ROYAL SOCIETY OPEN SCIENCE

(cc) BY

Research

Author for correspondence:
Amanda Bleichrodt
e-mail: ableichrodt1@student.gsu.edu

THE ROYAL SOCIETY PUBLISHING

Evaluating the forecasting performance of ensemble sub-epidemic frameworks and other time series models for the 2022–2023 mpox epidemic

Amanda Bleichrodt, Ruiyan Luo, Alexander Kirpich and Gerardo Chowell

Department of Population Health Sciences, School of Public Health, Georgia State University, Atlanta, GA, USA

AB, 0009-0007-2524-369X; GC, 0000-0003-2194-2251

During the 2022–2023 unprecedented mpox epidemic, near real-time short-term forecasts of the epidemic's trajectory were essential in intervention implementation and guiding policy. However, as case levels have significantly decreased, evaluating model performance is vital to advancing the field of epidemic forecasting. Using laboratory-confirmed mpox case data from the Centers for Disease Control and Prevention and Our World in Data teams, we generated retrospective sequential weekly forecasts for Brazil, Canada, France, Germany, Spain, the United Kingdom, the United States and at the global scale using an auto-regressive integrated moving average (ARIMA) model, generalized additive model, simple linear regression, Facebook's Prophet model, as well as the sub-epidemic wave and *n*-sub-epidemic modelling frameworks. We assessed forecast performance using average mean squared error, mean absolute error, weighted interval scores, 95% prediction interval coverage, skill scores and Winkler scores. Overall, the *n*-sub-epidemic modelling framework outcompeted other models across most locations and forecasting horizons, with the unweighted ensemble model performing best most frequently. The *n*-sub-epidemic and spatial-wave frameworks considerably improved in average forecasting performance relative to the ARIMA model (greater than 10%) for all performance metrics. Findings further support sub-epidemic frameworks for short-term forecasting epidemics of emerging and re-emerging infectious diseases.

1. Introduction

In May 2022, public health officials noted an unprecedented global surge in mpox (formally monkeypox) cases in multiple countries previously free of the disease [1]. Early on, little was known regarding critical transmission and control factors—vaccination, the role of asymptomatic individuals and transmission pathways—which shaped the variable epidemic trajectories noted in the over 111 impacted countries [2–4]. Therefore, as mpox cases increased in late July 2022, our team started to produce and make publicly available weekly short-term forecasts of new cases for the highest-affected countries (e.g. Brazil, Canada, France, Germany, Spain, the United Kingdom and the United States) and at the global level [5–7]. As part of our forecasting efforts, we employed semi-mechanistic growth models [8], which have shown previous forecasting success in the context of emerging infectious diseases [7,9,10].

For any emerging pathogen that rapidly transmits throughout a population, short-term forecasts of the epidemic's trajectory at different spatial scales can help guide policy and intervention strategies [11–16]. However, there is little opportunity to assess forecasting performance and improve models amid an ongoing public health crisis. Fortunately, as of September 2022, mpox cases have steadily declined worldwide [17–19], with non-endemic countries reporting 92 982 cases and 163 deaths overall as of 8 May 2024 [4]. Therefore, given the heterogeneous impact of the epidemic at different spatial scales and the substantial decline in mpox cases, a retrospective evaluation of the employed forecasting methodologies is vital to prepare for future public health events.

Several methodologies have been employed to forecast the trajectory of the 2022–2023 mpox epidemic in various geographical regions, including, but not limited to, models focused on human judgement [20], deep learning and artificial intelligence models [21,22], machine learning models [21–25], statistical models such as auto-regressive integrated moving average (ARIMA) models [21,23–27], compartmental models [28] and semi-mechanistic sub-epidemic models [7]. Performance metrics employed have mainly included variations of mean absolute error (MAE) [7,21,23–25,27], mean squared error (m.s.e.) [7,21–26] and mean absolute percentage error [21,23–25]. Only two studies used probabilistic measures of performance, such as the weighted interval score (WIS) [7,29]. Of the studies included here, four focused on the ascending phase of the epidemic (January–mid-August 2022) [20,24–26], three included forecasts of the epidemic's ascension and peak (February–September 2022) [21,22,29] and four focused on the ascending, peak and declining phases of the epidemic (January 2022–present) [7,23,27,28].

In this article, we conducted a comprehensive retrospective assessment of forecasting performance (one–four weeks ahead) of semi-mechanistic sub-epidemic modelling frameworks compared against other commonly employed statistical models, including ARIMA, generalized additive models (GAM), simple linear regression (SLR) and Facebook's Prophet model (Prophet) for the highest-impacted countries (e.g. Brazil, Canada, France, Germany, Spain, the United Kingdom and the United States) and on the global scale. We focused on producing and evaluating forecasts for the entirety of the epidemics in the countries of interest (21 July 2022 to 23 February 2023). We considered a comprehensive set of performance metrics in the field of epidemic forecasting, namely the MAE, m.s.e., 95% prediction interval (PI) coverage and WIS.

2. Methods

2.1. Data source and preparation

Our team obtained multiple daily confirmed mpox case series for the highest-impacted countries (e.g. Brazil, Canada, France, Germany, Spain, the United Kingdom and the United States) and on a global scale from the Centers for Disease Control and Prevention (CDC) [30] and the Our World in Data (OWID) GitHub page [18] on 9 and 15 August 2023, respectively. Although both data repositories employ slightly different collection and reporting methodologies, each defined a confirmed case as a laboratory-confirmed case (i.e. infection detected by polymerase chain reaction testing) [31,32]. Additional details regarding reporting methods and case obtainment for both sources can be found in electronic supplementary material, appendix S1, section A.

We aggregated cases to the weekly level to ensure an adequate number of cases for the forecasting process and to control daily reporting differences between locations. In this analysis, we defined a

week as Thursday to the following Wednesday to avoid calibrating the model with partial weeks of case data. We gathered data on the global level and for the countries that experienced the most significant outbreaks during the initial epidemic wave (e.g. Brazil, Canada, France, Germany, Spain, the United Kingdom and the United States) from the OWID team [18]. Our team also retrieved United States case data from the CDC [30] to explore differences in model performance dependent on the data source used.

2.2. Model calibration and forecasting strategy

Our analysis included 12 models in total: the *n*-sub-epidemic framework (e.g. top-ranked, second-ranked, weighted ensemble and unweighted ensemble), the sub-epidemic wave (spatial wave) framework (e.g. top-ranked, second-ranked, weighted ensemble and unweighted ensemble), a GAM, an SLR model, ARIMA model, each with an assumption of normality and a Prophet model. Descriptions of the *n*-sub-epidemic and spatial-wave frameworks, ARIMA model, GAM and Prophet model methodologies are given below. The SLR model included time as the only predictor; therefore, its description is not included in further detail.

For each location and model, we conducted and evaluated approximately 112 retrospective weekly sequential one- to four-week forecasts, with forecasts spanning from the week of 21 July 2022 to the week of 23 February 2023. Data posted in the week of 3 August 2023 were used to produce and evaluate all forecasts. We primarily employed 11-week calibration periods for each location, with the forecast date representing the first day of the week in which the forecast was produced. However, electronic supplementary material, appendix S2 contains a sensitivity analysis comparing the forecasting performance for 9- to 11-week calibration periods (electronic supplementary material, appendix S2).

For 14 July 2022 forecast date, we could not produce forecasts for Canada for the *n*-sub-epidemic framework owing to the low case counts during the early phase of the epidemic. We also could not produce forecasts for the weeks of 14 and 21 July 2022, using the *n*-sub-epidemic and spatial-wave frameworks for Brazil owing to low case counts. Similarly, some model calibration periods were between 8 and 10 weeks long for the *n*-sub-epidemic and spatial-wave frameworks, as initial zeros (i.e. any zeros at the start of the calibration period) were truncated until the first non-zero observation occurred.

2.3. The *n*-sub-epidemic modelling framework

The *n*-sub-epidemic framework employs multiple epidemic trajectories modelled as the aggregation of overlapping and asynchronous sub-epidemics [8,9]. For this analysis, a single sub-epidemic followed a three-parameter generalized logistic growth model, which has displayed competitive performance in the context of varying infectious diseases, including Zika, Ebola and COVID-19 [33–36]. Further details regarding the structure of the three-parameter generalized logistic growth model can be found in electronic supplementary material, appendix S1, section B.

An *n*-sub-epidemic trajectory comprises *n* overlapping sub-epidemics and is given by the following system of coupled differential equations:

$$\frac{dC_i(t)}{dt} = C_i'(t) = A_i(t) r_i C_i^{p_i}(t) \left(1 - \frac{C_i(t)}{K_{0_i}}\right). \tag{2.1}$$

The incidence curve of mpox cases is given by $\frac{dC_i(t)}{dt}$, where $C_i(t)$ tracks the cumulative number of mpox cases for sub-epidemic *i*. The parameters that characterize the shape of the *i*th sub-epidemic are given by (r_i, p_i, K_{0i}), for $i = 1, \ldots, n$. The parameter *r*, the growth rate per unit of time, remains positive, and parameter K_0 represents the final outbreak size. The 'scaling of growth' parameter $p \in [0,1]$ allows the model to capture early sub-exponential and exponential growth patterns. If $p = 0$, equation (2.1) describes a constant number of new cases over time, while $p = 1$ indicates that the early growth phase is exponential. Intermediate values of $p(0 < p < 1)$ describe early sub-exponential (e.g. polynomial) growth dynamics.

Here, *n* represents the number of sub-epidemics considered in the epidemic's trajectory. When $n = 1$, the sub-epidemic model is equivalent to the three-parameter generalized logistic growth model.

