Phase-Informed Bayesian Ensemble Models Improve Performance of COVID-19 Forecasts

Aniruddha Adiga*[1], Gursharn Kaur*[1], Lijing Wang[2], Benjamin Hurt[1], Przemyslaw Porebski[1], Srinivasan Venkatramanan[1], Bryan Lewis[1], Madhav Marathe[1, 3]

[1] Biocomplexity Institute, University of Virginia
[2] Boston Children's Hospital and Harvard Medical School
[3] Dept. of Computer Science, University of Virginia
aniruddha@virginia.edu, fug3aj@virginia.edu, marathe@virginia.edu

Abstract

Despite hundreds of methods published in the literature, forecasting epidemic dynamics remains challenging yet important. The challenges stem from multiple sources, including: the need for timely data, co-evolution of epidemic dynamics with behavioral and immunological adaptations, and the evolution of new pathogen strains. The ongoing COVID-19 pandemic highlighted these challenges; in an important article, Reich et al. did a comprehensive analysis highlighting many of these challenges.

In this paper, we take another step in critically evaluating existing epidemic forecasting methods. Our methods are based on a simple yet crucial observation — epidemic dynamics go through a number of phases (waves). Armed with this understanding, we propose a modification to our deployed Bayesian ensembling case time series forecasting framework. We show that ensembling methods employing the phase information and using different weighting schemes for each phase can produce improved forecasts. We evaluate our proposed method with both the currently deployed model and the COVID-19 forecasthub models. The overall performance of the proposed model is consistent across the pandemic but more importantly, it is ranked third and first during two critical rapid growth phases in cases, regimes where the performance of most models from the CDC forecasting hub dropped significantly.

Introduction

The COVID-19 pandemic has severely impacted global economic, social, and health. The Delta and Omicron variants have exceeded records for fatalities, case counts, and hospitalizations in the US and worldwide. The global economic impact is in trillions of dollars. More than 588M confirmed infections and 6.5M deaths have been reported worldwide, with the United States reporting over 1M deaths. We have seen different epidemic dynamic trajectories and mortality in various countries.

A highlight of the pandemic has been the near real-time availability of the incidence data, which has been crucial in understanding its dynamics. The availability of such data has been a breakthrough but has also raised the expectations to develop high-quality forecasts related to the pandemic. Several teams have engaged in forecasting these dynamics and have collectively worked towards refining and communicating their short-term forecasts (1−4 weeks) to public health agencies through initiatives such as CDC COVID-19 Forecast Hub[1] (Cramer et al. 2021) and the European COVID-19 Forecast Hub[2] (since the goal and setup of both the groups are alike, we refer to them as *The Hub* henceforth). *The Hub* allows teams to provide probabilistic forecasts for a range of *targets*, incident cases, deaths, cumulative deaths, and incident hospitalizations. We have been one of the teams, contributing forecasts since July 2020 and have focused on incident cases and hospitalizations at a granular level. We employ an ensemble (Adiga et al. 2021) of statistical, deep learning, and mechanistic models for forecasting multiple targets.

Over the last decade, as a part of developing forecasting methods for diseases such as influenza and dengue, it has been shown that ensemble-based forecasting methods tend to have better performance compared to individual model forecasts (Reich et al. 2019; Yamana, Kandula, and Shaman 2017). *The Hub* has also been aggregating using a simple median-based ensemble of all model forecasts and has experimented with a trained ensemble (Cramer et al. 2021). Simultaneously, evaluation reports prepared by *The Hub* indicate that the ensemble models have been among the top-performing models(Delphi 2021).

Despite providing a stable performance, compared to individual models, an ensemble model's forecast quality depends on its constituent models' quality. **Despite the efforts of the forecasting community, there needs to be a greater understanding of disease dynamics as teams have struggled to predict the onset, growth rate, peak size, and duration of the various waves.** Achieving good forecasting accuracy during the growth or surge phase is of high importance as it enables the effective allocation of medical resources which are strained during these times. Hence, the ensemble models have suffered from outlying forecasts and have failed to predict the local peaks (Ray et al. 2021), a trend we have also observed in our ensemble model fore-

*These authors contributed equally.

[1] https://viz.covid19forecasthub.org/
[2] https://covid19forecasthub.eu/

casts.

In these efforts, all the models face several challenges, some that have persisted since the beginning of the forecasting period. Important challenges include: (i) Irregular updates, (ii) Non-stationarity of the time series, and (iii) Co-evolution of pandemics with social, biological, and viral dynamics.

