A comparison of short-term probabilistic forecasts for the incidence of COVID-19 using mechanistic and statistical time series models

Nicolas Banholzer[1,2], Thomas Mellan[3], H Juliette T Unwin[3], Stefan Feuerriegel[4,5], Swapnil Mishra[6], and Samir Bhatt[3,7,*]

[1]ETH Zurich, Department of Management, Technology, and Economics, Zurich, Switzerland

[2]University of Bern, Institute of Social and Preventive Medicine, Bern, Switzerland

[3]Imperial College London, School of Public Health, London, United Kingdom

[4]LMU Munich, LMU Munich School of Management, Munich, Germany

[5]Munich Center for Machine Learning (MCML), Munich, Germany

[6]National University of Singapore and National University Health System, School of Public Health and Institute of Data Science, Singapore

[7]University of Copenhagen, Department of Public Health, Copenhagen, Denmark

[*]Corresponding author: samir.bhatt@sund.ku.dk

Abstract

Short-term forecasts of infectious disease spread are a critical component in risk evaluation and public health decision making. While different models for short-term forecasting have been developed, open questions about their relative performance remain. Here, we compare short-term probabilistic forecasts of popular mechanistic models based on the renewal equation with forecasts of statistical time series models. Our empirical comparison is based on data of the daily incidence of COVID-19 across six large US states over the first pandemic year. We find that, on average, probabilistic forecasts from statistical time series models are overall at least as accurate as forecasts from mechanistic models. Moreover, statistical time series models better capture volatility. Our findings suggest that domain knowledge, which is integrated into mechanistic models by making assumptions about disease dynamics, does not improve short-term forecasts of disease incidence. We note, however, that forecasting is often only one of many objectives and thus mechanistic models remain important, for example, to model the impact of vaccines or the emergence of new variants.

Keywords: probabilistic forecasting, mechanistic models, time series models, COVID-19

1 Introduction

Epidemic forecasting makes predictions about the development of observed quantities such as disease incidence that can inform health policy [1–6], e. g. by influencing decisions about resource allocation or the timing of public health interventions to control an outbreak. Forecasts can be made over long and short time horizons with different aims [7]. Long-term forecasts (months to years) often analyze hypothetical scenarios or, for example, how an epidemic develops with increased vaccination uptake or a new, more transmissible virus strain. By contrast, short-term forecasts (days to weeks) estimate what will happen in the near future. As a result, the performance of short-term forecasts is usually evaluated by comparing them with the observed outcome.

Different types of models can be used for epidemic forecasting, but these can be very broadly categorized into mechanistic and statistical time series models. *Mechanistic models* [8, 9] make assumptions about the data-generating process and are based on mathematically explicit expressions about disease dynamics. They establish a link between observed public health outcomes, such as the reported number of cases or deaths, and unobserved outcomes directly related to transmission, such as the number of infections or the instantaneous reproduction number. Formalizing these links requires disease-specific knowledge about epidemiological parameters such as the serial interval or incubation period. *Statistical time series models* [10–12] can be estimated without making assumptions about the data-generating process. Instead, these models detect statistical patterns in the observed time series, for example, through leveraging autocorrelation in the data. Since these patterns can be detected without domain knowledge, the statistical properties of these models are well characterized, and they can be used in a variety of forecasting applications.

Here, we compare short-term probabilistic forecasts between mechanistic and statistical time series models. Mechanistic models provide probabilistic forecasts by taking into account uncertainty about the epidemiological process. For example, mechanistic models use prior estimates of the generation or serial interval, which are subject to uncertainty, especially during early stages of an epidemic. Statistical time series models provide probabilistic forecasts via Bayesian inference where model parameters are considered random quantities to be integrated over. A comparison of the probabilistic forecasts from mechanistic and statistical time series models can inform about the added value of domain knowledge. If forecasts from mechanistic models were more accurate, this

would suggest that knowledge about disease dynamics and epidemiological parameters provide added value for short-term epidemic forecasting.

We compare the performance of short-term probabilistic forecasts from three statistical time series models (seasonal autoregressive integrated moving average (SARIMA) [10], Prophet [11, 13], Gaussian process (GP) [12]) with those from two popular mechanistic models based on the renewal equation (EpiEstim [8, 14] and EpiNow2 [9, 15]). We note that there are equivalences between mechanistic models based on renewal equations and the susceptible-infected-recovered model [16]. Our empirical comparison is based on data for the incidence of COVID-19 from six large US states for the first pandemic year between 15 March 2020 (the onset of the pandemic in the US) and 15 March 2021. Models are estimated repeatedly for each day using data from the last two months and, as in related work [7, 17, 18], probabilistic forecasts are made for the following 2 weeks. Forecasting performance is evaluated retrospectively by computing the difference between model-based forecasts and observed incidence.

2 Materials and methods

2.1 Data

We use daily data on the reported number of new cases of COVID-19 for six large US states (Arizona, California, Illinois, Maryland, New Jersey, and New York) from 19 January 2020 to 15 March 2021. Data are retrieved using the R package `covidcast` version 0.4.2 [19], which aggregates data on the daily number of confirmed cases of COVID-19 from the Johns Hopkins Coronavirus Resource Center [20] and computes the incidence (i. e. new cases per 100,000 people) using population data from the World Bank. Data are as of 1 January 2022, thus including revisions to real-time data. These revisions resulted in a negative number of new cases in a few instances, e. g. when a large number of duplicated records were removed. We replaced negative values (which are implausible) with the value from the day before during data preprocessing.

2.2 Forecasting task

The forecasting task is to make a daily 2-week forecast of the incidence of COVID-19 at day t in US state $s = 1, \ldots, 6$. Let $N_{s,t}$ denote the observed number of new cases of COVID-19 and let $\hat{N}_{s,t+1}$ to $\hat{N}_{s,t+14}$ denote the daily 2-week ahead forecasts of the number of new cases. Model-based forecasts

are generated for 51 weeks (i. e. 1 year) between 15 March 2020 and 15 March 2021. Forecasts are made on Sundays for the following 14 days starting on Monday. Each forecast is based on a model estimated using the historical time series up to day t. Specifically, the observed data two months prior to t (i. e. $N_{s,t-56}, \ldots, N_{s,t}$) are used as training data for model m to generate the probabilistic forecast $\hat{N}^{m,d}_{s,t+k}$, where $\hat{N}$ is the forecasted number of new cases and where $d = 1, \ldots, D$ denotes the draw (total number of D draws) from the posterior distribution $F^m_{s,t}$ of the model $m = 1, \ldots, M$ (total number of M models).

2.3 Summary of forecasting models

Table 1 shows a summary of the forecasting models considered in this work. The properties of each model and the implementation are described in greater detail in the following two sections. Note that we consider different modeling choices for each model by varying important model parameters or specifications. In the results, we first report for each model the modeling choice that achieved the best overall forecasting score. The models corresponding to these choices are then selected for comparison with each other.

2.4 Mechanistic models

We use the following mechanistic models based on the renewal equation: EpiEstim [8, 14] and EpiNow2 [9, 15].

(i) EpiEstim

EpiEstim builds upon the framework by Cori et al. [8] for estimating the time-varying reproduction number, which has been widely used for monitoring transmission of infectious diseases [5, 21–23]. The framework assumes that the number of new infections can be modeled with a Poisson renewal process where the number of newly infected individuals is proportional to the number of prevalent cases multiplied by their infectiousness [24]. Formally, the expected incidence on day t in state s is estimated as

$$E\left[I_{s,t}\right] = R_{s,t} \sum_{j=0}^{n} w_j I_{t-j,s}, \tag{1}$$

Table 1. Summary of probabilistic forecasting models and modeling choices.

Mechanistic models	
(i) EpiEstim	• Estimates the time-varying reproduction number over a fixed window τ. • Posterior distribution of reproduction number is derived analytically. • Assumes cases by date of infection with a constant time shift. • Uncertain generation time distribution. ⇒ Modeling choice: different values for τ.
(ii) EpiNow2	• Extension of EpiEstim. • Different methods to model the time-varying reproduction number. • Uncertain delay distributions for the time from infection to reporting. • Observational model with adjustment for weekday effects. ⇒ Modeling choice: different methods to model the time-varying reproduction number.
Statistical time series models	
(i) SARIMA	• Seasonal autoregressive moving average model. • Weekly differencing to account for weekday effects. ⇒ Modeling choice: different composition of the non-seasonal part of the model.
(ii) Prophet	• Additive time series model. • Trend component with linear growth model. • Growth rate changes at estimated change points determined by parameter τ. • Seasonal component for weekday effects. ⇒ Modeling choice: different values for τ.
(iii) GP	• Composition of kernels to model separate aspects of the time series. • Long-range squared exponential kernel for the trend. • Short-range squared exponential kernel for the short-term variation. • Periodic kernel for weekday effects. ⇒ Modeling choice: different length of the short-range kernel.

where $R_{s,t}$ is the time-varying reproduction number, w is the generation time (often assumed to be the serial interval distribution [25]), and $I_{t-j,s}$ are the number of new infections in the days $j = 0, \ldots, n$ before t. Cori et al. estimate $R_{s,t}$ with a Bayesian approach and derive the posterior distribution of $R_{s,t}$ analytically by assuming a conjugate Gamma prior distribution for $R_{s,t}$. The framework by Cori et al. is implemented in the R package `EpiEstim` (version 2.2-4) [14].

We apply EpiEstim by assuming a constant incubation period in which case the incidence of confirmed cases is exactly the incidence of infections, but shifted in time. We use an uncertain generation time distribution informed by prior estimates for the average (4.2, 95%-interval 3.3 to 5.3) and standard deviation (4.9, 95%-interval 3.0 to 8.3) of the generation time of SARS-CoV-2 [26]. Furthermore, we assume a Gamma prior for the basic reproduction number R_0 with a mean of 1.10 and standard deviation of 0.04 as in related work [27]. EpiEstim estimates the reproduction number $R_{s,t}$ over a time window $[t-\tau+1; t]$. The level of variation of the estimates for $R_{s,t}$ can be controlled by the parameter τ. Larger time windows (i. e. higher values for τ) reduce variation in the estimates of R_t and translate into changes in incidence that are more smooth. Empirically, we considered different modeling choices where we varied the time window from one day to two weeks.

EpiEstim returns the estimated mean and standard deviation of R_t in each time window. We use the mean and standard deviation from the last time window (i. e. the one right before the forecast is made) to compute the shape and scale parameter for the posterior Gamma distribution of $R_{s,t}$. Based on that, we sample posterior draws $R^1_{s,t}, \ldots, R^D_{s,t}$ for each sample of the generation time distribution, which are used together with the observed incidence $N_{s,0}, \ldots, N_{s,t}$ to project future incidence $\hat{N}^d_{s,t+1}$ to $\hat{N}^d_{s,t+21}$. Incidence is modeled with a Poisson distribution. Projections are made with the R package `projections` version 0.5.4 [28].

(ii) EpiNow2

EpiNow2 builds upon the framework by Abbott et al. [9] to estimate the time-varying reproduction number and has been used in recent work for forecasting the incidence of COVID-19 [7, 9, 29]. EpiNow2 uses a similar approach to EpiEstim, yet with multiple extensions. First, uncertainty about the estimated reproduction number can be modeled using, for example, a random walk or Gaussian process. Second, the latent number of new infections can be adjusted to account for the proportion of the population that is still susceptible to the virus. Third, EpiNow2 incorporates an observation

model where the expected number of new infections are mapped to the observed number of cases via uncertain delay distributions that adjust for the time from infection to reporting of a case. Fourth, the observed number of cases can be can be adjusted for weekday effects. The framework by Abbott et al. is implemented in the popular R package `EpiNow2` (version 1.3.2) [15].

We apply EpiNow2 by incorporating the observational model. Uncertain delay distributions are specified using prior estimates for the mean and standard deviation of the generation time [26] (same as for EpiEstim), the incubation period [30], and the reporting delay [31]. Furthermore, we assume a log-normal prior for the basic reproduction number R_0 parameterized to have the same mean and standard deviation as the Gamma distribution in EpiEstim. Empirically, we considered different models to estimate the time-varying reproduction number, using non-parametric backcalculation, a weekly random walk, or a Gaussian process parameterized as in related work [32]. In contrast to EpiEstim, observed cases are modeled with a negative binomial rather than a Poisson distribution, and are adjusted for weekday effects.