However, when $n > 1$, we employ the indicator variable, $A_i(t)$, to model the onset timing of the $(i + 1)$th sub-epidemic, where $(i + 1) \leq n$. Therefore, the $(i + 1)$th sub-epidemic is triggered when the cumulative curve of the ith sub-epidemic exceeds the case threshold value, C_{thr} (i.e. $C_{thr} \leq K_{0_i}$). Thus, we have

$$A_i(t) = \begin{cases} 1, & C_{i-1}(t) > C_{thr} \\ 0, & \text{Otherwise} \end{cases} \quad for\ i = 2, \dots n, \tag{2.2}$$

where $A_1(t) = 1$ for the first sub-epidemic. Therefore, $3n + 1$ parameters are needed to model a n-sub-epidemic trajectory when $n > 1$. This analysis considered a maximum of two sub-epidemics in the n-sub-epidemic trajectory ($n \leq 2$). The initial number of mpox cases is given by $C_1(0) = I_0$, where I_0 is the initial number of cases in the observed data. The cumulative curve of the n-sub-epidemic trajectory is given by

$$C_{tot}(t) = \sum_{i=1}^{n} C_i(t). \tag{2.3}$$

Overall, the n-sub-epidemic modelling framework can be applied to diverse epidemic patterns (e.g. multiple peaks and high-case level plateaus), such as the observed mpox epidemic trends (figure 1).

2.3.1. Parameter estimation and model selection

We employed a nonlinear least-squares method [8] to estimate the model parameters by fitting the model solution to the observed mpox data. Subsequently, we selected the top-ranked sub-epidemic models using the corrected Akaike information criterion (AIC_c) values of the set of best-fit models with different C_{thr} values. The AIC_c is given in [37,38],

$$AIC_c = n_d log(SSE) + 2m + \frac{2m(m+1)}{n_d - m - 1}, \tag{2.4}$$

where $SSE = \sum_{j=1}^{n_d} \left(f\left(t_j, \widehat{\Theta}\right) - y_{t_j} \right)^2$, m is the number of model parameters and n_d is the number of data points. Parameters from equation (2.4) were estimated from the nonlinear least-squares fit, which implicitly assumes normally distributed errors. Additional information regarding the parameter estimation process can be found in electronic supplementary material, appendix S1, section C.

2.3.2. Parametric bootstrapping

The n-sub-epidemic framework quantifies parameter uncertainty for the best-fit model, $f\left(t, \widehat{\Theta}\right)$, using a bootstrapping approach, which allows the computation of standard errors and related statistics without closed-form solutions [39]. Additionally, we ran the calibrated model forward in time to generate short-term forecasts (i.e. one–four weeks ahead) with quantified uncertainty, employing the same parametric bootstrapping method presented in [39] and discussed further in electronic supplementary material, appendix S1, section D. For this analysis, we used 300 bootstrap realizations to thoroughly characterize the parameter and forecasting uncertainty.

2.3.3. Constructing ensemble n-sub-epidemic models

We generated ensemble models from both the unweighted (equally weighted top-ranking models) and weighted combination of the two highest-ranking sub-epidemic models (i.e. top- and second-ranked) as deemed by the AIC_{c_i} for the ith ranked model, where $AIC_{c_1} \leq \cdots \leq AIC_{c_I}$ and $i = 1, \dots, I$. Details regarding the calculation of the weighted and unweighted ensemble models can be found in electronic supplementary material, appendix S1, section E. The PIs associated with the ensemble models were obtained using the bootstrapping approach discussed above.

2.4. Sub-epidemic wave framework (spatial-wave framework)

The sub-epidemic wave, or spatial-wave modelling framework, aggregates linked, overlapping, synchronous sub-epidemics (waves) to capture complex disease trajectories of differing shapes and sizes [10,40,41]. Like the n-sub-epidemic framework, the spatial-wave framework has shown past
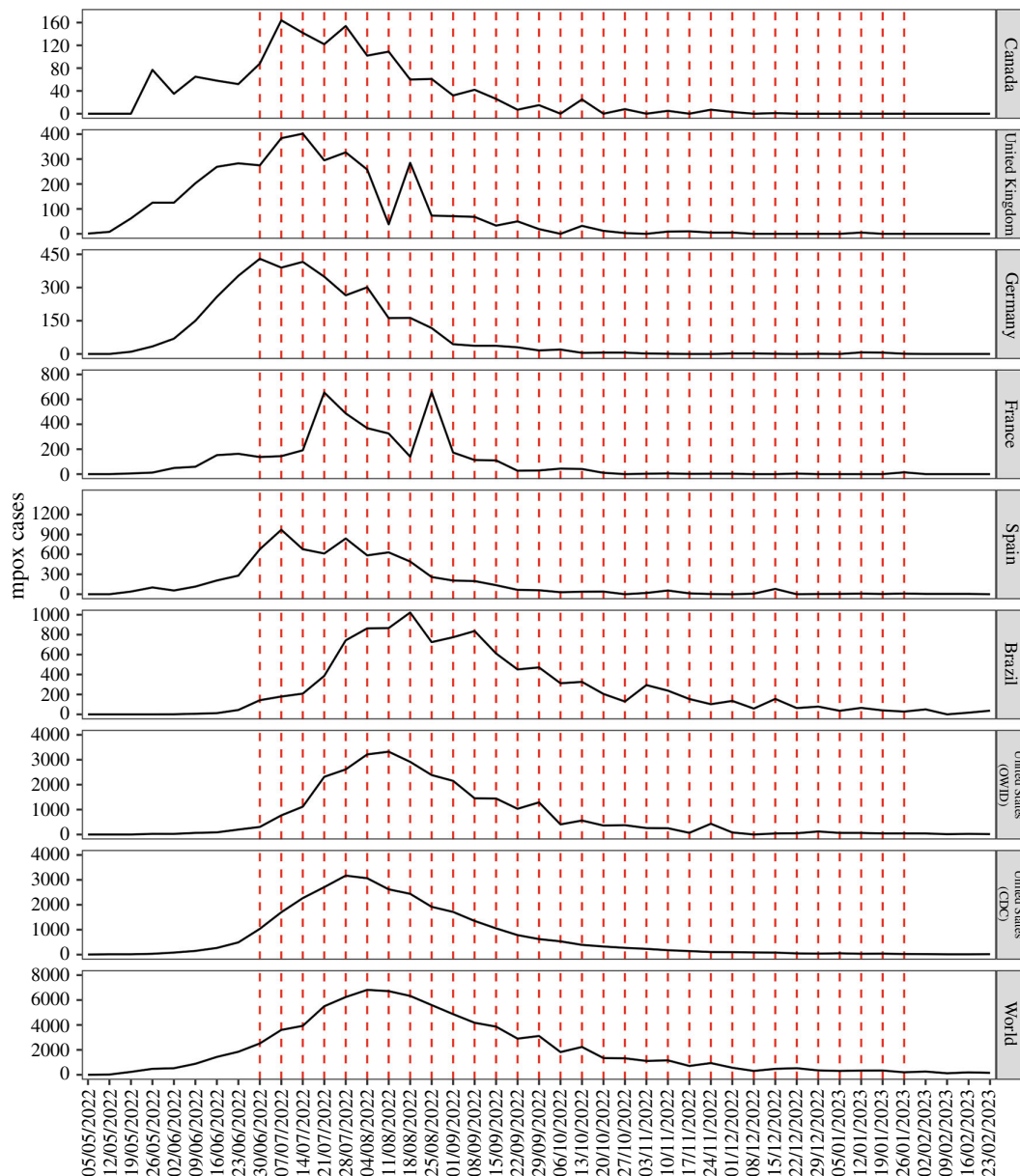
Figure 1. The epidemic trajectories for each study location of interest. The black solid line is the reported cases as of the week of 3 August 2023. The red-dashed lines are each of the forecasting periods included within the analysis (for all three calibration periods). Forecasts (one–four weeks) were produced through 23 February 2023, with the last forecast date being the week of 26 January 2023. Epidemic trajectories for Canada, the United Kingdom, Germany, France, Spain, Brazil, the United States (OWID) and the World are from the OWID team [18] and from the CDC for the United States (CDC) [30].

success in capturing diverse epidemic waves (e.g. high-level, prolonged epidemic plateaus and multi-peak trajectories), such as those seen throughout the mpox epidemic (figure 1) [10,40].

The mathematical building block for the spatial-wave model is the same three-parameter generalized logistic growth model discussed above. However, a sub-epidemic wave consists of n-overlapping sub-epidemics employing the following system of coupled differential equations [41]:

$$\frac{dC_i(t)}{dt} = rA_{i-1}(t)C_i(t)^p\left(1 - \frac{C_i(t)}{K_i}\right). \tag{2.5}$$

The incidence curve of mpox cases is given by $\frac{dC_i(t)}{dt}$, where $C_i(t)$ is the cumulative number of cases for sub-epidemic i at time t and K_i is the size of sub-epidemic i, where $i = 1, \ldots, n$. The indicator parameter, $A_i(t)$, is the same as given in equation (2.2). Starting from an initial sub-epidemic size, K_0, the size of subsequent sub-epidemics, K_i, decline at rate q following either an exponential or power-law function

as described in electronic supplementary material, appendix S1, section F. The growth rate per unit of time is represented by r, and the 'scaling of growth' parameter is represented by p. Unlike equation (2.1), parameters r and p do not vary across sub-epidemics, and the value of K_i depends on the decline rate q. Hence, a total of five parameters (r, p, C_{thr}, q, K_0) for $i = 1, \ldots, n$ are needed to characterize a sub-epidemic wave composed of two or more sub-epidemics $(n \geq 2)$ and four are needed when $n < 2$. This analysis considered a maximum of five sub-epidemics $(n \leq 5)$ in the epidemic wave trajectory.

The AIC_c used in the model selection process is given in equation (2.4). However, for the spatial-wave framework, $m = 5$ when the number of sub-epidemics is greater than one $(n > 1)$ and $m = 4$ when working with a single sub-epidemic $(n = 1)$. Like the n-sub-epidemic framework, the AIC_c for parameter estimation assumed nonlinear least-squares fit, which implicitly assumes the errors are normally distributed.

Parameter estimation and bootstrapping, construction of the ensemble models and forecasting methodologies followed the same procedures explained for the n-sub-epidemic modelling framework.