Summary of contributions. In this paper, we will provide a systematic evaluation of the currently deployed system described in (Adiga et al. 2021) and understand the influence of individual models on the ensemble's performance at different phases or contexts in the pandemic. Using this knowledge, we propose a training scheme that enables the model to leverage the context-specific historical performance of individual methods leading to improved forecast performance of the ensemble at critical phases. Although we analyze the efficacy of the proposed training method using our BMA model, the training scheme is fairly generic and can be applied to the likes of COVID-19 Forecasthub ensemble models. These new methods and insights have been obtained through popular artificial intelligence techniques such as Shapley value analysis and ensembling methods. Our contributions consist of two key components:

Analyzing pandemic dynamics and performance of methods (Section). Our deployed model (Adiga et al. 2021) in *The Hub* (listed as UVA-ENSEMBLE) employs Bayesian Model Averaging (BMA) ensembling technique to suitably combine probabilistic forecasts from autoregressive models, filtering methods, deep neural networks, and compartmental models. A retrospective analysis of our model's performance indicates that the BMA has a performance close to *The Hub*'s ensemble model. However, these evaluations only provide the relative performance of the methods and, in general, do not offer insights about absolute forecasting quality. By systematically evaluating the performance of models across the forecasting weeks, we are able to establish a model's relative influence during different phases of the pandemic. Specifically, we determine the influence of the model on the BMA's performance at different forecasting weeks – we observe that a method assigned high weights by the ensemble does not necessarily translate to improved performance.

Our analysis of the performance of models across different forecasting weeks indicates that: (i) compartmental models are useful during growth and decline phases but tend to over-predict during the surge and decline phases; (ii) purely data-driven models like LSTMs have a latency in picking up the change in phases, but can quickly learn the patterns and (iii) Statistical AR methods or Kalman filters based methods show superior performance during time of relative steady phase of the pandemic.

Development of new phase-informed methods.(Section) Based on these insights, we propose an improvement to our BMA model training technique – these improvements are based on assessing the current phase of the pandemic. First, we categorize the ground truth data of the pandemic into phases and then detect changes in those phases. The three important phases are *surge* (or growth), *plateau*, and *decline*. We then modify our ensembling technique to employ training data from historically similar phases experienced during the forecasting period. **We conduct detailed retrospective analysis to show that the new methods lead to improved performance at a critical phase, the surge phase, compared to existing methods. Our analysis shows that the proposed model has consistent performance across the pandemic. More importantly, it performs significantly well compared to *The Hub* models during the Delta wave's surge period (median ranking of 4) and the Omicron wave's surge phase (median ranking of 1), regime where the performance of most models from the CDC forecasting hub dropped significantly**.

Related Work

Epidemic forecasting has been a subject of active research over the last decade. Given the large number of papers written on this topic, we will provide references to surveys and papers that are most closely related – see (Nsoesie et al. 2014; Chakraborty et al. 2018; Wang et al. 2021, 2020; Rosenfeld and Tibshirani 2021; Bertozzi et al. 2020) and papers cited therein for recent work, challenges and progress. The compartmental models have been the most common framework during this pandemic, primarily because they are versatile enough to incorporate evolving aspects of disease dynamics. In *The Hub*, most models are variants of the Susceptible-Infectious-Recovered (SIR) (Anastassopoulou et al. 2020), and a detailed description of the compartmental model variants is provided in (Adiga et al. 2020). The details of the different models serving *The Hub* are available in (Hub 2020). Traditionally, statistical and other data-driven methods have shown to be effective in epidemic forecasting but also rely heavily on high-quality data. In COVID-19 forecasting, linear models have mainly been restricted to forecasting case trajectories at the national level (Hernandez-Matamoros et al. 2020; Kufel et al. 2020). Modelers have also considered complex systems, such as deep-learning models. Specifically, Long Short-Term Memory (LSTM) networks (Ramchandani, Fan, and Mostafavi 2020; Rodriguez et al. 2020) and Graph Neural Network (GNN) (Gao et al. 2020; Wang et al. 2021; Kapoor et al. 2020) have been employed in COVID-19 forecasting. During the early stages of the pandemic, despite the lack of data, LSTM models were explored and shown to provide reasonable forecasts by incorporating auxiliary data. Finally, in the framework of ensembling, Bayesian model averaging (BMA) is a well-studied, practical framework for probabilistic forecast averaging that, unlike the model selection, also takes into account the uncertainty in predictions. Its application to combining multiple weather models has been studied exhaustively by Raftery et al. (Raftery et al. 2005), while its effectiveness in weighting competing ILI forecasting models has been demonstrated in (Yamana, Kandula, and Shaman 2017).

A Retrospective Analysis of BMA Ensemble Forecasts

The constituent models in the currently deployed BMA ensemble forecasting method include several standard statisti-

cal, deep learning, and compartmental models. Specifically, we employ AR model and its variants (AR, AR_spatial, ARIMA), an LSTM model, an ensemble Kalman filter (EnKF), and a compartmental model (SEIR). The forecasts from these models are combined using a Bayesian model averaging technique. We have purposely employed simple models as their behavior is well studied and they also provide better explainability. A detailed description of the models is available in our work (Adiga et al. 2021).