2.5 Statistical time series models

We use the following statistical time series models: a seasonal autoregressive integrated moving average (SARIMA) model [10], Prophet [11], and a Gaussian process (GP) [12]. Different to mechanistic models, we directly model the log incidence $Y_{s,t}$ (i. e. the log of the confirmed number of daily new cases per 100,000 people). Forecasts from statistical time series models are transformed to yield the number of number of new cases $N_{s,t}$ for comparison with the forecasts from mechanistic models.

(i) SARIMA

SARIMA models [10] have been used by various studies to forecast the incidence of COVID-19 [33–37]. They are a linear combination of two components: the autoregressive (AR) and the moving average (MA) component. In the AR, the outcome $Y_{s,t}$ only depends on the lagged outcome $Y_{s,t-j}$, where j denotes the time lag. In the MA, $Y_{s,t}$ only depends on the lagged errors $\epsilon_{s,t-j} = Y_{s,t-j} - Y_{s,t-j-1}$, i. e. the difference between the observed and estimated outcome.

A SARIMA$(p = 1,d = 0,q = 1)(0,1,0)_7$ with one AR and one MA component can be written as

$$\Delta Y_{s,t} = \alpha + \beta \Delta Y_{s,t-1} + \phi \Delta \epsilon_{s,t}, \tag{2}$$

where $\Delta Y_{s,t} = Y_{s,t} - Y_{s,t-7}$ and $\Delta \epsilon_{s,t} = \epsilon_{s,t} - \epsilon_{s,t-7}$, respectively. Note that we use weekly differences of the log incidence in the seasonal part of the model to account for weekday effects. Empirically, we considered different non-seasonal parts by varying the orders of the AR (p), MA (q), and differencing (d). We use weakly informative priors for the model intercept (α), parameters of the AR (β), and the MA (ϕ)

$$\alpha \sim \text{Normal}(0, 2.5, 6), \quad \beta \sim \text{Normal}(0, 0.5), \quad \phi \sim \text{Normal}(0, 0.5). \tag{3}$$

The model is estimated with the R package `bayesforecast` version 1.0.1 [38]

(ii) Prophet

Prophet [11] is a popular open source software in many forecasting applications and has been used by various studies to model the incidence of COVID-19 [35,39,40]. It is an additive model with three components modeling the trend of the time series, seasonality, and holiday effects (not considered here). The crucial component is the trend. It can be modeled assuming linear or logistic growth and the growth rate is allowed to change at times s_j, $j = 1, \ldots, J$. The changes in growth rates δ_j are modeled with a hierarchical prior $\delta_j \sim \text{Laplace}(0, \tau)$. The number of times the growth rate can change can be controlled by adjusting the scale of the prior (τ). Greater values of τ translate into more frequent changes of the trend.

We use Prophet assuming linear growth for the trend component and account for weekday effects in the seasonality component. Empirically, we considered different values for the scale of the prior for the number of change points in the range of $\tau \in [0.10, 0.45]$. The implementation is based on the R package `prophet` version 1.0 [13].

(iii) Gaussian process (GP)

GPs [12] have been used in epidemic forecasting [41–43]. A GP is a stochastic process or collection of random variables such that the joint distribution of every finite subset is multivariate normal (Gaussian) distributed. A GP is completely defined by a function for the mean and covariance (commonly called the kernel function). The mean function can be specified arbitrarily but is often set to zero after the data is transformed, e. g. centered. The crucial component is therefore the kernel, which characterises the correlation between pairs of input data. Prior information is reflected in the choice of the kernel used to describe the behavior of the stochastic process. A popular kernel

is the squared exponential (SE) kernel k_{SE} defined as

$$k_{\text{SE}}(t_i, t_j) = \alpha^2 \exp\left(-\frac{(t_i - t_j)^2}{2\rho^2}\right), \tag{4}$$

where α^2 is the signal noise parameter determining the scale of the output and ρ is the length scale (or bandwidth) parameter determining the correlation between input pairs, i. e. two time points (t_i, t_j). Larger α allow for greater deviations from the mean and larger ρ result in functions with rougher paths. Often one of the two parameters ρ and α is fixed due to weak identifiability [12, 44]. For time series analysis, seasonal patterns can be modeled using a periodic kernel (PE)

$$k_{\text{PE}}(t_i, t_j) = \alpha^2 \exp\left(-\frac{2\sin(\pi |t_i - t_j|/p)^2}{2\rho^2}\right), \tag{5}$$

where p is an additional parameter for the periodicity.

We model $Y_{s,t}$ with a zero mean GP using a composition of three kernels

$$Y_{s,t} \sim \text{GP}\left(0, k_{\text{SE}}^{\text{long}} + k_{\text{SE}}^{\text{short}} + k_{\text{PE}}^{\text{week}} + \sigma^2 \delta_{ij}\right), \tag{6}$$

where $k_{\text{SE}}^{\text{long}}$ is a SE kernel with a long length scale and $k_{\text{SE}}^{\text{short}}$ is SE kernel with short length scale, $k_{\text{PE}}^{\text{week}}$ is PE kernel for weekday effects, and $\sigma^2 \delta_{ij}$ is a regularizing term with a Kronecker delta function ensuring σ^2 Gaussian noise is only added when $i = j$. The SE kernels model the long-term trend and the short-term variation using a long- and short-range correlation structure [44].

Prior specifications for each kernel are as follows. For $k_{\text{SE}}^{\text{long}}$, we set the length scale to $\rho = 56$ days (the length of the training period) and use a weakly informative prior for $\alpha \sim \text{Student-t}(\nu = 3, \mu = 0, \sigma = 5)$. For $k_{\text{PE}}^{\text{week}}$, we set the length scale to $\rho = 1$, the periodicity to $p = 7$ days, and use $\alpha \sim \text{Student-t}(\nu = 5, \mu = 0, \sigma = 2)$. Empirically, we considered different length scales $\rho \in (1, 7, 14)$ days in the short-range kernel $k_{\text{SE}}^{\text{short}}$, allowing the GP to react more or less quickly to short-term variation. A weakly informative prior is given by $\alpha \sim \text{Student-t}(\nu = 7, \mu = 0, \sigma = 2)$. We implemented the GPs in the probabilistic programming language Stan version 2.21.0 [45].

2.6 Evaluation measures

We evaluate probabilistic forecasts over a 2-week time window. We aggregate daily forecasts by week to eliminate potential weekday effects, which is not accounted for in the EpiEstim model. The following evaluation measures are separately computed for the 1- and 2-week ahead forecast.

Forecasting performance is primarily assessed based on the continuous ranked probability score (CRPS) [46]. It measures the difference between the forecasted and the observed value and is a strictly proper scoring rule. It is defined as

$$CRPS(F_f, x_t) = \int \left[F(y_t) - \mathbf{1}(y_t \geqslant x_t)\right]^2 \mathrm{d}y_t, \tag{7}$$

where F is the cumulative distribution function of the forecast y for observation x at time t. Note that the CRPS is a generalization of the absolute error between the forecasted and observed value. A lower CRPS therefore indicates a more accurate forecast.

We follow recommendations to compute CRPS for the log of the forecasted incidence [47]. This corresponds to evaluating the multiplicative rather than the absolute forecasting error. Importantly, taking the log mitigates the influence of extreme outliers in the forecasts often made by models during an exponentially growing epidemic.

CRPS is computed for each forecast and scores are summarized by model, state, and epidemic phase (labeled manually, see S1 Appendix Fig. **??**). In addition, we perform pairwise model comparisons [48], thereby measuring the forecasting performance of each model relative to others. This is accomplished by first computing the ratio of scores. The relative skill is then computed by taking the geometric mean of all score ratios. Lower score ratios (and lower relative skill) indicate better forecasting performance in pairwise comparisons.

Secondary performance measures include sharpness, calibration, and bias. Sharpness is measured as the normalized median of the absolute deviation from the median of the predictive samples [49]. Calibration is assessed based on the density of the probability integral transform (PIT) values [46] as well as the empirical coverage of the 50 % and 95 % predictive interval (PI) [7]. Bias is evaluated with the average proportion of posterior draws that are above the observed value.

Primary and secondary performance measures are computed using the R packages `scoringRules` version 1.0.2 [50] and `scoringutils` [48] version 1.0.1.

An important quality of epidemic forecasts is to anticipate exponential growth. This quality is not captured specifically by the CRPS. Therefore, we consider an additional metric by McDonald et al. [51] for detecting so-called "hotspots". A hotspot at time t in state s is defined as a weekly increase in incidence of more than 25%, i. e.

$$H_{s,t} = \begin{cases} 1, & \text{if } (I_{s,t} - I_{s,t-7})/(I_{s,t-7}) \geqslant 0.25, \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

The probability for a hotspot predicted by model m is then equal to the proportion of posterior draws where the weekly change in the forecast is larger than 25%. Performance is assessed with the area under the receiving operating characteristic curve (AUC). Similar to McDonald et al. [51], we exclude data where the weekly incidence is below 70 new cases per 100,000 people. This removes 35% of the observations. Of the remaining observations, 25% are classified as hotspots.

2.7 Software and model estimation

All analyses are performed in the software R (version 4.2.0) [52]. Probabilistic models are estimated using Markov chain Monte Carlo (MCMC) sampling (except for EpiEstim where the posterior is derived analytical). Each model is estimated with 4 Markov chains and 1,000 iterations of which the first 500 iterations were discarded as part of the warm-up. This results in 2,000 posterior samples.

2.8 Data and code availability

Data are from publicly available data sources (Johns Hopkins Coronavirus Resource Center for epidemiological data; World Bank for population data). Preprocessed data files together with reproducible code is available from `https://github.com/nbanho/covid_predict`.

3 Results

The specific modeling choice has only a marginal influence on the overall forecasting performance of each model (S1 Appendix Fig. **??**). Table 2 shows the modeling choice with the highest overall CRPS for each model. The models corresponding to these choices are used for the subsequent between-model comparisons of mechanistic and statistical time series models.

Table 2. Modeling choices with highest overall CRPS. Results for all choices are shown in S1 Appendix Fig. **??**.

Model	Modeling choice with highest overall CRPS
Mechanistic models	
EpiEstim	Time-varying reproduction number estimated over a 7-day time window.
EpiNow2	Time-varying reproduction number modeled with a Gaussian process.
Statistical time series models	
SARIMA	SARIMA(1,0,1)$(0,1,0)_7$ model with one AR and one MA component.
Prophet	Prior scale $\tau = 0.45$ allowing for high number of change points for the growth rate.
GP	Short-range kernel with a length scale of $\rho = 7$ days.

The 1- and 2-week ahead forecasts by each model for Arizona are shown as examples in Fig. 1. The 1-week ahead forecast is well-calibrated across all models and covers the observed weekly incidence most of the time, except for the peak in January 2021 for SARIMA and Prophet. Coverage becomes worse for the 2-week ahead forecast, especially the peak in January 2021 is overestimated by most models and the subsequent decline is not anticipated. Forecasts from EpiNow2 are wider than those from EpiEstim. Similarly, forecasts from GP are wider than those from SARIMA and Prophet.

Results for secondary performance measures (calibration, sharpness, and bias) across states are shown in Fig. 2. The distribution of the probability integral transform for EpiEstim, SARIMA, and Prophet is U-shaped (Fig. 2a), indicating that the predictive distribution is too narrow, whereas it is hump-shaped for EpiNow2 and GP, indicating that the predictive distribution is too wide. GP and EpiNow2 achieve good coverage of the 50%-PI and the 95%-PI, whereas EpiEstim, SARIMA, and Prophet cover both PIs less often than theoretically predicted (Fig. 2b). Forecasts from EpiNow2 are least sharp, especially the 2-week ahead forecasts exhibit by far the highest dispersion (Fig. 2c). EpiEstim tends to overestimate and GP tends to underestimate the true incidence (Fig. 2d). Other models do not show a similar behavior in terms of bias.