2.5. Auto-regressive integrated moving average models

ARIMA models are commonly employed in forecasting financial [42] and weather trends [43,44] and have become a standard benchmark in disease forecasting [9,21,23–26,45]. ARIMA models consist of three parts: (i) the auto-regression (AR) part involving regressing on the most recent values of the series, (ii) the moving average (MA) of error terms occurring contemporaneously and at previous times, and (iii) the integration (I) or differencing to account for the overall trend in the data and to make the time series stable. Mathematically, an ARIMA (p, d, q) process is given by

$$\phi(B)(1 - B)^d y_t = c + \theta(B)\epsilon_t, \tag{2.6}$$

where y_t denotes the number of mpox cases at time t and B denotes the backshift operator implying $By_t = y_{t-1}$ and $B(By_t) = B^2 y_t = y_{t-2}$, etc. The auto-regressive model is given by $\phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$, where p is the order of the model. The moving average model is given by $\theta(B) = 1 - \theta_1 B - \cdots - \theta_q B^q$ and q is the order of the moving average model. Finally, d is the degree of differencing. With this notation, $\phi(B)y_t = y_t - \phi_1 y_{t-1} - \ldots - \phi_p y_{t-p}$, $\theta(B)\epsilon_t = \epsilon_t - \theta_1 \epsilon_{t-1} + \ldots + \theta_q \epsilon_{t-q}$ and $(1 - B)^d$ means conducting the differencing d times. We used the auto.arima function from the R 'forecast' package to select orders and build the model [46], and the forecast function from the same package was used for forecasting [47]. Additional details regarding their application can be found in [48]. Any negative predicted values were truncated at zero.

2.6. Generalized additive models

GAMs extend generalized linear models to allow the fitting of nonlinear trends by including a sum of unknown smooth functions of some covariates while maintaining similar levels of model explainability and simplicity [49,50]. Specific to our study with time as the only covariate, assuming that y_t follows a normal distribution, our GAM is given as follows [48,49]:

$$y_t = \beta_0 + s(t) + \epsilon_t, \tag{2.7}$$

where $s(.)$ is an unknown smooth function of time and $\epsilon_t \sim N(0, \sigma^2)$. We used the GAM function in the R 'mgcv' package to fit this model [51], where the smooth function $s(.)$ is represented using basis functions (i.e. building blocks for creating more complex functions via linear combination). The default setting in 'mgcv' uses basis splines, which are piecewise polynomial functions. Specifically,

$$s(t) = \sum_{k=1}^{k} \beta_k b_k(t), \tag{2.8}$$

where $\{b_k(.)\}$ represent the basis functions, $\{\beta_k\}$ are the expansion coefficients to be estimated and k is the number of basis functions [49]. Here, k varied depending on the length of calibration data available for a given forecasting period. A discrete penalty was imposed on the basis coefficients to control the degree of smoothness, and the model was fitted by solving a penalized least square problem. The generalized cross-validation criterion selected the smoothness tuning parameter. A more

detailed description of the model fitting methodology can be found in [48,51], and the associated predict function was used for forecasting [52]. Any negative predicted values were truncated at zero.

2.7. Facebook's Prophet model

Facebook's Prophet model [53], initially designed for business-related forecasting, has recently been applied more frequently across multiple fields, including infectious disease modelling. Specifically, the model has produced both COVID-19 [45,54,55] and mpox forecasts [21]. The model's primary assumption is that it is 'decomposable'; therefore, $y(t)$, the main trend, can be decomposed into three pieces plus an error term. Thus, the model has the following form:

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t, \tag{2.9}$$

where $g(t)$ is a non-periodic component used for modelling the overall trend, $s(t)$ is a periodic component used for modelling the periodic changes over time and $h(t)$ is an irregular events component used for modelling the irregular changes (e.g. holidays or similar events). The component ϵ_t is an error of the model at time t.

We used the default setting of the R prophet function from the 'prophet' package as described in Taylor [56]. The predict function was used for forecasting from the model fit [52]. Any negative predicted values were truncated to zero. A more detailed description of the model fitting methodology can be found in [48,56].

2.8. Model evaluation

2.8.1. Performance metrics

Forecast performance was evaluated for each retrospectively generated one–four weeks ahead forecast using data downloaded the week of 3 August 2023. Descriptions of the employed metrics, MAE, m.s.e., 95% PI coverage and WIS, can be found in [8] and in electronic supplementary material, appendix S1, section G.

2.8.2. Skill scores and Winkler scores

We calculated skill scores to compare the proportion of improvement over the established ARIMA model (baseline model) for average m.s.e., MAE and WIS, using the n-sub-epidemic and spatial-wave frameworks as comparison models. Here, average refers to the average metric over all forecasting periods for each model, location and forecasting horizon. The formula is as follows [57]:

$$\frac{Baseline\ Model\ -\ Comparison\ Model}{Baseline\ Model} \times 100. \tag{2.10}$$

For 95% PI coverage, we first calculated Winkler scores to allow for the quantification of the proportion of improvement over the ARIMA model, using the n-sub-epidemic and spatial-wave frameworks as comparison frameworks. Winkler scores were calculated as follows [57]:

$$W_{\alpha, t} = \begin{cases} (u_{\alpha. t} - l_{\alpha, t}) + \frac{2}{\alpha}(l_{\alpha, t} - y_t), & if\ y_t < l_{\alpha, t} \\ (u_{\alpha. t} - l_{\alpha, t}), & if\ l_{\alpha, t} \leq y_t \leq u_{\alpha, t} \\ (u_{\alpha. t} - l_{\alpha, t}) + \frac{2}{\alpha}(y_t - u_{\alpha, t}) & if\ y_t > u_{\alpha, t} \end{cases} \tag{2.11}$$

where $u_{\alpha, t}$ is the upper bound of the 95% PI interval at time t, $l_{\alpha, t}$ is the lower bound at time t, y_t is the observed mpox case incidence at t, and $\alpha = 0.05$ as we were working with 95% PI intervals. We then calculated the skill scores (equation (2.10)) using the average Winkler scores, where scores were averaged across all forecasting periods for each model, location and forecasting horizon.

We selected the ARIMA model for the baseline model, as it has been frequently evaluated against other forecasting methodologies in the context of mpox [21,23–26]. Therefore, its inclusion in skill score calculations provides a more in-depth quantitative evaluation of the forecasting abilities of the n-sub-epidemic and spatial-wave frameworks against a well-vetted methodology.

R-studio (v. 4.3.1 (Beagle Scouts)) and MATLAB (v. R2022a) were used to produce and evaluate all forecasts.

3. Results

3.1. Overall and time-period-specific forecast performance

The n-sub-epidemic framework outperformed the established statistical (i.e. ARIMA, GAM, SLR and Prophet) and spatial-wave framework models across most locations. Specifically, the n-sub-epidemic unweighted ensemble was the most successful in Brazil, the United Kingdom, the United States (OWID) and the World. The model was followed in success by the top-ranked (Brazil) and weighted ensemble (Canada and Spain) n-sub-epidemic models. Nevertheless, the spatial-wave top-ranked model performed best overall in Germany, the ARIMA model in France and the GAM in the United States (CDC). Model performance varied when examining different epidemic phases with the statistical models performing best most frequently during the ascending period. The ensemble sub-epidemic frameworks performed best during the remainder of the epidemic. Overall, model performance improved as case levels declined (electronic supplementary material, appendix S3, figures A–D).

3.2. Country-level performance

3.2.1. Brazil

The n-sub-epidemic framework, specifically the top-ranked and unweighted ensemble models, performed best overall in Brazil, outperforming other models 44% of the time across performance metrics and horizons (figures 2–5). The n-sub-epidemic top-ranked and weighted ensemble models produced the lowest average m.s.e. (range: 16302.0–128915.5) and MAE (range: 86.4–135.0) across most forecasting horizons (figures 2 and 3). Both models improved between 8.5 and 29.0% over the ARIMA model regarding average m.s.e. and 21.0–33.2% for average MAE. The n-sub-epidemic top-ranked model most frequently produced the lowest average WIS (range: 57.4–97.5), improving 23.0–30.2% over the ARIMA model (figure 4). The n-sub-epidemic unweighted model produced the lowest average WIS (56.4), m.s.e. (14 240.9) and MAE (85.2) for the one-week forecasting horizon and performed best across all forecasting horizons regarding 95% PI coverage (range: 93.8–96.3%) (figures 2–5). When comparing average Winkler scores, the n-sub-epidemic unweighted ensemble improved 2.4–48.0% over the ARIMA model. Electronic supplementary material, tables A1–A5 contain the tabulation of skill scores and performance metrics across epidemic phases, models and forecasting horizons for Brazil (electronic supplementary material, appendix S4).

3.2.2. Canada

The n-sub-epidemic weighted ensemble and top-ranked models performed best overall in Canada, outperforming other models across forecasting horizons and metrics 69 and 56% of the time, respectively (figures 2–5). Across all forecasting horizons for average m.s.e. (range: 103.2–128.9) and MAE (range: 6.3–6.6), the n-sub-epidemic top-ranked and weighted ensemble models produced the lowest values (figures 2 and 3). Both models improved considerably over the ARIMA model regarding average m.s.e. (range: 83.6–88.2%) and MAE (range: 58.8–67.6%). The success of the n-sub-epidemic weighted ensemble model continued for average WIS, as the model produced the lowest WIS across most forecasting horizons (range: 4.3–4.5) and improved between 54.3 and 63.0% over the ARIMA model (figure 4). However, the ARIMA model performed best and most frequently regarding average 95% PI coverage (range: 94.25–96.55%). Electronic supplementary material, tables B1–B5 contain the tabulation of skill scores and performance metrics across epidemic phases, models and forecasting horizons for Canada (electronic supplementary material, appendix S4).

