Sensible Long-lead Forecast of COVID-19 Epidemic Outcomes

Wan Yang[1,2] and Jeffrey Shaman[3,4]

[1]Department of Epidemiology, Mailman School of Public Health, Columbia University, New York, NY, USA; [2]Herbert Irving Comprehensive Cancer Center, Columbia University Medical Center, New York, NY, USA; [3]Department of Environmental Health Sciences, Mailman School of Public Health, Columbia University, New York, NY, USA; Columbia Climate School, Columbia University, New York, NY, USA

Correspondence to: wy2202@cumc.columbia.edu

Short title: Long-lead COVID-19 forecast

One sentence summary: To support more proactive planning, we propose approaches that substantially improve long-lead COVID-19 forecast accuracy.

Abstract

Coronavirus disease 2019 (COVID-19) will likely remain a major public health burden; accurate forecast of COVID-19 epidemic outcomes several months into the future is needed to support more proactive planning. Here, we propose strategies to address three major forecast challenges, i.e., error growth, the emergence of new variants, and infection seasonality. Using these strategies in combination we generate retrospective predictions of COVID-19 cases and mortality 6 months in the future for 10 representative US states during July 2020 to September 2022. The optimized forecast approach using all three strategies consistently outperformed a baseline forecast approach across different variant waves and locations, for all forecast targets; overall, probabilistic forecast accuracy improved by 64% and 38% and point prediction accuracy by 133% and 87% for cases and deaths, respectively. Given this superior forecast skill, the strategies could support sensible long-lead forecast of COVID-19 and possibly other infectious diseases.

Keywords: COVID-19; long-term forecast; error growth; new variants; infection seasonality

INTRODUCTION

The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) emerged in late 2019, causing the coronavirus disease 2019 (COVID-19) pandemic. Since its onset, mathematical modeling has been widely applied to generate projections of potential pandemic trajectories, including for cases, hospitalizations, and deaths. These model projections are often based on specific assumptions (i.e., scenarios) regarding critical factors affecting transmission dynamics. For example, the scenarios often include a combination of public health policies [e.g., non-pharmaceutical interventions (NPIs) including lockdown/reopening and masking, and

vaccinations], population behavior (e.g., adherence to the policies and voluntary preventive measures), and anticipated changes in the epidemiological properties of SARS-CoV-2 variants (*1-4*). While such efforts have provided overviews of the potential outcomes under various scenarios, they do not assign likelihoods to the scenarios/projected trajectories, and the most likely trajectory is typically not known until the outcome is observed. That is, scenario projection is not equivalent to calibrated infectious disease *forecast*, which aims to reliably predict the most likely future outcomes during an epidemic. As COVID-19 will likely remain a major public health burden in the years to come, sensible forecast of the health outcomes several months in the future is needed to support more proactive planning.

Compared to forecast of epidemic infections (e.g., influenza), a number of additional challenges exist for long-lead COVID-19 forecast. First, SARS-CoV-2 new variants will likely continue to emerge and remain a major source of uncertainty when generating COVID-19 forecasts (*5, 6*). As has been observed for the major variants of concern (VOCs) reported to date (i.e., Alpha, Beta, Gamma, Delta, and Omicron), future new variants could arise at any time, could quickly displace other circulating variants, and could be more contagious than pre-existing variants, and/or erode prior infection- and/or vaccination-induced immunity to affect underlying population susceptibility. Further, multiple new variants could arise successively to cause multiple waves during a time span of, e.g., 6 months. Such frequent emergence and fast turnover of circulating variants is in stark contrast with epidemic infections. Second, many infections for other respiratory viruses tend to occur during a certain season of the year and as such, the seasonality can be incorporated to improve forecast accuracy (*7, 8*) as well as restrict the forecast window to the epidemic season (e.g., influenza during the winter in temperate regions). Potential seasonality for SARS-CoV-2 is still not well characterized. For instance, in the US, where larger waves have occurred in the winter during 2020-2022, smaller summer waves have also occurred (e.g. the initial Delta wave and Omicron subvariant waves; Fig 1). Third, given the unknown timing of new variant emergence, year-round, long-lead COVID-19 forecast will likely be needed. These unknowns also necessitate wider parameter ranges using a forecast ensemble to account for the uncertainty, which over long forecast horizons could lead to greater error growth and poorer predictive accuracy.

In this study, we aim to address the above challenges and develop sensible approaches that support long-lead prediction of COVID-19 epidemic outcomes. We propose three strategies for improving forecast accuracy and test the methods in combination by generating retrospective forecasts of COVID-19 cases and deaths in 10 representative states in the US (i.e., California, Florida, Iowa, Massachusetts, Michigan, New York, Pennsylvania, Texas, Washington, and Wyoming; Fig 1). Relative to a baseline approach, the optimized forecast approach based on our strategies largely improves forecast accuracy (64%/38% higher probabilistic log score for

cases/deaths, and 133%/87% higher point prediction accuracy for cases/deaths, tallied over 25,183 evaluations of forecasts made during July 2020 – Sep 2022, i.e., from the end of the initial wave to the time of this study). These results highlight strategies for developing and operationalizing long-lead COVID-19 forecasts with greater demonstrated accuracy and reliability.

RESULTS

Proposed strategies for long-lead COVID-19 forecast

To address the three challenges noted above, we propose three strategies in combination. Details are provided in Methods. Here, we summarize the idea behind each strategy. The first strategy is to constrain error growth during the extended forecast period. As noted above, the multiple uncertainties regarding SARS-CoV-2 (e.g., new variant properties) necessitate wider distributions of the state variables (e.g., population susceptibility) and parameters (e.g., virus transmissibility) at the point of forecast initiation. With a wider forecast ensemble, some ensemble members could predict earlier, large outbreaks, which if premature, would deplete modeled susceptibility and incorrectly preclude outbreaks later on. More generally, like other infections, COVID-19 epidemics often grow exponentially at first, triggering exponential changes in susceptibility and other state variables, which in turn feed back on the longer epidemic trajectory. Such infectious disease dynamics imply forecast error can also grow exponentially, leading to accelerated degradation of forecast accuracy after the first few weeks. To counteract this exponential error growth, we propose to apply a multiplicative factor $\gamma < 1$ to the covariance of the forecast state variables (e.g., susceptibility) while retaining the ensemble mean (see Methods). In so doing, we can represent initial uncertainty with a broad forecast ensemble while countering excessive error growth of the state variables. This strategy is similar to covariance inflation during system optimization (*9, 10*), where a $\gamma > 1$ is applied to the covariance to counteract over-narrowing of the model ensemble. Since here it acts in the opposite direction, we refer to this technique as deflation. Fig 2A shows example forecasts with deflation compared to without it.

The second strategy is to anticipate the impact of new variants. Genomic sequencing data can support prediction of the impact of a new variant a few weeks in the future (see Methods); however, variant displacement and competition dynamics can occur unexpectedly beyond those first few weeks, rendering historical data less relevant. Thus, for weeks farther in the future, instead of predicting specific new variants, we propose to use a set of heuristic rules to anticipate their likely emergence timing and impact on population susceptibility and virus transmissibility. Specifically, for the timing, we reason that new variants are more likely to emerge/circulate 1) after a large local wave when more infections could lead to more mutations, which could be timed based on local outbreak intensity during the preceding

months; and 2) during a time when a large part of the world is experiencing a large wave, which also could lead to new mutations. For instance, in the US, this could be during the winter when local large waves tend to occur, or the summer when places in the southern hemisphere are amid their winter waves. As such, this could be timed based on the calendar. For the new variant impact, we observe that, new variant circulation often results in gradual increases in susceptibility (e.g., a few percentages per week, based on estimates from New York City during VOC waves), possibly due to the substantial population immunity accumulated via infections and vaccinations. Similarly, changes in transmissibility also tend to occur gradually. As such, we propose to apply small increases to the model population susceptibility and transmissibility during those two plausible times. The rationale here is to anticipate the more common, non-major changes so as not to overpredict, because major VOCs causing dramatic changes are rarer and difficult to predict. Fig 2B shows example forecasts with these new variant settings compared to without them.

The third strategy accounts for seasonality. Instead of assuming a specific epidemic timing, we model the seasonal risk of SARS-CoV-2 infection based on plausible underlying drivers of infection seasonality for common respiratory viruses. Specifically, studies have shown that respiratory viruses including SARS-CoV-2 are sensitive to ambient humidity and temperature conditions, which could in turn modulate their survival and transmission (*11-15*). Accordingly, we developed a climate-forced model that includes both humidity and temperature to capture the reported virus response and parameterized the model based on long-term epidemic data observed for influenza (*16*). When modified and applied to account for SARS-CoV-2 seasonality in a model-inference framework, the estimated seasonal trends are able to capture the effects of different climate conditions (e.g., for UK, Brazil, India, and South Africa (*17-19*)). Here, we thus propose to apply this model and local climate data to estimate SARS-CoV-2 seasonality in each location (referred to as “fixed seasonality”; see estimates for the 10 states in Fig S1). In addition, as the model is parameterized based on influenza data, this estimated seasonality could differ from the true SARS-CoV-2 seasonality. To test this, we propose an alternative seasonality form that transforms the fixed seasonality trend to allow a more flexible phase timing and structure of seasonality (referred to as “transformed seasonality”; see Methods and examples for the 10 states in Fig S1). Fig 2C shows a comparison of the two seasonality forms and example forecasts with no seasonality and the two seasonality forms.

We test the above three strategies in combination, including three deflation settings (i.e., setting $\gamma = 1$, 0.95, and 0.9), two new variant settings (i.e., assuming no new variants vs. anticipating new variants per the rules noted above), and three seasonality settings (i.e., no, fixed, and transformed seasonality): in total, 12 (= 3 × 2 × 3) forecast approaches. To compare the performance of the 12 forecast approaches, we generated retrospective forecasts for 10

states, from July 2020 – August 2021 (Pre-Omicron period; 65 weeks in total) and December 2021 – September 2022 (Omicron period; ~37 weeks). For each week, a forecast for the following 26 weeks (~6 months) is generated after model training using data up to the week of forecast. We then evaluate the accuracy predicting the weekly number of cases and deaths during each of the 26 weeks (i.e. 1- to 26-week ahead prediction), as well as the peak timing (i.e. the week with the highest cases/deaths), peak intensity, cumulative number of cases and deaths over the 26 weeks.

For the forecast comparisons below, we evaluated probabilistic forecast accuracy using log score, i.e., the logarithm of the probability correctly assigned to the true target (see Methods). We also evaluated point prediction accuracy – assigning value 1 (i.e., accurate) to a forecast if the point prediction is within ±25% of the observed case/death count or within ±1 week of the observed peak week, and 0 (i.e., inaccurate) otherwise; as such, when averaged over all forecasts, this accuracy gives the percentage of forecasts a point prediction is accurate within these tolerances. Thus, both higher log score and higher point prediction accuracy indicate superior forecast performance.

The impact of deflation

Compared to forecasts with no deflation ($\gamma = 1$), applying a small deflation ($\gamma = 0.95$ or 0.9) consistently improves forecast performance (Fig 3, first two panels). This improvement is evident across all locations and combinations of the other two model settings (i.e., new variants and seasonality). The improvement is most pronounced for the long-lead (i.e., 17- to 26-week ahead) weekly forecasts and overall intensity-related targets (i.e., the totals accumulated over 26 forecast weeks and peak intensity; Figs S2-3), indicating deflation is able to effectively reduce error growth accumulated over time. We note that, as deflation works by reducing forecast spread, the forecast ensemble can also become under-dispersed, assign zero probability to the true target (most notably, for peak week), and in turn produce a lower log score (see Fig S2, the 3rd row of each heatmap for peak week). Given this trade-off, we did not test deflation factors <0.9.

Overall, relative to forecasts without deflation ($\gamma = 1$), log scores aggregated over all locations and targets were 18 – 20% higher for forecasts of cases (range of relative change across all combinations of the other two model settings; Table S1) and 7 – 8% higher for forecasts of mortality when a deflation factor γ of 0.95 was applied. The log scores further increased when using $\gamma = 0.9$, to 34 – 43% higher (than $\gamma = 1$) for forecasts of cases and 13 – 17% higher for forecasts of mortality. The improvement of point prediction accuracy was more pronounced. Aggregated over all locations and targets, point prediction accuracy was 33 – 63% higher for forecasts of cases and 24 – 40% higher for forecasts of mortality when using a deflation factor γ

of 0.9, relative to using no deflation (Table S1). As such, we use $\gamma = 0.9$ as the optimal setting for subsequent analyses.

Impact of the new variant settings

To ensure consistency and avoid over-fitting, we applied the same set of heuristic rules on new variant emergence timing and impact, as noted above, throughout the entire study period (i.e., including weeks before the emergence of SARS-CoV-2 VOCs). We expect the new variant settings to improve forecast performance during VOC waves but have a less pronounced or no effect during the 2nd wave (roughly, fall/winter 2020 – 2021), prior to VOC emergence. Indeed, overall, for the 2nd wave, forecast systems with the new variant settings (referred to as "new variant model") had similar performance as those assuming no new variants (relative changes in log score and accuracy: -2% to 0.6%, Table S2; Fig 4). However, for the VOC waves, applying the new variant settings improved forecast performance during the Alpha wave (roughly, spring 2021), Delta wave (roughly, summer/fall 2021), and Omicron wave (after December 2021; note no further desegregation was made for Omicron subvariant waves due to small sample sizes; Fig 4). For forecasts of cases, the improvement was consistently seen for all three VOC waves (relative changes in log score and accuracy all > 0, Table S2). The relative increases of log score were up to 119% during the Delta wave and up to 37% during the Omicron wave; the relative increases of point prediction accuracies were up to 89% during the Delta wave and up to 96% during the Omicron wave (Table S2 and Fig 4).

For forecasts of mortality, the new variant model had higher log scores during the Alpha and Delta waves, as well as higher point prediction accuracies during all three VOC waves (Table S2). However, log scores during the Omicron wave were similar for both models (e.g., overall log scores: -0.22 for the new variant model and -0.17 for the baseline, both assuming no seasonality); in addition, the log scores were slightly lower for the new variant model, likely due to the lower COVID-19-associated deaths during the Omicron wave.

Aggregated over all waves, targets, and locations, the new variant model had 17 – 34% and 3 – 8% higher log scores and 23 – 28% and 22 – 25% higher point prediction accuracies for forecasts of cases and deaths, respectively.

Impact of seasonality forms

Based on the above results, we focus on forecasts generated using the new variant model with a deflation factor of 0.9 to examine the three approaches to forecasting the effects of seasonality. As noted above, seasonality aims to capture changes in infection risk in response to environmental conditions (here, ambient humidity and temperature); for common respiratory viruses (e.g., influenza), infection risk in temperate regions is often higher during cold-dry

229 winter months (i.e., respiratory virus season, roughly mid-October to mid-April in the US), and
230 lower during the rest of the year (i.e., off season). Both the fixed and transformed seasonality
231 models consistently improved forecast performance during the respiratory virus season relative
232 to the no seasonality approach across all locations and targets (Fig 5A for log score and 5B for
233 point prediction accuracy, first two panels; Table S3). However, if only analyzing the off season,
234 both models had worse performance compared to the model assuming no seasonality (Fig 5, 4th
235 and 5th panels). Segregating the forecasts made for the off season by wave shows that both
236 seasonality models continued to outperform the no seasonality model during summer/fall 2020
237 (grouped with the 2nd wave); worse performance occurred during summer/fall 2021 (the Alpha
238 and Delta waves) and summer 2022 (Omicron subvariants; Table S3). These results suggest that
239 the seasonality models are able to capture the seasonal risk of SARS-CoV-2 infection, and that
240 the degraded performance during the off season may be due to challenges anticipating the
241 initial surge of VOCs occurring during those times.
242
243 Tallied over all time periods, in general, the seasonality models outperformed the no
244 seasonality model (Table S3 for all locations combined and Table S4 for each location).
245 Comparing the two seasonality forms, the transformed seasonality model outperformed the
246 fixed seasonality model overall (Table S3). Compared to the no seasonality model, the
247 transformed seasonality had 14% higher log score for forecasts of mortality, and 26% and 18%
248 higher point prediction accuracies for cases and mortality, respectively (Table S3). As noted
249 above, the improvement during the respiratory virus season was more substantial (Tables S3
250 and S4; Fig 5).
251
252 **Combined impact of deflation, new variant settings, and seasonality forms**
253 We now examine the forecast performance using the combined optimal approach (i.e.,
254 applying deflation with $\gamma = 0.9$, the new variant rules, and the transformed seasonality form),
255 compared to the baseline approach (i.e., no deflation, no new variants, and no seasonality). In
256 addition to the large uncertainties surrounding SARS-CoV-2 (e.g., new variants), there are also
257 large spatial heterogeneities. For example, across the 10 states included here, population
258 density ranged from 6 people per square mile in Wyoming to 884 per square mile in
259 Massachusetts (2020 data (*20*)); climate conditions span temperate (e.g., New York) and
260 subtropical climates (e.g., Florida; Fig S1). Given the uncertainties and spatial heterogeneities,
261 the robustness of any forecast approach is particularly important.
262
263 First, we examine the consistency of forecast performance over different variant periods.
264 During the pre-Omicron period (here, July 2020 – Dec 2021), the optimized forecast approach
265 consistently outperformed the baseline approach across all states, for both forecasts of cases
266 and deaths (Fig 6A); log scores improved by 85% overall for cases (range from 39% in

Washington to 134% in Florida) and by 62% overall for deaths (range from 24% in Washington to 117% in Florida; Table 1). During the Omicron period (here, Dec 2021 – September 2022), improvements were smaller but consistent across the 10 states for forecast of cases (Fig 6A; note that only 16-20 forecasts of long-lead targets were evaluated here, as observations are incomplete); as noted above, due to the much lower mortality during the Omicron period, both the optimized and baseline forecasts of mortality had similar log scores (e.g., median difference = - 0.02, Table 1). The overall consistency of the performance indicates that the optimized forecast approach is robust for forecasting long-lead COVID-19 epidemic outcomes for different variants.

Second, we examine the forecast performance during the US respiratory virus season (here, mid-October to mid-April) when larger COVID-19 waves have occurred. Tallied over all weeks during the respiratory virus season, the optimized approach outperformed the baseline approach for all 10 states (Fig 6B); log scores increased by 105% for cases (range from 32% in Washington to 213% in Massachusetts) and by 85% for mortality (range from 19% in Washington to 158% in Iowa; Table 1). During the off season, the optimized approach also generally outperformed the baseline approach (Table 1).

Third, we examine the accuracy predicting different epidemic targets. The optimized forecast approach consistently improved point prediction accuracy for all targets for all 10 states (Fig 7 and Table 2). In addition, the improvement was more substantial for long-lead targets (e.g., 9- to 16-week ahead and 17- to 26-week ahead forecasts, peak week, peak intensity, and the cumulative totals). For instance, the optimized approach increased accuracy from 24% to 42% (35% to 56%) predicting the peak week of cases (deaths), from 20% to 40% (30% to 50%) predicting the peak intensity of cases (deaths), roughly a 2-fold improvement for these two long-lead targets. The improvements were even more substantial for 17-26 week ahead forecasts and the cumulative totals over the entire 26 weeks (by 3- to 22-fold, Fig 7 and Table 2). We note the forecasts here were generated retrospectively with information that may not be available in real time and thus likely are more accurate as a result. Nonetheless, with the same information provided to both forecast approaches, the comparison here demonstrates the large improvement in forecast accuracy using the optimized forecast approach.