CRPS scores are shown in Fig. 3. The distribution of scores are right-skewed for all models, indicating a considerable number of large forecasting errors (Fig. 3a). For the 1-week ahead forecast, the scores are comparable between models. Mean and median CRPS are lowest (best) for GP and highest for Prophet. The latter also has the highest CRPS for the 95% quantile. CRPS scores are more dispersed for the 2-week ahead forecasts, but the differences between models remain similar, albeit more visible. GP achieves the best scores, and the distribution is more concentrated on low

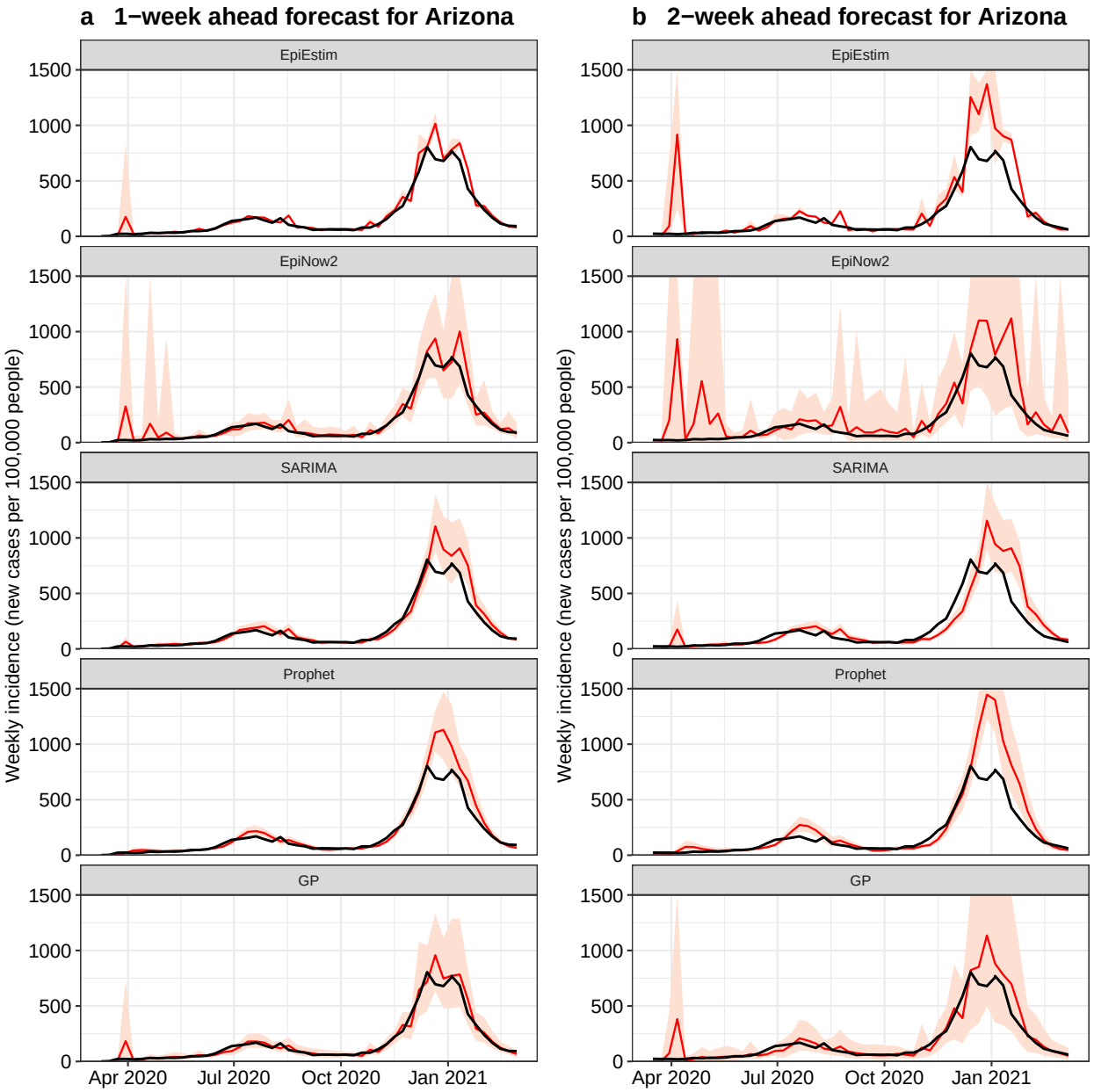


Fig 1. Probabilistic forecasts for Arizona by model. Probabilistic forecast (posterior mean as red line, 95%-PI as shaded area) for Arizona by each model. **(a)** 1-week ahead forecast. **(b)** 2-week ahead forecast. Observed incidence is shown with black lines. Probabilistic forecasts for all US states by model are shown in S1 Appendix **??-??**.

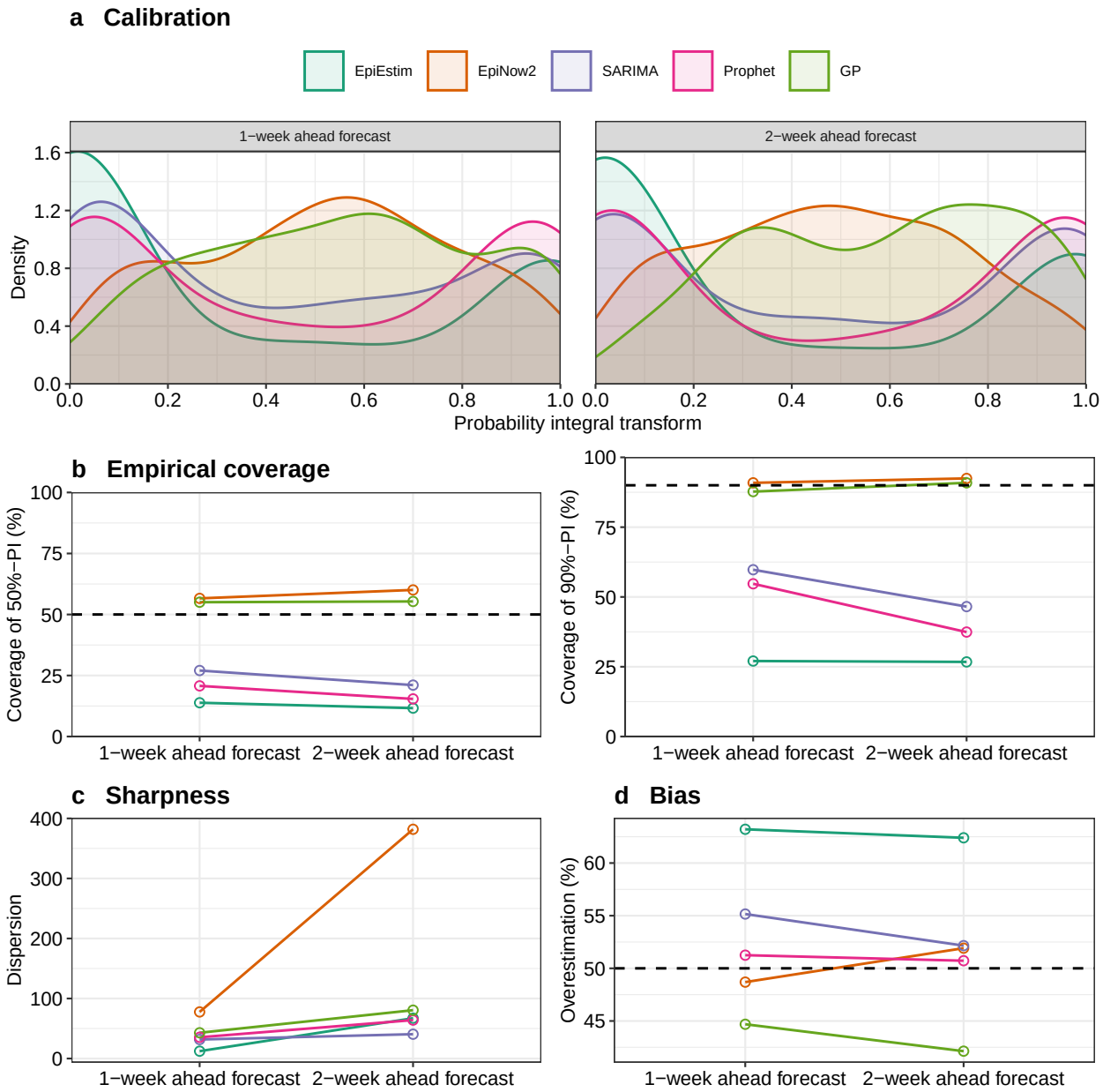


Fig 2. Evaluation of secondary performance measures overall. **(a)** Probabilistic calibration assessed with the density distribution of the probability integral transform. **(b-c)** Coverage of the 50 % (left panel) and 95 % (right panel) predictive interval [PI]. **(d)** Sharpness assessed with dispersion measured median of the absolute deviation from the median of the predictive samples. **(e)** Bias assessed with the average proportion of posterior draws above observed the incidence.

CRPS. By contrast, the distribution is more dispersed for Prophet and the CRPS for the 90% and 95% quantiles are higher. CRPS scores for SARIMA, EpiNow2, and EpiEstim are comparable and in between those of GP and Prophet. Differences in forecasting performance are confirmed by pairwise model comparisons. GP has the lowest (best) relative skill for both the 1- and 2-week ahead forecast (Fig. 3b) . Compared against other models, GP achieves the highest (best) score ratio, although the difference to SARIMA and EpiEstim are statistically not significant for the 2-week ahead forecast (Fig. 3c). In addition, GP has the lowest CRPS scores in all six US states (S1 Appendix Fig. **??**) and performs particularly well in phases of (sub)exponential decline (S1 Appendix Fig. **??**) because it anticipates sooner the peak and subsequent decline of incidence.

The CRPS evaluates the overall forecasting performance. Another important quality of forecasting models is to predict hotspots, i. e. large increases in weekly incidence. A comparison of the AUC for predicting hotspots shows that statistical time series models overall predict hotspots more accurately than mechanistic models (Fig. 4). Prophet has the highest AUC for both the 1- and 2-week ahead forecast, followed by GP and SARIMA.

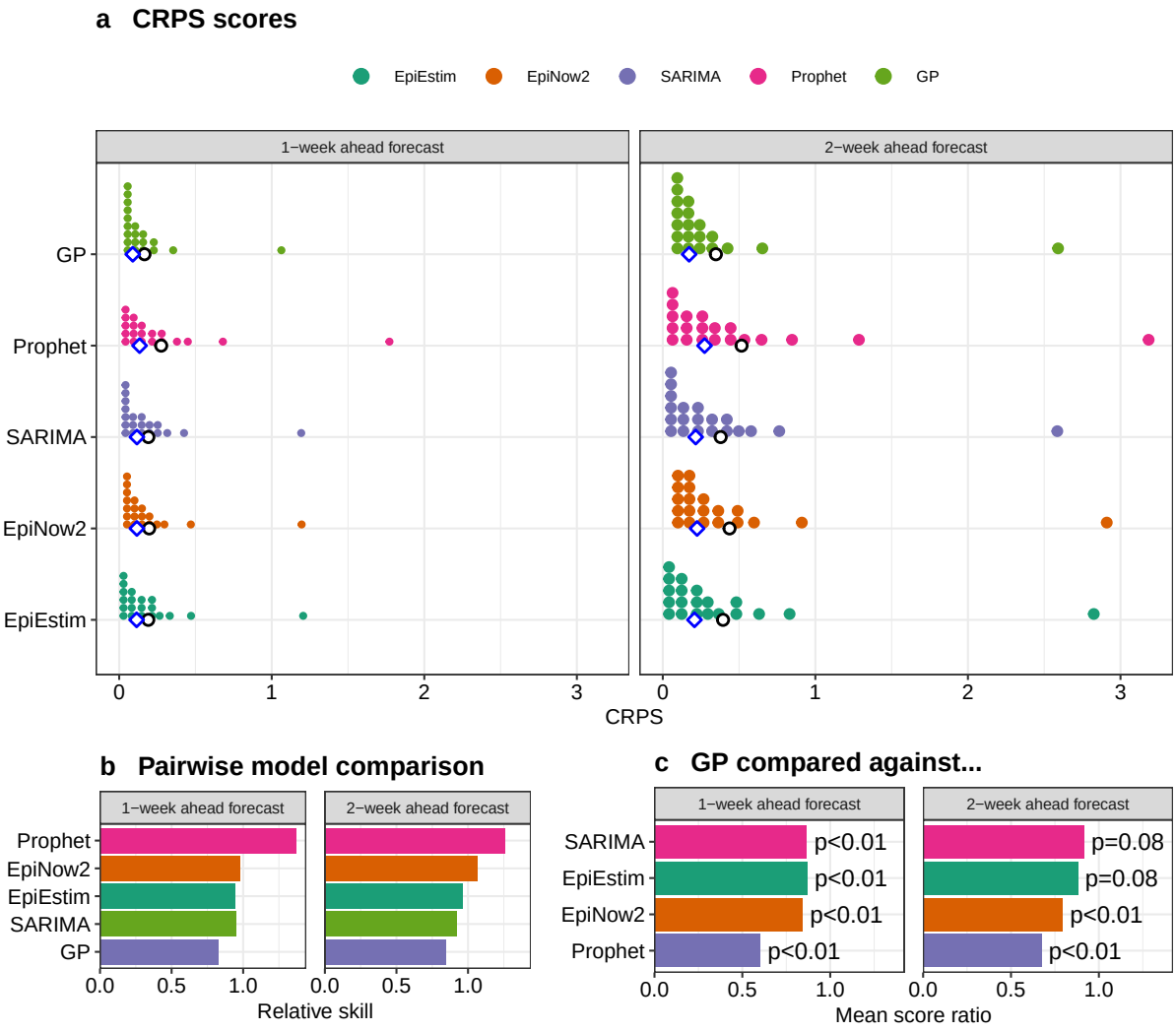
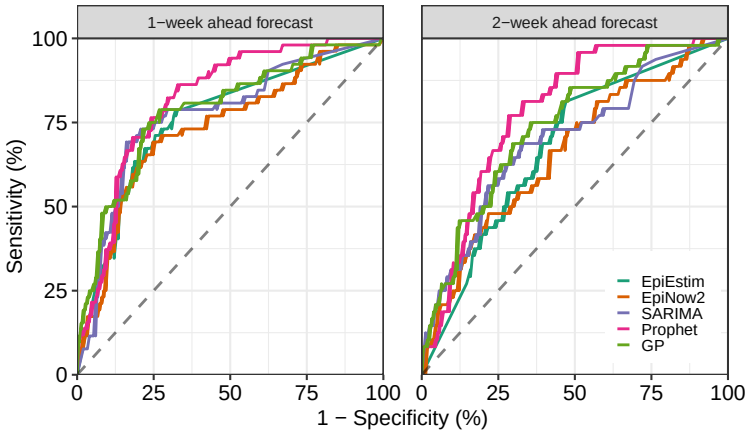