3.2.3. France

Unlike the other countries, the ARIMA model outperformed all other models 94% of the time across forecasting metrics and horizons, followed by the n-sub-epidemic top-ranked (6%) and weighted ensemble (6%) models (figures 2–5). The ARIMA model consistently produced the lowest average m.s.e. (range: 22986.7–24413.1), WIS (range: 61.1–62.3) and closest to 95% PI coverage on average (range: 89.7–93.1%) across forecasting horizons (figures 2, 3 and 5). Regarding average MAE, the
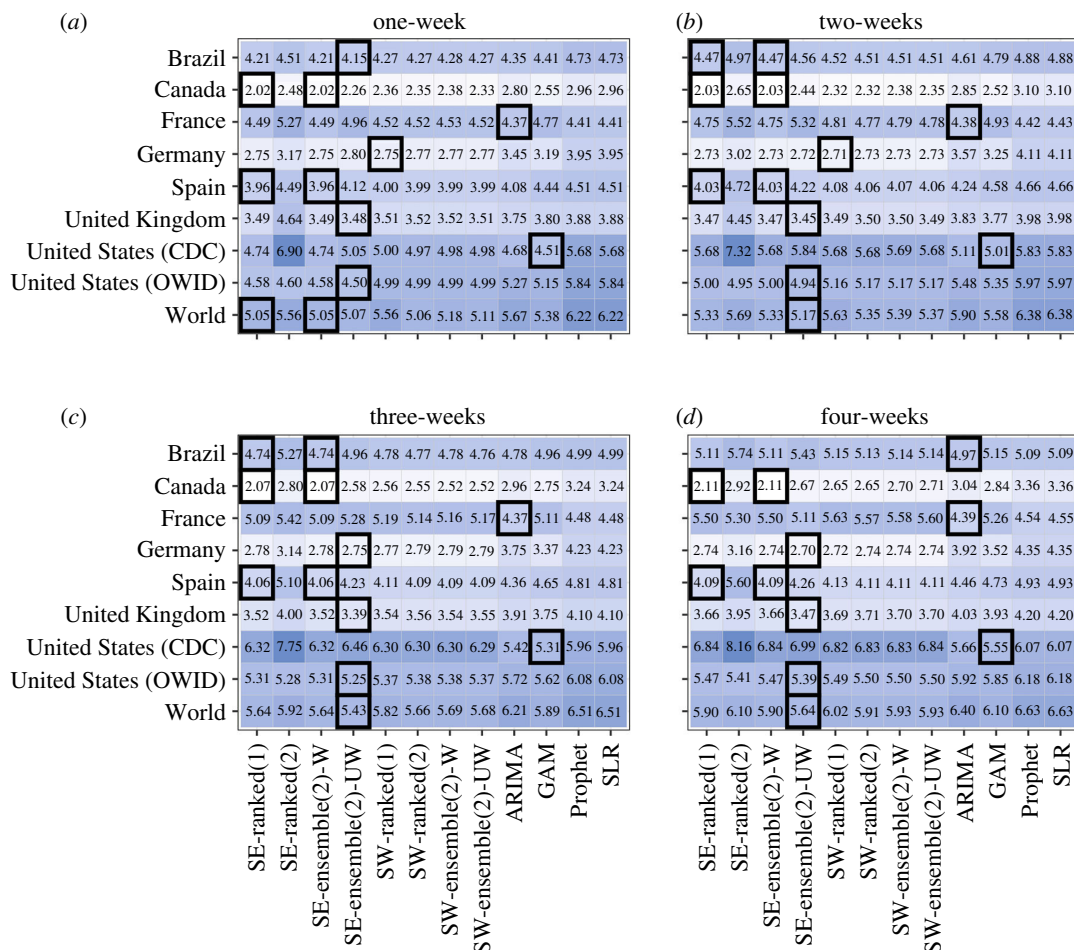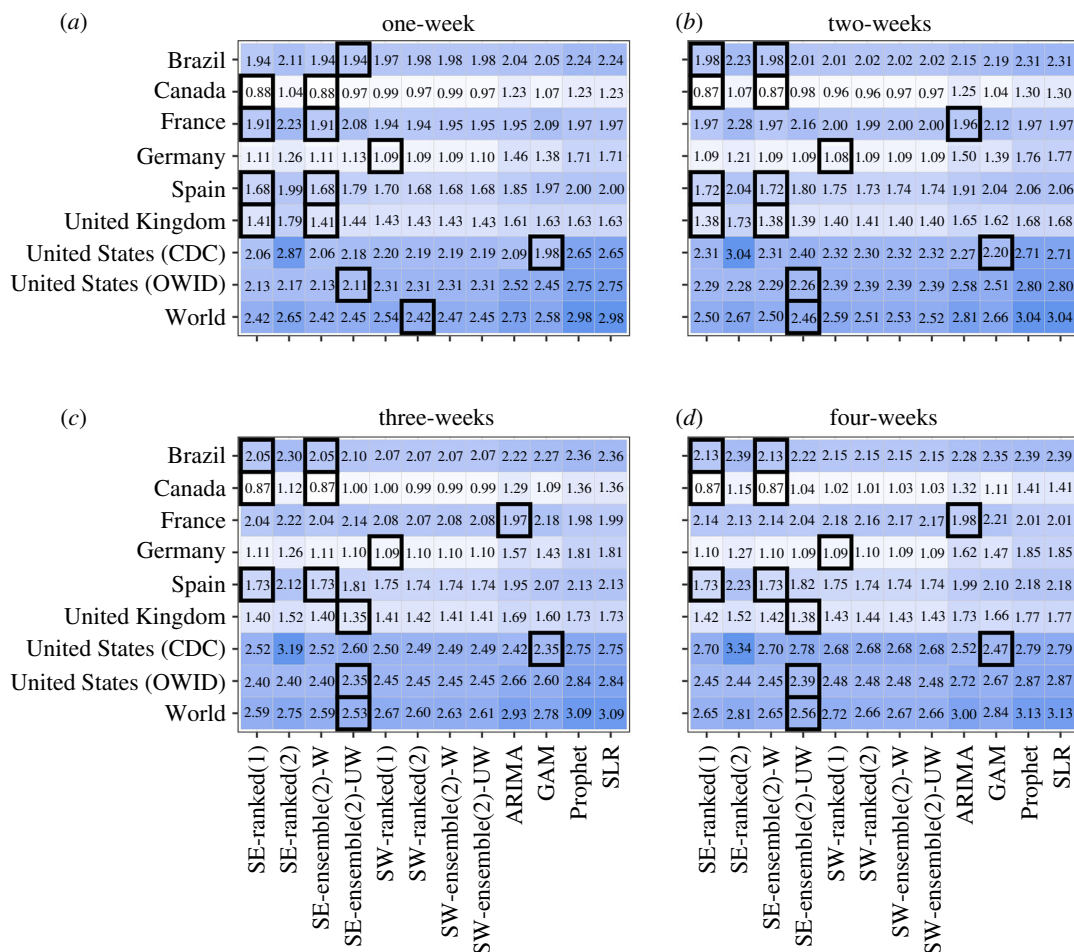
Figure 2. (*a*–*d*) Average m.s.e. for each location, model and forecasting horizon (weeks of 21 July 2022–23 February 2023). Approximately 1008 forecasts were produced for each model. The graph is shown on the $\log_{10}$ scale. Lighter colours indicate lower average m.s.e. scores, and darker blues indicate higher average m.s.e. scores. The black boxes indicate the best-performing model(s) for a given location and forecasting horizon. The *n*-sub-epidemic framework performed best most frequently across forecasting horizons and locations regarding average m.s.e.

ARIMA model performed best across most forecasting horizons (range: 90.5–93.8), albeit the one-week horizon, where the *n*-sub-epidemic top-ranked (80.9) and weighted ensemble performed best (80.9) (figure 3). However, the *n*-sub-epidemic top-ranked and weighted ensemble models saw minimal improvement over the ARIMA regarding average MAE (less than 10%). Electronic supplementary material, tables C1–C5 contain the tabulation of skill scores and performance metrics across epidemic phases, models and forecasting horizons for France (electronic supplementary material, appendix S4).

3.2.4. Germany

The spatial wave top-ranked model outperformed the other models 69% of the time across forecasting horizons and performance metrics, followed by the *n*-sub-epidemic (38%) unweighted ensemble model (figures 2–5). The spatial-wave top-ranked model produced the lowest average m.s.e. in the one- (559.9) and two-week (514.2) forecasting horizons and across all forecasting horizons regarding average MAE (range: 11.0–11.4) and WIS (range: 7.9–8.3) (figures 2–4). We also noted considerable improvement in average m.s.e. (range: 80.0–86.3%), MAE (range: 59.1–72.7%) and WIS (range: 55.1–70.8%) over the ARIMA model. The *n*-sub-epidemic unweighted ensemble produced the lowest average m.s.e. during the three- (563.06) and four-week (504.64) forecasting horizons, improving 90.1–94.0% over the ARIMA model (figure 2). Similarly, the *n*-sub-epidemic unweighted ensemble performed best most frequently regarding average 95% PI coverage (89.7%), improving 30.7–65.9% over the ARIMA model in average Winkler scores (figure 5). Electronic supplementary material, tables

Figure 3. (*a–d*) Average MAE for each location, model and forecasting horizon (weeks 21 July 2022–23 February 2023). Approximately 1008 forecasts were produced for each model. The graph is shown on the $\log_{10}$ scale. Lighter colours indicate lower average MAE scores, and darker blues indicate higher average MAE scores. The black boxes indicate the best-performing model(s) for a given location and forecasting horizon. The *n*-sub-epidemic framework performed best most frequently across forecasting horizons and locations regarding average MAE.

D1–D5 contain the tabulation of skill scores and performance metrics across epidemic phases, models and forecasting horizons for Germany (electronic supplementary material, appendix S4).