DISCUSSION

Given the uncertainties surrounding future SARS-CoV-2 transmission dynamics, it is immensely challenging to accurately predict long-lead COVID-19 epidemic outcomes. Here, we have proposed three strategies for sensibly improving long-lead COVID-19 forecast. Retrospective forecast accuracy is substantially improved using the three strategies in combination during both the pre-Omicron and Omicron periods, including for long-lead targets 6 months in the

305 future. This improvement is consistent among 10 representative states across the US, indicating
306 the robustness of the forecast method.
307
308 Our first strategy addresses the accumulation of forecast error over time. The simple deflation
309 method proposed here substantially improves forecast accuracy across different model settings
310 (here, different new variant and seasonality forms), time periods (pre-Omicron vs Omicron),
311 and locations (different states). This consistent improvement indicates that deflation is
312 effective in constraining outlier ensemble trajectories. Albeit possibly a severer issue for SARS-
313 CoV-2 due to larger uncertainties and less constrained parameter estimates, error growth is a
314 common challenge in forecasts of infectious diseases not limited to COVID-19 (*21, 22*). Future
315 work could examine the utility of deflation in improving forecast accuracy for other infectious
316 diseases.
317
318 Another challenge facing COVID-19 forecast derives from the uncertainty associated with new
319 variant emergence. Based on past epidemiological dynamics of and population response to
320 SARS-CoV-2 VOCs, we proposed a simple set of heuristic rules and applied them universally
321 across time periods and locations. Despite their simplicity, the results here show that these
322 heuristics substantially improved forecast accuracy compared to forecasts generated without
323 them. These findings suggest that, while it is challenging to forecast the emergence of specific
324 variants, the timing of future variant emergence and the impacts on key epidemiological
325 characteristics (i.e., population susceptibility and virus transmissibility) can be learned from
326 past VOC waves and used to support more accurate forecast. Much uncertainty remains
327 regarding future SARS-CoV-2 genomic evolution and population immunity; however, the
328 heuristics proposed here represent a first step anticipating the dynamic interplay of SARS-CoV-2
329 new variants and population immunity. Continued work to test the robustness of these
330 heuristics as SARS-CoV-2 and population immunity continue to co-evolve is thus warranted.
331
332 The third focus of this study is seasonality. Several prior studies have examined the potential
333 seasonality of COVID-19, using methods such as time series analyses, regression models, and
334 sinusoidal functions (*23-25*). However, the underlying mechanisms and likely nonlinear
335 response to seasonal drivers are not fully characterized. Several concurrent changes including
336 case ascertainment rate, NPIs and voluntary behavioral changes, and new variants further
337 complicate such characterization. Here, we used local weather data along with a mechanistic
338 model previously developed for influenza (*16, 17*) to capture the nonlinear response of
339 respiratory virus survival/transmission to humidity and temperature. In addition, we tested an
340 alternative seasonality form given the likely differences between SARS-CoV-2 and influenza
341 (e.g., likely higher infection risk for SARS-CoV-2 than influenza during the summer). When
342 incorporated in a model-inference framework and forecast system, forecasts with both

seasonality forms outperformed their counterpart without seasonality. Importantly, improvements during the respiratory season were consistent throughout the pandemic, including for the VOC waves, as well as for all 10 states with diverse climate conditions (Table S4 and Fig S1). These findings indicate the robustness of the seasonality functions and the importance of incorporating seasonality in COVID-19 forecast. More fundamentally, the results support the idea that a common set of seasonal environmental/climate conditions influence the transmission dynamics of respiratory viruses, including but not limited to SARS-CoV-2 and influenza. Among the 10 states tested here, the transformed seasonality function tended to outperform the fixed seasonality function based on parameters estimated for influenza, except for Massachusetts and Michigan (Table S4). This difference in performance suggests there are likely nuances in the seasonality of different respiratory viruses despite shared general characteristics.

To focus on the above three challenges, in our retrospective forecasts, we used data/estimates to account for several other factors shaping COVID-19 dynamics. These included behavioral changes (including those due to NPIs), vaccination uptake, changing detection rates and hence case ascertainment rate, as well as changes in infection fatality risk due to improvement of treatment, vaccination, prior infection, and differences in the innate virulence of circulating variants. For real-time forecast, such data and estimates would likely not be available and thus forecast accuracy would likely be degraded. Nonetheless, as societies emerge from the acute pandemic phase, many of these factors would likely reach certain norms (e.g., a relative stable fraction of the population may continue to adopt preventive measures and detection rates may stay low), reducing these uncertainties. Thus, though demonstrated retrospectively, the superior skill of the forecast methods developed here demonstrate means for generating more accurate and sensible long-lead COVID-19 forecasts.

METHODS

Data sources and processing

For model training, we used reported COVID-19 case and mortality data to capture transmission dynamics, mobility data to represent concurrent NPIs, and vaccination data to account for changes in population susceptibility due to vaccination. State level COVID-19 case and mortality data were sourced from the New York Times (NYT) (*26*) and included all variants. In our previous studies, overall case/mortality data were sufficient to estimate key epidemiological parameters when different VOC waves were separated in time (*17-19*); however, in the US, the Omicron BA.1 wave overlapped substantially with the Delta wave during November 2021 – January 2022, making inference challenging. Thus, here we separated the forecasts into two periods (i.e., a pre-Omicron period combining all non-Omicron variants, and an Omicron period combining all Omicron subvariants) and used variant-specific case and mortality data for model

training and forecast evaluation for each period. Specifically, we used variant proportion data sourced from GISAID (*27*) and compiled by CoVariants.org (*28*) to compute the weekly number of cases and deaths due to non-Omicron variants and Omicron, separately (for simplicity, loosely referred to as variant-specific case and mortality data). Because only biweekly variant proportion data at the state level were available from CoVariants.org, we used a spline function to impute weekly variant proportion. To compute weekly variant-specific cases, we multiplied the NYT weekly case data by the estimated weekly variant proportion for the same week. To compute weekly variant-specific deaths, we multiplied the NYT weekly mortality data by the estimated weekly variant proportion three weeks later (i.e., assuming a 3-week lag from case detection to death; note the 3-week lag was based on the approximate time lag between the peaks of incidence and mortality time series).

Mobility data were derived from Google Community Mobility Reports (42); we aggregated all business-related categories (i.e., retail and recreational, grocery and pharmacy, transit stations, and workplaces) in all locations in each state to weekly intervals. State level COVID-19 vaccination data were sourced from Our World in Data (*29, 30*). For models including seasonality, weather data (i.e., temperature and humidity) were used to estimate infection seasonality trends. Hourly surface station temperature and relative humidity came from the Integrated Surface Dataset (ISD) maintained by the National Oceanic and Atmospheric Administration (NOAA) and are accessible using the "stationaRy" R package (*31, 32*). We computed specific humidity using temperature and relative humidity per the Clausius-Clapeyron equation (*33*). We then aggregated these data for all weather stations in each state with measurements since 2000 and calculated the average for each week of the year during 2000-2020.

Model calibration before forecast generation (i.e. inference)

To generate a forecast, we first calibrated the model to reported weekly COVID-19 case and mortality data using an inference algorithm. The model-inference system is similar to systems we developed to estimate changes in transmissibility and immune erosion for SARS-CoV-2 VOCs including Alpha, Beta, Gamma, Delta, and Omicron (*17-19*). However, to account for the fast waning of vaccine protection against infection and differential vaccine effectiveness (VE) against different variants, here we additionally accounted for variant-specific VE and waning vaccine protection against infection. Below we describe each component of the model-inference system implemented here.

Epidemic model

The epidemic model follows an SEIRSV-ImmLoss (susceptible-exposed-infectious-recovered-susceptible-vaccination-immune-loss) construct per Eqn 1:

$$\begin{cases} \frac{dS}{dt} = \frac{R}{L_t} - \frac{b_t e_t m_t \beta_t IS}{N} - \varepsilon + \sum_{\tau=0}^{\tau=T} \rho_\tau V_{t-\tau} - \sum_{k=1}^{k=K} v_k \\ \frac{dE}{dt} = \frac{b_t e_t m_t \beta_t IS}{N} - \frac{E}{Z_t} + \varepsilon \\ \frac{dI}{dt} = \frac{E}{Z_t} - \frac{I}{D_t} \\ \frac{dR}{dt} = \frac{I}{D_t} - \frac{R}{L_t} \\ \frac{dV}{dt} = \sum_{k=1}^{k=K} v_k - \sum_{\tau=0}^{\tau=T} \rho_\tau V_{t-\tau} \end{cases} \quad \text{(Eqn 1)}$$

where S, E, I, R are the number of susceptible, exposed (but not yet infectious), infectious, and recovered/deceased individuals; V is the number of individuals vaccinated and protected from infection; N is the population size; and ε is the number of travel-imported infections. In addition, the model includes the following key components:

1) Virus-specific properties, including a time-varying variant-specific transmission rate β_t, latency period Z_t, infectious period D_t, and immunity period L_t. Of note, the immunity period L_t and the term R/L_t in Eqn 1 are used to model the waning of infection-induced immune protection against infection. Also note that all parameters are estimated for each week (t) as described below.
2) The impact of NPIs. Specifically, we used relative population mobility from Google Community Mobility Reports (*34*) to adjust the transmission rate via the term m_t, as the overall impact of NPIs (e.g., reduction of the time-varying effective reproduction number) has been reported to be highly correlated with population mobility during the COVID-19 pandemic (*35-37*). To further account for potential changes in effectiveness, the model additionally includes a parameter, e_t, to scale NPI effectiveness.
3) Vaccination including boosters, using the term $\sum_{k=1}^{k=K} v_k$. Here, v_k is the number of individuals immunized after the k-th dose (k = 1,…,3 for 2 primary and 1 booster dose here, excluding those immunized after previous doses). We computed v_k using vaccination data and adjusted for vaccine effectiveness (VE)(*38-42*). Note that while the 2nd booster dose has been administered for a subset of the population, such data have not been made publicly available and thus not included in our model. Further, given the 2nd dose and 3rd dose (i.e. 1st booster) were administered ~6 months apart (i.e., beyond the estimated VE duration against infection (*41*)), here we combined data for these two doses. In doing so, we have simplified the model and implicitly assumed that the 3rd dose resumed VE against infection to a level similar to the 2nd dose, as the same VE was applied (Table S5).
4) Waning of vaccine protection against infection, using the term $\sum_{\tau=0}^{\tau=\mathrm{T}} \rho_\tau V_{t-\tau}$. We computed the total number who were vaccinated τ days ago and lost protection on day-t ($V_{t-\tau}$) per the VE waning probability (ρ_τ). The probabilities ρ_τ for time τ =0, …, T (T= the maximum

duration from the earliest vaccination rollout; Table S5) were calculated using VE duration data (*41*) and $V_{t-\tau}$ was computed per line 5 of Eqn 1.

5) Infection seasonality, computed using temperature and specific humidity data (*17*). See details in section "Seasonality settings" below.

Observation model to account for under-detection and time-lags in COVID-19 outcomes

Using the model-simulated number of infections occurring each day, we further computed the number of cases and deaths each week to match with the observations, as done in Yang et al. (*43*). Briefly, we included 1) a time-lag from infectiousness to detection (i.e., an infection being diagnosed as a case), drawn from a gamma distribution with a mean of $T_{d,mean}$ days and a standard deviation of $T_{d,\,sd}$ days, to account for delays in detection; 2) an infection-detection rate (r_t), i.e. the fraction of infections (including subclinical or asymptomatic infections) reported as cases, to account for under-detection; 3) a time-lag from infectiousness to death; and 4) an infection-fatality risk (IFR_t). Each week, the infection-detection rate (r_t), infection-fatality risk (IFR_t), and the two time-to-detection parameters ($T_{d,\,mean}$ and $T_{d,\,sd}$) were estimated along with other parameters (see below). The time-lag from infectiousness to death during the pre-Omicron period was drawn from a gamma distribution with a mean of 14 days and a standard deviation of 14 days, roughly based on data from New York City (unpublished work). For the Omicron period, many deaths were identified posthumously; thus, it is difficult to estimate the time-lag from infection to death for Omicron infections. Here, based on the slightly longer time-lag between the peaks of case and mortality time series during the Omicron period, we assumed a gamma distribution with a mean of 24 days (i.e., assuming an additional 10-day lag) and a standard deviation of 14 days.

To compute the model-simulated number of new cases each week, we multiplied the model-simulated number of new infections per day by the infection-detection rate, and further distributed these simulated cases in time per the distribution of time-from-infectiousness-to-detection. Similarly, to compute the model-simulated deaths per week and account for delays in time to death, we multiplied the simulated-infections by the IFR and then distributed these simulated deaths in time per the distribution of time-from-infectious-to-death. We then aggregated these daily numbers to weekly totals to match with the weekly case and mortality data for model-inference.

Model inference and parameter estimation

The inference system uses the ensemble adjustment Kalman filter (EAKF (*44*)), a Bayesian statistical method, to estimate model state variables (i.e., S, E, I, R from Eqn 1) and parameters (i.e., β_t, Z_t, D_t, L_t, e_t, from Eqn 1, as well as r_t, IFR_t $T_{d,mean}$, and $T_{d,\,sd}$ from the observation model). Briefly, the EAKF uses an ensemble of model realizations (n=500 here), each with initial

parameters and variables randomly drawn from a *prior* range (Table S5). After model initialization, the system integrates the model ensemble forward in time for a week (per Eqn 1) to compute the prior distribution for each model state variable, as well as the model-simulated number of cases and deaths for that week. The system then combines the prior estimates with the observed case and death data for the same week to compute the posterior per Bayes' theorem (*44*). In addition, as in (*17-19*), during the filtering process, we applied space-reprobing (*45*), i.e., random replacement of parameter values for a small fraction of the model ensemble, to explore a wider range of parameter possibilities (Table S5). The space-reprobing algorithm, along with the EAKF, allows the system to capture potential changes over time (e.g., increased detection for variants causing more severe disease, or increases in population susceptibility and transmission rate due to a new variant).

Of note, except for the transmission rate β_t, the same prior range was used for all states included here (Table S5). For β_t, initial testing indicated the transmission rate was likely lower in less densely populated states; thus, we set the prior range of β_t for each state, based on local population density (see Table S5).

Variations in forecast systems (deflation, new variant, and seasonality settings)

In total, 12 forecast approaches were tested (3 deflation levels × 2 new variants settings × 3 seasonality forms). Below we describe each setting.

Deflation during the forecast period

The deflation algorithm is patterned after covariance inflation, as used in filtering methods (*9, 10*). However, unlike inflation applied during filtering (i.e., the model training period via data assimilation), deflation is applied during the forecast period. As the state variables change dynamically per the epidemic model (i.e., Eqn 1), the error of some epidemic trajectories can amplify exponentially over time. Thus, here we applied deflation only to the state variables (i.e., not to the model parameters), per:

$$x_i^{def} = \gamma(x_i - \bar{x}) + \bar{x}, i = 1, \dots, n \quad \text{(Eqn 2)}$$

where x_i is i-th ensemble member of a given state variable (here, S, E, or I) at each time step during the forecast period, before the deflation; x_i^{def} is the corresponding "deflated" value; γ is the deflation factor; and $\bar{x}$ is the ensemble mean of the state variable. Per Eqn 2, deflation retains the ensemble mean, while reducing the ensemble spread to constrain error accumulation. In this study, we tested three levels of deflation, by setting γ to 1 (i.e., no deflation), 0.95, and 0.9, separately.

Settings for anticipating the impact of new variants during the forecast period

As noted above, the uncertainty due to the possible emergence or surge of new variants in the future is a major challenge for long-lead COVID-19 forecast. To anticipate and account for potential surges and their impact on COVID-19 epidemic outcomes, we tested two approaches. The first, baseline approach simply assumes there are no changes in the circulating variant during the forecast period. For this approach, the forecasts were generated using the latest population susceptibility and transmission parameter estimates at the point of forecast initiation.

For the second, new variant approach, we devised a set of heuristics to anticipate the likely timing and impact of new variant emergence during the forecast period. For the very near future (1- to 5 weeks), we used available genomic sequencing data (see "Data sources and processing"). Specifically, we first estimated the growth rate for each circulating variant based on variant proportion 6- to 2 weeks prior to the week of forecast initiation (i.e., assuming a 2-week lag for genomic data collection); for simplicity, we used a log-linear model [i.e., log(variant proportion during week-*t*) ~ week-*t*]. If any variant had a high growth rate (here, arbitrarily set to 10% per week), we deemed it a rising variant that could further affect the population susceptibility and overall virus transmissibility. To anticipate its impact, we then used a smoothing spline to 1) project when the variant would reach a 100% proportion and 2) project the number of weeks for the variant to grow from 0% to 100%. The first estimate was then used to set the timing of the continued impact and the second estimate was used to scale the increases in population susceptibility. Here, arbitrarily, we assumed a baseline of 1.5 – 4.5% increase in susceptibility for each week the new variant increased in proportion; however, if the estimated growth rate (g) was >20%, to account for the faster growth, we scaled that baseline by a factor of $(1+g)/1.2$. While the growth advantage of a new variant could also come from increased transmissibility, for simplicity, here we opted to solely adjust for population susceptibility. In addition, given the fast displacement of new variants, we opted not to use projected estimates 5 weeks beyond the forecast week, i.e. these changes were only applied to the first 5 weeks of a forecast.

When genomic data could not be used (i.e. beyond the first 5 forecast weeks or when genomic data were not available), we used the following heuristics to anticipate the likely timing and impact of a new variant surge:

i) New variants tend to emerge after a recent large wave (here, defined arbitrarily as a 25% attack rate over 3 months for the non-Omicron period, and a 33% attack rate over 2 months for the Omicron period). We identified these times using estimated/forecasted infection rates during the preceding months and the forecast period.
ii) New variants tend to emerge and/or become widespread during northern hemisphere winter (December – February), southern hemisphere winter (June – August), and/or the

monsoon season (e.g. June – September in India) and could be introduced to the US during these months. We identified these times using calendar month.

iii) During the above times with potential new variant emergence, the population susceptibility and virus transmissibility could increase. Accordingly, to account for susceptibility changes, we resampled half of the model ensemble to increase the population susceptibility by 2-9% for weeks flagged per the conditions described in i and ii. However, this susceptibility increase was only triggered when the mean population susceptibility was below 40% to avoid over-adjustment. Similarly, to avoid the system being trapped in an outbreak-begets-outbreak cycle, no further adjustments were made to susceptibility if a wave had been forecast the prior weeks or the cumulative adjustment had exceeded a threshold (here, set to 40% of the population over 26 weeks for pre-Omicron period and 60% for the Omicron period). To account for transmissibility changes, we expanded the variance of the transmission rate (i.e., β_t in Eqn 1) by applying an inflation factor of 1.3 (pre-Omicron period) or 1.1 (Omicron period) to ensemble members falling between the 50th and 95th/90th (pre-Omicron/Omicron period) percentiles (i.e., the ones with higher but not too extreme values) for weeks identified per the conditions described in i and ii.

Seasonality settings

Seasonality is incorporated in the epidemic model (Eqn 1) and applied throughout the model calibration (i.e., inference) and forecast periods. Here, we tested three seasonality settings. The first assumes no changes in seasonal risk of infection, by setting b_t in Eqn 1 to 1 for all weeks (referred to as "no seasonality").

The second seasonality form represents the dependency of respiratory virus survival, including that of SARS-CoV-2, to temperature and humidity (*12, 46*) per the following equations:

$$R_0(t) = [a_0 q^2(t) + a_1 q(t) + a_2][\frac{T_c}{T(t)}]^{T_{exp}} \quad \text{(Eqn 3a)}$$

$$b_t = \frac{R_0(t)}{\overline{R_0(t)}} \quad \text{(Eqn 3b)}$$

As described previously (*17, 18*), the seasonality function in Eqn 3a assumes that humidity has a bimodal effect on seasonal risk of infection, with both low and high humidity conditions favoring transmission [i.e., the parabola in the 1st set of brackets, where $q(t)$ is weekly specific humidity measured by local weather stations and t = 1,…,52, i.e., week 1 to week 52 of the year]; this effect is further modulated by temperature, with low temperatures promoting transmission and temperatures above a certain threshold limiting transmission [i.e., the 2nd set of brackets, where $T(t)$ is weekly temperature measured by local weather stations and T_c is the threshold]. As SARS-CoV-2 specific parameters (a_0, a_1, a_2, T_c, and T_{exp} in Eqn 3a) are not available, we used parameters estimated for influenza (*16*) and scaled the weekly outputs [i.e., $R_0(t)$] by the annual mean (i.e., $\overline{R_0}$) per Eqn 3b, as done in Yang and Shaman (*17*). In doing so, the scaled outputs (b_t) are no longer specific to influenza; rather, they represent the *relative*,

seasonality-related transmissibility by week, general to viruses sharing similar seasonal responses. The estimated relative seasonal trend, b_t, is then used to adjust the relative transmission rate at time t in Eqn 1. As the relative seasonality trend is computed deterministically using local humidity and temperature data, we refer to this setting as "fixed seasonality".

The third seasonality setting (referred to as "transformed seasonality") transforms the b_t estimates from Eqn 3b to allow flexibility in the seasonal trend. To do so, we include three parameters to fine tune the peak of the seasonal trend (p_{shift}; i.e., the number of weeks earlier or later than the peak estimated for influenza), the number of weeks during a year with b_t >1 (δ; i.e., the duration with elevated infection risk), and another parameter $b_{t,lwr}$ that adjusts the lowest b_t value. Specifically, the transformation first adjusts values of b_t greater than 1, by shifting the timing by p_{shift} weeks and adjusting the duration with elevated infection risk to δ, per

$$b'_{t\{b_t>1\}} = \frac{b_{t\{b_t>1\}+p_{shift}}}{\frac{n_{b_t>1}}{\delta}} \quad \text{(Eqn 4a)}$$

where $n_{b_t>1}$ is the number of weeks with b_t >1 during the 1-year cycle. For weeks with b_t ≤1, the transformation adjusts the values, by shifting the timing by p_{shift} weeks and adjusting the duration with lower infection risk to $52-\delta$, per

$$b'_{t\{b_t\leq 1\}} = \frac{b_{t\{b_t\leq 1\}+p_{shift}}}{n_{b_t\leq 1}/(52-\delta)} \quad \text{(Eqn 4b)}$$

The approach then further scales $b'_{t\{b_t\leq 1\}}$ to increase the relative infection risk, per

$$b''_{t\{b_t\leq 1\}} = 1-(1-b'_{t\{b_t\leq 1\}})(min\left\{1, \frac{\min\{b'_{t\{b_t\leq 1\}}\}}{b_{t,lwr}}\right\}) \text{ (Eqn 4c)}$$

$b''_{t\{b_t>1\}} \equiv b'_{t\{b_t>1\}}$ and $b''_{t\{b_t\leq 1\}}$ are then pooled together and scaled to have a mean of 1 over the 1-year cycle, per

$$b'''_t = \frac{b''_t}{\overline{b''_t}} \text{ (Eqn 4d)}$$

There could be multiple combinations of the three parameters (i.e., p_{shift}, δ, and $b_{t,lwr}$). Due to the lack of SARS-CoV-2 data to inform the parameter estimates, here we opted to optimize the range for each parameter (as opposed to estimate specific best-fit parameters). Briefly, we tested 2 levels (low vs. high) for each parameter and thus 8 in combination for each state (Fig S4). We then identified the best range for each state based on forecast performance during the 2nd wave, i.e., before the surge of SARS-CoV-2 VOCs to minimize potential confounding. The best parameter ranges (Fig S4) were then used in the transformed seasonality model in the main analysis.