Fig 3. Evaluation of forecasting performance. **(a)** Empirical quantile distribution of continuous ranked probability score (CRPS) by model. Each dot represents one of the 20 quantiles of the empirical distribution. Median CRPS is shown with the blue square and mean CRPS with the black dot. Lower CRPS indicate better forecasting performance. **(b)** Relative skill for pairwise model comparisons according to CRPS scores. Lower values indicate better relative forecasting performance. **(c)** Mean score ratio for GP compared against other models. Score ratios closer to 1 indicate forecasting performances more similar to GP.



Model	1−week AUC	2−week AUC
EpiEstim	76	68
EpiNow2	73	66
SARIMA	77	70
Prophet	82	78
GP	79	74

Fig 4. Evaluation of hotspot prediction. Area under the receiver operating characteristic curve (AUC) for predicting hotspot, defined as a weekly increase in incidence of more than 25% to more than 10 new cases per 100,000 people). Higher AUC indicate greater ability to detect hotspots.

4 Discussion

4.1 Summary

We compared the short-term forecasts from mechanistic and statistical time series models for the incidence of COVID-19 in six large US states during the first year of pandemic. Forecasts from statistical time series models were at least as accurate as those from mechanistic models. GP achieved the best overall forecasting performance, but the difference to other models such as SARIMA and EpiEstim were not large, especially for the 2-week ahead forecast. The forecasts from GP and EpiNow2 were less sharp than those from other models (EpiEstim, SARIMA, and Prophet), but their empirical coverage was better. Finally, statistical time series models predicted large increases in incidence more accurately than mechanistic models.

4.2 Findings in context

Our work shows that statistical time series models achieve similar if not better short-term forecasting performance than mechanistic models. This is in line with previous work showing that statistical time series models with various known applications can also provide reasonably accurate forecasts for disease spread [35, 36, 41]. Nevertheless, epidemic forecasts remain difficult for both types of models. The reason is that forecasts are better understood as projections that are entirely based on the historical time series of the outcome. As such, these projections assume that nothing is changing

and that the historical trend continues into the future. However, there can be factors influencing disease spread that are not captured by the historical time series, e. g. voluntary behavioral changes or non-pharmaceutical interventions [25, 53, 54].

A related study compared human- and model-based forecasts for the cases and deaths of COVID-19 up to four weeks ahead [32]. Human-based forecasts were elicited using a web application where participants could select a predictive distribution (e. g. log-normal), and then visualize and interactively adjust their numerical forecasts. Individual forecasts were aggregated with a quantile-wise mean and the crowd forecasts were compared to the probabilistic forecasts from mechanistic models. The study showed that crowd forecasts were more accurate than forecasts from mechanistic models. In particular, mechanistic models often overestimated the incidence at the epidemic peaks. Similarly, we found in our study that the peak incidence was often overestimated by mechanistic models, but also statistical time series models. An explanation for why model-based forecasts overestimate the peak compared to human-based forecasts may be that only the latter can factor in the possible impact of non-pharmaceutical interventions that are implemented right at the time when forecasts have to be made.

Model-based forecasts of disease incidence could be further improved by using ensembles or including auxiliary indicators. Previous studies showed that the combined forecast of multiple models can be more accurate than the forecasts from each individual model [7, 55, 56], and that additional data streams (e. g. online surveys or mobility) can improve forecasting performance [51, 57–59]. Ensemble forecasts and auxiliary indicators were not the focus of this work, but we acknowledge their potential.

4.3 Limitations

Our study has limitations. First, we evaluated the multiplicative error of forecasts by comparing the log incidence. We chose the multiplicative error as it is less sensitive to outliers and ensures comparability across states and epidemic phases [32]. A drawback of the multiplicative error is that it is the same regardless of the absolute level of incidence. This can be unfavorable to models with large errors during phases of low incidence, which are less concerning. Second, we used the CRPS as our primary performance measure to evaluate forecasts. The CRPS is a proper scoring rule that is used across multiple domains to evaluate numerical forecasts [46]. However, policymakers may

evaluate forecasts differently using a utility function which, for example, penalizes underestimation of disease incidence more than overestimation. Third, although we compared models that are commonly used to forecast incidence, they still only represent a subset of all possible models. We therefore acknowledge that superior performance may be achieved with extensions of the models considered in this work or other types of models that are not considered in this study, e. g. compartmental models [60], gradient boosted trees [36], or deep learning [59]. Fourth, forecasts were made and analyzed retrospectively, which may affect their evaluation. By contrast, forecasts made in real-time are often based on incomplete incidence data [51] or forecasters may make manual adjustments before submission [7]. Fifth, we evaluated our forecasts based on the reported incidence of COVID-19. We therefore acknowledge that the actual number of cases of COVID-19 may be higher than reported number of cases [61], i. e. ascertainment bias is not addressed during our analysis. Sixth, we evaluated forecasts on the incidence of COVID-19 in six large US states over one year with carefully labeled epidemic phases. Future research may extend our analysis, e. g. considering different geographical regions, other epidemiological outcomes (e. g. deaths [7, 55] or hospitalizations [37]), and or even epidemics of other pathogens (e. g. influenza [1, 3]).

4.4 Concluding remarks and outlook

Mechanistic models incorporate knowledge about disease dynamics. Such domain knowledge is important, for example, when modeling the effects of vaccination uptake or the emergence of new virus variants. However, the added value for short-term forecasts of disease incidence is less clear. We empirically compared short-term probabilistic forecasts of disease from mechanistic and statistical time series models. We found that statistical time series models achieved similar overall performance than mechanistic models, suggesting that disease incidence can be predicted in the short term without modeling disease dynamics. Future research may evaluate whether forecasts can be improved further by other means such as considering ensemble models or auxiliary indicators.

References

1. Biggerstaff M, Johansson M, Alper D, Brooks LC, Chakraborty P, Farrow DC, et al. Results from the second year of a collaborative effort to forecast influenza seasons in the United States. Epidemics. 2018;24:26–33. doi:10.1016/j.epidem.2018.02.003.

2. Meltzer MI, Atkins CY, Santibanez S, Knust B, Petersen BW, Ervin ED, et al. Estimating the future number of cases in the Ebola epidemic – Liberia and Sierra Leone, 2014–2015. Morbidity and Mortality Weekly Report. 2014;63(3). Available from: https://stacks.cdc.gov/view/cdc/24901.

3. Bogoch II, Brady OJ, Kraemer MUG, German M, Creatore MI, Kulkarni MA, et al. Anticipating the international spread of Zika virus from Brazil. Lancet. 2016;387(10016):335–336. doi:10.1016/S0140-6736(16)00080-5.

4. Nsoesie E, Mararthe M, Brownstein J. Forecasting Peaks of Seasonal Influenza Epidemics. PLoS Curr. 2013;5. doi:10.1371/currents.outbreaks.bb1e879a23137022ea79a8c508b030bc.

5. Bhatia S, Parag KV, Wardle J, Imai N, Van Elsland SL, Lassmann B, et al.. Global predictions of short- to medium-term COVID-19 transmission trends: A retrospective assessmen. medRxiv [Preprint]; 2021 [cited 2022 March 3]. Available from: https://www.medrxiv.org/content/10.1101/2021.07.19.21260746.abstract. doi:10.1101/2021.07.19.21260746.

6. Massonnaud C, Roux J, Crépey P. COVID-19: Forecasting short term hospital needs in France. medRxiv [Preprint]; 2020 [cited 2022 March 3]. Available from: https://www.medrxiv.org/content/10.1101/2020.03.16.20036939v1. doi:10.1101/2020.03.16.20036939.

7. Bracher J, Wolffram D, Deuschel J, Görgen K, Ketterer JL, Ullrich A, et al. A pre-registered short-term forecasting study of COVID-19 in Germany and Poland during the second wave. Nat Commun. 2021;12:5173. doi:10.1038/s41467-021-25207-0.

8. Cori A, Ferguson NM, Fraser C, Cauchemez S. A new framework and software to estimate time-varying reproduction numbers during epidemics. Am J Epidemiol. 2013;178(9):1505–1512. doi:10.1093/aje/kwt133.

9. Abbott S, Hellewell J, Thompson RN, Sherratt K, Gibbs HP, Bosse NI, et al. Estimating the time-varying reproduction number of SARS-CoV-2 using national and subnational case counts. Welcome Open Research. 2020;(5):112. doi:10.12688/wellcomeopenres.16006.2.

10. Hyndman RJ, Athanasopoulos G. Chapter 8: ARIMA models. In: Forecasting: principles and practice. OTexts; 2018. p. 221–274.

11. Taylor SJ, Letham B. Forecasting at Scale. Am Stat. 2018;72(1):37–45. doi:10.1080/00031305.2017.1380080.

12. Rasmussen CE. Gaussian Processes for Machine Learning. MIT Press; 2006.

13. Taylor S, Letham B. Prophet: Automatic Forecasting Procedure; 2021. R package version 1.0. Available from: https://cran.r-project.org/web/packages/prophet.

14. Cori A, Kamvar Z, Stockwin J, Jombart T, Dahlqwist E, FitzJohn R, et al.. EpiEstim: A tool to estimate time varying instantaneous reproduction number during epidemics; 2021. R package version 2.2-4. Available from: https://cran.r-project.org/web/packages/EpiEstim.

15. Abbott S, Hellewell J, Sherrat K, Gostic KM, Hickson J, Badr HS, et al.. EpiNow2: Estimate real-time case counts and time-varying epidemiological parameters; 2021. R package version 1.3.2. Available from: https://cran.r-project.org/web/packages/EpiNow2.

16. Champredon D, Dushoff J, Earn DJD. Equivalence of the Erlang-distributed SEIR epidemic model and the renewal equation. SIAM J Appl Math. 2018;78(6):3258–3278.

17. Jewell NP, Lewnard JA, Jewell BL. Predictive mathematical models of the COVID-19 pandemic: Underlying principles and value of projections. JAMA. 2020;323(19):1893–1894. doi:10.1001/jama.2020.6585.