3.2.5. Spain

The *n*-sub-epidemic framework performed superiorly in Spain, with the most success noted for the weighted ensemble model (69%) across forecasting periods and metrics (figures 2–5). Regarding average m.s.e. (range: 9020.3–12286.1) and MAE (range: 46.3–52.4), the *n*-sub-epidemic top-ranked and weighted ensemble models produced the lowest values across all forecasting horizons (figures 2 and 3). Similarly, both models improved considerably over the ARIMA model regarding average m.s.e. (range: 25.7–57.4%) and MAE (range: 33.2–45.9%). The *n*-sub-epidemic weighted ensemble produced the lowest average WIS (range: 33.7–35.4) most frequently across forecasting horizons, improving 19.9–41.6% over the ARIMA model (figure 4). The *n*-sub-epidemic second-ranked model performed well regarding average 95% PI coverage, producing the closest to 95% coverage across most forecasting horizons (range: 94.3–94.8%) (figure 5). However, the ARIMA model outperformed the *n*-sub-epidemic second-ranked model across all forecasting horizons regarding Winkler scores. Electronic supplementary material, tables E1–E5 contain the tabulation of skill scores and performance metrics across epidemic phases, models and forecasting horizons for Spain (electronic supplementary material, appendix S4).
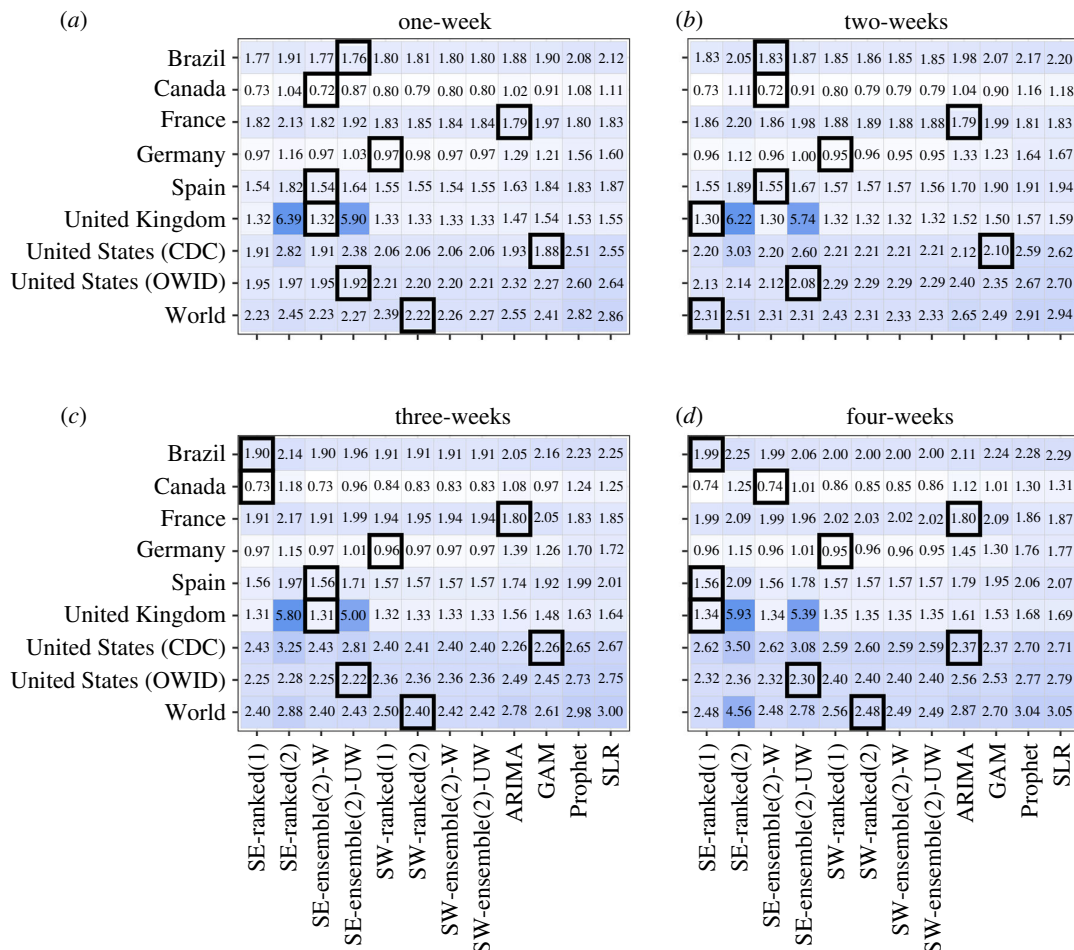
Figure 4. (*a*–*d*) Average WIS for each location, model and forecasting horizon (weeks of 21 July 2022–23 February 2023). Approximately 1008 forecasts were produced for each model. The graph is shown on the $\log_{10}$ scale. Lighter colours indicate lower average WIS, and darker blues indicate higher average WIS. The black boxes indicate the best-performing model(s) for a given location and forecasting horizon. The *n*-sub-epidemic framework performed best most frequently across forecasting horizons and locations regarding average WIS.

3.2.6. United Kingdom

Overall, the *n*-sub-epidemic unweighted ensemble model outperformed other models 56% of the time across forecasting horizons and metrics, followed in success by the top-ranked (25%) and weighted (25%) ensemble models (figures 2–5). The *n*-sub-epidemic unweighted ensemble model produced the lowest average m.s.e. (range: 2454.9–3012.8) across all forecasting horizons and for the three- (21.5) and four-week (22.8) horizons regarding average MAE (figures 2 and 3). However, the model saw considerable success over the ARIMA model across all horizons in average m.s.e. (range: 45.9–72.1%) and average MAE (range: 32.1–56.6%). The *n*-sub-epidemic top-ranked and weighted ensemble produced the lowest average MAE in the one- (24.7) and two-week (22.9) horizons and saw split success across forecasting horizons regarding average WIS (figures 3 and 4). The *n*-sub-epidemic unweighted ensemble performed superiorly across most forecasting horizons for average 95% PI coverage (figure 5). Electronic supplementary material, tables F1–F5 contain the tabulation of skill scores and performance metrics across epidemic phases, models and forecasting horizons for the United Kingdom (electronic supplementary material, appendix S4).

3.2.7. United States (CDC)

Unlike the other countries, the GAM model performed best in the United States (CDC), outperforming the other models 69% of the time, followed in success by the spatial-wave framework models (25%) across forecasting horizons and metrics (figures 2–5). The GAM produced the lowest average m.s.e.

(a) one-week

	SE-ranked(1)	SE-ranked(2)	SE-ensemble(2)-W	SE-ensemble(2)-UW	SW-ranked(1)	SW-ranked(2)	SW-ensemble(2)-W	SW-ensemble(2)-UW	ARIMA	GAM	Prophet	SLR
Brazil	85.2	96.3	85.2	96.3	85.7	82.1	85.7	82.1	82.8	79.3	65.5	55.2
Canada	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	96.6	89.7	89.7	69.0
France	86.2	82.8	86.2	86.2	82.8	86.2	86.2	86.2	89.7	72.4	86.2	65.5
Germany	89.7	89.7	89.7	89.7	89.7	89.7	89.7	89.7	79.3	72.4	62.1	34.5
Spain	89.7	93.1	89.7	93.1	93.1	89.7	89.7	89.7	96.6	75.9	75.9	65.5
United Kingdom	86.2	93.1	86.2	89.7	86.2	86.2	86.2	86.2	89.7	79.3	86.2	72.4
United States (CDC)	86.2	86.2	86.2	93.1	93.1	93.1	93.1	93.1	89.7	75.9	44.8	17.2
United States (OWID)	82.8	93.1	79.3	100.0	51.7	55.2	51.7	55.2	82.8	75.9	69.0	48.3
World	93.1	86.2	93.1	93.1	89.7	93.1	93.1	93.1	72.4	79.3	65.5	41.4

(b) two-weeks

	SE-ranked(1)	SE-ranked(2)	SE-ensemble(2)-W	SE-ensemble(2)-UW	SW-ranked(1)	SW-ranked(2)	SW-ensemble(2)-W	SW-ensemble(2)-UW	ARIMA	GAM	Prophet	SLR
Brazil	87.0	90.7	87.0	94.4	87.5	85.7	87.5	87.5	84.5	79.3	60.3	51.7
Canada	100.0	98.2	100.0	100.0	100.0	100.0	100.0	100.0	96.6	89.7	82.8	67.2
France	86.2	84.5	86.2	87.9	84.5	87.9	87.9	87.9	91.4	74.1	81.0	65.5
Germany	87.9	87.9	87.9	89.7	87.9	87.9	87.9	89.7	84.5	74.1	44.8	31.0
Spain	91.4	94.8	91.4	96.6	91.4	91.4	91.4	91.4	96.6	79.3	69.0	62.1
United Kingdom	89.7	93.1	89.7	94.8	89.7	89.7	89.7	89.7	89.7	81.0	82.8	70.7
United States (CDC)	84.5	82.8	84.5	93.1	94.8	94.8	94.8	94.8	89.7	77.6	31.0	17.2
United States (OWID)	75.9	86.2	77.6	87.9	50.0	50.0	51.7	51.7	81.0	75.9	60.3	48.3
World	94.8	91.4	93.1	96.6	89.7	93.1	93.1	93.1	70.7	77.6	51.7	32.8

(c) three-weeks

	SE-ranked(1)	SE-ranked(2)	SE-ensemble(2)-W	SE-ensemble(2)-UW	SW-ranked(1)	SW-ranked(2)	SW-ensemble(2)-W	SW-ensemble(2)-UW	ARIMA	GAM	Prophet	SLR
Brazil	87.7	91.4	87.7	93.8	88.1	86.9	88.1	88.1	85.1	78.2	56.3	48.3
Canada	100.0	96.4	100.0	100.0	100.0	100.0	100.0	100.0	94.3	89.7	75.9	65.5
France	88.5	85.1	87.4	89.7	85.1	88.5	88.5	88.5	92.0	77.0	78.2	66.7
Germany	87.4	88.5	87.4	89.7	86.2	86.2	86.2	86.2	85.1	75.9	43.7	29.9
Spain	92.0	94.3	92.0	96.6	93.1	92.0	93.1	92.0	97.7	81.6	63.2	58.6
United Kingdom	90.8	93.1	90.8	95.4	90.8	90.8	90.8	90.8	88.5	80.5	78.2	69.0
United States (CDC)	83.9	81.6	83.9	93.1	95.4	95.4	95.4	95.4	87.4	75.9	24.1	14.9
United States (OWID)	75.9	87.4	75.9	88.5	50.6	54.0	52.9	51.7	78.2	74.7	57.5	47.1
World	90.8	89.7	90.8	96.6	86.2	88.5	89.7	89.7	70.1	75.9	41.4	28.7