Retrospective forecast

We tested the above 12 model-inference and forecast approaches (3 deflation levels × 2 new variants settings × 3 seasonality forms) for 10 states, i.e., California, Florida, Iowa, Massachusetts, Michigan, New York, Pennsylvania, Texas, Washington, and Wyoming. The 10 states span the 10 Health and Human Services (HHS) regions across the US, representing a wide range of population characteristics and COVID-19 pandemic dynamics (Fig 1). For all states, retrospective forecasts of weekly cases and deaths 26 weeks (i.e., 6 months) into the future were generated for the non-Omicron period and the Omicron period, separately. For the non-Omicron period, we initiated forecasts each week from the week of July 5, 2020 (i.e., after the initial wave) through the week of August 15, 2021. We stopped initiating the non-Omicron forecasts in mid-August 2021 to allow at least a few weeks of Delta-related data to calibrate the model before forecasting the Delta wave. However, a 6-month forecast initiated during mid-June – mid-August 2021 would extend to mid-December 2021 – mid-Feb 2022, when Omicron BA.1 had become predominant, depending on location; this overlap would lead to lower forecast accuracy, since here we did not account for the emergence of Omicron BA.1 and fast displacement of Delta. For the Omicron period, we initiated forecasts starting 5 weeks after local detection of Omicron BA.1 (roughly in early December 2021, depending on local data) through the week of September 25, 2022 (i.e., the last week with available data at the time of this study).

To generate each forecast, we ran the model-inference system until the week of forecast, halted the inference, and used the population susceptibility and transmissibility of the circulating variant estimated at that time to predict cases and deaths for the following 26 weeks (i.e., 6 months). Because the infection detection rate and IFR are linked to observations of cases and deaths, changes of these quantities during the forecast period could obscure the underlying infection rate and forecast accuracy. Thus, for these two parameters specifically, we used available model-inference estimates for corresponding forecast period weeks to allow comparison of model-forecast cases and deaths with the data while focusing on testing the accuracy of different model settings (e.g., seasonality and new variant settings). For the same reason, we used all available mobility and vaccination data including those for the forecast period, which would not be available in real time. For weeks in the future without data/estimates, we used the latest estimates instead.

Both the model-inference and forecast were run with n = 500 model realizations (i.e., ensemble members). The ensemble and its distribution provided probabilistic forecasts for 4 types of targets here, i.e., 1-to 26-week ahead prediction, peak intensity, peak week, and cumulative totals over the entire 26-week forecast period. For example, for the 1-week ahead prediction (c_{t+1}), the fraction of ensemble members falling in a given bin $[c_i, c_{i+1})$ can be used to

represent the forecast probability density, i.e., $\Pr(c_{t+1} \in [c_i, c_{i+1})) = n_{\{c_{t+1} \geq c_i \,\&\, c_{t+1} < c_{i+1}\}}/n$. Similarly, for the peak week prediction, predicted peak week by individual ensemble members (p_w = 1, 2, …, 26) can be aggregated and the distribution can be used to represent the probability distribution of the forecast, i.e., $\Pr(p_w = w) = n_{\{p_w = w\}}/n$. To account for model and filter stochasticity, we repeated each forecast 10 times, each time with initial parameters and state variables randomly drawn from the same prior ranges. We then averaged results over the 10 runs for each evaluation.

Forecast evaluation

To evaluate forecast performance, we examined the following metrics: 1) 1- to 26-week ahead prediction, and 2) peak week, 3) peak intensity, and 4) cumulative total over the 26-week forecast period, for cases and deaths, separately. For each metric, we computed both the log score based on the probabilistic forecast and the accuracy of point prediction.

Specifically, to compute the log score, we first binned the forecast ensemble to generate the forecast probability distribution $\Pr(x)$, e.g., $\Pr(c_{t+1})$ for 1-week ahead prediction and $\Pr(p_w)$ for peak week, as noted above. Here, for cases, bins of the weekly targets were set to $[0, 0.05\%)$, $[0.05\%, 0.1\%)$, …, $[0.95, 1\%)$, and $[1\%, 100\%]$ (i.e., increments of 0.05%, or 500 per million people, up to 1% of the population; and the rest combined in the last bin); bins of cumulative cases over 26 weeks were set to $[0, 2\%)$, $[2\%, 4\%)$, …, $[8\%, 10\%)$, $[10\%, 15\%)$, $[15\%, 20\%)$,…,$[45\%, 50\%)$, and $[50\%, 100\%]$ (i.e., increments of 2% up to 10%, then increments of 5% up to 50% of the population; and the rest combined in the last bin). For mortality, bins of the weekly targets were set to $[0, 0.001\%)$, $[0.001, 0.002)$, …, $[0.019\%, 0.02\%)$, and $[0.02\%, 100\%]$ (i.e., increments of 0.001%, or 10 per million people, up to 0.02% of the population; and the rest combined in the last bin); bins of cumulative deaths over 26 weeks were set to $[0, 0.02\%)$, $[0.02\%, 0.04\%)$, …,$[0.08\%, 0.1\%)$, $[0.1\%, 0.15\%)$, $[0.15\%, 0.2\%)$,…, $[0.45\%, 0.5\%)$, and $[0.5\%, 100\%]$ (i.e., increments of 0.02% up to 0.1%, then increments of 0.05% up to 0.5% of the population; and the rest combined in the last bin). For the peak week of both cases and deaths, the bin size was set to 1 week. The log score was then computed as:

$$\log score = \log[\Pr(x)_{x \in bin^*} + \Pr(x)_{x \in bin^{*-1}} + \Pr(x)_{x \in bin^{*+1}}] \quad \text{(Eqn 5)}$$

where $\Pr(x)$ is the forecast probability for target x; bin^* is the bin that contains the observed value for that target (see bin specifications above) and bin^{*-1} and bin^{*+1} are the two adjacent bins. Note that, here we used smaller bins and deemed ensemble members falling within the bin covering the observation and its two adjacent bins accurate, which is equivalent to using a single larger bin spanning all those smaller bins. However, as the probabilistic forecasts (i.e.,

probabilities in each bin) were generated and stored before the final evaluation, using smaller bins allowed more flexible post processing and evaluation if needed (e.g., the log score can be computed based on a single small bin if preferred).

For the accuracy of point prediction, we deemed a forecast accurate (assigned a value of 1) if the median of the forecast ensemble is within ±1 week of the observed peak week or within ±25% of the observed case/death count, and inaccurate (assigned a value of 0) otherwise. As noted above, when aggregated over multiple forecasts, the average would represent the percentage of time a point prediction is accurate within these tolerances.

We compared the performance of each forecast approach (i.e., each of the 12 combinations of deflation, new variant setting, and seasonality form) overall or by forecast target, segregated by time period or respiratory virus season. To compute the overall score for each stratum, we took the arithmetic mean of the log score or point prediction accuracy of forecasts generated by each forecast system, either across all forecast targets or for each target, over 1) all forecast weeks during the entire study period (i.e., July 2020 – September 2022), 2) the pre-Omicron period and Omicron period, separately, 3) the respiratory virus season (mid-October to mid-April, 6 months) and off season (the remaining 6 months), separately. Of note, for the non-Omicron period, the last few forecasts extended to early 2022, when Delta and Omicron cocirculated. Given the low number of Delta-associated cases/deaths in 2022, weekly targets (i.e., 1- to 26- week ahead prediction) for weeks in 2022 were excluded from the evaluation; however, as Delta was the main circulating variant during the 6-month period for these forecasts, all the overall targets (i.e., peak week, peak intensity, and cumulative total) were evaluated based on Delta-specific data and included in the analysis.

For pairwise comparison of forecast approaches, we computed the difference of log score or accuracy by simple subtraction of the two arithmetic-means. Relative difference was also computed. For the log score, the percent relative difference was computed as:

$$\%\, relative\ difference\ in \log score = \frac{\exp\,(mean\ \log score\ of\ system\ a) - \exp\,(mean \log score\ of\ system\ b)}{\exp\,(mean \log score\ of\ system\ b)} \times 100\% \quad \text{(Eqn 6)}$$

As noted in (*47*), the exponent of the mean log score (Eqn 5) can be interpreted as the probability correctly assigned to the bins containing the observations; thus Eqn 6 gives the relative difference in the correctly assigned forecast probability. The percent relative difference in accuracy of point prediction was computed as:

$$\%\, relative\ difference\ in\ accuracy = \frac{mean\ accuracy\ of\ system\ a - mean\ accuracy\ of\ system\ b}{mean\ accuracy\ of\ system\ b} \times 100\% \quad \text{(Eqn 7)}$$

In addition, we also computed the pair-wised difference of log score or accuracy (paired by forecast week for each target and location) and used boxplots to examine the distributions (see, e.g., Fig 6). We used the Wilcoxon rank sum test, a non-parametric statistical method, to test whether there is a difference in the median of the pair-wised differences (*48*). All inference, forecast, and statistical analyses were carried out using the R language (https://www.r-project.org).

Acknowledgements: This study was supported by the National Institute of Allergy and Infectious Diseases (AI145883 and AI163023), the Centers for Disease Control and Prevention (CDC) and the Council of State and Territorial Epidemiologists (CSTE; contract no.: NU38OT00297), and the CDC Center for Forecasting and Outbreak Analytics (contract no.: 75D30122C14289).

Competing interests: JS and Columbia University disclose partial ownership of SK Analytics. JS discloses consulting for BNI.

References:

1. R. K. Borchering, C. Viboud, E. Howerton, C. P. Smith, S. Truelove, M. C. Runge, N. G. Reich, L. Contamin, J. Levander, J. Salerno, W. van Panhuis, M. Kinsey, K. Tallaksen, R. F. Obrecht, L. Asher, C. Costello, M. Kelbaugh, S. Wilson, L. Shin, M. E. Gallagher, L. C. Mullany, K. Rainwater-Lovett, J. C. Lemaitre, J. Dent, K. H. Grantz, J. Kaminsky, S. A. Lauer, E. C. Lee, H. R. Meredith, J. Perez-Saez, L. T. Keegan, D. Karlen, M. Chinazzi, J. T. Davis, K. Mu, X. Xiong, Y. P. A. Pastore, A. Vespignani, A. Srivastava, P. Porebski, S. Venkatramanan, A. Adiga, B. Lewis, B. Klahn, J. Outten, J. Schlitt, P. Corbett, P. A. Telionis, L. Wang, A. S. Peddireddy, B. Hurt, J. Chen, A. Vullikanti, M. Marathe, J. M. Healy, R. B. Slayton, M. Biggerstaff, M. A. Johansson, K. Shea, J. Lessler, Modeling of Future COVID-19 Cases, Hospitalizations, and Deaths, by Vaccination Rates and Nonpharmaceutical Intervention Scenarios - United States, April-September 2021. *MMWR Morb Mortal Wkly Rep* **70**, 719-724 (2021).
2. N. G. Reich, J. Lessler, S. Funk, C. Viboud, A. Vespignani, R. J. Tibshirani, K. Shea, M. Schienle, M. C. Runge, R. Rosenfeld, E. L. Ray, R. Niehus, H. C. Johnson, M. A. Johansson, H. Hochheiser, L. Gardner, J. Bracher, R. K. Borchering, M. Biggerstaff, Collaborative Hubs: Making the Most of Predictive Epidemic Modeling. *American Journal of Public Health* **112**, 839-842 (2022).
3. S. Truelove, C. P. Smith, M. Qin, L. C. Mullany, R. K. Borchering, J. Lessler, K. Shea, E. Howerton, L. Contamin, J. Levander, J. Kerr, H. Hochheiser, M. Kinsey, K. Tallaksen, S. Wilson, L. Shin, K. Rainwater-Lovett, J. C. Lemairtre, J. Dent, J. Kaminsky, E. C. Lee, J. Perez-Saez, A. Hill, D. Karlen, M. Chinazzi, J. T. Davis, K. Mu, X. Xiong, Y. P. A. Pastore, A. Vespignani, A. Srivastava, P. Porebski, S. Venkatramanan, A. Adiga, B. Lewis, B. Klahn, J. Outten, M. Orr, G. Harrison, B. Hurt, J. Chen, A. Vullikanti, M. Marathe, S. Hoops, P. Bhattacharya, D. Machi, S. Chen, R. Paul, D. Janies, J. C. Thill, M. Galanti, T. K. Yamana, S. Pei, J. L. Shaman, J. M. Healy, R. B. Slayton, M. Biggerstaff, M. A. Johansson, M. C. Runge, C. Viboud, Projected resurgence of COVID-19 in the United States in July-December

2021 resulting from the increased transmissibility of the Delta variant and faltering vaccination. *Elife* **11**, (2022).

4. K. Nixon, S. Jindal, F. Parker, N. G. Reich, K. Ghobadi, E. C. Lee, S. Truelove, L. Gardner, An evaluation of prospective COVID-19 modelling studies in the USA: from data to science translation. *Lancet Digit Health* **4**, e738-e747 (2022).
5. K. Koelle, M. A. Martin, R. Antia, B. Lopman, N. E. Dean, The changing epidemiology of SARS-CoV-2. *Science* **375**, 1116-1121 (2022).
6. P. V. Markov, A. Katzourakis, N. I. Stilianakis, Antigenic evolution will lead to new SARS-CoV-2 variants with unpredictable severity. *Nat Rev Microbiol* **20**, 251-252 (2022).
7. J. Shaman, S. Kandula, W. Yang, A. Karspeck, The use of ambient humidity conditions to improve influenza forecast. *PLOS Computational Biology* **13**, e1005844 (2017).
8. S. C. Kramer, J. Shaman, Development and validation of influenza forecasting for 64 temperate and tropical countries. *Plos Computational Biology* **15**, (2019).
9. J. L. Anderson, S. L. Anderson, A Monte Carlo Implementation of the Nonlinear Filtering Problem to Produce Ensemble Assimilations and Forecasts. *Mon. Weather Rev.* **127**, 2741-2758 (1999).
10. J. L. Anderson, An adaptive covariance inflation error correction algorithm for ensemble filters. *Tellus A* **59**, 210-224 (2007).
11. M. Moriyama, W. J. Hugentobler, A. Iwasaki, Seasonality of Respiratory Viral Infections. *Annu Rev Virol* **7**, 83-101 (2020).
12. D. H. Morris, K. C. Yinda, A. Gamble, F. W. Rossine, Q. Huang, T. Bushmaker, R. J. Fischer, M. J. Matson, N. Van Doremalen, P. J. Vikesland, L. C. Marr, V. J. Munster, J. O. Lloyd-Smith, Mechanistic theory predicts the effects of temperature and humidity on inactivation of SARS-CoV-2 and other enveloped viruses. *Elife* **10**, (2021).
13. A. C. Lowen, S. Mubareka, J. Steel, P. Palese, Influenza Virus Transmission Is Dependent on Relative Humidity and Temperature. *PLoS Pathog* **3**, e151 (2007).
14. W. Yang, S. Elankumaran, L. C. Marr, Relationship between humidity and influenza A viability in droplets and implications for influenza's seasonality. *PLoS One* **7**, e46789 (2012).
15. Y. Li, X. Wang, H. Nair, Global Seasonality of Human Seasonal Coronaviruses: A Clue for Postpandemic Circulating Season of Severe Acute Respiratory Syndrome Coronavirus 2? *J Infect Dis* **222**, 1090-1097 (2020).
16. H. Yuan, S. C. Kramer, E. H. Y. Lau, B. J. Cowling, W. Yang, Modeling influenza seasonality in the tropics and subtropics. *PLoS Comput Biol* **17**, e1009050 (2021).
17. W. Yang, J. Shaman, Development of a model-inference system for estimating epidemiological characteristics of SARS-CoV-2 variants of concern. *Nature Communications* **12**, 5573 (2021).
18. W. Yang, J. Shaman, COVID-19 pandemic dynamics in India, the SARS-CoV-2 Delta variant and implications for vaccination. *J R Soc Interface* **19**, 20210900 (2022).
19. W. Yang, J. L. Shaman, COVID-19 pandemic dynamics in South Africa and epidemiological characteristics of three variants of concern (Beta, Delta, and Omicron). *Elife* **11**, (2022).

20. statista, Population density in the U.S. by federal states including the District of Columbia in 2020. https://www.statista.com/statistics/183588/population-density-in-the-federal-states-of-the-us/
21. S. Pei, M. A. Cane, J. Shaman, Predictability in process-based ensemble forecast of influenza. *PLoS Comput Biol* **15**, e1006783 (2019).
22. S. Pei, J. Shaman, Counteracting structural errors in ensemble forecast of influenza outbreaks. *Nat Commun* **8**, 925 (2017).
23. Y. W. Choi, A. Tuel, E. A. B. Eltahir, On the Environmental Determinants of COVID-19 Seasonality. *Geohealth* **5**, e2021GH000413 (2021).
24. C. Merow, M. C. Urban, Seasonality and uncertainty in global COVID-19 growth rates. *Proc Natl Acad Sci U S A* **117**, 27456-27464 (2020).
25. X. Liu, J. Huang, C. Li, Y. Zhao, D. Wang, Z. Huang, K. Yang, The role of seasonality in the spread of COVID-19 pandemic. *Environ Res* **195**, 110874 (2021).
26. The New York Times, Coronavirus (Covid-19) Data in the United States. 10/18/2022. https://github.com/nytimes/covid-19-data
27. Global Initiative on Sharing All Influenza Data (GISAID), Tracking of Variants. https://www.gisaid.org/hcov19-variants/
28. CoVariants. https://covariants.org
29. Our World in Data, Data on COVID-19 (coronavirus) vaccinations by Our World in Data. 10/18/2022. https://github.com/owid/covid-19-data/blob/master/public/data/vaccinations/us_state_vaccinations.csv
30. E. Mathieu, H. Ritchie, E. Ortiz-Ospina, M. Roser, J. Hasell, C. Appel, C. Giattino, L. Rodés-Guirao, A global database of COVID-19 vaccinations. *Nature human behaviour* **5**, 947-953 (2021).
31. R. Iannone, Package 'stationaRy'. https://cran.r-project.org/web/packages/stationaRy/stationaRy.pdf
32. R. Iannone, stationaRy. https://github.com/rich-iannone/stationaRy
33. J. Wallace, P. Hobbs, *Atmospheric Science: An Introductory survey*. (Academic Press, New York, ed. 2nd Edition, 2006), pp. 504.
34. Google Inc., Community Mobility Reports. https://www.google.com/covid19/mobility/
35. W. Yang, J. Shaff, J. Shaman, Effectiveness of non-pharmaceutical interventions to contain COVID-19: a case study of the 2020 spring pandemic wave in New York City. *J R Soc Interface* **18**, 20200822 (2021).
36. A. Lasry, D. Kidder, M. Hast, J. Poovey, G. Sunshine, K. Winglee, N. Zviedrite, F. Ahmed, K. A. Ethier, CDC Public Health Law Program, New York City Department of Health and Mental Hygiene, Louisiana Department of Health, Public Health Seattle and King County, San Francisco COVID-Response Team, Alameda County Public Health Department, San Mateo County Health Department, Marin County Division of Public Health, Timing of Community Mitigation and Changes in Reported COVID-19 and Community Mobility - Four U.S. Metropolitan Areas, February 26-April 1, 2020. *MMWR. Morbidity and mortality weekly report* **69**, 451-457 (2020).
37. M. U. G. Kraemer, C. H. Yang, B. Gutierrez, C. H. Wu, B. Klein, D. M. Pigott, C.-D. W. G. Open, L. du Plessis, N. R. Faria, R. Li, W. P. Hanage, J. S. Brownstein, M. Layan, A.

Vespignani, H. Tian, C. Dye, O. G. Pybus, S. V. Scarpino, The effect of human mobility and control measures on the COVID-19 epidemic in China. *Science* **368**, 493-497 (2020).

38. F. P. Polack, S. J. Thomas, N. Kitchin, J. Absalon, A. Gurtman, S. Lockhart, J. L. Perez, G. Pérez Marc, E. D. Moreira, C. Zerbini, R. Bailey, K. A. Swanson, S. Roychoudhury, K. Koury, P. Li, W. V. Kalina, D. Cooper, R. W. Frenck, L. L. Hammitt, Ö. Türeci, H. Nell, A. Schaefer, S. Ünal, D. B. Tresnan, S. Mather, P. R. Dormitzer, U. Şahin, K. U. Jansen, W. C. Gruber, Safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine. *New Engl J Med*, (2020).