18. Krymova E, Béjar B, Thanou D, Sun T, Manetti E, Lee G, et al. Trend estimation and short-term forecasting of COVID-19 cases and deaths worldwide. Proc Natl Acad Sci U S A. 2022;119(32):e2112656119. doi:10.1073/pnas.2112656119.

19. Arnold T, Bien J, Brooks L, Colquhoun S, Farrow D, Grabman J, et al.. Covidcast: Client for delphi's COVIDcast epidata API; 2021. R package version 0.4.2. Available from: https://cran.r-project.org/web/packages/covidcast.

20. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect Dis. 2020;20(5):533–534. doi:10.1016/S1473-3099(20)30120-1.

21. Riccardo F, Ajelli M, Andrianou XD, Bella A, Manso MD, Fabiani M, et al. Epidemiological characteristics of COVID-19 cases and estimates of the reproductive numbers 1 month into the epidemic, Italy, 28 January to 31 March 2020. Euro Surveill. 2020;25(49):2000790. doi:10.2807/1560-7917.ES.2020.25.49.2000790.

22. Al Wahaibi A, Al Manji A, Al Maani A, Al Rawahi B, Al Harthy K, Alyaquobi F, et al. COVID-19 epidemic monitoring after non-pharmaceutical interventions: The use of time-varying reproduction number in a country with a large migrant population. Int J Infect Dis. 2020;99:466–472. doi:10.1016/j.ijid.2020.08.039.

23. Huisman JS, Scire J, Angst DC, Li J, Neher RA, Maathuis MH, et al. Estimation and worldwide monitoring of the effective reproductive number of SARS-CoV-2. eLife. 2022;11:e71345. doi:10.7554/eLife.71345.

24. Fraser C. Estimating individual and household reproduction numbers in an emerging epidemic. PLoS One. 2007;2(8):e758. doi:10.1371/journal.pone.0000758.

25. Flaxman S, Mishra S, Gandy A, Unwin HJT, Mellan TA, Coupland H, et al. Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. Nature. 2020;584(7820):257–261. doi:10.1038/s41586-020-2405-7.

26. Ganyani T, Kremer C, Chen D, Torneri A, Faes C, Wallinga J, et al. Estimating the generation interval for coronavirus disease (COVID-19) based on symptom onset data, March 2020. Euro Surveill. 2020;25(17):2000257. doi:10.2807/1560-7917.ES.2020.25.17.2000257.

27. Bosse NI, Abbott S, Bracher J, Hain H, Quilty BJ, Jit M, et al.. Comparing human and model-based forecasts of COVID-19 in Germany and Poland. medRxiv [Preprint]; 2021 [cited 2022 March 3]. Available from: https://www.medrxiv.org/content/10.1101/2021.12.01.21266598v1. doi:10.1101/2021.12.01.21266598.

28. Jombart T, Nouvellet P, Bhatia S, Kamvar ZN, Taylor T, Ghozzi S. Projections: Project Future Case incidence; 2021. R package version 0.5.4. Available from: https://cloud.r-project.org/web/packages/projections.

29. Davies NG, Abbott S, Barnard RC, Jarvis CI, Kucharski AJ, Munday JD, et al. Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. Science. 2021;372(6538):eabg3055. doi:10.1126/science.abg3055.

30. Lauer SA, Grantz KH, Bi Q, Jones FK, Zheng Q, Meredith HR, et al. The incubation period of Coronavirus Disease 2019 (COVID-19) from publicly reported confirmed cases: Estimation and application. Ann Intern Med. 2020;172(9):577–582. doi:10.7326/M20-0504.

31. Cereda D, Tirani M, Rovida F, Demicheli V, Ajelli M, Poletti P, et al.. The early phase of the COVID-19 outbreak in Lombardy, Italy. arXiv [Preprint]; 2020 [cited 2022 March 3]. Available from: https://arxiv.org/abs/2003.09320. doi:10.48550/ARXIV.2003.09320.

32. Bosse NI, Abbott S, Bracher J, Hain H, Quilty BJ, Jit M, et al. Comparing human and model-based forecasts of COVID-19 in Germany and Poland. PLoS Comput Biol. 2022;18(9):e1010405. doi:10.1371/journal.pcbi.1010405.

33. Alabdulrazzaq H, Alenezi MN, Rawajfih Y, Alghannam BA, Al-Hassan AA, Al-Anzi FS. On the accuracy of ARIMA based prediction of COVID-19 spread. Results Phys. 2021;27:104509. doi:10.1016/j.rinp.2021.104509.

34. Kufel T. ARIMA-based forecasting of the dynamics of confirmed COVID-19 cases for selected European countries. Equilibrium. 2020;15(2):181–204. doi:10.24136/eq.2020.009.

35. Wang Y, Yan Z, Wang D, Yang M, Li Z, Gong X, et al. Prediction and analysis of COVID-19 daily new cases and cumulative cases: times series forecasting and machine learning models. BMC Infect Dis. 2022;22(1):495. doi:10.1186/s12879-022-07472-6.

36. Rahman MS, Chowdhury AH, Amrin M. Accuracy comparison of ARIMA and XGBoost forecasting models in predicting the incidence of COVID-19 in Bangladesh. PLoS Global Public Health. 2022;2(5):e0000495. doi:10.1371/journal.pgph.0000495.

37. Perone G. Comparison of ARIMA, ETS, NNAR, TBATS and hybrid models to forecast the second wave of COVID-19 hospitalizations in Italy. Eur J Health Econ. 2022;23(6):917–940. doi:10.1007/s10198-021-01347-4.

38. Matamoros AA, Torres CC, Dala A, Hyndman R, O'Hara-Wild M. Bayesforecast: Bayesian Time Series Modeling with Stan; 2021. R package version 1.0.1. Available from: https:

//cran.r-project.org/web/packages/bayesforecast.

39. Battineni G, Chintalapudi N, Amenta F. 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. 2020;doi:10.1108/ACI-09-2020-0059.

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

41. Johnson LR, Gramacy RB, Cohen J, Mordecai E, Murdock C, Rohr J, et al. Phenomenological forecasting of disease incidence using heteroskedastic Gaussian processes: A dengue case study. Ann Appl Stat. 2018;12(1):27–66. doi:10.1214/17-AOAS1090.

42. Arias Velásquez RM, Mejía Lara JV. Forecast and evaluation of COVID-19 spreading in USA with reduced-space Gaussian process regression. Chaos Solitons Fractals. 2020;136:109924. doi:10.1016/j.chaos.2020.109924.

43. Qian Z, Alaa AM, van der Schaar M. In: Advances in Neural Information Processing Systems. vol. 33; 2020. p. 10729–10740. Available from: https://proceedings.neurips.cc/paper/2020/hash/79a3308b13cd31f096d8a4a34f96b66b-Abstract.html.

44. Hawryluk I, Hoeltgebaum H, Mishra S, Miscouridou X, Schnekenberg RP, Whittaker C, et al. Gaussian process nowcasting: Application to COVID-19 mortality reporting. Proceedings of the 37th Conference on Uncertainty in Artificial Intelligence. 2021;161:1258–1268. Available from: https://proceedings.mlr.press/v161/hawryluk21a.html.

45. Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, Betancourt M, et al. Stan: A probabilistic programming language. J Stat Softwb. 2017;76:1–32. doi:10.18637/jss.v076.i01.

46. Gneiting T, Balabdaoui F, Raftery AE. Probabilistic forecasts, calibration and sharpness. J R Stat Soc Series B. 2007;69(2):243–268. doi:10.1111/j.1467-9868.2007.00587.x.

47. Bosse NI, Abbott S, Cori A, van Leeuwen E, Bracher J, Funk S. Transformation of forecasts for evaluating predictive performance in an epidemiological context. medRxiv [Preprint]; 2023 [cited 2022 March 3]. Available from: https://doi.org/10.1101/2023.01.23.23284722. doi:10.1101/2023.01.23.23284722.

48. Bosse N, Abbott S, Gruson H, Bracher J, Funk S. Scoringutils: Utilities for Scoring and Assessing Predictions; 2022. R package version 1.0.1. Available from: https://cran.r-project.org/web/packages/scoringutils.

49. Funk S, Camacho A, Kucharski AJ, Lowe R, Eggo RM, Edmunds WJ. Assessing the performance of real-time epidemic forecasts: A case study of Ebola in the Western Area region of Sierra Leone, 2014-15. PLoS Comput Biol. 2019;15(2):e1006785. doi:10.1371/journal.pcbi.1006785.

50. Jordan A, Krueger F, Lerch S. ScoringRules: Scoring Rules for Parametric and Simulated Distribution Forecasts; 2022. R package version 1.0.2. Available from: https://cran.r-project.org/web/packages/scoringRules.

51. McDonald DJ, Bien J, Green A, Hu AJ, DeFries N, Hyun S, et al. Can auxiliary indicators improve COVID-19 forecasting and hotspot prediction? Proc Natl Acad Sci U S A. 2021;118(51):e2111453118. doi:10.1073/pnas.2111453118.

52. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing; 2022. Version 4.2.1. Available from: https://www.R-project.org/.

53. Brauner JM, Mindermann S, Sharma M, Johnston D, Salvatier J, Gavenčiak T, et al. Inferring the effectiveness of government interventions against COVID-19. Science. 2021;371(6531):eabd9338. doi:10.1126/science.abd9338.

54. Banholzer N, Weenen Ev, Lison A, Cenedese A, Seeliger A, Kratzwald B, et al. Estimating the effects of non-pharmaceutical interventions on the number of new infections with COVID-19 during the first epidemic wave. PLoS One. 2021;16(6):e0252827. doi:10.1371/journal.pone.0252827.

55. Cramer EY, Ray EL, Lopez VK, Bracher J, Brennen A, Castro Rivadeneira AJ, et al. Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the United States. Proc Natl Acad Sci U S A. 2022;119(15):e2113561119. doi:10.1073/pnas.2113561119.

56. Sherratt K, Gruson H, Grah R, Johnson H, Niehus R, Prasse B, et al.. Predictive performance of multi-model ensemble forecasts of COVID-19 across European nations. medRxiv [Preprint]; 2022 [cited 2022 March 3]. Available from: https://www.medrxiv.org/content/10.1101/2022.06.16.22276024v1. doi:10.1101/2022.06.16.22276024.

57. Reinhart A, Brooks L, Jahja M, Rumack A, Tang J, Agrawal S, et al. An open repository of real-time COVID-19 indicators. Proc Natl Acad Sci U S A. 2021;118(51):e2111452118. doi:10.1073/pnas.2111452118.

58. Schwabe A, Persson J, Feuerriegel S. Predicting COVID-19 spread from large-scale mobility data. In: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining; 2021. p. 3531–3539. Available from: https://doi.org/10.1145/3447548.3467157. doi:10.1145/3447548.3467157.

59. Rodríguez A, Tabassum A, Cui J, Xie J, Ho J, Agarwal P, et al. DeepCOVID: An Operational Deep Learning-driven Framework for Explainable Real-time COVID-19 Forecasting. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35; 2021. p. 15393–15400. Available from: https://ojs.aaai.org/index.php/AAAI/article/view/17808. doi:10.1609/aaai.v35i17.17808.

60. Reiner RC, Barber RM, Collins JK, Zheng P, Adolph C, Albright J, et al. Modeling COVID-19 scenarios for the United States. Nat Med. 2021;27(1):94–105. doi:10.1038/s41591-020-1132-9.

61. Noh J, Danuser G. Estimation of the fraction of COVID-19 infected people in U.S. states and countries worldwide. PLoS One. 2021;16(2):e0246772. doi:10.1371/journal.pone.0246772.

Acknowledgments

We thank Anne Cori for her support with the implementation of EpiEstim. We thank Nikos Bosse for his suggestions on evaluating probabilistic forecasts.

Ethics approval

Ethics approval was not required for this study.

Competing interests

SF reports grants from the Swiss National Science Foundation outside of the submitted work. All other authors declare no competing interests.