(d) four-weeks

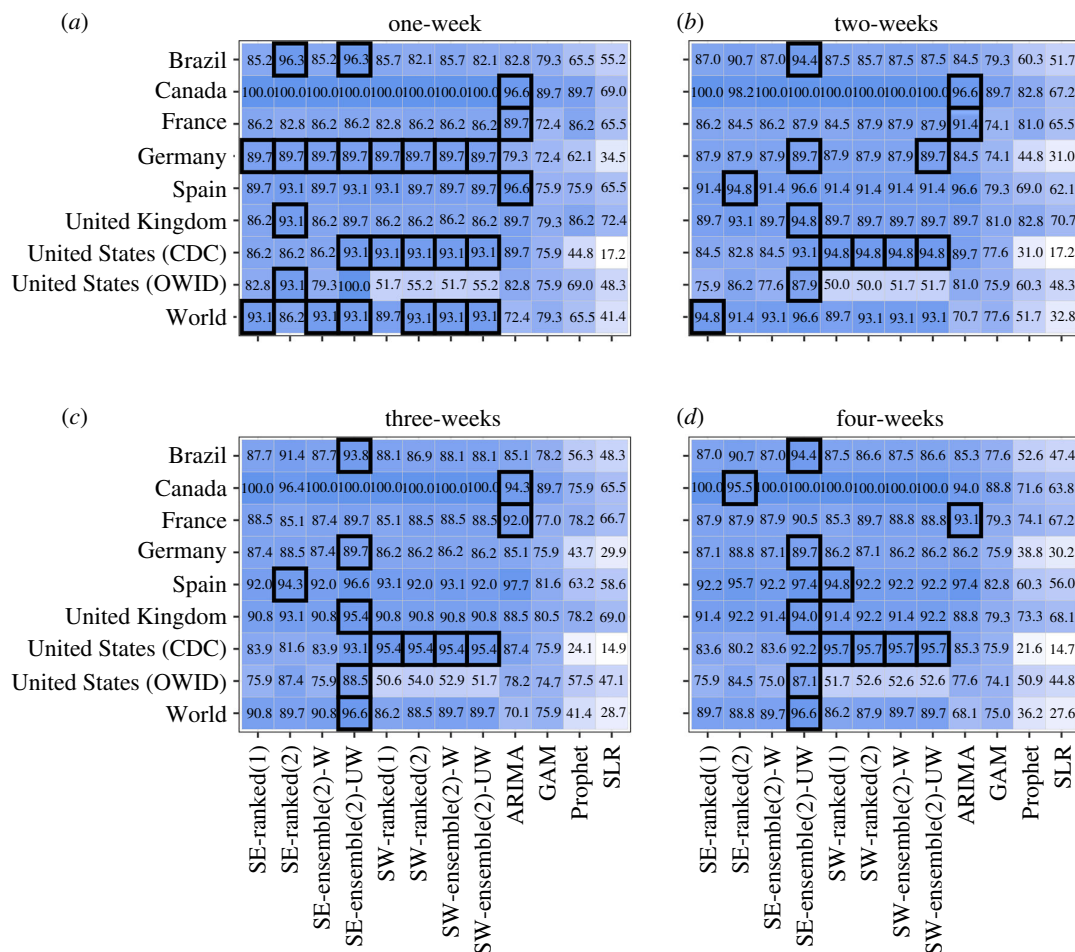|	SE-ranked(1)	SE-ranked(2)	SE-ensemble(2)-W	SE-ensemble(2)-UW	SW-ranked(1)	SW-ranked(2)	SW-ensemble(2)-W	SW-ensemble(2)-UW	ARIMA	GAM	Prophet	SLR
Brazil	87.0	90.7	87.0	94.4	87.5	86.6	87.5	86.6	85.3	77.6	52.6	47.4
Canada	100.0	95.5	100.0	100.0	100.0	100.0	100.0	100.0	94.0	88.8	71.6	63.8
France	87.9	87.9	87.9	90.5	85.3	89.7	88.8	88.8	93.1	79.3	74.1	67.2
Germany	87.1	88.8	87.1	89.7	86.2	87.1	86.2	86.2	86.2	75.9	38.8	30.2
Spain	92.2	95.7	92.2	97.4	94.8	92.2	92.2	92.2	97.4	82.8	60.3	56.0
United Kingdom	91.4	92.2	91.4	94.0	91.4	92.2	91.4	92.2	88.8	79.3	73.3	68.1
United States (CDC)	83.6	80.2	83.6	92.2	95.7	95.7	95.7	95.7	85.3	75.9	21.6	14.7
United States (OWID)	75.9	84.5	75.0	87.1	51.7	52.6	52.6	52.6	77.6	74.1	50.9	44.8
World	89.7	88.8	89.7	96.6	86.2	87.9	89.7	89.7	68.1	75.0	36.2	27.6

Figure 5. (a–d) Average 95% PI coverage for each location, model and forecasting horizon (weeks of 21 July 2022–23 February 2023). Approximately 1008 forecasts were produced for each model. Lighter colours indicate lower average 95% PI coverage, and darker blues indicate higher 95% PI coverage. The black boxes indicate the best-performing model(s) for a given location and forecasting horizon. The n-sub-epidemic unweighted ensemble performed best and most frequently across forecasting horizons and locations regarding average 95% PI coverage.

(range: 32070.5–352246.3) and MAE (range: 95.0–291.3) across all forecasting horizons and the one-week to three-week forecasting horizons for average WIS (range: 74.9–179.0) (figures 2–4). However, all spatial-wave framework models performed superiorly across all forecasting horizons for average 95% PI coverage (range: 93.1–95.7%), albeit the one-week horizon, where the n-sub-epidemic unweighted ensemble model performed equally well (93.1%) (figure 5). Nevertheless, the ARIMA model outperformed the spatial-wave framework and the n-sub-epidemic unweighted ensemble model across all forecasting horizons regarding average Winkler scores. Electronic supplementary material, tables G1–G5 contain the tabulation of skill scores and performance metrics across epidemic phases, models and forecasting horizons for the United States (CDC) (electronic supplementary material, appendix S4).

3.2.8. United States (OWID)

The n-sub-epidemic unweighted ensemble model performed best across all forecasting horizons regarding average m.s.e., MAE and WIS and across most horizons regarding average 95% PI coverage, outperforming other models 94% of the time (figures 2–5). Regarding average m.s.e., the model produced values ranging from 31 944.2 to 246 467.4, a 66.0–82.9% improvement over the ARIMA model (figure 2). We also observed considerable improvements over the ARIMA model in average MAE (range: 50.6–60.9%) and WIS (range: 45.4–60.8%). The n-sub-epidemic unweighted ensemble model produced superior average 95% PI coverage across most forecasting horizons (range: 87.1–88.5%), improving 27.5–70.6% over the ARIMA model in average Winkler scores (figure 5). Electronic supplementary material, tables H1–H5 contain the tabulation of skill scores and performance metrics

across epidemic phases, models and forecasting horizons for the United States (OWID) (electronic supplementary material, appendix S4).

3.3. World

The *n*-sub-epidemic unweighted ensemble model, followed by the spatial-wave second-ranked model, outperformed other models across forecasting horizons and metrics 56 and 31% of the time, respectively (figures 2–5). The *n*-sub-epidemic unweighted ensemble model produced the lowest average m.s.e. (range: 147 046.3–435 707.3) and MAE (range: 286.3–362.3) across the two- to four-week forecasting horizons, improving considerably over the ARIMA model (greater than 47%) (figures 2 and 3). Similarly, the model performed superior regarding average 95% PI coverage most frequently across forecasting horizons (range: 93.1–96.6%), albeit the two-week forecasting horizon where the *n*-sub-epidemic top-ranked model performed best (94.8%) (figure 5). The *n*-sub-epidemic unweighted ensemble model improved 42.4–56.3% over the ARIMA model regarding average Winkler scores. However, the spatial-wave second-ranked model performed best across most forecasting horizons regarding average WIS (range: 166.5–297.6), improving 53.5–59.7% over the ARIMA model (figure 4). Electronic supplementary material, tables I1–I5 contain the tabulation of skill scores and performance metrics across epidemic phases, models and forecasting horizons for the World (electronic supplementary material, appendix S4).

4. Discussion

In the context of the global emergency posed by the 2022–2023 mpox epidemic, we have systematically investigated the short-term forecasting performance of multiple models that have shown competitive performance during epidemics and pandemics, and rely on minimal data of the epidemic's trajectory [7,9,10]. Overall, the *n*-sub-epidemic framework, specifically the unweighted ensemble model, followed by the spatial-wave framework, ARIMA model and GAM, performed best most frequently across locations and forecasting horizons compared with the other established modelling techniques (e.g. SLR and Prophet).

The *n*-sub-epidemic framework outperformed other models in Brazil, Canada, Spain, the United Kingdom, the United States (OWID) and the World, whereas the spatial-wave framework did best in Germany. However, France and the United States (CDC) behaved uniquely compared with other countries with the ARIMA model and GAM, respectively, performing best across forecasting horizons. Excluding France and the United States (CDC), the spatial-wave and *n*-sub-epidemic models frequently improved considerably in each metric, including in average Winkler scores, over the ARIMA model. Overall, our findings further support the sub-epidemic and ensemble frameworks for forecasting emerging infectious diseases, especially in the face of limited epidemiological data [9,10].

Although each included location experienced unique epidemic trajectories throughout the mpox epidemic, the consistent success of the *n*-sub-epidemic and spatial-wave frameworks highlights the utility of the aggregated sub-epidemic approach in capturing and producing short-term forecasts for a wide variety of epidemic trends. For example, the epidemic trajectories of the mpox epidemic in Canada, the United Kingdom and the United States (OWID) differ considerably (figure 1). Canada saw multiple peaks in cases throughout the ascending and declining phases of the epidemic. In contrast, the United Kingdom experienced a relatively smooth ascending phase with a prominent peak during the descending phase, and the United States (OWID) saw a mostly smooth epidemic trajectory without drastic peaks during the ascending or descending phases. Although different, the *n*-sub-epidemic unweighted ensemble model performed best across forecasting horizons for all three locations. Similar observations hold for Germany, Brazil, Spain and the World.

France and the United States (CDC) behaved differently from the other locations, with the ARIMA performing best in France and the GAM in the United States (CDC). The mpox epidemic in France did not follow a unimodal epidemic disease trajectory as seen in other study areas; instead, they experienced two prominent case peaks with slow ascending and descending phases (figure 1). However, throughout the outbreak, large stretches of the epidemic's trajectory appear to form linear trends (i.e. ascending, between peaks and tail-end). As the ARIMA model inherently assumes linearity, the linear-like trajectory of mpox in France may explain the success of the ARIMA model in capturing and forecasting the disease's trajectory [58]. Unlike France, the United States (CDC) had minimal

week-to-week fluctuation, forming a smooth, unimodal epidemic trajectory (figure 1). The GAM, which performed best in the United States (CDC), captures common nonlinear trends, such as seen for the United States (CDC) epidemic trajectory [49]. Therefore, similar to France, epidemic data following distinctive trajectories may explain the success seen by the ARIMA and GAM models in France and the United States (CDC).