Our primary effort over the pandemic has been focused on making accurate high geographical resolution *probabilistic* forecasts, i.e., at the US county level. We are one of the long-standing models (team name: UVA-Ensemble) in *The Hub*, providing forecasts for incident cases at the county-level, and the model has provided relatively good performance. We refer the reader to Section for a detailed comparison of forecasting performance with other state-of-the-art models in *The Hub*. As mentioned previously, most models have failed at important points in time, including our model and this behavior warrants a thorough analysis. We focus on our ensemble model and provide a detailed investigation of its performance across different forecasting weeks.

Pandemic phases. Despite heterogeneity in the COVID-19 time series, we broadly observe three phases and classify the observed time period based on the rate of change of reported cases: Surge (period of steep growth in cases), Decline, and Plateau. We want to note that the definitions of phases are subjective (several exist[3]) and can be user annotated or obtained through standard time-series change point detection algorithms (Aminikhanghahi and Cook 2017). We discuss a standard phase identification technique in Section . Our primary purpose of phase classification is to capture distinct trends in the time series and leverage that information to better train the BMA model.

Influence of Methods at Different Phases of the Pandemic

Our preliminary analysis (Adiga et al. 2021), showed that based on the average performance computed across all counties and all forecasting weeks, all methods were important. We also observed that the dominant methods (methods assigned the highest weight) for a given location changed across forecasting weeks. In the following sections, we investigate the influence of methods at different phases of the pandemic.

Model ablation analysis Inspired by the concept of Shapley value (Winter 2002) in game theory, we measure the contribution of each method in the BMA ensemble using the ablation analysis. Let $N = \{1, \ldots, n\}$ be the set of methods in the BMA ensemble. A method i influence in the ensemble is governed by the performance of other methods feeding into the BMA. For example, i, in the presence of a subset of methods with historical performance inferior to it, will receive high weights. On the other hand, its effect on the BMA can be insignificant in the presence of a better performing set of models. In order to understand the influence of i across all such groups, we consider all possible subsets that can be generated from the $N-1$ methods and train the BMA on each subset. The influence of i is determined by comparing the performance with and without the inclusion of i in the subsets. Since each county is trained independently in our framework, we determine the influence of i for each county c and define the payoff set function at time t as $v^{c,t}(S) = \frac{|y^{c,t} - f^{c,t}(S)|}{y^{c,t}}, \quad \text{for } S \subseteq N$, where $y^{c,t}$ is the ground truth for c at time t and $f^{c,t}(S)$ is the forecast obtained from the BMA ensemble when S set of methods are incorporated into BMA. Then the expected marginal contribution of a method i can be defined by the average change in the $v^{c,t}$ using the set of methods S, if i is included with S in the BMA ensemble. The average ablation score is defined as: $\phi_i^{c,t} = \frac{1}{2^{n-1}-1} \sum_{S(\neq\emptyset)\subset N\setminus\{i\}} \left[v^{c,t}(S\cup\{i\}) - v^{c,t}(S)\right]$. This equation is similar to the Shapley values used in cooperative game theory. However, the Shapley values place certain constraints that are violated by our payoff function.

Results In Figure 1, we present the observed ϕ_i^t values for $i \in \{$AR, ARIMA, AR_spatial, EnKF, LSTM and SEIR$\}$. Note that negative ϕ values for a method i indicate an overall reduction in the MAPE, when i is included with a subset of methods S in the ensemble. In particular, ARIMA and LSTM get the most significant negative values throughout the observed time period. The SEIR has variable performance from around July 2021 to September 2021. During the beginning of the surge phase, the SEIR models are typically able to capture the rapid increase in cases, but in the subsequent weeks, tend to overestimate the growth. Due to their superior performance during the past weeks they still get assigned high weights thus leading to inferior performance of the BMA. Owing to their construction, the SEIR model can match the exponential growth observed during the surge phase.

Proposed Method - Phase Informed Ensemble

Motivated by the fact that the performance of the ensemble is influenced by different models in different *phases*, we propose a method to supply the phase information during the BMA ensemble training. In the BMA framework, we independently train a single BMA to calibrate forecast ensembles per county. Considering K models per county, the BMA assumes that the conditional density of y given the forecasts $f_1, \ldots, f_K$ generated from models $M_1, M_2, \ldots, M_K$ is given by

$$p(y|f_1, f_2, \cdots, f_K) = \sum_{k=1}^{K} w_k g_k(y|f_k), \qquad (1)$$

where w_k is the posterior probability of the k^{th} model's forecast being the best one, and $g_k(y|f_k)$ is the conditional density of y given f_k. With normal approximation for the conditional density i.e. $y|f_k \sim \mathcal{N}(f_k, \sigma_k^2)$, (1) is a finite mixture of Gaussians, and we proceed to determine the weights w_k and σ_k. Given the distribution (1), the weights and variance parameters are obtained as the maximum likelihood

[3]https://www.cdc.gov/flu/pandemic-resources/planning-preparedness/global-planning-508.html