39. L. R. Baden, H. M. El Sahly, B. Essink, K. Kotloff, S. Frey, R. Novak, D. Diemert, S. A. Spector, N. Rouphael, C. B. Creech, J. McGettigan, S. Khetan, N. Segall, J. Solis, A. Brosz, C. Fierro, H. Schwartz, K. Neuzil, L. Corey, P. Gilbert, H. Janes, D. Follmann, M. Marovich, J. Mascola, L. Polakowski, J. Ledgerwood, B. S. Graham, H. Bennett, R. Pajon, C. Knightly, B. Leav, W. Deng, H. Zhou, S. Han, M. Ivarsson, J. Miller, T. Zaks, C. S. Group, Efficacy and Safety of the mRNA-1273 SARS-CoV-2 Vaccine. *N Engl J Med* **384**, 403-416 (2021).

40. E. J. Haas, F. J. Angulo, J. M. McLaughlin, E. Anis, S. R. Singer, F. Khan, N. Brooks, M. Smaja, G. Mircus, K. Pan, J. Southern, D. L. Swerdlow, L. Jodar, Y. Levy, S. Alroy-Preis, Impact and effectiveness of mRNA BNT162b2 vaccine against SARS-CoV-2 infections and COVID-19 cases, hospitalisations, and deaths following a nationwide vaccination campaign in Israel: an observational study using national surveillance data. *The Lancet* **397**, 1819-1829 (2021).

41. UK Heath Security Agency, COVID-19 vaccine surveillance report (Week 17, 28 April 2022). https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1072064/Vaccine-surveillance-report-week-17.pdf

42. F. C. M. Kirsebom, N. Andrews, J. Stowe, S. Toffa, R. Sachdeva, E. Gallagher, N. Groves, A.-M. O'Connell, M. Chand, M. Ramsay, J. L. Bernal, COVID-19 vaccine effectiveness against the omicron (BA.2) variant in England. *The Lancet Infectious Diseases*.

43. W. Yang, S. Kandula, M. Huynh, S. K. Greene, G. Van Wye, W. Li, H. T. Chan, E. McGibbon, A. Yeung, D. Olson, A. Fine, J. Shaman, Estimating the infection-fatality risk of SARS-CoV-2 in New York City during the spring 2020 pandemic wave: a model-based analysis. *The Lancet. Infectious diseases* **21**, 203-212 (2021).

44. J. L. Anderson, An ensemble adjustment Kalman filter for data assimilation. *Mon. Weather Rev.* **129**, 2884-2903 (2001).

45. W. Yang, J. Shaman, A simple modification for improving inference of non-linear dynamical systems. *arXiv*, 1403.6804 (2014).

46. J. Biryukov, J. A. Boydston, R. A. Dunning, J. J. Yeager, S. Wood, A. L. Reese, A. Ferris, D. Miller, W. Weaver, N. E. Zeitouni, A. Phillips, D. Freeburger, I. Hooper, S. Ratnesar-Shumate, J. Yolitz, M. Krause, G. Williams, D. G. Dawson, A. Herzog, P. Dabisch, V. Wahl, M. C. Hevey, L. A. Altamura, M. B. Frieman, Increasing Temperature and Relative Humidity Accelerates Inactivation of SARS-CoV-2 on Surfaces. *mSphere* **5**, e00441-00420 (2020).

47. N. G. Reich, L. C. Brooks, S. J. Fox, S. Kandula, C. J. McGowan, E. Moore, D. Osthus, E. L. Ray, A. Tushar, T. K. Yamana, M. Biggerstaff, M. A. Johansson, R. Rosenfeld, J. Shaman, A

collaborative multiyear, multimodel assessment of seasonal influenza forecasting in the United States. *Proc Natl Acad Sci U S A* **116**, 3146-3154 (2019).

48. F. X. Diebold, R. S. Mariano, Comparing Predictive Accuracy. *J Bus Econ Stat* **13**, 253-263 (1995).

Figure and Table Captions

Fig 1. Geospatial distribution of the 10 states and overall COVID-19 outcomes. Heatmaps show reported cumulative COVID-19 incidence rates (A) and COVID-19-associated mortality rates (B) in the 10 states included in this study. Line plots show reported weekly number of COVID-19 cases (C) and COVID-19-associated deaths (D) during the study period, for each state.

Fig 2. Example forecasts. Vertical dashed lines indicate the week of forecast. Dots show reported weekly cases per 1 million people; only those to the left of the vertical lines are used to calibrate the model and those to the right of the vertical lines are plotted for comparison. Blue lines and blue areas (darker blue = 50% CI; lighter blue = 80% CI) show model training estimates. Red lines and red areas (dark red = 50% CI; lighter red = 80% CI) show model forecasts using model settings as labeled in the subtitles.

Fig 3. Impact of deflation on forecast performance. Heatmaps show the differences in mean log score (A) or point prediction accuracy (B) between all forecast approaches with different deflation settings (deflation factor $\gamma = 0.95$ vs none in the 1st column, 0.9 vs none in the 2nd column, and 0.9 vs 0.95 in the 3rd column; see panel subtitles). Results are aggregated for each forecast approach (see specific settings of new variants and seasonality in the y-axis labels) and location (x-axis) over all forecast targets and forecast weeks, for cases (1st row) and deaths (2nd row), separately. For each pairwise comparison (e.g., 0.95 vs none), a positive difference in log score or point prediction accuracy indicates the former approach (e.g., 0.95) outperforms the latter (e.g., none).

Fig 4. Impact of new variant settings on forecast performance. Heatmaps show the differences in mean log score (A) or point prediction accuracy (B) between forecast approaches with vs without anticipation of new variant emergence. All forecasts here were generated using a deflation factor of 0.9. Results are aggregated for each forecast approach (see specific setting of seasonality in panel subtitles), variant wave (y-axis), and location (x-axis) over all forecast targets and forecast weeks for cases (1st row) and deaths (2nd row), separately. A positive difference indicates superior performance of the forecast approach with anticipation of new variant emergence.

Fig 5. Impact of seasonality settings on forecast performance. Heatmaps show the differences in mean log score (A) or point prediction accuracy (B), between pairs of forecast approaches with different seasonality settings (see panel subtitles). All forecasts here were generated using a deflation factor of 0.9 and the new variant setting. Results are aggregated for each forecast target (y-axis) and location (x-axis), over either the respiratory virus season (first 3 columns) or the off season (last 3 columns), for cases (1st row) and deaths (2nd row), separately. For each pairwise comparison (e.g., fixed vs no seasonality), a positive difference in log score or point prediction accuracy indicates the former approach (e.g., with fixed seasonality) outperforms the latter (e.g., with no seasonality).

Fig 6. Probabilistic forecast accuracy of the optimized and baseline forecast approaches. Boxplots show the distributions of pair-wise difference in log score by variant period (A) or respiratory virus season (B; see panel subtitles). Results are aggregated by location (color-coded for each state) and forecast target (x-axis), for cases and deaths (see panel subtitles), separately. The numbers show the range of number of evaluations of each forecast target (e.g., 59 predictions of peak week during the pre-Omicron period, for each state; 16-20 predictions of peak week during the Omicron period, depending on the timing of Omicron detection in each state). A positive difference indicates superior log score of the optimized forecast approach.

Fig 7. Point prediction accuracy of the optimized and baseline forecast systems. Points show the average accuracy over all forecast weeks (A) or respiratory virus season (B). Results are aggregated by location (x-axis) and forecast target (panel subtitles) for cases (1st row) and deaths (2nd row, see panel subtitles) separately. Filled dots show the mean accuracy of forecasts generated using the baseline system; filled triangles show the accuracy of forecasts generated using the optimized forecast system. The lines linking the two accuracies show the changes (mostly increases, as the triangles are more often above the dots), due to the combined application of the three proposed strategies (deflation, new variants, and transformed seasonality settings). Note all forecasts were generated retrospectively; to enable comparison of the model settings, mobility and vaccination data and estimates of infection detection rate and infection fatality risk during the forecast period were used (see main text for detail).

Table 1. Comparison of probabilistic forecast accuracy by the optimized and the baseline forecast approaches. Numbers show the relative difference in mean log score computed using Eqn 6, the median of pairwise difference in log score (95% CI of the median); asterisk (*) indicates if the median is significantly >0 or <0 at the $\alpha = 0.05$ level, per a Wilcoxon rank sum test. Positive numbers indicate superior performance of the optimized forecast approach.

Table 2. Comparison of point prediction accuracy by the optimized and the baseline forecast approaches. Numbers show the mean point prediction accuracy of forecasts generated using the baseline v. the optimized forecast approach; asterisk (*) indicates if the median of pairwise accuracy difference is significantly >0 or <0 at the $\alpha = 0.05$ level, per a Wilcoxon rank sum test. Note all forecasts were generated retrospectively; to enable comparison of forecast approaches, mobility and vaccination data and estimates of infection detection rate and infection fatality risk during the forecast period were used (see main text for detail).

Supplemental Figure and Table Captions

Fig S1. Comparison of seasonality forms. For each state (each panel), the blue line shows the estimated trend of seasonal infection risk using Eqns 3a-b and location weather data (temperature and humidity). Grey lines show 100 examples of the transformed seasonal trends per Eqns 4a-d with parameters randomly sampled from the best parameter ranges (Table S6); the black line shows the mean of the 100 example trends.

Fig S2. Impact of deflation on probabilistic forecast of different targets. Heatmaps show differences in mean log score for cases (A) and deaths (B), between each forecast approach with different deflation settings (deflation factor γ = 0.95 vs none in the 1st row, 0.9 vs none in the 2nd row, and 0.9 vs 0.95 in the 3rd row; see panel subtitles). Results are aggregated over all forecast weeks for each type of target (y-axis), forecast approach (see specific settings of new variants and seasonality in subtitles), and location (x-axis). For each pairwise comparison (e.g., 0.95 vs none), a positive difference indicates the former approach (e.g., 0.95) outperforms the latter (e.g., none).

Fig S3. Impact of deflation on point estimate accuracy of different targets. Heatmaps show differences in forecast accuracy of point estimates for cases (A) and deaths (B), between each forecast approach with different deflation settings (deflation factor γ = 0.95 vs none in the 1st row, 0.9 vs none in the 2nd row, and 0.9 vs 0.95 in the 3rd row; see panel subtitles). Results are aggregated over all forecast weeks for each type of target (y-axis), forecast approach (see specific settings of new variants and seasonality in subtitles), and location (x-axis). For each pairwise comparison (e.g., 0.95 vs none), a positive difference indicates the former approach (e.g., 0.95) outperforms the latter (e.g., none).

Fig S4. Comparison of forecast performance using the transformed seasonality function, with different parameter ranges. The parameter ranges are shown in x-axis labels for the three parameters in Eqn4a-d (from bottom to top: p_{shift}, δ, and $b_{t,\,lwr}$). 'x's indicate the best parameter ranges for the corresponding state.

Table S1. Impact of deflation. Numbers show the relative difference in mean log score computed using Eqn 6, or relative difference in mean point prediction accuracy computed using Eqn 7. For each pairwise comparison (e.g., 0.95 vs none), a positive difference indicates the former approach (e.g., 0.95) outperforms the latter (e.g., none).

Table S2. Impact of new variants settings. Numbers show the relative difference in mean log score computed using Eqn 6, or relative difference in mean point prediction accuracy computed using Eqn 7, by variant wave. A positive number indicates superior performance of the forecast approach with anticipation of new variant emergence.

Table S3. Impact of seasonality, aggregated over all 10 states. Numbers show the relative difference in mean log score or point prediction accuracy, the median of pair-wise difference in log score (95% CI of the median); asterisk (*) indicates if the median is significantly >0 or <0 at the α = 0.05 level, per a Wilcoxon rank sum test. A positive difference indicates superior log score or point prediction accuracy of the first listed approach; a negative difference indicates superior log score or point prediction accuracy of the second listed approach.

Table S4. Impact of seasonality, by state. Numbers show the relative difference in mean log score or point prediction accuracy, the median of pair-wise difference in log score (95% CI of

the median); asterisk (*) indicates if the median is significantly >0 or <0 at the $\alpha = 0.05$ level, per a Wilcoxon rank sum test. A positive difference indicates superior log score or point prediction accuracy of the first listed approach; a negative difference indicates superior log score or point prediction accuracy of the second listed approach.

Table S5. Prior ranges for the parameters and variables used in the model-inference system. Parameters/state variables are initialized by drawing from uniform distributions specified in the rows labeled "Initialization". During the filtering process, space-reprobing is applied to explore the state space, i.e., a small fraction of the ensemble members are randomly replaced with values drawn from the uniform distributions specified in the rows labeled with "Space-reprobing".

Figures

Fig 1. Geospatial distribution of the 10 states and overall COVID-19 outcomes. Heatmaps show reported cumulative COVID-19 incidence rates (A) and COVID-19-associated mortality rates (B) in the 10 states included in this study. Line plots show reported weekly number of COVID-19 cases (C) and COVID-19-associated deaths (D) during the study period, for each state.

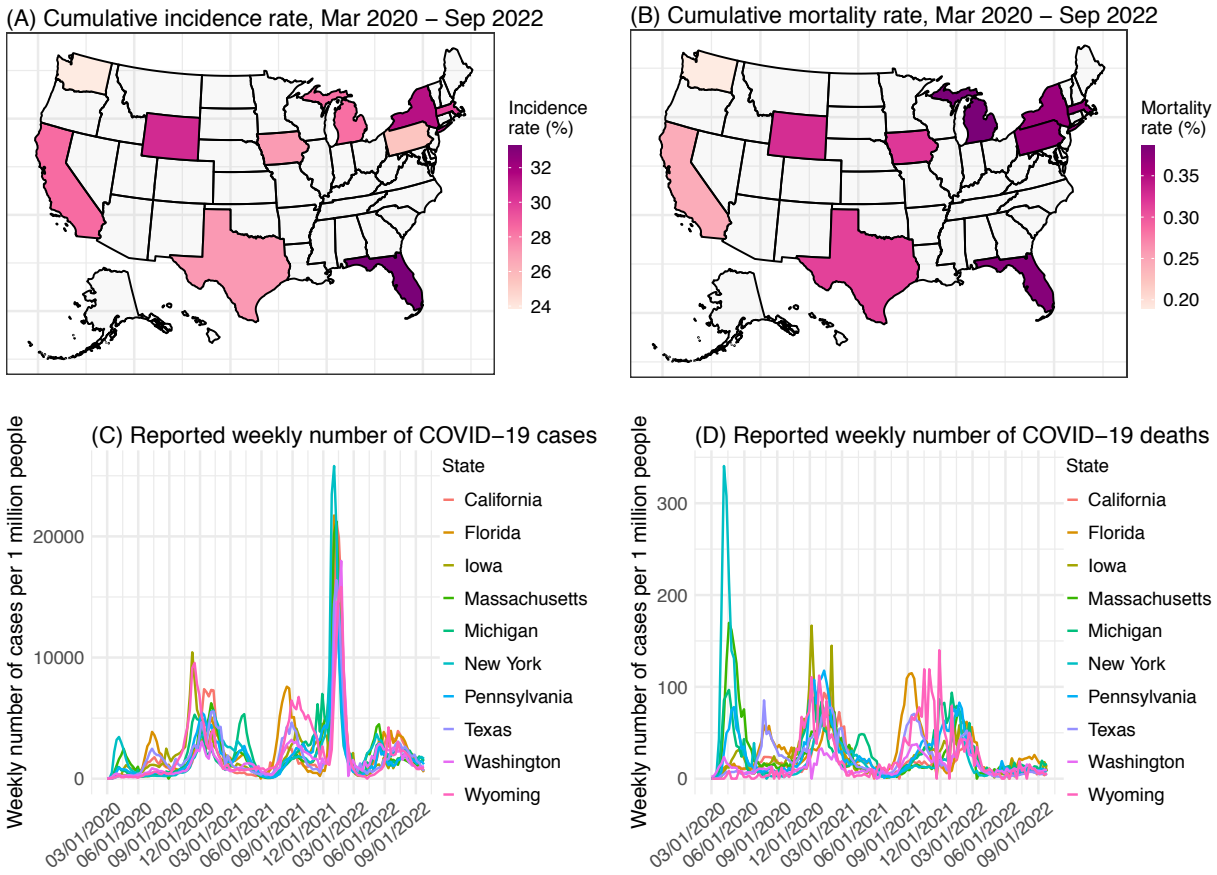


Fig 2. Example forecasts. Vertical dashed lines indicate the week of forecast. Dots show reported weekly cases per 1 million people; only those to the left of the vertical lines are used to calibrate the model and those to the right of the vertical lines are plotted for comparison. Blue lines and blue areas (darker blue = 50% CI; lighter blue = 80% CI) show model training estimates. Red lines and red areas (dark red = 50% CI; lighter red = 80% CI) show model forecasts using model settings as labeled in the subtitles.

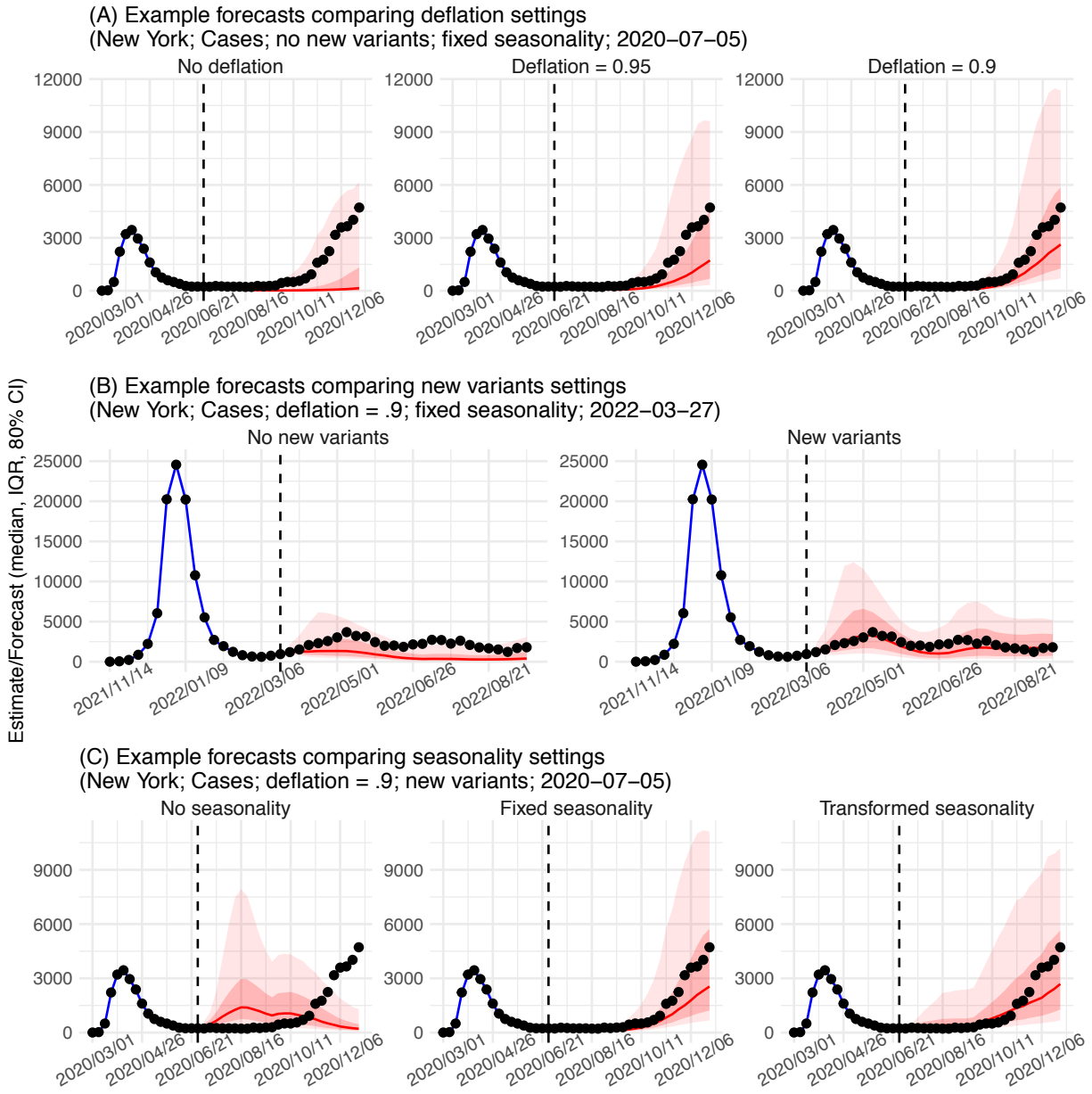


Fig 3. Impact of deflation on forecast performance. Heatmaps show the differences in mean log score (A) or point prediction accuracy (B) between all forecast approaches with different deflation settings (deflation factor γ = 0.95 vs none in the 1st column, 0.9 vs none in the 2nd column, and 0.9 vs 0.95 in the 3rd column; see panel subtitles). Results are aggregated for each forecast approach (see specific settings of new variants and seasonality in the y-axis labels) and location (x-axis) over all forecast targets and forecast weeks, for cases (1st row) and deaths (2nd row), separately. For each pairwise comparison (e.g., 0.95 vs none), a positive difference in log score or point prediction accuracy indicates the former approach (e.g., 0.95) outperforms the latter (e.g., none).