Funding

NB and SF acknowledge funding from the Swiss National Science Foundation (SNSF) as part of the Eccellenza grant 186932 on "Data-driven health management". TM, SM, HJTU and SB acknowledge funding from the MRC Centre for Global Infectious Disease Analysis (reference MR/R015600/1), jointly funded by the UK Medical Research Council (MRC) and the UK Foreign, Commonwealth and Development Office (FCDO), under the MRC/FCDO Concordat agreement and is also part of the EDCTP2 programme supported by the European Union; and acknowledges funding by Community Jameel. The funding bodies had no control over design, conduct, data, analysis, review, reporting, or interpretation of the research conducted. SB is funded by the National Institute for Health Research (NIHR) Health Protection Research Unit in Modelling and Health Economics, a partnership between the UK Health Security Agency, Imperial College London and LSHTM (grant code NIHR200908). Disclaimer: “The views expressed are those of the author(s) and not necessarily those of the NIHR, UK Health Security Agency or the Department of Health and Social Care.” SB acknowledges financial support from the Novo Nordisk Foundation via The Novo Nordisk Young Investigator Award (NNF20OC0059309) which also supports SM. SB acknowledges financial support from the Danish National Research Foundation via a chair grant. SB acknowledges financial support from The Eric and Wendy Schmidt Fund For Strategic Innovation via the Schmidt Polymath Award (G-22-63345).

Author Contributions

NB, TM, SM, and SB conceptualized the study. NB collected and analyzed the data. NB, TM, SM, and SB contributed to methodology. All authors contributed to interpretation and visualization of data and results. NB wrote the original draft. All authors contributed to writing – review & editing.

Supporting information

S1 Appendix. Supplementary material. Fig S1. Labeling of the epidemic trajectories. Fig S2. Forecasting performance for different modeling choices. Fig S3. Probabilistic forecasts for Arizona by model. Fig S4. Probabilistic forecasts for Illinois by model. Fig S5. Probabilistic forecasts for Maryland by model. Fig S6. Probabilistic forecasts for New Jersey by model. Fig S7. Probabilistic forecasts for New York by model. Fig S8. Evaluation of forecasting performance by US state. Fig S9. Evaluation of forecasting performance by epidemic phase.

S1 Appendix

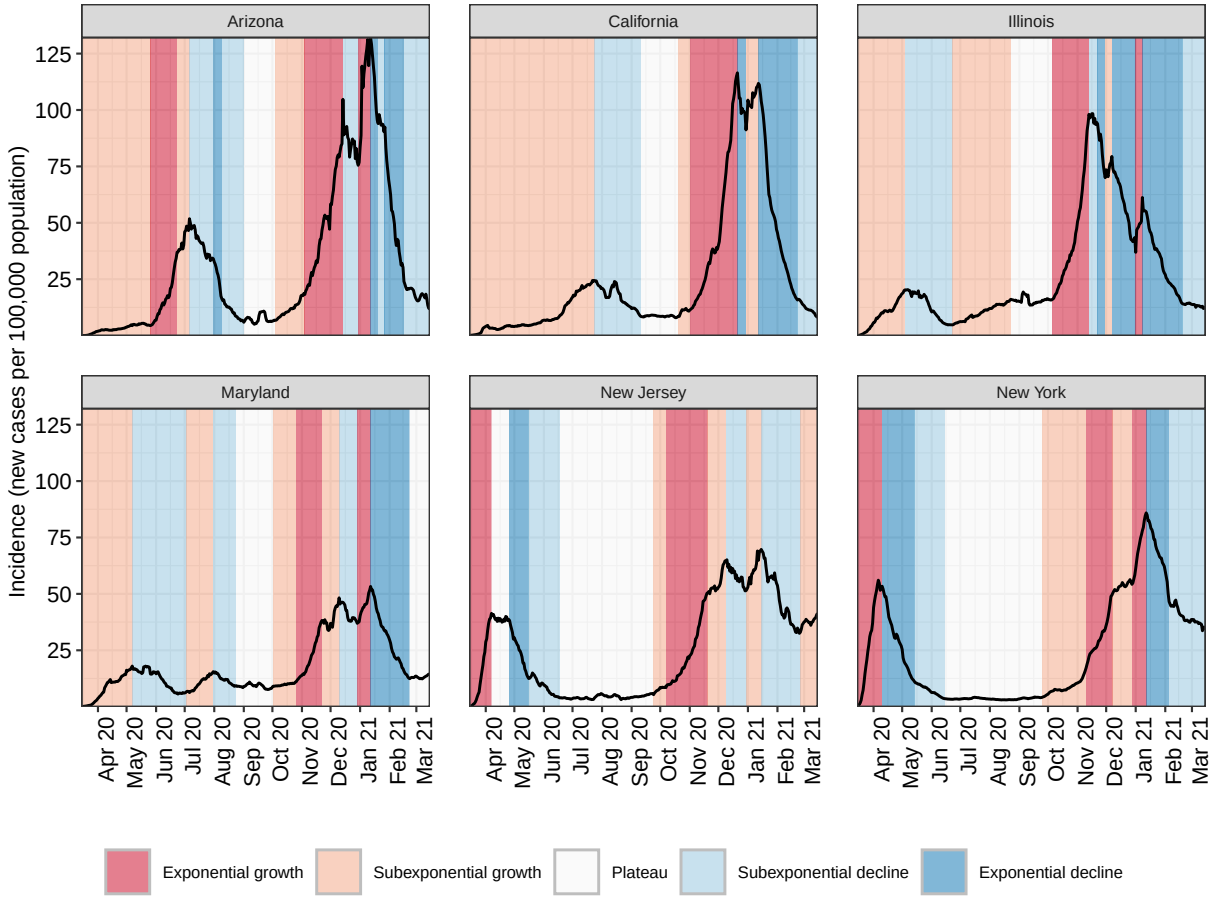


Fig S1. Labeling of the epidemic trajectories. Epidemic trajectories in US states manually labeled according to five phases: exponential increase, subexponential increase, plateau, subexponential decline, exponential decline.

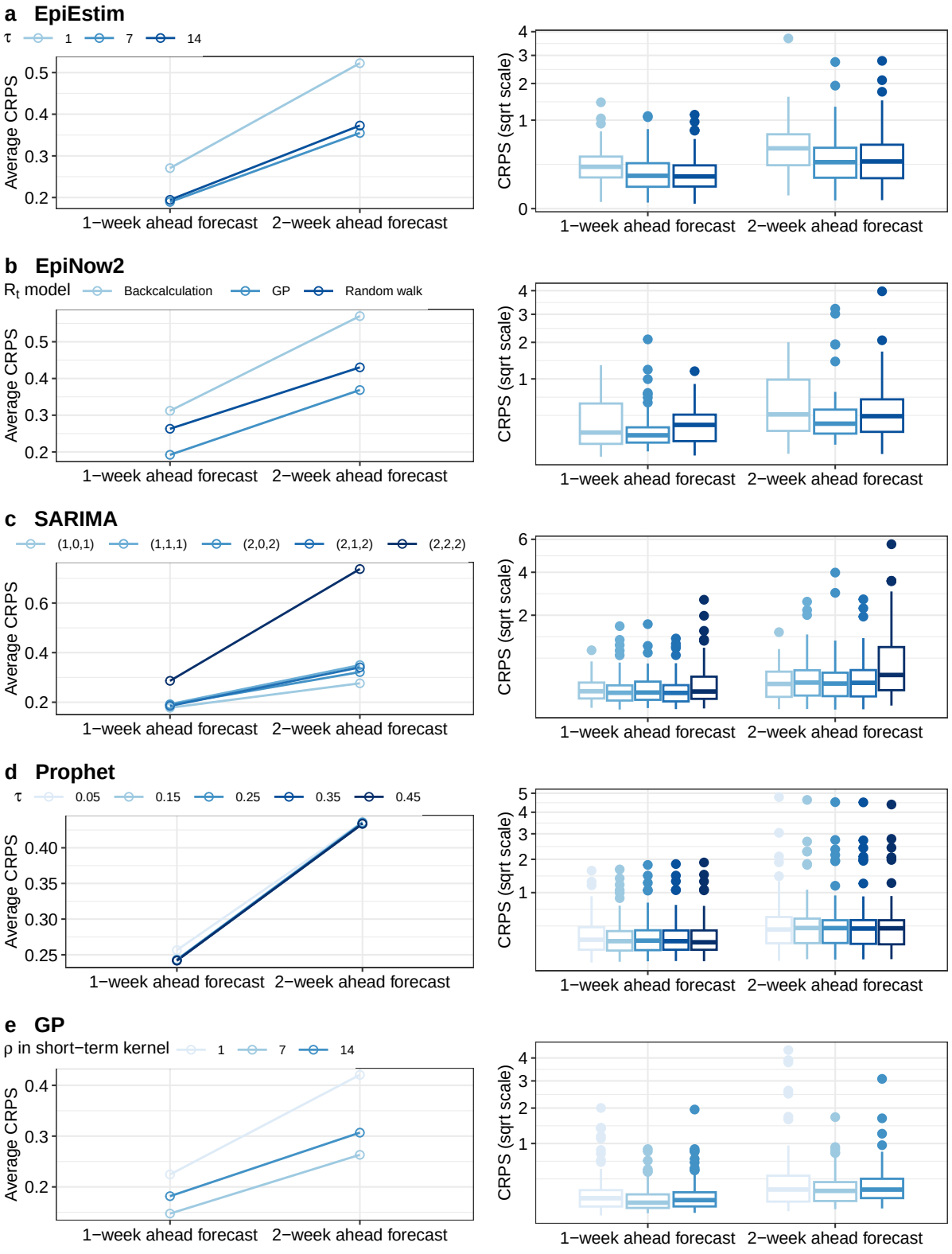


Fig S2. Forecasting performance for different modeling choices. CRPS (left average scores, right boxplot of scores) for the 1- and 2-week ahead forecast by model specification. **(a)** EpiEstim. **(b)** EpiNow2. **(c)** SARIMA. **(d)** Prophet. **(e)** GP.

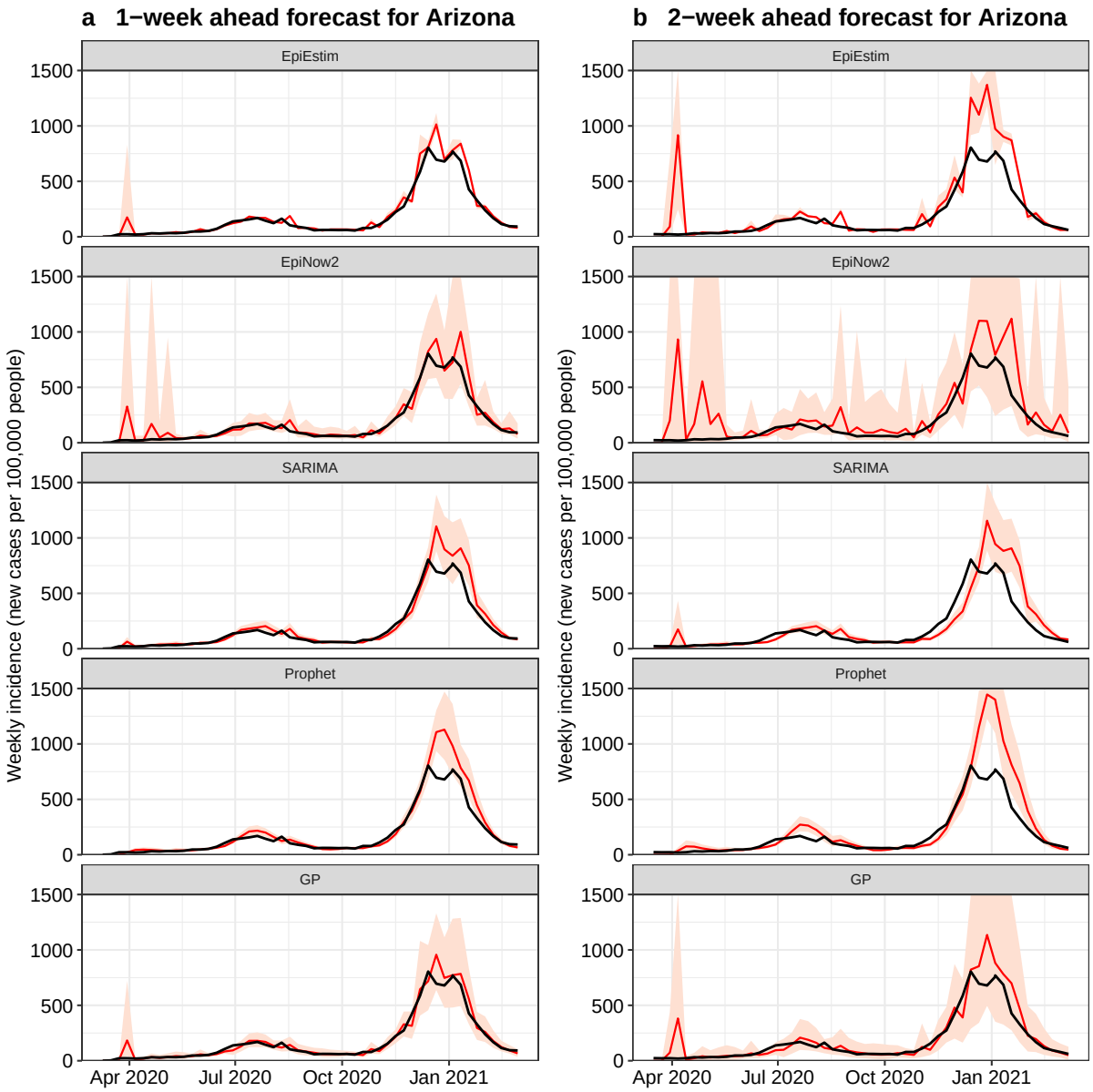


Fig S3. Probabilistic forecasts for Arizona by model. Probabilistic forecast (posterior mean as red line, 95%-PI as shaded area) for Arizona by each model. **(a)** 1-week ahead forecast. **(b)** 2-week ahead forecast. Observed incidence is shown with black lines.