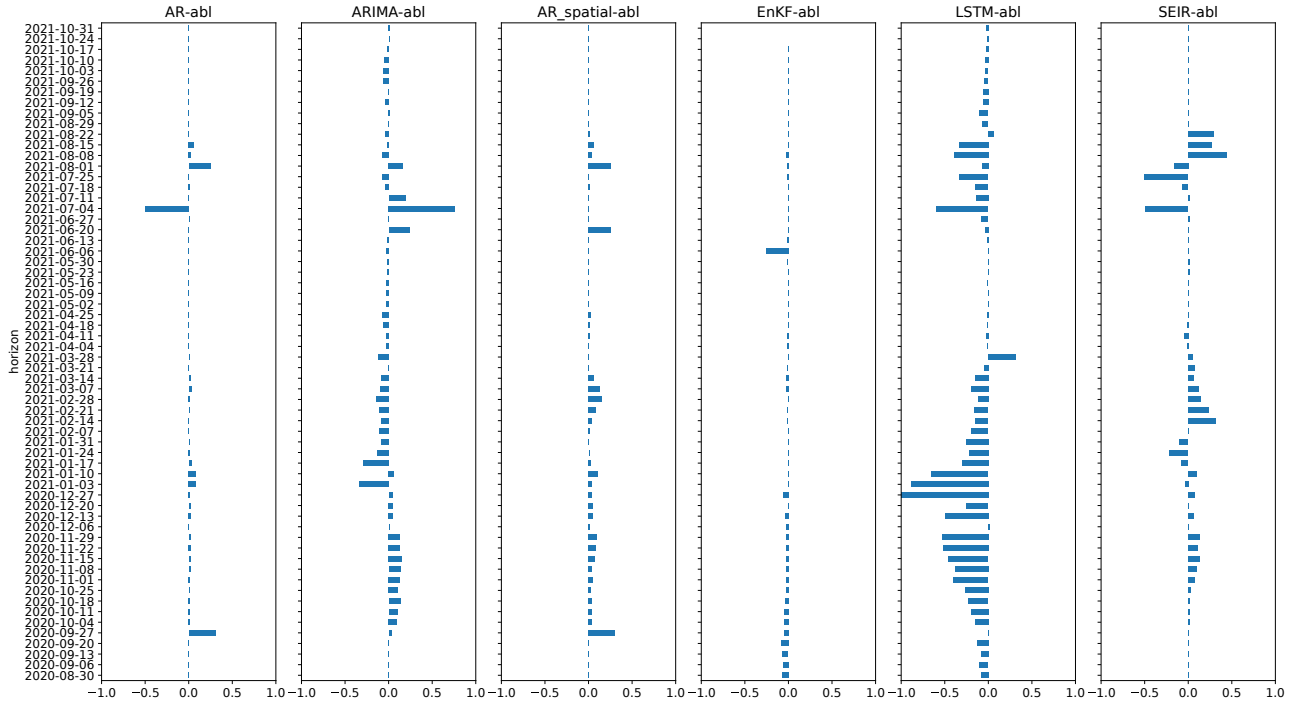


Figure 1: Understanding the influence of models through ablation analysis.

estimate using the standard expectation-maximization (EM) algorithm (Raftery et al. 2005), which alternates between the E-step and the M-step with the updates for w_k and σ_k in the j^{th} iteration given by the (E-step)

$$z_{k,t}^{(j)} = \frac{w_k^{(j-1)} g(y_t | f_{k,t}, \sigma_k^{(j-1)})}{\sum_{i=1}^{K} w_i^{(j-1)} g(y_t | f_{i,t}, \sigma_i^{(j-1)})},$$

and (M-step)

$$w_k^{(j)} = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} z_{k,t}^{(j)}, \; \sigma_k^{(j)2} = \frac{\sum_{t \in \mathcal{T}} z_{k,t} (y_t - f_{k,t})^2}{\sum_{t \in \mathcal{T}} z_{k,t}}. \quad (2)$$

In the existing framework (Adiga et al. 2021), $\mathcal{T}$ corresponds to the previous N contiguous weeks of training samples, that is, for a forecast week T, $\mathcal{T} = \{T-1, T-2, \cdots, T-N\}$. Given the highly nonstationary data, it is to be noted that in our training, in order to ensure that the most recent trend is captured, we consider only the most recent N weeks and not the entire set of historical forecasts.

Through the analysis in Section we observed that different methods influence the ensemble's performance at different phases. The issue of picking the best performing method is particularly more pronounced during a surge or decline phase. A major drawback of the existing approach where the BMA weights (w_k) are determined by the individual model performance over the past N weeks, without considering the phase, is the latency in picking the best performing method.