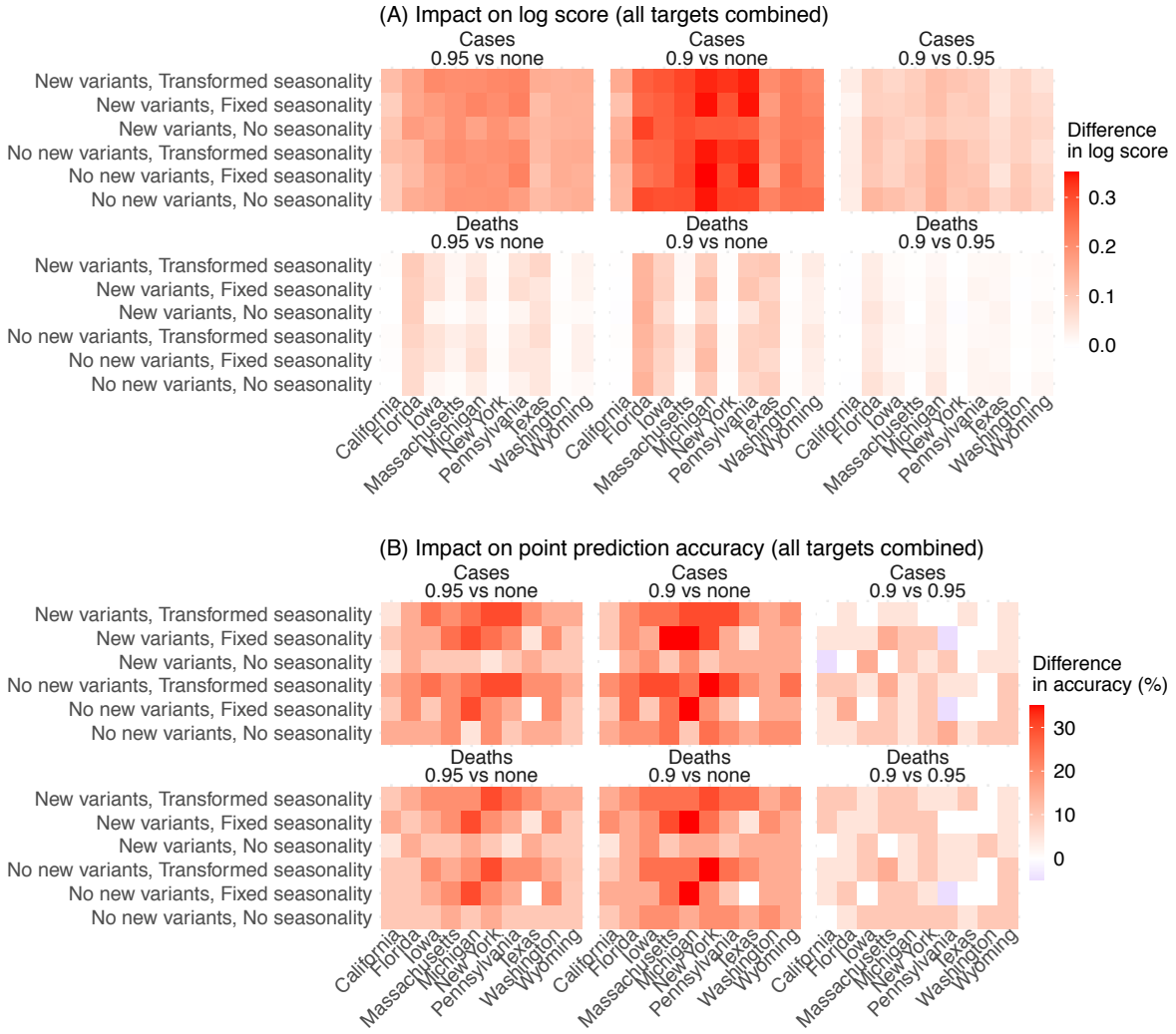


Fig 4. Impact of new variant settings on forecast performance. Heatmaps show the differences in mean log score (A) or point prediction accuracy (B) between forecast approaches with vs without anticipation of new variant emergence. All forecasts here were generated using a deflation factor of 0.9. Results are aggregated for each forecast approach (see specific setting of seasonality in panel subtitles), variant wave (y-axis), and location (x-axis) over all forecast targets and forecast weeks for cases (1st row) and deaths (2nd row), separately. A positive difference indicates superior performance of the forecast approach with anticipation of new variant emergence.

Fig 5. Impact of seasonality settings on forecast performance. Heatmaps show the differences in mean log score (A) or point prediction accuracy (B), between pairs of forecast approaches with different seasonality settings (see panel subtitles). All forecasts here were generated using a deflation factor of 0.9 and the new variant setting. Results are aggregated for each forecast target (y-axis) and location (x-axis), over either the respiratory virus season (first 3 columns) or the off season (last 3 columns), for cases (1st row) and deaths (2nd row), separately. For each pairwise comparison (e.g., fixed vs no seasonality), a positive difference in log score or point prediction accuracy indicates the former approach (e.g., with fixed seasonality) outperforms the latter (e.g., with no seasonality).

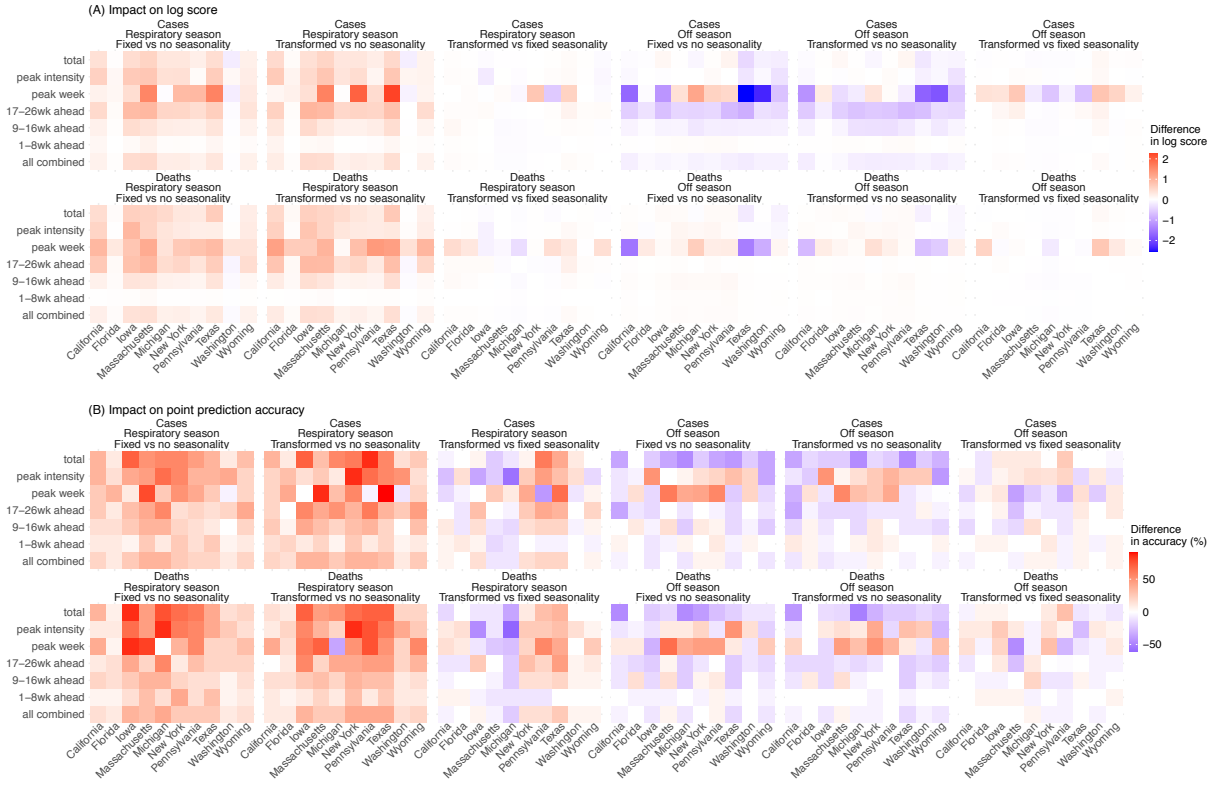


Fig 6. Probabilistic forecast accuracy of the optimized and baseline forecast approaches. Boxplots show the distributions of pair-wise difference in log score by variant period (A) or respiratory virus season (B; see panel subtitles). Results are aggregated by location (color-coded for each state) and forecast target (x-axis), for cases and deaths (see panel subtitles), separately. The numbers show the range of number of evaluations of each forecast target (e.g., 59 predictions of peak week during the pre-Omicron period, for each state; 16-20 predictions of peak week during the Omicron period, depending on the timing of Omicron detection in each state). A positive difference indicates superior log score of the optimized forecast approach.

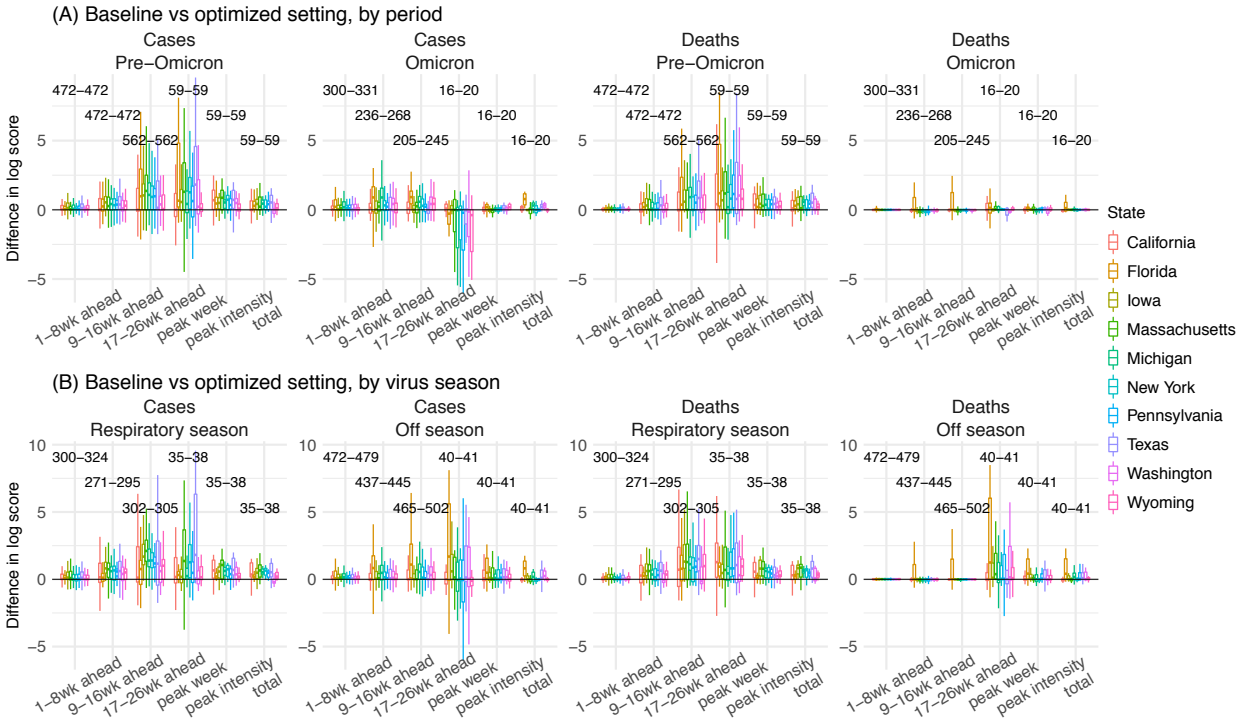
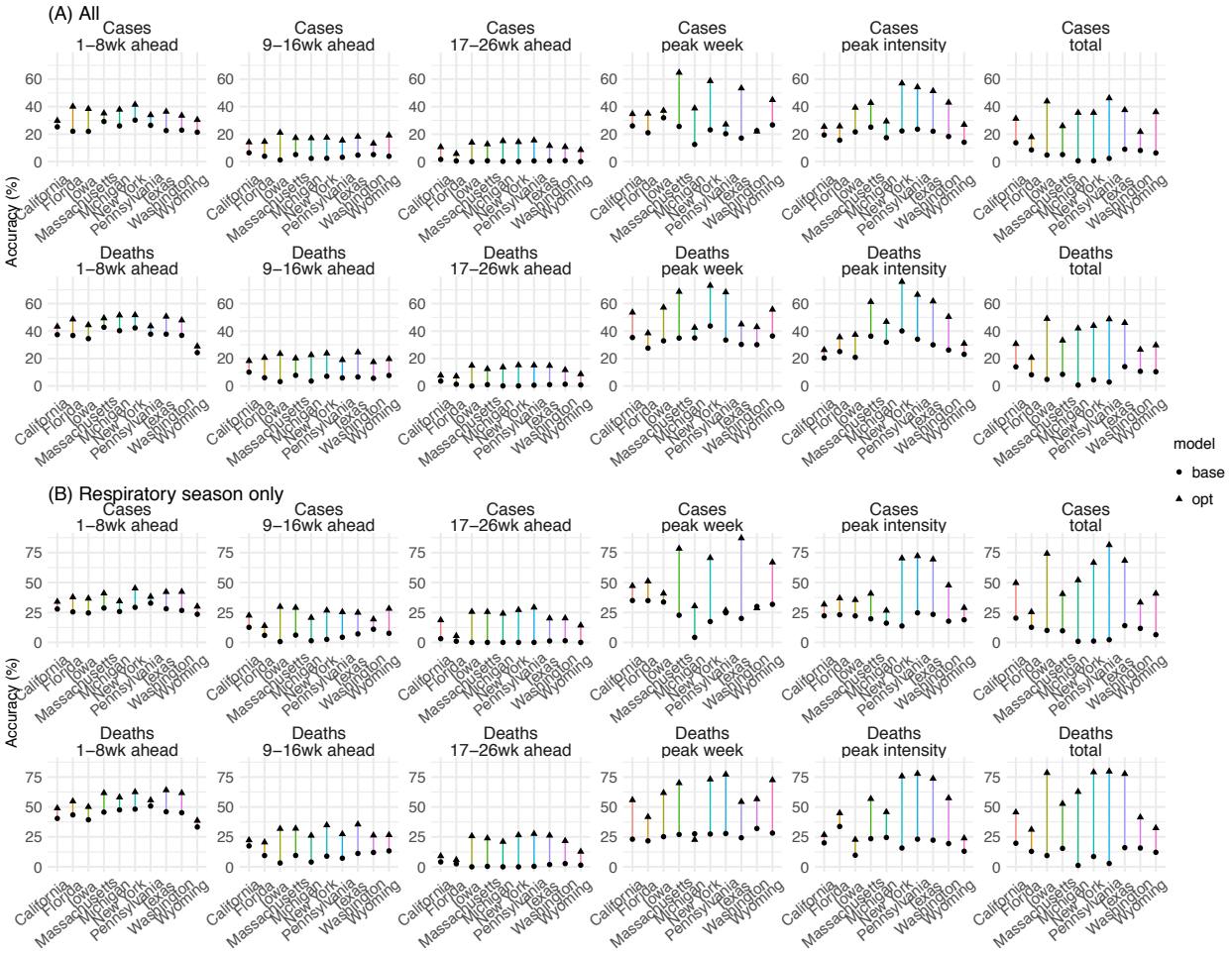


Fig 7. Point prediction accuracy of the optimized and baseline forecast systems. Points show the average accuracy over all forecast weeks (A) or respiratory virus season (B). Results are aggregated by location (x-axis) and forecast target (panel subtitles) for cases (1st row) and deaths (2nd row, see panel subtitles) separately. Filled dots show the mean accuracy of forecasts generated using the baseline system; filled triangles show the accuracy of forecasts generated using the optimized forecast system. The lines linking the two accuracies show the changes (mostly increases, as the triangles are more often above the dots), due to the combined application of the three proposed strategies (deflation, new variants, and transformed seasonality settings). Note all forecasts were generated retrospectively; to enable comparison of the model settings, mobility and vaccination data and estimates of infection detection rate and infection fatality risk during the forecast period were used (see main text for detail).



Tables

Table 1. Comparison of probabilistic forecast accuracy by the optimized and the baseline forecast approaches. Numbers show the relative difference in mean log score computed using Eqn 6, the median of pairwise difference in log score (95% CI of the median); asterisk (*) indicates if the median is significantly >0 or <0 at the α = 0.05 level, per a Wilcoxon rank sum test. Positive numbers indicate superior performance of the optimized forecast approach.

State	Measure	All	Pre-Omicron period	Omicron period	Respiratory season	Off season
All	Cases	63.6%, 0.34 (0.33, 0.35) *	85.3%, 0.43 (0.42, 0.44) *	27.2%, 0.21 (0.2, 0.22) *	105%, 0.53 (0.51, 0.54) *	40.6%, 0.22 (0.21, 0.23) *
All	Deaths	38%, 0.07 (0.06, 0.08) *	61.7%, 0.25 (0.24, 0.27) *	0.184%, -0.018 (-0.02, -0.016) *	84.5%, 0.42 (0.41, 0.44) *	13.7%, 0 (0, 0) *
California	Cases	46.3%, 0.18 (0.16, 0.21) *	65.8%, 0.26 (0.22, 0.32) *	15.9%, 0.11 (0.08, 0.14) *	93.7%, 0.35 (0.27, 0.43) *	21.2%, 0.13 (0.1, 0.15) *
California	Deaths	29.9%, 0.01 (0, 0.01) *	50.6%, 0.06 (0.04, 0.1) *	-1.26%, -0.028 (-0.032, -0.024) *	91.4%, 0.34 (0.26, 0.42) *	0.225%, -0.0095 (-0.013, -0.006) *
Florida	Cases	119%, 0.51 (0.46, 0.55) *	134%, 0.42 (0.37, 0.49) *	90.7%, 0.63 (0.57, 0.68) *	48.9%, 0.23 (0.2, 0.27) *	183%, 0.77 (0.71, 0.83) *
Florida	Deaths	84.5%, 0.23 (0.2, 0.26) *	117%, 0.28 (0.24, 0.33) *	33.6%, 0.11 (0.06, 0.18) *	58.5%, 0.27 (0.24, 0.31) *	104%, 0.17 (0.12, 0.26) *
Iowa	Cases	69.8%, 0.4 (0.37, 0.44) *	108%, 0.59 (0.55, 0.63) *	10.6%, 0.09 (0.06, 0.13) *	172%, 0.84 (0.77, 0.89) *	24.7%, 0.19 (0.16, 0.22) *
Iowa	Deaths	50.6%, 0.19 (0.16, 0.22) *	86.4%, 0.43 (0.39, 0.47) *	-4.49%, -0.024 (-0.03, -0.018) *	158%, 0.71 (0.66, 0.77) *	5.74%, 0 (0, 0.01) *
Massachusetts	Cases	84.1%, 0.42 (0.38, 0.46) *	131%, 0.66 (0.61, 0.72) *	17.6%, 0.17 (0.13, 0.2) *	213%, 0.91 (0.83, 0.99) *	29%, 0.16 (0.14, 0.19) *
Massachusetts	Deaths	40.1%, 0.04 (0.03, 0.06) *	68.9%, 0.28 (0.23, 0.34) *	-3.32%, -0.031 (-0.041, -0.022) *	130%, 0.61 (0.55, 0.67) *	0.439%, -0.005 (-0.009, -0.002) *
Michigan	Cases	76.4%, 0.48 (0.45, 0.51) *	86.9%, 0.49 (0.45, 0.52) *	55.9%, 0.46 (0.4, 0.51) *	119%, 0.62 (0.57, 0.68) *	53.2%, 0.39 (0.35, 0.42) *
Michigan	Deaths	36.9%, 0.12 (0.09, 0.14) *	60.9%, 0.32 (0.28, 0.35) *	-2.89%, -0.029 (-0.036, -0.022) *	85.6%, 0.48 (0.44, 0.52) *	12.2%, -0.0044 (-0.0095, -0.00048) *
New York	Cases	60.7%, 0.37 (0.34, 0.4) *	79.6%, 0.47 (0.44, 0.5) *	28.8%, 0.22 (0.19, 0.25) *	117%, 0.63 (0.57, 0.69) *	31.1%, 0.21 (0.18, 0.24) *
New York	Deaths	20.6%, 0.01 (0.01, 0.02) *	35.5%, 0.09 (0.06, 0.15) *	-4.31%, -0.025 (-0.032, -0.019) *	61.1%, 0.37 (0.33, 0.4) *	-0.7%, -0.0056 (-0.009, -0.0029) *
Pennsylvania	Cases	49.2%, 0.32 (0.3, 0.35) *	72.4%, 0.45 (0.42, 0.48) *	12.1%, 0.16 (0.13, 0.18) *	94.9%, 0.56 (0.52, 0.6) *	24.7%, 0.18 (0.16, 0.2) *

Pennsylvania	Deaths	30.5%, 0.12 (0.09, 0.15) *	51.7%, 0.3 (0.27, 0.33) *	-3.29%, -0.024 (-0.034, -0.016) *	71.9%, 0.41 (0.38, 0.44) *	8.45%, 0.01 (0, 0.01) *
Texas	Cases	74.3%, 0.32 (0.29, 0.36) *	120%, 0.51 (0.46, 0.57) *	9.88%, 0.07 (0.04, 0.12) *	158%, 0.64 (0.56, 0.71) *	34%, 0.18 (0.15, 0.21) *
Texas	Deaths	59.3%, 0.17 (0.14, 0.21) *	103%, 0.47 (0.42, 0.52) *	-1.88%, -0.012 (-0.015, -0.009) *	141%, 0.67 (0.6, 0.74) *	20.5%, 0.01 (0.01, 0.02) *
Washington	Cases	38.5%, 0.26 (0.23, 0.28) *	38.8%, 0.24 (0.21, 0.27) *	37.8%, 0.28 (0.25, 0.32) *	31.7%, 0.28 (0.24, 0.32) *	43.3%, 0.24 (0.2, 0.26) *
Washington	Deaths	13.8%, 0 (-0.0015, 0)	24.1%, 0.04 (0.02, 0.08) *	-4.22%, -0.028 (-0.034, -0.022) *	19.1%, 0.07 (0.03, 0.13) *	10.4%, -0.001 (-0.0029, 0.00093)
Wyoming	Cases	34.5%, 0.21 (0.19, 0.23) *	45.7%, 0.25 (0.23, 0.28) *	13.4%, 0.11 (0.07, 0.15) *	73.8%, 0.42 (0.38, 0.46) *	13.7%, 0.09 (0.07, 0.12) *
Wyoming	Deaths	26.8%, 0.05 (0.04, 0.08) *	42.7%, 0.19 (0.17, 0.22) *	-1.53%, -0.014 (-0.019, -0.01) *	73.9%, 0.36 (0.33, 0.41) *	3.06%, 0 (0.00094, 0.01) *

Table 2. Comparison of point prediction accuracy by the optimized and the baseline forecast approaches. Numbers show the mean point prediction accuracy of forecasts generated using the baseline v. the optimized forecast approach; asterisk (*) indicates if the median of pairwise accuracy difference is significantly >0 or <0 at the α = 0.05 level, per a Wilcoxon rank sum test. Note all forecasts were generated retrospectively; to enable comparison of forecast approaches, mobility and vaccination data and estimates of infection detection rate and infection fatality risk during the forecast period were used (see main text for detail).