The model performance also differed when looking at specific epidemic phases (e.g. ascending, peak, descending and tail-end of the epidemic). For example, across locations, the established statistical models outperformed other models most frequently during the ascending phase of the epidemic. However, there were multiple epidemic phases with few forecasting periods available, limiting the ability to examine country-specific epidemic phase performance in greater detail.

Finally, the significant success of the n-sub-epidemic unweighted and weighted ensemble models highlights the continued utility of ensemble modelling in short-term forecasting [59–66]. Ensemble modelling has also shown considerable success against individual top-ranking sub-epidemic and other statistical models in past mpox short-term forecasting efforts [7], along with that of COVID-19 [9,61,67], seasonal influenza [62,64] and Zika [68]. Each ensemble model (e.g. weighted and unweighted) included within the analysis performed competitively at various forecasting horizons and for different areas. However, the unweighted ensemble n-sub-epidemic model outperformed all other included models for most locations, demonstrating the continued competitiveness of ensemble modelling in short-term mpox forecasting.

Our study is not exempt from limitations. We used two sources for weekly time-series laboratory-confirmed mpox cases, each using varying approaches to compile and disseminate case information. Similarly, the reporting patterns for countries varied as well. Therefore, the trends noted above may be a function of reporting delays rather than true case fluctuations. However, to account for the effect of reporting differences on the observed epidemic trajectories, we aggregated cases weekly to adjust for within-week reporting delays. Owing to the unprecedented nature of the epidemic and the limited epidemic data available, we examined forecasts derived from the second-ranked spatial wave and n-sub-epidemic models. Frequently, there was little statistical support for the second-ranked models relative to the top-ranked model for both frameworks. The spatial-wave and ensemble n-sub-epidemic frameworks are semi-mechanistic and provide insight into the natural processes that formed the observed epidemic trends from the aggregated sub-epidemics. Nevertheless, both frameworks do not account for explicit mechanisms of reactive behavioural modification and interventions, which probably played a significant role in controlling the epidemic at local and global levels [13,69]. Finally, current models are not responsive to seasonal or event-specific risk behaviours, which limits their application to short-term forecasts.

5. Conclusion

In conclusion, our findings further support the competitive performance of the aggregated sub-epidemic methodologies in producing short-term forecasts within the context of mpox, along with the utility of ensemble modelling therein producing short-term forecasts. However, the frameworks could be expanded to other disease types as they have shown utility in short-term forecasting for epidemiologically different diseases (e.g. COVID-19 and Ebola). Finally, short-term forecasting tools requiring minimal data inputs are essential in the face of unprecedented pandemics and epidemics. Therefore, further evaluation and refinements of the ensemble sub-epidemic frameworks could be achieved via comparison with other models since all the forecasting results are publicly available to the community.

References

1. World Health Organization. 2023 Multi-country outbreak of mpox. *Emerg. Sit. updates.* **27**, 1–17. https://www.who.int/publications/m/item/multi-country-outbreak-of-mpox--external-situation-report-27---14-august-2023

2. 2022 Monkeypox declared a global health emergency by the World Health Organization. In *UN news*. See https://news.un.org/en/story/2022/07/1123152.

3. Centers for Disease Control and Prevention. 2023 Technical report 1: multi-national mpox outbreak, United States, 2022. In *Mpox 1*. https://archive.cdc.gov/www_cdc_gov/poxvirus/mpox/cases-data/technical-report/report-1.html.

4. Centers for Disease Control and Prevention. 2023 2022-2023 Mpox outbreak global map. In *2022-2023 outbreak cases & data*. https://www.cdc.gov/poxvirus/mpox/response/2022/world-map.html.

5. Chowell G. 2022 *GitHub*. monkeypox-usa. See https://github.com/gchowell/monkeypox-usa.

6. Bleichrodt A. 2023 *GitHub*. Summer-23-Mpox-Forecasts. See https://github.com/bleicham/Summer-23-Mpox-Forecasts.

7. Bleichrodt A, Dahal S, Maloney K, Casanova L, Luo R, Chowell G. 2023 Real-time forecasting the trajectory of monkeypox outbreaks at the national and global levels, July–October 2022. *BMC Med.* **21**, 19. (doi:10.1186/s12916-022-02725-2)

8. Chowell G, Dahal S, Bleichrodt A, Tariq A, Hyman JM, Luo R. 2024 *SubEpiPredict:* a tutorial-based primer and toolbox for fitting and forecasting growth trajectories using the ensemble *n*-sub-epidemic modeling framework. *Infect. Dis. Model.* **9**, 411–436. (doi:10.1016/j.idm.2024.02.001)

9. Chowell G, Dahal S, Tariq A, Roosa K, Hyman JM, Luo R. 2022 An ensemble *n*-sub-epidemic modeling framework for short-term forecasting epidemic trajectories: application to the COVID-19 pandemic in the USA. *PLoS Comput. Biol.* **18**, e1010602. (doi:10.1371/journal.pcbi.1010602)

10. Chowell G, Tariq A, Hyman JM. 2019 A novel sub-epidemic modeling framework for short-term forecasting epidemic waves. *BMC Med.* **17**, 164. (doi:10.1186/s12916-019-1406-6)

11. World Health Organization. 2023 An mpox resurgence in the European region this spring and summer? To prevent that, key measures must continue. In *News*. https://www.who.int/europe/news/item/17-05-2023-an-mpox-resurgence-in-the-european-region-this-spring-and-summer--to-prevent-that--key-measures-must-continue.

12. Centers for Disease Control and Prevention. 2023 About flu forecasting. In *FluSight: Flu Forecasting*. https://www.cdc.gov/flu/weekly/flusight/how-flu-forecasting.htm#:~:text=Flu%20forecasts%20can%20inform%20messaging,hospital%20beds%20and%20treatment%20resources.

13. Centers for Disease Control and Prevention. 2022 Technical report 4: multi-national mpox outbreak,United States, 2022. In *Mpox 4*. https://archive.cdc.gov/#/details?url=https://www.cdc.gov/poxvirus/mpox/cases-data/technical-report/report-4.html.

14. Centers for Disease Control and Prevention. 2023 COVID-19 forecasting and mathematical modeling. In *COVID19*. https://www.cdc.gov/coronavirus/2019-ncov/science/forecasting/forecasting-math-modeling.html.

15. Panovska-Griffiths J *et al*. 2023 Responsive modelling of the mpox epidemic in England as a cross-government and academia collaboration: the process, challenges, and lessons learned. *Lancet. Microbe.* **4**, e658–e660. (doi:10.1016/S2666-5247(23)00151-9)

16. Alshanbari HM, Iftikhar H, Khan F, Rind M, Ahmad Z, El-Bagoury AAAH. 2023 On the implementation of the artificial neural network approach for forecasting different healthcare events. *Diagnostics* **13**, 1310. (doi:10.3390/diagnostics13071310)

17. Kupferschmidt K. 2022 Monkeypox cases are plummeting. Scientists are debating why. *Science* (doi:10.1126/science.adf5133)

18. Mathieu E, Spooner F, Dattani S, Ritchie H, Roser M. 2022 *GitHub*. Mpox (monkeypox) data. See https://github.com/owid/monkeypox.

19. Reardon S. 2022 What does the future look like for monkeypox? *Nature New Biol.* **610**, 250–252. (doi:10.1038/d41586-022-03204-7)

20. McAndrew T *et al*. 2022 Early human judgment forecasts of human monkeypox, May 2022. *Lancet Digit. Health* **4**, e569–e571. (doi:10.1016/S2589-7500(22)00127-3)

21. Long B, Tan F, Newman M. 2023 Forecasting the Monkeypox outbreak using ARIMA, Prophet, NeuralProphet, and LSTM models in the United States. *Forecasting* **5**, 127–137. (doi:10.3390/forecast5010005)

22. Manohar B, Das R. 2022 Artificial neural networks for the prediction of monkeypox outbreak. *Trop. Med. Infect. Dis.* **7**, 424. (doi:10.3390/tropicalmed7120424)

23. Iftikhar H, Khan M, Khan MS, Khan M. 2023 Short-term forecasting of monkeypox cases using a novel filtering and combining technique. *Diagnostics* **13**, 1923. (doi:10.3390/diagnostics13111923)

24. Qureshi M, Khan S, Bantan RAR, Daniyal M, Elgarhy M, Marzo RR, Lin Y. 2022 Modeling and forecasting monkeypox cases using stochastic models. *J. Clin. Med. Res.* **11**, 6555. (doi:10.3390/jcm11216555)

25. Yasmin F, Hassan MM, Zaman S, Aung ST, Karim A, Azam S. 2022 A forecasting prognosis of the monkeypox outbreak based on a comprehensive statistical and regression analysis. *Computation* **10**, 177. (doi:10.3390/computation10100177)

26. Pramanik A, Sultana S, Rahman MS. 2022 Time series analysis and forecasting of monkeypox disease using ARIMA and SARIMA model. In *2022 13th Int. Conf. on Computing Communication and Networking Technologies (ICCCNT)*, Kharagpur, India, vol. **13**, pp. 1–7, (doi:10.1109/ICCCNT54827.2022.9984345)

27. Iftikhar H, Daniyal M, Qureshi M, Tawaiah K, Ansah RK, Afriyie JK. 2023 A hybrid forecasting technique for infection and death from the mpox virus. *Digit. Health* **9**, 20552076231204748. (doi:10.1177/20552076231204748)

28. Brand SPC *et al*. 2023 The role of vaccination and public awareness in forecasts of mpox incidence in the United Kingdom. *Nat. Commun*. **14**, 4100. (doi:10.1038/s41467-023-38816-8)

29. Overton CE, Abbott S, Christie R, Cumming F, Day J, Jones O, Paton R, Turner C, Ward T. 2023 Nowcasting the 2022 mpox outbreak in England. *PLoS Comput. Biol.* **19**, e1011463. (doi:10.1371/journal.pcbi.1011463)

30. Centers for Disease Control and Prevention. 2023 U.S. mpox case trends reported to CDC. In *2022-2023 Outbreak Cases & Data*. https://www.cdc.gov/poxvirus/mpox/response/2022/mpx-trends.html.