We identify and address this issue by designing a BMA ensemble that uses the knowledge of the relevant phase to get improved weights. On that note, for a weekly case counts time series, we first segment the ground truth week indices into surge (S), decline (D), and plateau (P) phases. Let $\mathcal{T}_S$, $\mathcal{T}_D$, and $\mathcal{T}_P$ be the set of all week indices corresponding to the surge, decline, and plateau, respectively. The phase-informed BMA then considers all the historical forecasts made by individual methods during the specified phase for training the weights. That is, for a particular phase $r \in \{S, D, P\}$, estimation of weights and variance in (2) (M-Step) can be modified as

$$w_{k,r}^{(j)} = \frac{1}{|\mathcal{T}_r|} \sum_{t \in \mathcal{T}_r} z_{k,t}^{(j)}, \; \sigma_{k,r}^{(j)2} = \frac{\sum_{t \in \mathcal{T}_r} z_{k,t} (y_t - f_{k,t})^2}{\sum_{t \in \mathcal{T}_r} z_{k,t}}. \quad (3)$$

We next discuss the phase segmentation technique that enables us to determine $\mathcal{T}_r$.

Real-Time Phase Segmentation To segment the whole time series into different phases, we first approximate the nonlinear time series with a piece-wise linear function. We use a standard R package `segmented` (Muggeo 2008) to estimate multiple break-points.

Note that, in real-time forecasting, since we obtain a new data point each week, the phase segments must be reestimated. Given the new data point, we would want to refine our estimates of phases and ensure that they stay the same. Hence, we apply the segmentation each week, only on data starting from the most recent two break-points. The algorithm is described in Algorithm 1. Using the estimated break-points $\{b_1, \ldots, b_m\}$ with Algorithm 1 for the ground truth $y_1, \ldots, y_t$, we classify the time interval between any two consecutive break-points as a surge (S), decline (D), or plateau (P) phase. We employ a simple criterion for the phase classification, which defines a time interval $(b_k, b_{k+1}]$

Algorithm 1: Recursive Piece-wise linear fit

Input: Ground truth $y_1, \ldots, y_T$
Output: Set of break-points $\{b_1, b_2, \ldots, b_m\}$

Start with $w = 15$
Get a piece-wise fit for $y_1, y_2, \ldots, y_w$ with break-points $b_1^{(w)} \leq b_2^{(w)} \leq \cdots \leq b_{k_w}^{(w)}$
$\mathcal{B}(w) := \{b_1^{(w)}, \ldots, b_{k_w}^{(w)}\}$
while $w + 1 \leq t \leq T$ **do**
    Get a piece-wise fit for $\{y_s : b_{k_{t-1}-2}^{(t-1)} \leq s \leq t\}$ with break-points $b_1^{(t)} \leq b_2^{(t)} \leq \cdots \leq b_{k_t}^{(t)}$
    $b_1^\star := \max \mathcal{B}(t-1)$, $b_2^\star := \max \mathcal{B}(t-1) \setminus b_1^\star$
    $\mathcal{B}(t) \leftarrow (\mathcal{B}(t-1) \setminus \{b_1^\star, b_2^\star\}) \cup \{b_1^{(t)}, \ldots, b_{k_t}^{(t)}\}$
end while
return the final set of break-points $\mathcal{B}(T)$

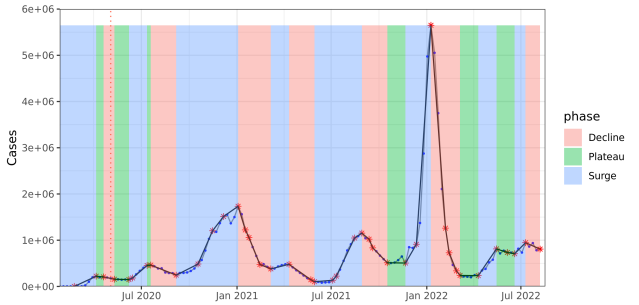


Figure 2: A piece-wise linear fit and phase classification for USA case counts obtained with Algorithm 1 and phase segmentation scheme.

as a surge (or decline) phase if there is at least a 10% increment (or at least a 10% reduction) in the case count from the start of the time interval b_k to end of the interval b_{k+1} and plateau phase otherwise. Figure 2 shows an example of the estimated break-points and the phase identification for the national-level case counts for January 2020 to July 2022.

Results

Sample Forecasts of Phase-Informed Forecasts In Figure 3, we show forecasts provided by the phase-informed BMA (PI-BMA) and the non-phase-informed BMA (BMA) for two US counties. In these forecasts, we observe that during the inflection points (change from surge to decline), the BMA typically over-predicts (cf. Figures 3a and 3b) while the PI-BMA is able to forecast the trajectory relatively well. These are examples but a comparison of performance between BMA and PI-BMA is discussed next.