State	Measure	1-8wk ahead	9-16wk ahead	17-26wk ahead	peak week	peak intensity	total
All	Cases	26% v 38% *	4% v 20% *	1% v 16% *	24% v 42% *	20% v 40% *	7% v 33% *
All	Deaths	39% v 48% *	7% v 25% *	1% v 16% *	35% v 56% *	30% v 50% *	9% v 36% *
California	Cases	27% v 36% *	8% v 17% *	3% v 12% *	26% v 34%	20% v 28% *	13% v 30% *
California	Deaths	38% v 45% *	10% v 23% *	4% v 11% *	37% v 56% *	23% v 29%	13% v 26% *
Florida	Cases	25% v 43% *	4% v 16% *	1% v 8% *	23% v 35% *	19% v 28% *	9% v 18% *
Florida	Deaths	40% v 51% *	6% v 25% *	1% v 8% *	27% v 31%	28% v 41% *	8% v 22% *
Iowa	Cases	24% v 37% *	1% v 25% *	0% v 21% *	32% v 41% *	22% v 38% *	6% v 44% *
Iowa	Deaths	37% v 47% *	4% v 26% *	0% v 23% *	34% v 58% *	22% v 41% *	6% v 46% *
Massachusetts	Cases	30% v 36% *	5% v 20% *	1% v 16% *	23% v 61% *	24% v 43% *	5% v 25% *
Massachusetts	Deaths	44% v 50% *	8% v 24% *	1% v 15% *	34% v 67% *	36% v 60% *	8% v 34% *
Michigan	Cases	28% v 40% *	2% v 21% *	0% v 19% *	15% v 37% *	20% v 30% *	1% v 27% *
Michigan	Deaths	43% v 52% *	4% v 26% *	0% v 18% *	38% v 43%	33% v 43% *	3% v 28% *
New York	Cases	30% v 42% *	2% v 24% *	0% v 18% *	26% v 59% *	24% v 59% *	3% v 39% *
New York	Deaths	44% v 51% *	7% v 28% *	0% v 20% *	44% v 79% *	40% v 78% *	7% v 47% *
Pennsylvania	Cases	27% v 37% *	4% v 22% *	0% v 23% *	24% v 32% *	20% v 59% *	4% v 50% *
Pennsylvania	Deaths	40% v 48% *	9% v 25% *	1% v 23% *	33% v 70% *	35% v 67% *	9% v 53% *
Texas	Cases	24% v 38% *	5% v 22% *	1% v 16% *	20% v 50% *	20% v 51% *	8% v 44% *
Texas	Deaths	39% v 49% *	7% v 26% *	1% v 19% *	34% v 46% *	33% v 61% *	13% v 50% *
Washington	Cases	24% v 37% *	6% v 17% *	1% v 17% *	22% v 24%	17% v 46% *	7% v 30% *
Washington	Deaths	39% v 51% *	7% v 25% *	2% v 18% *	32% v 45% *	24% v 57% *	9% v 31% *
Wyoming	Cases	23% v 32% *	5% v 19% *	0% v 10% *	28% v 42% *	15% v 20% *	8% v 25% *
Wyoming	Deaths	26% v 32% *	9% v 22% *	1% v 10% *	38% v 64% *	23% v 28% *	13% v 23% *

Supplemental Figures and Tables

Fig S1. Comparison of seasonality forms. For each state (each panel), the blue line shows the estimated trend of seasonal infection risk using Eqns 3a-b and location weather data (temperature and humidity). Grey lines show 100 examples of the transformed seasonal trends per Eqns 4a-d with parameters randomly sampled from the best parameter ranges (Table S6); the black line shows the mean of the 100 example trends.

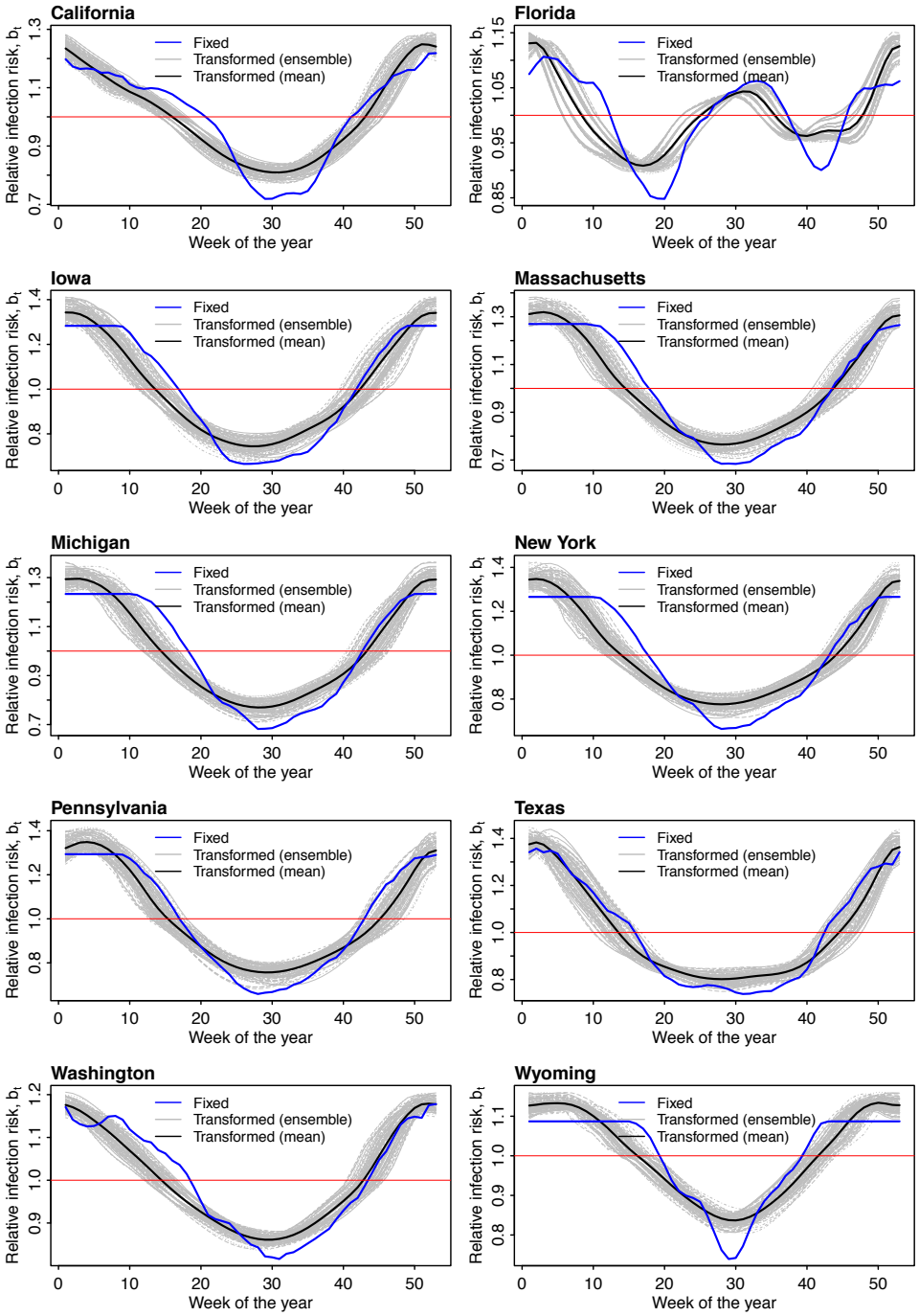


Fig S2. Impact of deflation on probabilistic forecast of different targets. Heatmaps show differences in mean log score for cases (A) and deaths (B), between each forecast approach with different deflation settings (deflation factor γ = 0.95 vs none in the 1st row, 0.9 vs none in the 2nd row, and 0.9 vs 0.95 in the 3rd row; see panel subtitles). Results are aggregated over all forecast weeks for each type of target (y-axis), forecast approach (see specific settings of new variants and seasonality in subtitles), and location (x-axis). For each pairwise comparison (e.g., 0.95 vs none), a positive difference indicates the former approach (e.g., 0.95) outperforms the latter (e.g., none).

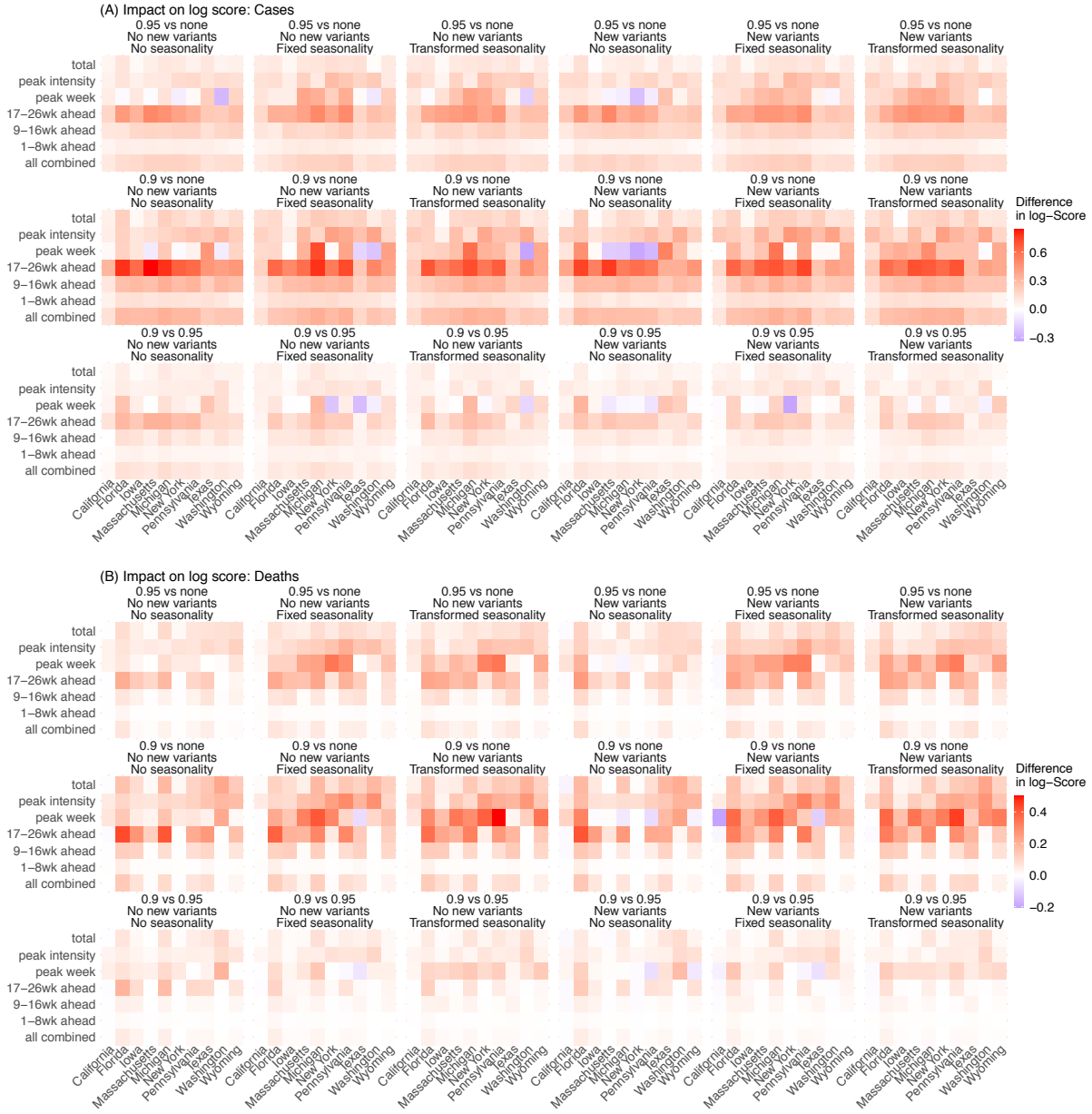


Fig S3. Impact of deflation on point estimate accuracy of different targets. Heatmaps show differences in forecast accuracy of point estimates for cases (A) and deaths (B), between each forecast approach with different deflation settings (deflation factor γ = 0.95 vs none in the 1st row, 0.9 vs none in the 2nd row, and 0.9 vs 0.95 in the 3rd row; see panel subtitles). Results are aggregated over all forecast weeks for each type of target (y-axis), forecast approach (see specific settings of new variants and seasonality in subtitles), and location (x-axis). For each pairwise comparison (e.g., 0.95 vs none), a positive difference indicates the former approach (e.g., 0.95) outperforms the latter (e.g., none).

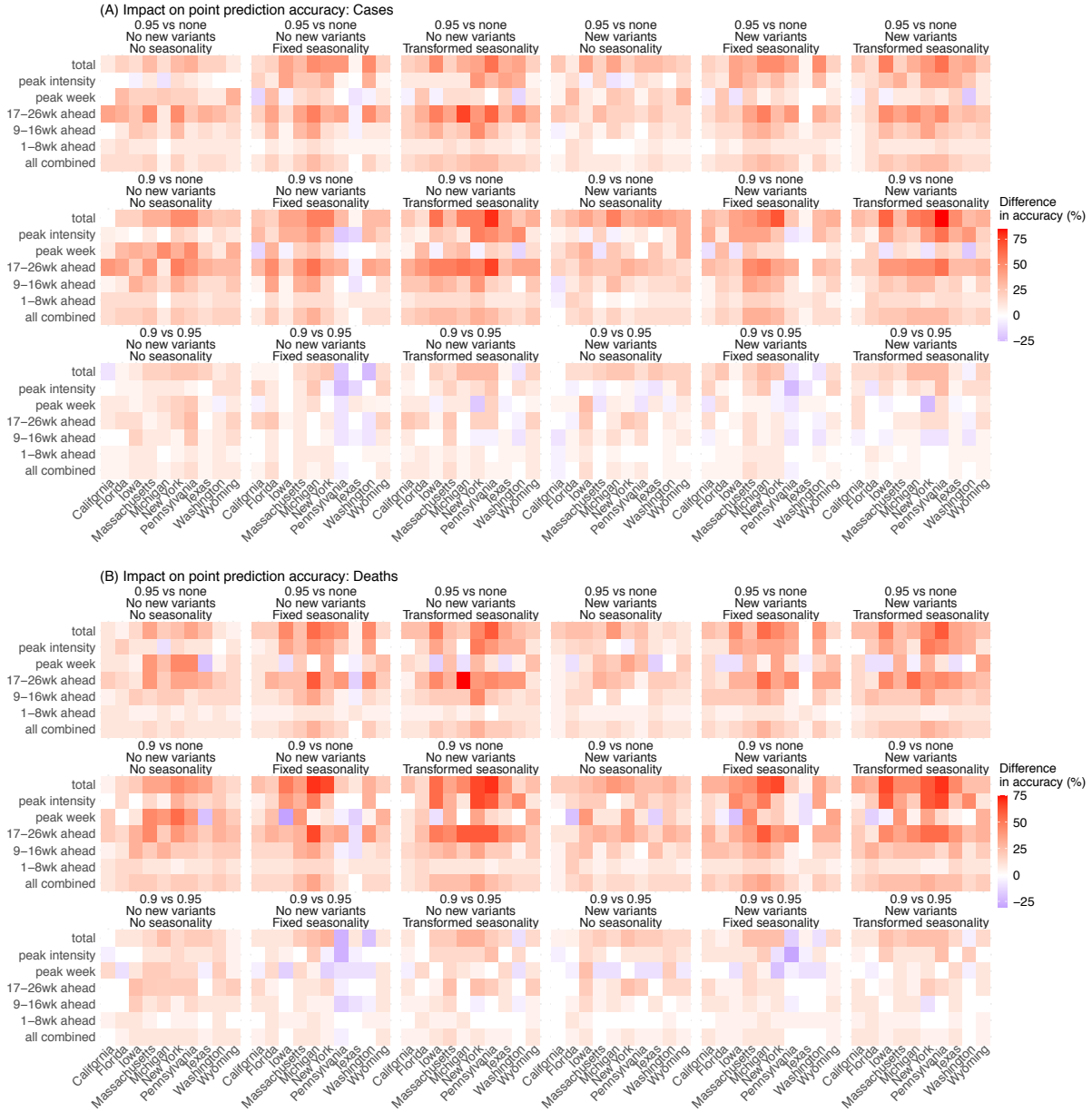


Fig S4. Comparison of forecast performance using the transformed seasonality function, with different parameter ranges. The parameter ranges are shown in x-axis labels for the three parameters in Eqn4a-d (from bottom to top: p_{shift}, δ, and $b_{t,\ lwr}$). 'x's indicate the best parameter ranges for the corresponding state.

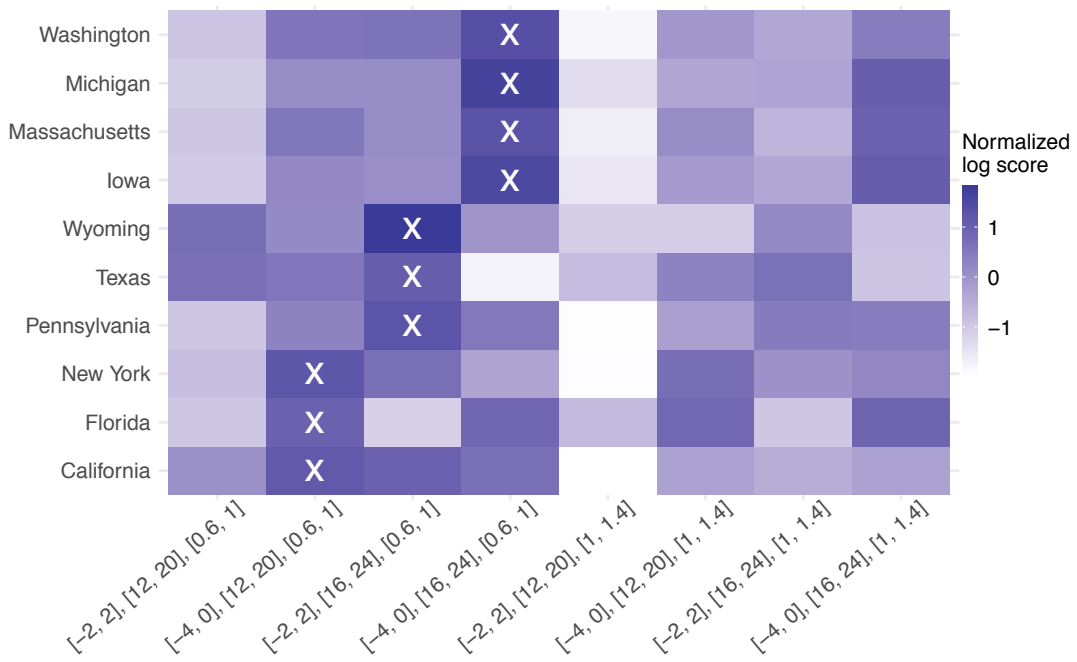


Table S1. Impact of deflation. Numbers show the relative difference in mean log score computed using Eqn 6, or relative difference in mean point prediction accuracy computed using Eqn 7. For each pairwise comparison (e.g., 0.95 vs none), a positive difference indicates the former approach (e.g., 0.95) outperforms the latter (e.g., none).

