensembles, support early detection of outbreaks, efficient resource allocation, and targeted interventions. The ability to capture complex epidemic dynamics ensures reliable forecasts that inform public health strategies, enabling authorities to respond proactively and minimize the impact of infectious diseases.

4.2.1 Limitations

- **Data Quality and Availability:** Influenza case data relies on reporting systems that may have inconsistencies, delays, or under reporting, potentially affecting the accuracy of forecasts. Similarly, wastewater data may vary in sampling frequency, data collection methods, and regional coverage, limiting the generalizability of the findings.
- **Temporal Scope:** The study's time frame (26 weeks) limits its ability to assess long-term trends, seasonal variations, or the impact of evolving public health measures.

4.2.2 Future Work

- **Extended Time Periods:** Future studies should consider extended time periods to capture a broader range of influenza dynamics, including seasonal and inter-annual variations.
- **Error Structures:** This analysis only utilized a normal error structure. Future work will explore alternative error structures, such as Poisson and negative binomial errors, to better model the variability and distribution of influenza case data.

4.3 Conclusion

This study explored advanced predictive models using the *n-sub epidemic model* and the *STAT MOD PREDICT toolbox* to improve influenza forecasting across the United States. The analysis identified the *ARIMA model* as the most effective, consistently achieving the lowest error metrics and reliable prediction intervals, particularly at the national and regional levels. Models within the *n-sub epidemic model* framework, such as Weighted Ensembles, also performed well, excelling in regions with complex epidemic dynamics.

The findings emphasize the importance of accurate forecasting for public health decision-making, supporting early interventions, efficient resource allocation, and tailored regional strategies. By combining robust statistical modeling with innovative surveillance techniques, this work offers a reliable framework for mitigating influenza's public health impact.

References

[1] Centers for Disease Control and Prevention, “Key Facts About Influenza (Flu),” Available: https://www.cdc.gov/flu/about/keyfacts.htm, Accessed: Apr. 2025.

[2] J. I. Tokars, S. J. Olsen, and C. Reed, “Seasonal Incidence of Symptomatic Influenza in the United States,” *Clinical Infectious Diseases*, vol. 66, pp. 1511–1518, 2018. doi: https://doi.org/10.1093/cid/cix1060

[3] Centers for Disease Control and Prevention, “Overview of Influenza Surveillance in the United States,” Available: https://www.cdc.gov/flu/weekly/overview.htm, Accessed: Apr. 2025.

[4] Centers for Disease Control and Prevention, ”Influenza Surveillance Reports,” Available: https://www.cdc.gov/flu/weekly/overview.htm, Accessed: Nov. 2024.

[5] Centers for Disease Control and Prevention, ”National Wastewater Surveillance System: Data Methods,” Available: https://www.cdc.gov/nwss/about-data.html#data-method, Accessed: Nov. 2024.

[6] Centers for Disease Control and Prevention, “CDC FluSight: Predict the Influenza Season Challenge,” Available: https://archive.cdc.gov/#/details?url=https://www.cdc.gov/flu/news/predict-flu-challenge.htm, Accessed: Apr. 2025.

[7] A. M. Bleich, ”STAT MOD PREDICT Toolbox,” Available: https://github.com/bleicham/StatModPredict, Accessed: Nov. 2024.

[8] G. Chowell, ”Ensemble n-Subepidemic Framework for Forecasting Epidemics,” Available: https://github.com/gchowell/ensemble_n-subepidemic_framework, Accessed: Nov. 2024.

[9] World Health Organization, “An mpox resurgence in the European region this spring and summer? To prevent that, key measures must continue,” *News*, May 17, 2023. Available: https://www.who.int/europe/news/item/17-05-2023-an-mpox-resurgence-in-the-european-region-this-spring-and-s Accessed: Apr. 2025.

[10] G. Chowell, A. Bleichrodt, and R. Luo, ”Parameter Estimation and Forecasting with Quantified Uncertainty for Ordinary Differential Equation Models Using QuantDiffForecast: A MATLAB Toolbox and Tutorial,” *Statistics in Medicine*, vol. 43, pp. 1826–1848, 2024. DOI: https://doi.org/10.1002/sim.10036.

[11] G. Chowell, A. Bleichrodt, S. Dahal et al., ”GrowthPredict: A Toolbox for Fitting and Forecasting Growth Trajectories Using Phenomenological Models,” *Scientific Reports*, vol. 14, no. 1630, 2024. DOI: https://doi.org/10.1038/s41598-024-51852-8.

[12] G. Chowell, S. Dahal, A. Bleichrodt et al., ”SubEpiPredict: A Toolbox for Fitting and Forecasting Growth Trajectories Using the Ensemble n-Sub-epidemic Framework,” *Infectious Disease Modelling*, vol. 9, pp. 411-436, 2024. DOI: https://doi.org/10.1016/j.idm.2024.02.001.

[13] A. Shafi, “What are generalized additive models?,” *Medium*, 2021. Available: https://towardsdatascience.com/generalised-additive-models-6dfbedf1350a, Accessed: Apr. 2025.

[14] T. Dimri, S. Ahmad, and M. Sharif, “Time series analysis of climate variables using seasonal ARIMA approach,” *Journal of Earth System Science*, vol. 129, article 149, 2020. doi: `https://doi.org/10.1007/s12040-020-01408-x`, Accessed: Apr. 2025.

[15] A. Bleichrodt, A. Phan, R. Luo, A. Kirpich, and G. Chowell-Puente, “StatModPredict: A user-friendly R-shiny interface for fitting and forecasting with statistical models,” *SSRN*, 2024. doi: `https://doi.org/10.2139/ssrn.4849702`, Accessed: Apr. 2025.

[16] M. Tektaş, “Weather forecasting using ANFIS and ARIMA models: case study for Istanbul,” *Environmental Research, Engineering and Management (EREM)*, vol. 51, pp. 5–10, 2010. doi: `https://doi.org/10.5755/j01.erem.51.1.58`, Accessed: Apr. 2025.

[17] R. Hyndman, “auto.ARIMA: Fit best ARIMA model to univariate time series,” *RDocumentation*, Available: `https://www.rdocumentation.org/packages/forecast/versions/8.21.1/topics/auto.arima`, Accessed: Apr. 2025.

[18] S. J. Taylor and B. Letham, “Forecasting at scale,” *The American Statistician*, vol. 72, no. 1, pp. 37–45, 2018. doi: `https://doi.org/10.1080/00031305.2017.1380080`

[19] A. Kirpich, A. Shishkin, A. Thomas, A. Weppelmann, P. Perez Tchernov, and Y. Skums, “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*, vol. 12, p. 5475, 2022. doi: `https://doi.org/10.1038/s41598-022-09345-z`

[20] S. Taylor, “prophet: Prophet forecaster,” *RDocumentation*, Available: `https://www.rdocumentation.org/packages/prophet/versions/1.0/topics/prophet`, Accessed: Apr. 2025.