A comparison of BMA and PI-BMA In all our analysis, we consider aggregate performance across three regimes, (i) Overall-80 forecasts weeks (1 August 2020 $-$ 1 January 2022), (ii) Delta wave surge region (15 July 2021 $-$ 15 August 2021), and (iii) Omicron wave (15 December 2021 $-$ 15 January 2022). The latter two regimes are specifically considered as these correspond to the surge phases where most models failed to forecast the rapid increase in cases

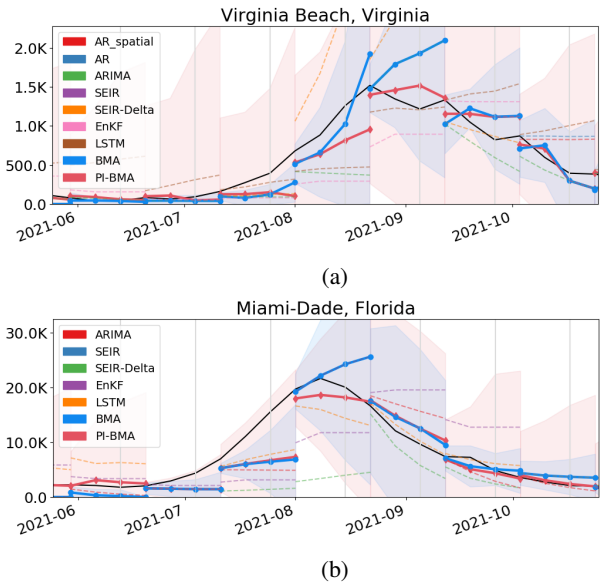


Figure 3: Examples comparing PI-BMA forecasts (red line) with the deployed BMA forecasts (green line) for two US counties (a) Virginia Beach (Virginia) and (b) Miami-Dade (Florida). A visual inspection of these examples show that PI-BMA when compared to BMA is able to capture inflection regions better.

(Ray et al. 2021). In order to assess the performance of the PI-BMA, we compare the median MAE obtained with BMA and PI-BMA in different time periods. In Figures 4, we compare the mean of the median (over all counties) MAE taken across all three regimes. We observe that the PI-BMA model has slightly smaller or comparable MAE in all three cases, for both short term (1-week ahead) and long term (4-week ahead) forecasts. We mostly consider median performance as opposed to mean performance as the forecasts quality of few counties with high levels of noise can be significantly poor and can affect the aggregate performance of a model.

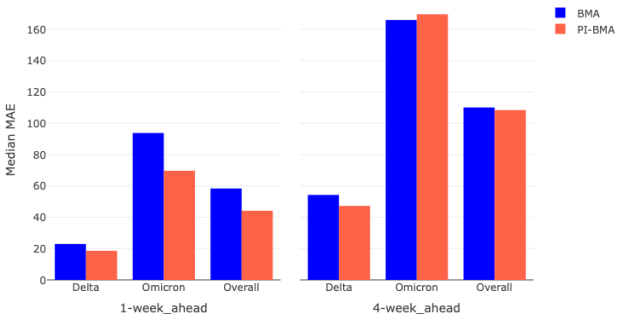


Figure 4: A comparison of BMA and PI-BMA based on the median MAE across all counties.

Retrospective Evaluation: A Comparison with *The Hub* Models

So far. over 100 models from dozens of teams have submitted forecasts to *The Hub*, with the numbers varying every week. Among the many teams, only a handful have been consistently providing county-level forecasts. In order to make a fair comparison, we only consider teams that have been providing consistent forecasts across most counties and targets since August 2020. It should be noted that across the 80 forecasting weeks, 15 models have provided a significant number of forecasts. The model details are available in (Hub 2020). As the pandemic progressed, we observed that the number of models started to drop after July 2021. As mentioned, the teams provide probabilistic forecasts in the quantile format. In order to compare the forecast quantiles of the different models, we use the Weighted Interval Score (WIS), the *de facto* standard in the epidemiological forecasting community for probabilistic forecast evaluation (Bracher et al. 2020):

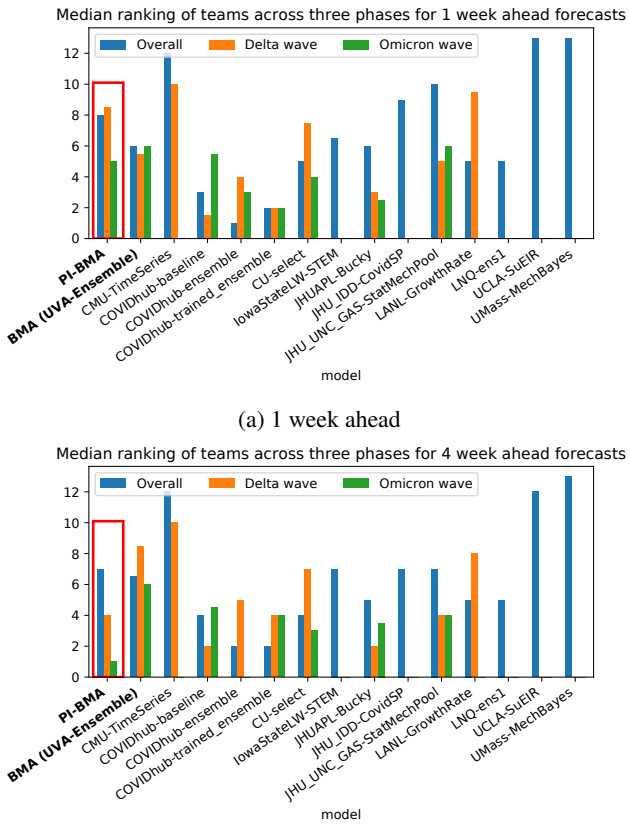
$$WIS_{\alpha_{0:k}}(F,y) = \frac{1}{K+0.5}\sum_{k=0}^{K}\frac{\alpha_k}{2}(u_k - l_k) + \frac{2}{\alpha_k}(l_k - y)\mathbb{1}(y < l_k) + \frac{2}{\alpha_k}(y - u_k)\mathbb{1}(y > u_k), \quad (4)$$