31. Centers for Disease Control and Prevention. 2023 Case definitions for use in the 2022 mpox response. In *Mpox*. https://www.cdc.gov/poxvirus/mpox/clinicians/case-definition.html.

32. World Health Organization. 2022 Surveillance, case investigation and contact tracing for mpox (monkeypox): interim guidance. See https://www.who.int/publications/i/item/WHO-MPX-Surveillance-2022.4 (accessed 22 December 2022).

33. Chowell G *et al*. 2016 Using phenomenological models to characterize transmissibility and forecast patterns and final burden of zika epidemics. *PLoS Curr*. **8**, f14b2217c902f453d9320a43a35b9583. (doi:10.1371/currents.outbreaks.f14b2217c902f453d9320a43a35b9583)

34. Pell B, Kuang Y, Viboud C, Chowell G. 2018 Using phenomenological models for forecasting the 2015 Ebola challenge. *Epidemics* **22**, 62–70. (doi:10.1016/j.epidem.2016.11.002)

35. Roosa K, Lee Y, Luo R, Kirpich A, Rothenberg R, Hyman JM, Yan P, Chowell G. 2020 Real-time forecasts of the COVID-19 epidemic in China from February 5th to February 24th, 2020. *Infect. Dis. Model*. **5**, 256–263. (doi:10.1016/j.idm.2020.02.002)

36. Roosa K, Lee Y, Luo R, Kirpich A, Rothenberg R, Hyman JM, Yan P, Chowell G. 2020 Short-term forecasts of the COVID-19 epidemic in Guangdong and Zhejiang, China: February 13–23, 2020. *J. Clin. Med.* **9**, 596. (doi:10.3390/jcm9020596)

37. Hurvich CM, Tsai CL. 1989 Regression and time series model selection in small samples. *Biometrika* **76**, 297–307. (doi:10.1093/biomet/76.2.297)

38. Sugiura N. 1978 Further analysis of the data by Akaike's information criterion and the finite corrections. Commun. Stat. – Theory Methods **7**, 13–26. (doi:10.1080/03610927808827599)

39. Hastie T, Tibshirani R, Friedman J. 2009 *The elements of statistical learning*. New York, NY: Springer. (doi:10.1007/978-0-387-84858-7)

40. Chowell G, Rothenberg R, Roosa K, Tariq A, Hyman JM, Luo R. 2022 Sub-epidemic model forecasts during the first wave of the COVID-19 pandemic in the USA and European hotspots. In *Mathematics of public health*, pp. 85–137. Cham, Switzerland: Springer International Publishing. (doi:10.1007/978-3-030-85053-1)

41. Chowell G, Tariq A, Dahal S, Bleichrodt A, Luo R, Hyman JM. 2024 SpatialWavePredict: a tutorial-based primer and toolbox for forecasting growth trajectories using the ensemble spatial wave sub-epidemic modeling framework. *BMC Med. Res. Methodol.* **24**, 131. (doi:10.1186/s12874-024-02241-2)

42. Mondal P, Shit L, Goswami S. 2014 Study of effectiveness of time series modeling (ARIMA) in forecasting stock prices. *IJCSEA* **4**, 13–29. (doi:10.5121/ijcsea.2014.4202)

43. Dimri T, Ahmad S, Sharif M. 2020 Time series analysis of climate variables using seasonal ARIMA approach. *J. Earth Syst. Sci.* **129**, 149. (doi:10.1007/s12040-020-01408-x)

44. Tektaş M. 2010 Weather forecasting using ANFIS and ARIMA MODELS: case study for Istanbul. *EREM* **51**, 5–10. (doi:10.5755/j01.erem.51.1.58)

45. Aditya Satrio CB, Darmawan W, Nadia BU, Hanafiah N. 2021 Time series analysis and forecasting of coronavirus disease in Indonesia using ARIMA model and PROPHET. *Procedia Comput. Sci*. **179**, 524–532. (doi:10.1016/j.procs.2021.01.036)

46. Hyndman R. *RDocumentation*. auto.ARIMA: fit best ARIMA model to univariate time series. See https://www.rdocumentation.org/packages/forecast/versions/8.21.1/topics/auto.arima.

47. forecast. *RDocumentation*. Forecasting time series. See https://www.rdocumentation.org/packages/forecast/versions/8.4/topics/forecast.

48. Bleichrodt A, Phan A, Luo R, Kirpich A, Chowell-Puente G. 2024 StatModPredict: a user-friendly R-shiny interface for fitting and forecasting with statistical models. *SSRN*. (doi:10.2139/ssrn.4849702)

49. Shafi A. 2021 *Medium*. What are generalized additive models?. See https://towardsdatascience.com/generalised-additive-models-6dfbedf1350a.

50. Wood SN. 2017 *Generalized additive models: an introduction with R*, 2nd edn. Boca Raton, FL: CRC Press.

51. Wood SN. *RDocumentation*. gam: generalized additive models with integrated smoothness estimation. See https://www.rdocumentation.org/packages/mgcv/versions/1.9-1/topics/gam.

52. RDocumentation. predict: model predictions. See https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/predict.

53. Taylor SJ, Letham B. 2018 Forecasting at scale. *Am. Stat*. **72**, 37–45. (doi:10.1080/00031305.2017.1380080)

54. Battineni G, Chintalapudi N, Amenta F. 2020 Forecasting of COVID-19 epidemic size in four high hitting nations (USA, Brazil, India and Russia) by Fb-Prophet machine learning model. *Appl. Comput. Inform*. (doi:10.1108/ACI-09-2020-0059)

55. Kirpich A, Shishkin A, Thomas A, Weppelmann A, Perez Tcher- Nov P, Skums Y. 2022 Excess mortality in Belarus during the COVID-19 pandemic as the case study of a country with limited non-pharmaceutical interventions and limited reporting. *Sci. Rep.* **12**, 5475. (doi:10.1038/s41598-022-09345-z)

56. Taylor S. *RDocumentation*. prophet: Prophet forecaster. See https://www.rdocumentation.org/packages/prophet/versions/1.0/topics/prophet.

57. Hyndman RJ, Athanasopoulos G. 2018 *Forecasting: principles and practice*. O.Texts. See https://otexts.com/fpp3.

58. Sarkar S. 2020 *Medium*. Time series forecasting and analysis: ARIMA and seasonal-ARIMA. See https://medium.com/analytics-vidhya/time-series-forecasting-and-analysis-arima-and-seasonal-arima-cacaf61ae863.

59. Arieli I, Babichenko Y, Smorodinsky R. 2018 Robust forecast aggregation. *Proc. Natl Acad. Sci. USA* **115**, E12135–E12143. (doi:10.1073/pnas.1813934115)

60. Johansson MA *et al*. 2019 An open challenge to advance probabilistic forecasting for dengue epidemics. *Proc. Natl Acad. Sci. USA* **116**, 24268–24274. (doi:10.1073/pnas.1909865116)

61. Ray EL *et al*. 2020 Ensemble forecasts of Coronavirus disease 2019 (COVID-19) in the U.S. *medRxiv*. (doi:10.1101/2020.08.19.20177493)

62. Reich NG *et al*. 2019 A collaborative multiyear, multimodel assessment of seasonal influenza forecasting in the United States. *Proc. Natl Acad. Sci. USA* **116**, 3146–3154. (doi:10.1073/pnas.1812594116)

63. Reich NG *et al*. 2022 Collaborative hubs: making the most of predictive epidemic modeling. *Am. J. Public Health* **112**, 839–842. (doi:10.2105/AJPH.2022.306831)

64. Reich NG *et al*. 2019 Accuracy of real-time multi-model ensemble forecasts for seasonal influenza in the U.S. *PLoS Comput. Biol.* **15**, e1007486. (doi:10.1371/journal.pcbi.1007486)

65. Viboud C *et al*. 2018 The RAPIDD ebola forecasting challenge: synthesis and lessons learnt. *Epidemics* **22**, 13–21. (doi:10.1016/j.epidem.2017.08.002)

66. Webb CT *et al*. 2017 Ensemble modelling and structured decision-making to support emergency disease management. *Prev. Vet. Med*. **138**, 124–133. (doi:10.1016/j.prevetmed.2017.01.003)

67. Cramer EY *et al*. 2022 Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the United States. *Proc. Natl Acad. Sci. USA* **119**, e2113561119. (doi:10.1073/pnas.2113561119)

68. Oidtman RJ *et al*. 2021 Trade-offs between individual and ensemble forecasts of an emerging infectious disease. *Nat. Commun.* **12**, 5379. (doi:10.1038/s41467-021-25695-0)

69. Spicknall IH *et al*. 2022 Modeling the impact of sexual networks in the transmission of monkeypox virus among gay, bisexual, and other men who have sex with men — United States, 2022. *MMWR. Morb. Mortal. Wkly. Rep.* **71**, 1131–1135. (doi:10.15585/mmwr.mm7135e2)

70. Bleicham. 2024 Evaluating-the-forecasting-performance-of-ensemble-sub-epidemic-frameworks-and-other-time-series. *Github*. See https://github.com/bleicham/Evaluating-the-forecasting-performance-of-ensemble-sub-epidemic-frameworks-and-other-time-series.

71. Bleicham. 2024 Bleicham/evaluating-the-forecasting-performance-of-ensemble-sub-epidemic-frameworks-and-other-time-series: Aluating-the-forecasting-performance-of-ensemble-sub-epidemic-frameworks-and-other-time-series. Zenodo. See https://zenodo.org/records/11504600.

72. Bleichrodt A, Luo R, Kirpich A, Chowell G. 2024 Supplementary material from: Evaluating the forecasting performance of ensemble sub-epidemic frameworks and other time series models for the 2022-2023 Mpox epidemic. Figshare. (doi:10.6084/m9.figshare.c.7296248)