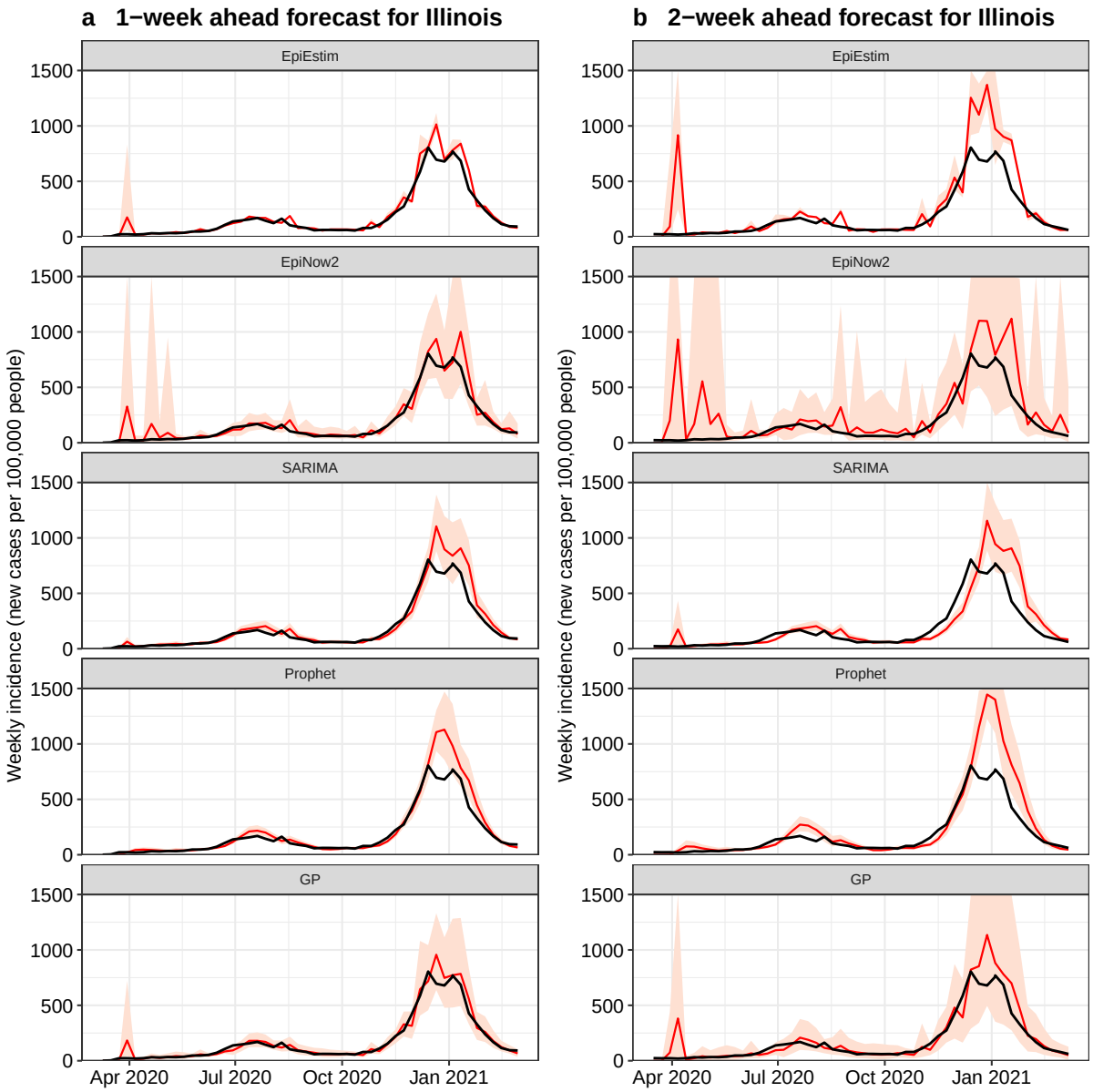


Fig S4. Probabilistic forecasts for Illinois by model. Probabilistic forecast (posterior mean as red line, 95%-PI as shaded area) for Illinois by each model. **(a)** 1-week ahead forecast. **(b)** 2-week ahead forecast. Observed incidence is shown with black lines.

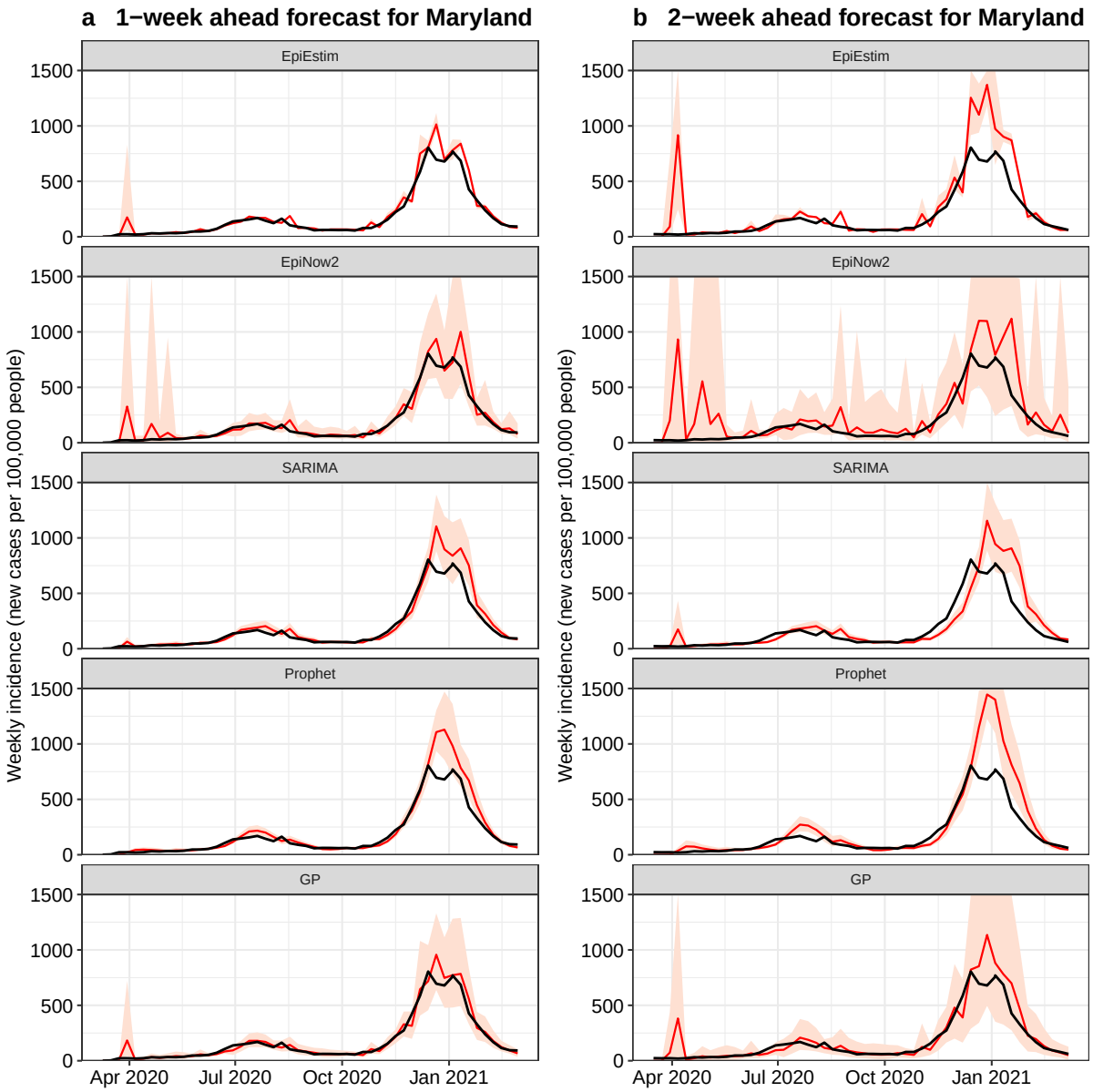


Fig S5. Probabilistic forecasts for Maryland by model. Probabilistic forecast (posterior mean as red line, 95%-PI as shaded area) for Maryland by each model. **(a)** 1-week ahead forecast. **(b)** 2-week ahead forecast. Observed incidence is shown with black lines.

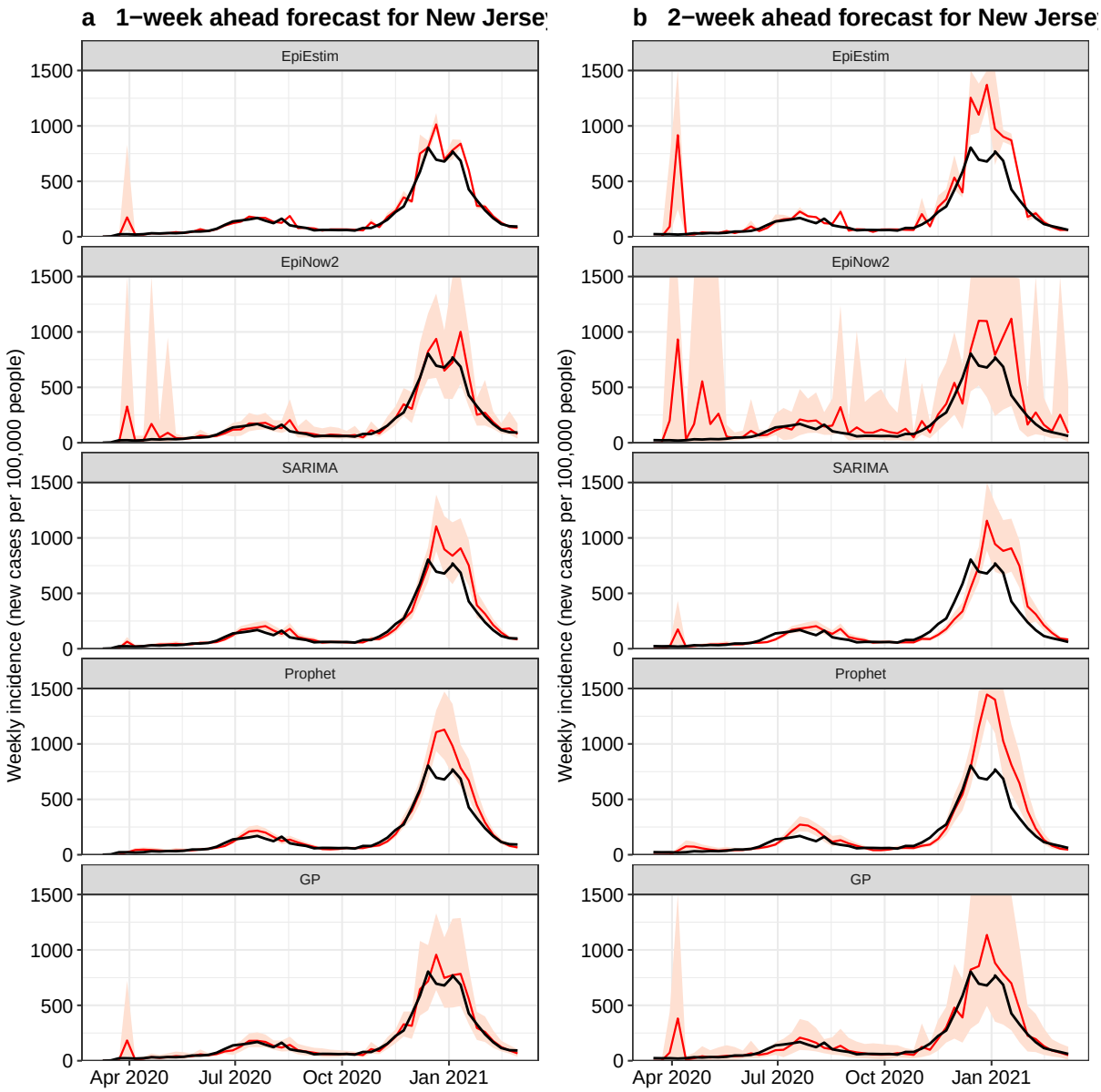


Fig S6. Probabilistic forecasts for New Jersey by model. Probabilistic forecast (posterior mean as red line, 95%-PI as shaded area) for New Jersey by each model. **(a)** 1-week ahead forecast. **(b)** 2-week ahead forecast. Observed incidence is shown with black lines.