where y is the observed value (ground truth case count corresponding to a week) for a given location and date, F is the forecast defined in terms of the median m, upper quantiles u_k, and lower quantiles l_k of the predictive distribution, respectively. $K = 3$ is the number of intervals considered. $\mathbb{1}$ is the indicator function.

We first rank the performance of a model for each forecast week and target horizon by considering its median WIS score across all the counties (the model having the lowest median score is ranked one). We next determine the median ranking of different models during different regimes, and the results are shown in Figures 5a and 5b for 1 week ahead and 4 weeks ahead forecast horizons, respectively. The blue bars, which correspond to the median ranking computed across all forecasting weeks, indicate that both BMA (UVA-Ensemble) and PI-BMA are ranked around 6−7. Focusing on the more challenging target of 4-weeks ahead, we observe that the PI-BMA is one the top-ranked models during the critical phases of Delta wave surge (median ranking of 4 out of 9) and Omicron wave surge (median ranking of 1 out of 6). The PI-BMA's performance indicates that the model can effectively incorporate the phase information and provide considerably better forecasts during critical phases when compared to both BMA (UVA-Ensemble) and the rest of the forecast hub models. It should be noted that the *COVIDhub ensemble* and *COVIDhub-trained_ensemble* use forecasts from highly tuned individual models but our model is able to out perform them during the critical phases. This validates the use of selective sampling of training data by ensembling methods.

Conclusions

The paper undertakes a critical and comprehensive review of several well-studied methods for forecasting COVID-19 dynamics. Based on the analysis, we proposed a phase-informed Bayesian ensembling method that significantly improves forecast skills at important critical phases.



(a) 1 week ahead

(b) 4 week ahead

Figure 5: A comparison of several *The Hub* models performance. The median ranking of models for (a) 1 week ahead forecasts and (b) 4 week ahead forecasts computed across different regimes. Blue bars show the median ranking of models computed across all the forecasting weeks, orange bars correspond to the median ranking of models computed for the Delta wave's surge phase, and green bars correspond to the median ranking of models during the Omicron wave's surge phase. Rankings across different phases indicate that the PI-BMA (red box) can provide significantly better forecasts, especially 4 weeks ahead, for critical surge phases corresponding to the *Delta wave* (median ranking of 4) and the *Omicron wave* (median ranking of 1).

Acknowledgments

Funding agencies: NSF Rapid 2142997, CSTE/CDC 5 NU38OT000297, NIH Grant 1R01GM109718, NSF BIG DATA Grant IIS-1633028, NSF: OAC-1916805, NSF Expeditions in Computing Grant CCF-1918656, CCF-1917819, NSF RAPID CNS-2028004, NSF RAPID OAC-2027541, US-CDC 75D30119C05935, Google, UVA Strategic Investment Fund award number SIF160, DTRA Contract No. HDTRA1-19-D-0007, and VDH Grant VDH-21-501-0141

References

Adiga, A.; Dubhashi, D.; Lewis, B.; Marathe, M.; Venkatramanan, S.; and Vullikanti, A. 2020. Mathematical models for covid-19 pandemic: a comparative analysis. *Journal of the Indian Institute of Science*, 100(4): 793–807.

Adiga, A.; Wang, L.; Hurt, B.; Peddireddy, A.; Porebski, P.; Venkatramanan, S.; Lewis, B. L.; and Marathe, M. 2021. All models are useful: Bayesian ensembling for robust high resolution covid-19 forecasting. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2505–2513.

Aminikhanghahi, S.; and Cook, D. J. 2017. A survey of methods for time series change point detection. *Knowledge and information systems*, 51(2): 339–367.

Anastassopoulou, C.; Russo, L.; Tsakris, A.; and Siettos, C. 2020. Data-based analysis, modelling and forecasting of the COVID-19 outbreak. *PloS one*, 15(3): e0230405.

Bertozzi, A. L.; Franco, E.; Mohler, G.; Short, M. B.; and Sledge, D. 2020. The challenges of modeling and forecasting the spread of COVID-19. *Proceedings of the National Academy of Sciences*, 117(29): 16732–16738.

Bracher, J.; Ray, E. L.; Gneiting, T.; and Reich, N. G. 2020. Evaluating epidemic forecasts in an interval format. *arXiv preprint arXiv:2005.12881*.

Chakraborty, P.; Lewis, B.; Eubank, S.; Brownstein, J. S.; Marathe, M.; and Ramakrishnan, N. 2018. What to know before forecasting the flu. *PLOS Computational Biology*, 14(10): e1005964.