				Pairwise comparison of deflation settings		
Metric	Measure	New variant setting	Seasonality setting	0.95 vs none	0.9 vs none	0.9 vs 0.95
Log score	Cases	No new variants	No seasonality	20.1%	43.3%	19.4%
Log score	Cases	No new variants	Fixed seasonality	18.1%	34.2%	13.6%
Log score	Cases	No new variants	Transformed seasonality	19%	36.8%	15%
Log score	Cases	New variants	No seasonality	20.2%	38.7%	15.4%
Log score	Cases	New variants	Fixed seasonality	19.7%	35%	12.8%
Log score	Cases	New variants	Transformed seasonality	20%	35.9%	13.3%
Log score	Deaths	No new variants	No seasonality	7.67%	17.2%	8.85%
Log score	Deaths	No new variants	Fixed seasonality	7.8%	13.3%	5.1%
Log score	Deaths	No new variants	Transformed seasonality	7.94%	14.2%	5.75%
Log score	Deaths	New variants	No seasonality	7.43%	15.1%	7.19%
Log score	Deaths	New variants	Fixed seasonality	8.23%	13%	4.45%
Log score	Deaths	New variants	Transformed seasonality	8.4%	13.2%	4.45%
Accuracy	Cases	No new variants	No seasonality	28%	46.5%	14.4%
Accuracy	Cases	No new variants	Fixed seasonality	30.5%	38.6%	6.23%
Accuracy	Cases	No new variants	Transformed seasonality	47.8%	63.2%	10.4%
Accuracy	Cases	New variants	No seasonality	20.5%	32.5%	10%
Accuracy	Cases	New variants	Fixed seasonality	38.3%	47.2%	6.45%
Accuracy	Cases	New variants	Transformed seasonality	47%	55%	5.44%
Accuracy	Deaths	No new variants	No seasonality	15.5%	27%	9.98%
Accuracy	Deaths	No new variants	Fixed seasonality	20%	24.6%	3.84%
Accuracy	Deaths	No new variants	Transformed seasonality	26.8%	36%	7.22%
Accuracy	Deaths	New variants	No seasonality	14.7%	24.1%	8.18%
Accuracy	Deaths	New variants	Fixed seasonality	26.1%	34.1%	6.36%
Accuracy	Deaths	New variants	Transformed seasonality	31.2%	39.9%	6.62%

Table S2. Impact of new variants settings. Numbers show the relative difference in mean log score computed using Eqn 6, or relative difference in mean point prediction accuracy computed using Eqn 7, by variant wave. A positive number indicates superior performance of the forecast approach with anticipation of new variant emergence.

Metric	Measure	Seasonality setting	2nd wave	Alpha	Delta	Omicron
Log score	Cases	No seasonality	-2.09%	0.0271%	66.9%	10.5%
Log score	Cases	Fixed seasonality	-0.573%	-0.33%	119%	36.6%
Log score	Cases	Transformed seasonality	-1.07%	-1.03%	93.7%	34.1%
Log score	Deaths	No seasonality	-1.7%	0.00306%	19.1%	-5.31%
Log score	Deaths	Fixed seasonality	-0.387%	-0.291%	34.1%	-2.27%
Log score	Deaths	Transformed seasonality	-0.918%	-0.651%	27.8%	-2.73%
Accuracy	Cases	No seasonality	-1.52%	-0.11%	89.1%	37.1%
Accuracy	Cases	Fixed seasonality	0.578%	2.16%	16.8%	95.8%
Accuracy	Cases	Transformed seasonality	-1.33%	1.96%	26.1%	77.8%
Accuracy	Deaths	No seasonality	-1.45%	0.121%	69.9%	40.2%
Accuracy	Deaths	Fixed seasonality	-0.278%	2.73%	15.9%	67.7%
Accuracy	Deaths	Transformed seasonality	-1.31%	2.47%	22.5%	63.2%

Table S3. Impact of seasonality, aggregated over all 10 states. Numbers show the relative difference in mean log score or point prediction accuracy, the median of pair-wise difference in log score (95% CI of the median); asterisk (*) indicates if the median is significantly >0 or <0 at the α = 0.05 level, per a Wilcoxon rank sum test. A positive difference indicates superior log score or point prediction accuracy of the first listed approach; a negative difference indicates superior log score or point prediction accuracy of the second listed approach.

Wave	Season	Metric	Measure	Fixed vs no seasonality	Transformed vs no seasonality	Transformed vs fixed seasonality
All	All	Log score	Cases	-4.45%, -0.0075 (-0.012, -0.0034)*	-2.8%, -0.006 (-0.01, -0.002)*	1.73%, 0.01 (0, 0.01)*
All	All	Log score	Deaths	12.7%, 0.04 (0.04, 0.04)*	14.2%, 0.04 (0.04, 0.04)*	1.32%, -0.0015 (-0.002, -0.00054)*
All	All	Accuracy	Cases	20.4%, 10% (10%, 10%)	26.3%, 10% (10%, 10%)	4.92%, 0% (0%, 5%)*
All	All	Accuracy	Deaths	15.8%, 10% (9.99%, 10%)	18.2%, 10% (10%, 10%)	2.09%, 0% (0%, 0%)*
All	Respiratory season	Log score	Cases	39.8%, 0.19 (0.18, 0.2)	42.3%, 0.2 (0.18, 0.21)	1.8%, 0.01 (0, 0.01)*
All	Respiratory season	Log score	Deaths	37.4%, 0.16 (0.15, 0.17)	41.1%, 0.17 (0.15, 0.18)	2.64%, 0.01 (0, 0.01)*
All	Respiratory season	Accuracy	Cases	60.3%, 20% (20%, 25%)	75.3%, 25% (25%, 30%)	9.38%, 5.01% (5%, 5%)
All	Respiratory season	Accuracy	Deaths	47.3%, 25% (20%, 25%)	54.3%, 25% (25%, 30%)	4.76%, 5% (0%, 5%)
All	Off season	Log score	Cases	-25.9%, -0.14 (-0.14, -0.13)*	-24.6%, -0.14 (-0.16, -0.13)*	1.68%, 0 (0, 0.01)*
All	Off season	Log score	Deaths	-1.3%, 0.02 (0.02, 0.03)*	-0.86%, 0.02 (0.02, 0.02)*	0.444%, -0.0031 (-0.0036, -0.0025)*
All	Off season	Accuracy	Cases	-5.95%, -5% (-5%, 0%)*	-6.04%, 0% (-4.99%, 0%)*	-0.0987%, 0% (0%, 0.01%)*
All	Off season	Accuracy	Deaths	-6.77%, -5% (-5%, -4.99%)*	-7.64%, -4.99% (-5%, -5%)*	-0.94%, 0% (0%, 0.01%)*
2nd wave	Off season	Log score	Cases	32%, 0.22 (0.2, 0.24)	23.5%, 0.16 (0.14, 0.17)	-6.42%, -0.056 (-0.063, -0.051)*
2nd wave	Off season	Log score	Deaths	23.9%, 0.17 (0.16, 0.19)	17.1%, 0.13 (0.12, 0.14)	-5.5%, -0.048 (-0.052, -0.043)*
2nd wave	Off season	Accuracy	Cases	37.6%, 15% (15%, 20%)	35.6%, 20% (15%, 20%)	-1.43%, -0.01% (-5%, 0%)*

2nd wave	Off season	Accuracy	Deaths	45.6%, 25% (20%, 30%)	28.1%, 20% (15%, 20%)	-12%, -10% (-15%, -10%)*
Alpha	Off season	Log score	Cases	-9.35%, 0.01 (0, 0.01)*	-24.7%, -0.052 (-0.11, 0)*	-16.9%, -0.087 (-0.11, -0.052)*
Alpha	Off season	Log score	Deaths	-3.33%, 0.01 (0.01, 0.01)*	-10.8%, 0.01 (0.01, 0.02)*	-7.74%, 0 (0, 0)*
Alpha	Off season	Accuracy	Cases	-42.4%, -30% (-35%, -25%)*	-39.8%, -30% (-35%, -25%)*	4.52%, 5% (-0.01%, 5.01%)
Alpha	Off season	Accuracy	Deaths	-28.5%, -25% (-25%, -20%)*	-33.8%, -30% (-30%, -25%)*	-7.53%, -5.01% (-10%, 0%)*
Delta	Off season	Log score	Cases	-49.6%, -0.57 (-0.61, -0.55)*	-35.4%, -0.32 (-0.35, -0.3)*	28.2%, 0.25 (0.23, 0.26)
Delta	Off season	Log score	Deaths	-12%, 0.02 (0.02, 0.02)*	-6.52%, 0.02 (0.02, 0.02)*	6.28%, -0.0024 (-0.003, -0.0015)*
Delta	Off season	Accuracy	Cases	-55.8%, -25% (-25%, -20%)*	-50%, -20% (-25%, -20%)*	13.1%, 4.99% (0%, 5%)*
Delta	Off season	Accuracy	Deaths	-50.3%, -25% (-25%, -20%)*	-43.2%, -20% (-20%, -15%)*	14.5%, 5% (4.99%, 10%)
Omicron	Off season	Log score	Cases	-19.5%, -0.15 (-0.16, -0.14)*	-24%, -0.16 (-0.17, -0.15)*	-5.68%, -0.013 (-0.017, -0.0085)*
Omicron	Off season	Log score	Deaths	4.77%, 0.03 (0.03, 0.03)*	3.84%, 0.02 (0.02, 0.03)*	-0.887%, -0.0035 (-0.0041, -0.003)*
Omicron	Off season	Accuracy	Cases	4.46%, 0.01% (0%, 5%)*	2.52%, 0% (0%, 5%)*	-1.86%, -0.01% (0%, 0%)*
Omicron	Off season	Accuracy	Deaths	-1.59%, 0% (0%, 0%)*	-1.78%, 0% (0%, 0%)*	-0.191%, -0.01% (0%, 0%)*

Table S4. Impact of seasonality, by state. Numbers show the relative difference in mean log score or point prediction accuracy, the median of pair-wise difference in log score (95% CI of the median); asterisk (*) indicates if the median is significantly >0 or <0 at the α = 0.05 level, per a Wilcoxon rank sum test. A positive difference indicates superior log score or point prediction accuracy of the first listed approach; a negative difference indicates superior log score or point prediction accuracy of the second listed approach.

State	Season	Metric	Measure	Fixed vs no seasonality	Transformed vs no seasonality	Transformed vs fixed seasonality
California	All	Log score	Cases	-7.14%, -0.042 (-0.062, -0.024)*	-1.53%, -0.028 (-0.045, -0.013)*	6.04%, 0.02 (0.01, 0.02)*
California	All	Log score	Deaths	16.7%, 0.04 (0.04, 0.05)*	21.4%, 0.04 (0.03, 0.04)*	3.96%, 0 (0, 0.01)*
California	All	Accuracy	Cases	13.1%, 5% (0%, 9.99%)	11.3%, 5% (0%, 10%)	-1.6%, 0% (0%, 0%)*
California	All	Accuracy	Deaths	4.74%, 0% (0.01%, 5%)*	1.86%, 0% (-0.01%, 5%)*	-2.75%, 0% (-4.99%, 0%)*
California	Respiratory season	Log score	Cases	42.5%, 0.17 (0.13, 0.22)	56.7%, 0.24 (0.19, 0.3)	9.96%, 0.05 (0.04, 0.06)*
California	Respiratory season	Log score	Deaths	48.4%, 0.21 (0.15, 0.3)	60%, 0.27 (0.19, 0.37)	7.8%, 0.03 (0.02, 0.04)*
California	Respiratory season	Accuracy	Cases	51.9%, 20% (15%, 20%)	58.9%, 20% (15%, 25%)	4.61%, 4.99% (-0.01%, 5%)*
California	Respiratory season	Accuracy	Deaths	27.9%, 15% (10%, 20%)	22.4%, 10% (5%, 15%)	-4.33%, -5% (-10%, 0%)*
California	Off season	Log score	Cases	-30.3%, -0.17 (-0.2, -0.14)*	-27.9%, -0.19 (-0.23, -0.16)*	3.49%, -0.0024 (-0.013, 0)*
California	Off season	Log score	Deaths	-0.598%, 0.03 (0.02, 0.03)*	0.856%, 0.02 (0.02, 0.03)*	1.46%, -1e-05 (-0.0014, 0.00098)*
California	Off season	Accuracy	Cases	-11%, -5% (-10%, 0%)*	-18.4%, -9.99% (-15%, -5.01%)*	-8.21%, -5% (-10%, 0%)*
California	Off season	Accuracy	Deaths	-8.26%, -5% (-10%, 0%)*	-9.65%, -5% (-10%, -5%)*	-1.52%, 0% (-4.99%, 5%)*
Florida	All	Log score	Cases	-12.6%, -0.02 (-0.032, -0.01)*	-6.11%, 0.001 (-0.0085, 0.01)*	7.47%, 0.03 (0.02, 0.04)*
Florida	All	Log score	Deaths	0.697%, 0.02 (0.01, 0.02)*	4.42%, 0.02 (0.02, 0.03)*	3.7%, 0.01 (0.01, 0.01)*
Florida	All	Accuracy	Cases	8.52%, 5% (5%, 10%)	11.5%, 5% (5%, 10%)	2.74%, 0% (0%, 5%)*
Florida	All	Accuracy	Deaths	3.73%, 5% (0%, 5.01%)	4.9%, 5% (0%, 5.01%)	1.14%, 0% (0%, 0%)*
Florida	Respiratory season	Log score	Cases	-2.01%, 0.03 (0.02, 0.05)*	1.72%, 0.04 (0.02, 0.05)*	3.81%, 0.02 (0.01, 0.04)*

Florida	Respiratory season	Log score	Deaths	2.77%, 0.05 (0.04, 0.06)	8.52%, 0.06 (0.04, 0.07)	5.59%, 0.03 (0.02, 0.04)*
Florida	Respiratory season	Accuracy	Cases	19%, 10% (9.99%, 15%)	20.7%, 15% (10%, 20%)	1.44%, 0% (-5%, 5%)*
Florida	Respiratory season	Accuracy	Deaths	11.8%, 10% (9.99%, 15%)	12.9%, 10% (5%, 15%)	0.958%, 0% (-5%, 4.99%)*
Florida	Off season	Log score	Cases	-19.1%, -0.066 (-0.088, -0.046)*	-11%, -0.02 (-0.037, -0.0059)*	10%, 0.03 (0.02, 0.05)*
Florida	Off season	Log score	Deaths	-0.675%, 0.01 (0.01, 0.01)*	1.75%, 0.01 (0.01, 0.02)*	2.44%, 0 (0.00049, 0)*
Florida	Off season	Accuracy	Cases	2.1%, 0% (0%, 5%)*	5.84%, 5% (0%, 5%)	3.66%, 0% (0.01%, 5%)*
Florida	Off season	Accuracy	Deaths	-1.82%, 0% (-5%, 0%)*	-0.565%, 0% (0%, 4.99%)*	1.28%, 0.01% (0%, 4.99%)*
Iowa	All	Log score	Cases	1.02%, 0.02 (0.01, 0.03)*	6.96%, 0.03 (0.02, 0.04)*	5.89%, 0.02 (0.01, 0.03)*
Iowa	All	Log score	Deaths	23.6%, 0.07 (0.06, 0.08)	25.7%, 0.06 (0.05, 0.07)	1.64%, -0.001 (-0.0031, 5e-04)*
Iowa	All	Accuracy	Cases	12.9%, 5% (0%, 10%)	28.3%, 10% (10%, 15%)	13.7%, 10% (5%, 10%)
Iowa	All	Accuracy	Deaths	17.8%, 10% (5%, 10%)	22.7%, 10% (10%, 15%)	4.19%, 5% (0.01%, 5.01%)
Iowa	Respiratory season	Log score	Cases	77.8%, 0.47 (0.4, 0.53)	77.4%, 0.48 (0.42, 0.54)	-0.246%, 0.01 (-0.003, 0.02)*
Iowa	Respiratory season	Log score	Deaths	72.9%, 0.38 (0.31, 0.46)	76.2%, 0.44 (0.36, 0.5)	1.95%, 0.01 (-0.001, 0.02)*
Iowa	Respiratory season	Accuracy	Cases	71.5%, 25% (20%, 25%)	112%, 35% (30%, 40%)	23.4%, 15% (9.99%, 15%)
Iowa	Respiratory season	Accuracy	Deaths	81%, 35% (30%, 35%)	88.7%, 35% (30%, 40%)	4.26%, 5% (0.01%, 10%)
Iowa	Off season	Log score	Cases	-30.2%, -0.11 (-0.14, -0.076)*	-23.2%, -0.076 (-0.1, -0.05)*	10.1%, 0.02 (0.01, 0.03)*
Iowa	Off season	Log score	Deaths	-0.723%, 0.03 (0.03, 0.04)*	0.71%, 0.03 (0.03, 0.03)*	1.44%, -0.0035 (-0.0055, -0.0025)*
Iowa	Off season	Accuracy	Cases	-17.6%, -10% (-15%, -5%)*	-14.9%, -10% (-10%, -4.99%)*	3.29%, 0% (0%, 5%)*
Iowa	Off season	Accuracy	Deaths	-18.8%, -10% (-15%, -10%)*	-15.5%, -10% (-15%, -5.01%)*	4.1%, 0% (0%, 5%)*

Massachusetts	All	Log score	Cases	8.31%, 0.05 (0.03, 0.07)	4.64%, 0.02 (0, 0.03)*	-3.39%, -0.03 (-0.038, -0.023)*
Massachusetts	All	Log score	Deaths	30%, 0.11 (0.1, 0.12)	28.8%, 0.08 (0.07, 0.09)	-0.937%, -0.015 (-0.018, -0.012)*
Massachusetts	All	Accuracy	Cases	49.1%, 20% (15%, 25%)	33.5%, 15% (10%, 15%)	-10.4%, -10% (-10%, -5%)*
Massachusetts	All	Accuracy	Deaths	34.7%, 20% (15%, 20%)	28.1%, 15% (10%, 20%)	-4.87%, -5% (-5%, 0%)*
Massachusetts	Respiratory season	Log score	Cases	83.6%, 0.5 (0.44, 0.57)	78%, 0.45 (0.39, 0.52)	-3.05%, -0.034 (-0.044, -0.023)*
Massachusetts	Respiratory season	Log score	Deaths	72.7%, 0.42 (0.34, 0.49)	71%, 0.41 (0.35, 0.48)	-0.976%, -0.021 (-0.028, -0.015)*
Massachusetts	Respiratory season	Accuracy	Cases	121%, 35% (35%, 40%)	99.7%, 35% (30%, 40%)	-9.73%, -10% (-10%, -5%)*
Massachusetts	Respiratory season	Accuracy	Deaths	68.3%, 30% (25%, 35%)	56.7%, 25% (25%, 30%)	-6.9%, -5% (-10%, -5%)*
Massachusetts	Off season	Log score	Cases	-24%, -0.074 (-0.11, -0.046)*	-26.8%, -0.18 (-0.24, -0.14)*	-3.62%, -0.026 (-0.036, -0.017)*
Massachusetts	Off season	Log score	Deaths	7.39%, 0.05 (0.04, 0.06)*	6.42%, 0.03 (0.03, 0.04)*	-0.91%, -0.012 (-0.014, -0.0096)*
Massachusetts	Off season	Accuracy	Cases	3.66%, 0% (0%, 5%)*	-8.13%, -4.99% (-10%, 0%)*	-11.4%, -5% (-10%, -5%)*
Massachusetts	Off season	Accuracy	Deaths	7.54%, 5% (0.01%, 9.99%)	5.05%, 5% (0%, 5%)	-2.31%, 0% (-5%, 0%)*
Michigan	All	Log score	Cases	0.571%, -0.0055 (-0.024, 0.01)*	-9.27%, -0.034 (-0.054, -0.015)*	-9.79%, -0.03 (-0.038, -0.023)*
Michigan	All	Log score	Deaths	16.8%, 0.06 (0.05, 0.08)	7.52%, 0.04 (0.03, 0.05)*	-7.95%, -0.032 (-0.036, -0.028)*
Michigan	All	Accuracy	Cases	26.8%, 10% (9.99%, 15%)	20.5%, 10% (5%, 10%)	-5.01%, -5% (-5%, 0%)*
Michigan	All	Accuracy	Deaths	21.6%, 10% (9.99%, 15%)	10.2%, 5% (5%, 9.99%)	-9.42%, -5% (-9.99%, -5%)*
Michigan	Respiratory season	Log score	Cases	38.8%, 0.24 (0.19, 0.29)	29.1%, 0.19 (0.15, 0.24)	-7.03%, -0.036 (-0.044, -0.027)*
Michigan	Respiratory season	Log score	Deaths	38.7%, 0.19 (0.15, 0.24)	33%, 0.16 (0.12, 0.2)	-4.09%, -0.036 (-0.043, -0.028)*
Michigan	Respiratory season	Accuracy	Cases	116%, 35% (30%, 40%)	84.3%, 25% (25%, 30%)	-14.6%, -10% (-15%, -5%)*