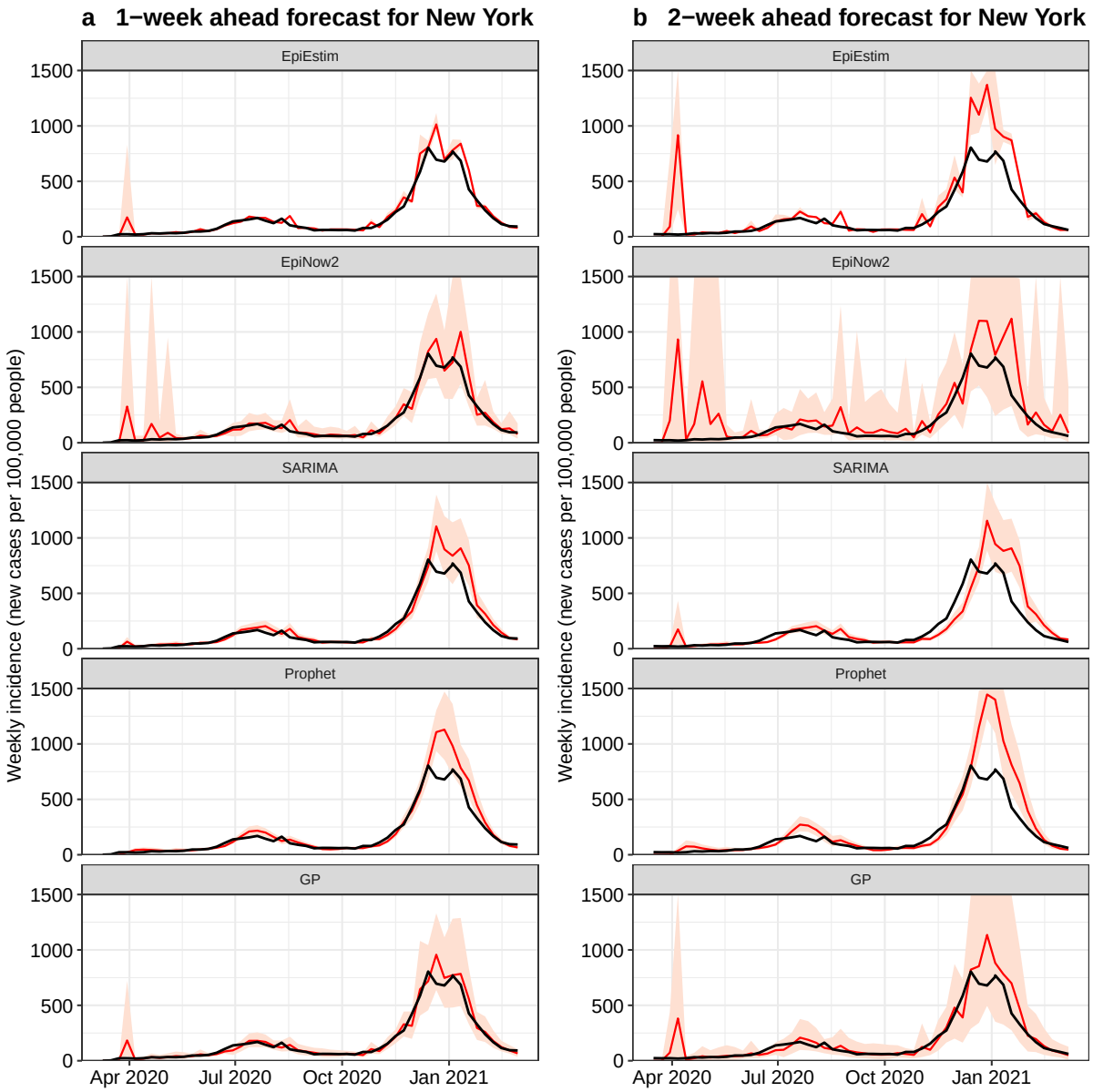


Fig S7. Probabilistic forecasts for New York by model. Probabilistic forecast (posterior mean as red line, 95%-PI as shaded area) for New York by each model. **(a)** 1-week ahead forecast. **(b)** 2-week ahead forecast. Observed incidence is shown with black lines.

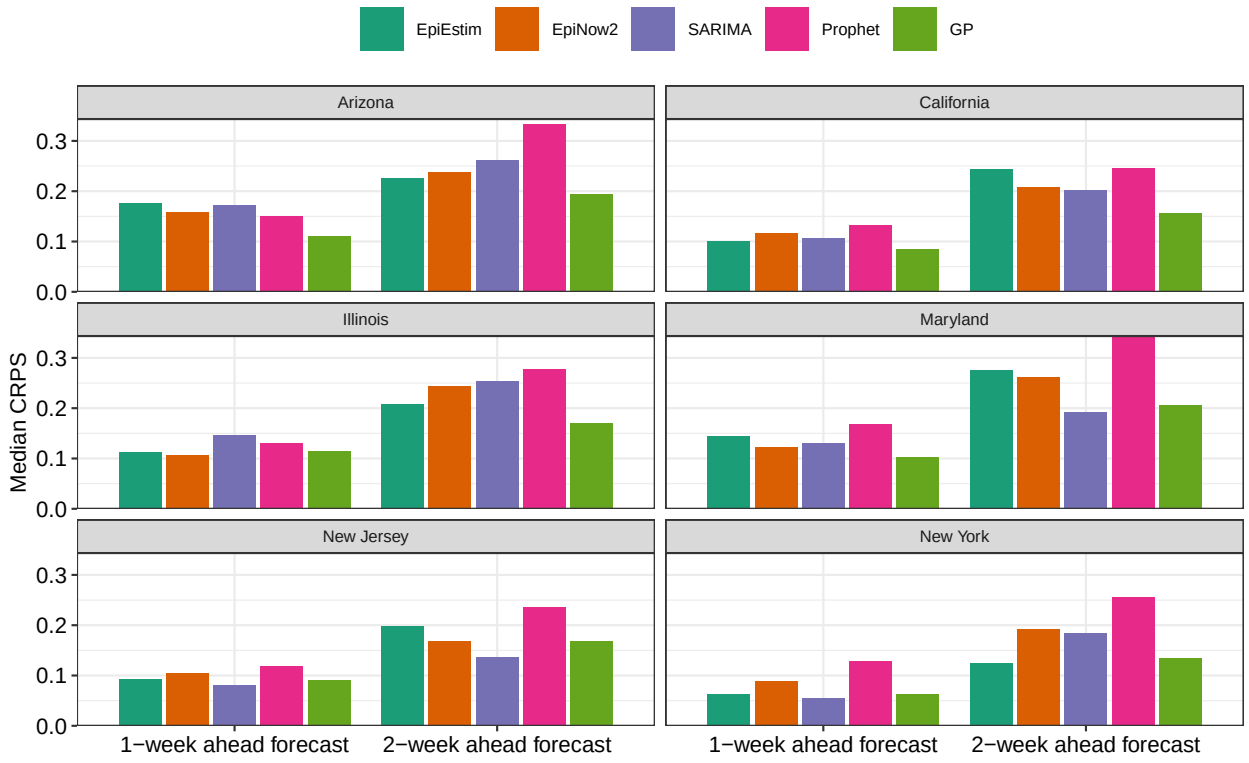


Fig S8. Evaluation of forecasting performance by US state. Median CRPS by US state for the 1- and 2-week ahead forecast by model.

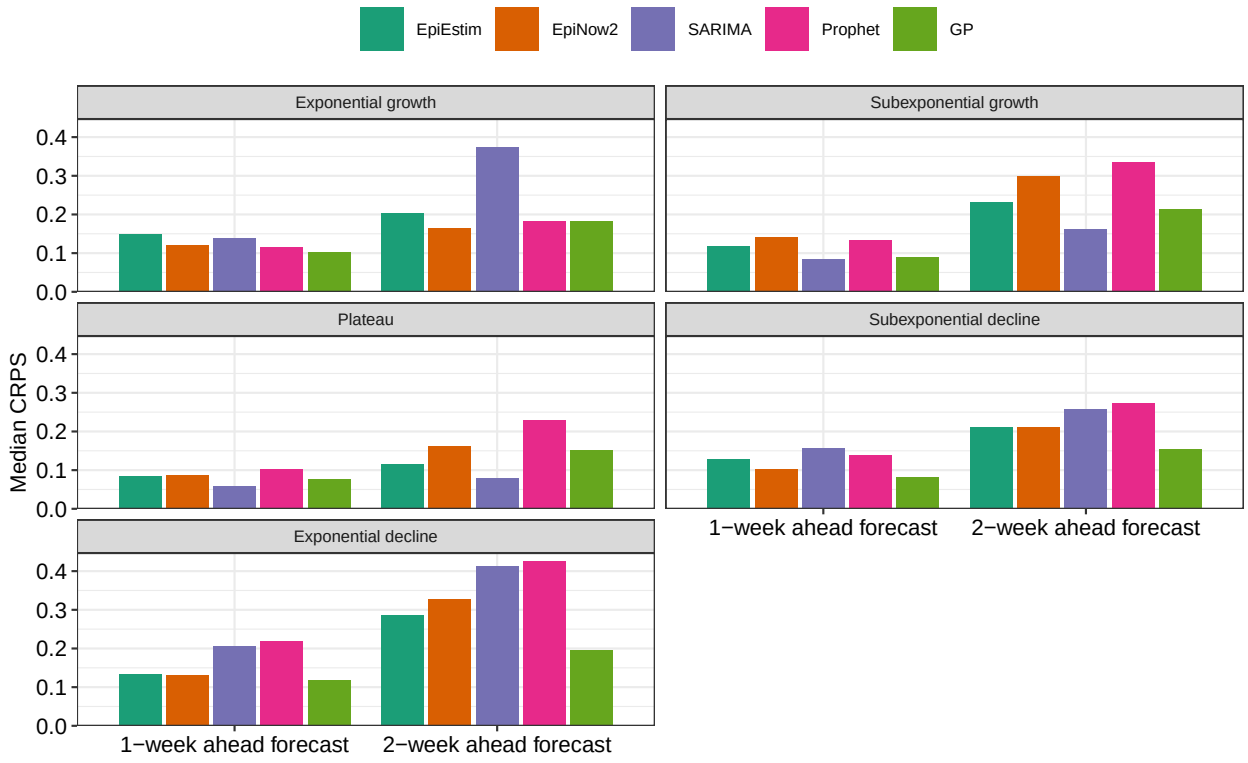


Fig S9. Evaluation of forecasting performance by epidemic phase. Median CRPS by different epidemic phases (see labeling in Figure S1) for the 1- and 2-week ahead forecast by model.