Cramer, E. Y.; Huang, Y.; Wang, Y.; Ray, E. L.; Cornell, M.; Bracher, J.; Brennen, A.; Castro Rivadeneira, A. J.; Gerding, A.; House, K.; Jayawardena, D.; Kanji, A. H.; Khandelwal, A.; Le, K.; Niemi, J.; Stark, A.; Shah, A.; Wattanachit, N.; Zorn, M. W.; Reich, N. G.; and Consortium, U. C.-. F. H. 2021. The United States COVID-19 Forecast Hub dataset. *Scientific Data*.

Delphi. 2021. Forecast Evaluation Dashboard. https://delphi.cmu.edu/forecast-eval/. Accessed: 2022-12-09.

Gao, J.; Sharma, R.; Qian, C.; Glass, L. M.; Spaeder, J.; Romberg, J.; Sun, J.; and Xiao, C. 2020. STAN: Spatio-Temporal Attention Network for Pandemic Prediction Using Real World Evidence. *arXiv preprint arXiv:2008.04215*.

Hernandez-Matamoros, A.; Fujita, H.; Hayashi, T.; and Perez-Meana, H. 2020. Forecasting of COVID19 per regions using ARIMA models and polynomial functions. *Applied Soft Computing*, 96: 106610.

Hub, C. F. 2020. Home - COVID 19 forecast hub. https://covid19forecasthub.org/. Accessed: 2022-12-09.

Kapoor, A.; Ben, X.; Liu, L.; Perozzi, B.; Barnes, M.; Blais, M.; and O'Banion, S. 2020. Examining COVID-19 Forecasting using Spatio-Temporal Graph Neural Networks. *arXiv preprint arXiv:2007.03113*.

Kufel, T.; et al. 2020. ARIMA-based forecasting of the dynamics of confirmed Covid-19 cases for selected European countries. *Equilibrium. Quarterly Journal of Economics and Economic Policy*, 15(2): 181–204.

Muggeo, V. 2008. segmented: An R package to Fit Regression Models with Broken-Line Relationships. *R NEWS*, 8/1: 20–25.

Nsoesie, E. O.; Brownstein, J. S.; Ramakrishnan, N.; and Marathe, M. V. 2014. A systematic review of studies on forecasting the dynamics of influenza outbreaks. *Influenza and other respiratory viruses*, 8(3): 309–316.

Raftery, A. E.; Gneiting, T.; Balabdaoui, F.; and Polakowski, M. 2005. Using Bayesian model averaging to calibrate forecast ensembles. *Monthly weather review*, 133(5): 1155–1174.

Ramchandani, A.; Fan, C.; and Mostafavi, A. 2020. DeepCOVIDNet: An Interpretable Deep Learning Model for Predictive Surveillance of COVID-19 Using Heterogeneous Features and Their Interactions. *arXiv preprint arXiv:2008.00115*.

Ray, E.; et al. 2021. Challenges in training ensembles to forecast covid-19 cases and deaths in the united states. *Int. Inst. Forecasters*.

Reich, N. G.; Brooks, L. C.; Fox, S. J.; Kandula, S.; McGowan, C. J.; Moore, E.; Osthus, D.; Ray, E. L.; Tushar, A.; Yamana, T. K.; et al. 2019. A collaborative multiyear, multimodel assessment of seasonal influenza forecasting in the United States. *Proceedings of the National Academy of Sciences*, 116(8): 3146–3154.

Rodriguez, A.; Tabassum, A.; Cui, J.; Xie, J.; Ho, J.; Agarwal, P.; Adhikari, B.; and Prakash, B. A. 2020. DeepCOVID: An Operational Deep Learning-driven Framework for Explainable Real-time COVID-19 Forecasting. *medRxiv*.

Rosenfeld, R.; and Tibshirani, R. J. 2021. Epidemic tracking and forecasting: Lessons learned from a tumultuous year. *Proceedings of the National Academy of Sciences*, 118(51).

Wang, L.; Adiga, A.; Venkatramanan, S.; Chen, J.; Lewis, B.; and Marathe, M. 2020. Examining deep learning models with multiple data sources for covid-19 forecasting. In *2020 IEEE International Conference on Big Data (Big Data)*, 3846–3855. IEEE.

Wang, L.; Ben, X.; Adiga, A.; Sadilek, A.; Tendulkar, A.; Venkatramanan, S.; Vullikanti, A.; Aggarwal, G.; Talekar, A.; Chen, J.; et al. 2021. Using Mobility Data to Understand and Forecast COVID19 Dynamics. *IJCAI 2021 Workshop on AI for Social Good*.

Winter, E. 2002. The shapley value. *Handbook of game theory with economic applications*, 3: 2025–2054.

Yamana, T. K.; Kandula, S.; and Shaman, J. 2017. Individual versus superensemble forecasts of seasonal influenza outbreaks in the United States. *PLoS computational biology*, 13(11): e1005801.