Michigan	Respiratory season	Accuracy	Deaths	83.9%, 40% (35%, 40%)	44.1%, 25% (20%, 25%)	-21.6%, -20% (-20%, -15%)*
Michigan	Off season	Log score	Cases	-18.6%, -0.14 (-0.18, -0.11)*	-28%, -0.21 (-0.24, -0.16)*	-11.6%, -0.029 (-0.042, -0.017)*
Michigan	Off season	Log score	Deaths	4.38%, 0.04 (0.03, 0.05)*	-6.47%, 0.02 (0.01, 0.03)*	-10.4%, -0.028 (-0.032, -0.024)*
Michigan	Off season	Accuracy	Cases	-16.4%, -10% (-10%, -5%)*	-10.6%, -5.01% (-9.99%, 0%)*	7.01%, 5% (0%, 10%)
Michigan	Off season	Accuracy	Deaths	-18%, -10% (-15%, -10%)*	-11.4%, -5% (-10%, 0%)*	8.01%, 5.01% (0%, 10%)
New York	All	Log score	Cases	-3.26%, 0.01 (5.9e-06, 0.02)*	-4.56%, -0.002 (-0.02, 0.01)*	-1.34%, -0.0065 (-0.014, 6.4e-05)*
New York	All	Log score	Deaths	19%, 0.08 (0.07, 0.09)	17.9%, 0.06 (0.05, 0.07)	-0.852%, -0.011 (-0.014, -0.0086)*
New York	All	Accuracy	Cases	30.7%, 15% (10%, 15%)	51%, 20% (20%, 25%)	15.6%, 10% (5%, 15%)
New York	All	Accuracy	Deaths	21.7%, 10% (10%, 15%)	32.1%, 20% (15%, 20%)	8.58%, 10% (5%, 10%)
New York	Respiratory season	Log score	Cases	44.9%, 0.24 (0.21, 0.28)	44.3%, 0.24 (0.2, 0.28)	-0.435%, -0.007 (-0.019, 0)*
New York	Respiratory season	Log score	Deaths	37.8%, 0.21 (0.18, 0.25)	37.8%, 0.21 (0.17, 0.24)	-0.00206%, -0.0056 (-0.014, 0)*
New York	Respiratory season	Accuracy	Cases	62.7%, 25% (20%, 30%)	100%, 35% (30%, 40%)	22.9%, 15% (10%, 20%)
New York	Respiratory season	Accuracy	Deaths	51.3%, 25% (20%, 30%)	80.5%, 40% (35%, 40%)	19.3%, 15% (15%, 20%)
New York	Off season	Log score	Cases	-26.3%, -0.11 (-0.14, -0.085)*	-27.7%, -0.2 (-0.25, -0.15)*	-1.94%, -0.0074 (-0.017, 2.2e-05)*
New York	Off season	Log score	Deaths	7.78%, 0.05 (0.05, 0.06)	6.25%, 0.04 (0.03, 0.04)*	-1.42%, -0.012 (-0.014, -0.01)*
New York	Off season	Accuracy	Cases	6.51%, 5% (-0.01%, 10%)	14%, 5% (0%, 10%)	7.02%, 5% (0%, 10%)
New York	Off season	Accuracy	Deaths	-0.706%, -0.01% (-4.99%, 5%)*	-4.44%, -5% (-10%, 0%)*	-3.76%, -5% (-5%, 0%)*
Pennsylvania	All	Log score	Cases	-5.3%, -0.008 (-0.026, 0.01)*	-5.72%, 0.01 (-0.0085, 0.02)*	-0.438%, -0.0039 (-0.011, 0)*
Pennsylvania	All	Log score	Deaths	13.8%, 0.07 (0.06, 0.08)	12.9%, 0.06 (0.05, 0.07)	-0.791%, -0.012 (-0.016, -0.009)*

Pennsylvania	All	Accuracy	Cases	29.7%, 10% (10%, 15%)	47.5%, 20% (15%, 20%)	13.7%, 10% (5%, 10%)
Pennsylvania	All	Accuracy	Deaths	29.4%, 15% (10%, 20%)	36.4%, 20% (15%, 20%)	5.39%, 5% (0%, 9.99%)
Pennsylvania	Respiratory season	Log score	Cases	38.5%, 0.18 (0.14, 0.22)	38.6%, 0.2 (0.17, 0.23)	0.0339%, -0.003 (-0.016, 0.01)*
Pennsylvania	Respiratory season	Log score	Deaths	36.2%, 0.15 (0.12, 0.18)	35.4%, 0.14 (0.11, 0.16)	-0.636%, -0.017 (-0.028, -0.006)*
Pennsylvania	Respiratory season	Accuracy	Cases	48.9%, 20% (15%, 25%)	99.6%, 35% (30%, 40%)	34.1%, 20% (15%, 25%)
Pennsylvania	Respiratory season	Accuracy	Deaths	50.6%, 25% (20%, 30%)	76.7%, 40% (35%, 40%)	17.4%, 15% (10%, 20%)
Pennsylvania	Off season	Log score	Cases	-26.7%, -0.18 (-0.23, -0.14)*	-27.2%, -0.19 (-0.24, -0.14)*	-0.754%, -0.0046 (-0.012, 0)*
Pennsylvania	Off season	Log score	Deaths	0.845%, 0.05 (0.04, 0.06)*	-0.0572%, 0.04 (0.04, 0.05)*	-0.895%, -0.011 (-0.014, -0.0085)*
Pennsylvania	Off season	Accuracy	Cases	13%, 5% (0.01%, 9.99%)	2.05%, 0% (-0.01%, 4.99%)*	-9.71%, -5.01% (-10%, -4.99%)*
Pennsylvania	Off season	Accuracy	Deaths	11.1%, 5% (0%, 10%)	1.46%, 0% (-4.99%, 5%)*	-8.66%, -5% (-10%, -5%)*
Texas	All	Log score	Cases	-9.81%, -0.02 (-0.043, -0.00054)*	1.82%, 0.01 (-0.0066, 0.02)*	12.9%, 0.06 (0.05, 0.07)
Texas	All	Log score	Deaths	6.34%, 0.03 (0.02, 0.03)*	15.6%, 0.03 (0.02, 0.03)*	8.68%, 0.01 (0.01, 0.01)*
Texas	All	Accuracy	Cases	11.1%, 5% (0.01%, 5%)	33.5%, 15% (10%, 15%)	20.2%, 10% (10%, 15%)
Texas	All	Accuracy	Deaths	9.13%, 5% (0%, 5%)	27.6%, 15% (10%, 15%)	16.9%, 10% (9.99%, 15%)
Texas	Respiratory season	Log score	Cases	59.4%, 0.34 (0.26, 0.42)	77.3%, 0.43 (0.36, 0.51)	11.3%, 0.05 (0.04, 0.07)
Texas	Respiratory season	Log score	Deaths	51.9%, 0.3 (0.23, 0.37)	68%, 0.42 (0.35, 0.49)	10.6%, 0.07 (0.05, 0.08)
Texas	Respiratory season	Accuracy	Cases	50.1%, 20% (15%, 25%)	96.9%, 30% (25%, 35%)	31.2%, 20% (15%, 25%)
Texas	Respiratory season	Accuracy	Deaths	47%, 25% (20%, 25%)	91.5%, 40% (35%, 40%)	30.2%, 25% (20%, 25%)
Texas	Off season	Log score	Cases	-38.5%, -0.22 (-0.27, -0.18)*	-29.9%, -0.13 (-0.17, -0.096)*	14%, 0.06 (0.05, 0.08)
Texas	Off season	Log score	Deaths	-16.3%, 0.02 (0.02, 0.02)*	-10.1%, 0.01 (0.01, 0.02)*	7.43%, -1.8e-06 (-0.00099, 0.00049)*

Texas	Off season	Accuracy	Cases	-16%, -10% (-15%, -5.01%)*	-10.5%, -5% (-10%, -5%)*	6.57%, 5% (0.01%, 5%)
Texas	Off season	Accuracy	Deaths	-18.8%, -15% (-15%, -10%)*	-19.6%, -15% (-15%, -10%)*	-0.909%, 0% (-0.01%, 0%)*
Washington	All	Log score	Cases	-14%, -0.092 (-0.11, -0.078)*	-13.5%, -0.098 (-0.11, -0.084)*	0.577%, 0.01 (0, 0.01)*
Washington	All	Log score	Deaths	-4.44%, -0.002 (-0.007, 0.00092)*	-1.65%, 0 (-0.0015, 0)*	2.92%, 0.01 (0.01, 0.01)*
Washington	All	Accuracy	Cases	16.5%, 5% (4.99%, 10%)	13.5%, 5% (0.01%, 10%)	-2.54%, 0% (0%, 0.01%)*
Washington	All	Accuracy	Deaths	11%, 5.01% (0%, 10%)	11.3%, 5.01% (5%, 9.99%)	0.291%, 0% (-0.01%, 0%)*
Washington	Respiratory season	Log score	Cases	0.837%, -0.026 (-0.046, -0.0079)*	3.75%, -0.0025 (-0.018, 0.01)*	2.89%, 0.03 (0.02, 0.03)*
Washington	Respiratory season	Log score	Deaths	0.291%, -0.05 (-0.068, -0.032)*	2.95%, -0.032 (-0.046, -0.017)*	2.65%, 0.02 (0.01, 0.02)*
Washington	Respiratory season	Accuracy	Cases	27%, 10% (5%, 15%)	25.6%, 10% (5.01%, 15%)	-1.03%, 0% (-0.01%, 0%)*
Washington	Respiratory season	Accuracy	Deaths	23.9%, 10% (10%, 15%)	29.2%, 15% (10%, 20%)	4.31%, 4.99% (0%, 5%)*
Washington	Off season	Log score	Cases	-22.7%, -0.15 (-0.17, -0.13)*	-23.4%, -0.19 (-0.22, -0.17)*	-0.949%, -0.01 (-0.018, -0.0035)*
Washington	Off season	Log score	Deaths	-7.49%, 0.01 (0.01, 0.01)*	-4.62%, 0.01 (0.01, 0.01)*	3.1%, 0 (0, 0)*
Washington	Off season	Accuracy	Cases	5.51%, 0.01% (0%, 5.01%)*	0.819%, 0.01% (0%, 5%)*	-4.45%, 0% (-5%, 0%)*
Washington	Off season	Accuracy	Deaths	-1.29%, -0.01% (-5%, 0%)*	-5.72%, -5% (-5%, 0%)*	-4.5%, -5% (-5%, 0%)*
Wyoming	All	Log score	Cases	0.44%, -0.00092 (-0.007, 0)*	1.31%, 0 (-0.0045, 0.01)*	0.869%, 0.01 (0.01, 0.01)*
Wyoming	All	Log score	Deaths	8.84%, 0.03 (0.02, 0.03)*	12.6%, 0.03 (0.02, 0.03)*	3.47%, 0 (0, 0.01)*
Wyoming	All	Accuracy	Cases	6.6%, 4.99% (0%, 5%)*	14.4%, 5% (5%, 10%)	7.28%, 5% (0.01%, 5.01%)
Wyoming	All	Accuracy	Deaths	3.37%, 0.01% (0%, 5%)*	7.98%, 5% (0%, 5%)	4.46%, 0% (-0.01%, 5%)*
Wyoming	Respiratory season	Log score	Cases	40.7%, 0.19 (0.16, 0.22)	42.9%, 0.17 (0.15, 0.2)	1.51%, -0.0036 (-0.013, 0)*
Wyoming	Respiratory season	Log score	Deaths	34.1%, 0.18 (0.14, 0.21)	39.6%, 0.16 (0.13, 0.19)	4.1%, 8.3e-05 (-0.0084, 0.01)*
Wyoming	Respiratory season	Accuracy	Cases	57.9%, 20% (15%, 25%)	70.8%, 25% (20%, 30%)	8.14%, 5% (0.01%, 10%)

Wyoming	Respiratory season	Accuracy	Deaths	29.6%, 15% (10%, 20%)	41.1%, 20% (15%, 25%)	8.87%, 5% (0%, 10%)
Wyoming	Off season	Log score	Cases	-19.5%, -0.15 (-0.18, -0.12)*	-19.1%, -0.13 (-0.16, -0.093)*	0.454%, 0.02 (0.01, 0.02)*
Wyoming	Off season	Log score	Deaths	-5.06%, 0.01 (0.01, 0.01)*	-2.16%, 0.01 (0.01, 0.02)*	3.05%, 0.01 (0, 0.01)*
Wyoming	Off season	Accuracy	Cases	-20%, -10% (-15%, -5%)*	-14.9%, -5% (-10%, 0%)*	6.39%, 5% (0%, 10%)
Wyoming	Off season	Accuracy	Deaths	-14.2%, -9.99% (-10%, -5.01%)*	-14.2%, -9.99% (-10%, -5%)*	0%, 0% (-0.01%, 5%)*

Table S5. Prior ranges for the parameters and variables used in the model-inference system. Parameters/state variables are initialized by drawing from uniform distributions specified in the rows labeled "Initialization". During the filtering process, space-reprobing is applied to explore the state space, i.e., a small fraction of the ensemble members are randomly replaced with values drawn from the uniform distributions specified in the rows labeled with "Space-reprobing".

Type	Parameters/ variables	Symbol	Range	Note
Initialization	Initial susceptible	S(t=0)	All locations: non-Omicron period, U[99%, 100%] population; Omicron period, U[50%, 90%] population	n/a
Initialization	Initial exposed	E(t=0)	All locations: U[5, 50] × no. cases during 1st week	n/a
Initialization	Initial infectious	I(t=0)	All locations: U[5,50] × no. cases during 1st week	n/a
Initialization	Infectious period	D	All locations: U[2, 5] days	n/a
Initialization	Latency period	Z	All locations: U[2, 5] days	n/a
Initialization	Duration of immunity (from prior infection)	L	All locations: non-Omicron period, U[2, 3] years; Omicron period: U[1, 3] years	n/a
Initialization	Time-to-detection, mean	Td, mean	All locations: U[5, 8] days	n/a
Initialization	Time-to-detection, sd	Td, sd	All locations: U[1, 3] days	To allow variation in time to diagnosis/reporting
Initialization	Scaling of NPI effectiveness	e	All locations: U[0.5, 1.5]	Around 1, with a large bound to be flexible
Initialization	Vaccine efficacy (VE)	n/a	All locations: before Delta, VE1=85%, VE2 = 95%; Delta, VE1 = 50%, VE2 = 80%; Omicron, VE1 = 10%, VE2 (combined 2nd and 3rd doses) = 70%	Used higher VE values, as the observations included both cases/infections and

				deaths; i.e., here VE is for both infections and mortality
Initialization	VE waning	ρ	rho(t) = 1/(1+exp(-k * (t - tm.imm/2); for wildtype: k = 0.026; tm.imm = 322; for Delta: k = 0.025; tm.imm = 280; for Omicron: k = 0.024; tm.imm = 256	Parameter in the logistic function fitted based on data from UKHSA
Initialization	Infection-detection rate	r	all locations: U [0.01, 0.05]	n/a
Initialization	Infection-fatality risk	IFR	all locations: U [0.005, 0.015]	n/a
Initialization	Transmission rate	β	Wyoming: U [0.41, 0.6]; Iowa, Texas, Washington: U [0.44, 0.75]; California, Florida, Massachusetts, Michigan, New York, Pennsylvania: U [0.45, 0.75]	n/a
Space-reprobing	Transmission rate	β	Alpha (2021-02-14 to 2021-06-19): Wyoming, U [0.54, 0.9]; Iowa, Texas, Washington, U [0.57, 1.12]; California, Florida, Massachusetts, Michigan, New York, Pennsylvania, U [0.585, 1.125]; Delta (2021-06-06 to 2021-12-18): Wyoming, U [0.5, 1.02]; Iowa, Texas, U [0.52, 1.27]; Washington, U [0.53, 1.27]; California, Florida, Michigan, New York, Pennsylvania, U [0.54, 1.275]; Delta_holiday (2021-07-04 to 2021-07-31): Massachusetts, U [0.855, 1.275]; Delta1 (2021-06-13 to 2021-07-03): Massachusetts, U [0.54, 1.275]; Delta2 (2021-08-01 to 2021-08-21): Massachusetts, U [0.54, 1.02]; Delta3 (2021-08-22 to 2021-12-18): Massachusetts, U [0.54, 1.275]; Omicron_BA.1 (2021-12-05 to 2022-03-26): Wyoming, U [0.62, 1.2]; Texas, U [0.65, 1.5]; Iowa, Washington, U [0.66, 1.5]; California, Florida, Massachusetts, Michigan, New York, Pennsylvania, U [0.675, 1.5]; Omicron_BA.2 (2022-03-06 to 2022-05-07): Wyoming, U [0.75, 1.68]; Iowa, Texas, Washington, U [0.79, 2.1]; California, Florida, Massachusetts, Michigan, New York, Pennsylvania, U [0.81, 2.1]; Omicron_BA.2.12.1 (2022-04-03 to 2022-06-18): Wyoming, U [0.75, 1.68]; Iowa, Texas, Washington, U [0.79, 2.1]; California, Florida, Massachusetts, Michigan, New York, Pennsylvania, U [0.81, 2.1];	n/a

			Omicron_nonBA.1o2 (2022-06-05 to 2022-12-31): Wyoming, U [0.75, 1.68]; Iowa, Texas, Washington, U [0.79, 2.1]; California, Florida, Massachusetts, Michigan, New York, Pennsylvania, U [0.81, 2.1]; wave1 (2020-01-01 to 2020-09-30): Wyoming, U [0.41, 0.6]; Iowa, Texas, Washington, U [0.44, 0.75]; California, Florida, Massachusetts, Michigan, New York, Pennsylvania, U [0.45, 0.75]; wave2 (2020-10-01 to 2021-03-20): Wyoming, U [0.41, 0.66]; Iowa, Texas, Washington, U [0.44, 0.83]; California, Florida, Massachusetts, Michigan, New York, Pennsylvania, U [0.45, 0.825]	
Space-reprobing	Infection-detection rate	*r*	2020summer (2020-06-01 to 2020-10-03): Massachusetts, New York, Wyoming, U [0.05, 0.25]; Michigan, Pennsylvania, U [0.05, 0.3125]; California, Florida, Iowa, Texas, Washington, U [0.05, 0.375]; 2021summer (2021-06-01 to 2021-06-30): California, Iowa, Massachusetts, Michigan, New York, Pennsylvania, U [0.02, 0.1]; California, Florida, Iowa, Massachusetts, Michigan, New York, Pennsylvania, Texas, Washington, U [0.03, 0.25]; Alpha (2021-02-14 to 2021-05-31): Florida, Iowa, Massachusetts, Michigan, Pennsylvania, Texas, U [0.2, 0.4]; California, New York, Washington, Wyoming, U [0.2, 0.48]; Delta (2021-05-31 to 2022-01-01): Florida, Texas, Washington, Wyoming, U [0.1, 0.5]; Delta1 (2021-06-21 to 2021-07-17): California, Iowa, Massachusetts, Michigan, New York, Pennsylvania, U [0.1, 0.5]; Delta2 (2021-06-27 to 2021-08-08): California, Iowa, Massachusetts, Michigan, New York, Pennsylvania, U [0.3, 0.5]; Delta3 (2021-07-19 to 2022-01-08): California, Iowa, Massachusetts, Michigan, New York, Pennsylvania, U [0.1, 0.5]; massvax (2021-05-14 to 2021-05-31): California, Iowa, Massachusetts, Michigan, New York, Pennsylvania, Washington, U [0.05, 0.3]; Omicron0 (2021-10-15 to 2021-12-11): all locations, U [0.001, 0.05]; Omicron1a (2021-12-05 to 2021-12-18): all locations, U [0.1, 0.5]; Omicron1b (2021-12-05 to 2021-12-25): all locations, U [0.1, 0.6]; Omicron2 (2021-12-26 to 2022-06-18): all locations, U [0.05, 0.2]; Omicron3 (2022-06-05 to 2022-12-31): all locations, U [0.02, 0.18]; wave1 (2020-03-08 to 2020-05-31): all locations, U [0.08, 0.35]; wave2 (2020-09-06 to 2021-03-20): Florida, Iowa, Massachusetts, Michigan,	n/a

			Pennsylvania, Texas, U [0.12, 0.4]; New York, Washington, Wyoming, U [0.144, 0.48]; California, U [0.144, 0.54]	
Space-reprobing	Infection-fatality risk	IFR	2020summer (2020-06-01 to 2020-09-30): all locations, U [1e-04, 0.005]; Alpha (2021-03-07 to 2021-06-19): all locations, U [1e-04, 0.015]; Delta (2021-06-13 to 2022-01-01): all locations, U [1e-04, 0.015]; massvax (2021-05-14 to 2021-06-26): California, Florida, Iowa, Massachusetts, Michigan, New York, Pennsylvania, Texas, Washington, U [1e-04, 0.01]; Omicron_BA.1 (2021-12-12 to 2022-06-18): all locations, U [4e-05, 0.004]; Omicron_nonBA.1 (2022-06-05 to 2022-12-31): all locations, U [8e-06, 0.0032]; wave1early (2020-03-16 to 2020-04-15): all locations, U [0.005, 0.025]; wave1late (2020-04-16 to 2020-05-31): all locations, U [0.001, 0.015]; wave2 (2020-10-01 to 2021-04-24): all locations, U [2e-04, 0.0125]	n/a