Enhancing COVID-19 Ensemble Forecasting Model Performance Using Auxiliary Data Sources

Aniruddha Adiga*§, Gursharn Kaur*§, Benjamin Hurt*, Lijing Wang†, Przemyslaw Porebski*,
Srinivasan Venkatramanan*, Bryan Lewis*, Madhav Marathe*‡
*Biocomplexity Institute, University of Virginia, Charlottesville, Virginia
† New Jersey Institute of Technology, New Jersey
‡Dept. of Computer Science, University of Virginia, Charlottesville, Virginia
Email: aniruddha@virginia.edu, marathe@virginia.edu

Abstract—**Real-time forecasting of non-stationary time series is a challenging problem, especially, when the time series evolves rapidly. For such cases, it has been observed that ensemble models consisting of a diverse set of model classes can perform consistently better than individual models. In order to account for the nonstationarity of the data and the lack of availability of training examples, the models are retrained in real-time using the most recent observed data samples. Motivated by the robust performance properties of ensemble models, we developed a Bayesian model averaging ensemble technique consisting of statistical, deep learning, and compartmental models for forecasting epidemiological signals, specifically, COVID-19 signals. We observed the epidemic dynamics go through a number of phases (waves). In our ensemble model, we observed that different model classes performed differently during the various phases. Armed with this understanding, in this paper, we propose a modification to the ensembling method to employ these phase information and use different weighting schemes for each phase to produce improved forecasts. However, predicting the phases of such time series is a significant challenge, especially when the evolution of the time series is governed by behavioral and immunological adaptations. We explore multiple datasets that can serve as leading indicators of trend changes and employ *transfer entropy* techniques to capture the relevant indicator. We propose a phase prediction algorithm to estimate the phases using the leading indicators. Using the knowledge of the estimated phase, we selectively sample the training data from similar phases. We evaluate our proposed methodology on our currently deployed COVID-19 forecasting model and the COVID-19 forecasthub models. The overall performance of the proposed model is consistent across the pandemic. More importantly, it is ranked *second* during two critical rapid growth phases in cases, regimes where the performance of most models from CDC forecasting hub dropped significantly.**

1. Introduction

An ensemble of models have been studied extensively and have shown to produce superior performance when compared to individual models [1]. In forecasting, ensemble models, specifically, Bayesian model averaging (BMA) [2] which are suitable for ensembling probabilistic forecasts have been shown to produce superior performance over individual models in varied applications such as epidemilogy [3], [4], [5], weather predictions [2], hydrology [6], political sciences [7], etc. Given, its wide-spread use in probabilistic forecasting, we developed a multi-class BMA ensemble model for COVID-19 case forecasting [8] in order to provide weekly forecasts to the CDC's COVID-19 ForecastHub (also referred to as *The Hub*) [9]. We are one of the teams contributing forecasts (listed as UVA-ENSEMBLE) since July 2020 and focused on incident cases and hospitalizations at a gran- ular level. The constituent models in the UVA-ENSEMBLE forecasting method include several standard statistical, deep learning and compartmental models. Specifically, we employ autoregressive models and its variants (AR with spatial exogenous regressor, ARIMA), an LSTM model, an ensemble Kalman filter (EnKF), and a compartmental model (SEIR). A detailed description of the models is available in our work [8].In order to account for nonstationarity of the case time series, we retrain the models every week.

An ensemble model, despite providing robust performances, is dependent on the quality of its constituent models. Despite the efforts of the forecasting community, there is considerable lack of 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 effective allocation of medical resources which are strained during these times. Hence, the ensemble models have suffered from outlying forecasts and have failed to catch the local peaks [10], a trend we have also observed in our own ensemble model forecasts.

Key observations from our real-time forecasting effort

- In the COVID-19 cases time series, we observe three distinct phases characterized by periods of rapid

§. Equal contribution

growth, plateau, and steady decline. The rapid growth phases are particularly critical as they often lead to severe burden on the medical resources.
- An ensemble model, despite providing robust performances, is dependent on the quality of its constituent models. Despite the best effort of the modelers, due to a lack of understanding of the disease dynamics of the novel virus, most teams struggled to forecast the waves, including our models.
- A retrospective evaluation of our deployed model [8] in *The Hub* across the forecasting weeks indicates that the BMA has a performance close to *The Hub*'s ensemble model. Further, we performed model ablation analysis to understand the influence of individual models on the ensemble. Our analysis indicates that: (i) compartmental models are useful during growth and decline phases but tend to over-predict during 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.

These observations have prompted us to modify the ensemble model to incorporate the phase information in the training and propose a phase-informed BMA model. These new methods employ big data analytics and machine learning techniques such as transfer entropy [11] to improve on forecasting performance. The contribution of this paper is three fold and is summarized as follows:

1.1. Summary of contributions.

Development of a new phase-informed BMA. We propose a modification to the BMA training where only training samples corresponding to a particular phase are employed to determine the weights. This selective sampling helps influence the BMA to assign higher weights to models with superior performance during similar phases observed historically.

Identifying leading indicators from multiple data sources In the phase prediction, a major challenge is to determine leading indicators to the case time series. We propose a transfer entropy technique [11] to obtain the leading indicators and then predict the phase for the target time series as a function of the phases of the leading indicators (sources).

Phase inference We present a new, simple method to address the problem of phase classification of a given time series. The phase classification involves assigning each time point to one of three phases {*Surge*, *Plateau*, *Decline*}. This is achieved by approximating the time series as a piece-wise linear signal and then inferring the phase based on the slope the linear fits.

Using the phase information, the BMA model is able to leverage the context-specific historical performance of individual methods thus 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 real-time forecasting tasks, for example to the likes of COVID-19 Forecasthub ensemble models.

An outline of the steps included in the pipeline is given in Figure 1.

1.2. Related Works

In the context of epidemic forecasting, auxiliary data sources have been employed extensively for improving forecast accuracy. In influenza-like illness and dengue forecasting additional sources such as Google health search trends, medical health records, weather data have been extensively s in forecasting models have shown to improve forecasting accuracy [12], [13], [14]. Similarly, in COVID-19 forecasting, mobility data, digital thermometers, medical records, etc. [15], [16], [17], [18]. The COVIDcast [19] data repository, has been able to aggregate data from mulitple providers and Faceook surveys. The utility of the dataset in forecasting is discussed in [20].

Multiple measures exist for determining the information flow between two signals, a popular method is the Granger causality test [21] which determines the linear relationship between signals, which is rarely the case. Under such circumstances the use information theoretic approaches such as transfer entropy [11], mutual information [22] and symbolic transfer entropy [23] have been shown to capture the source-target information flows.

2. Phase-Informed Bayesian Model Averaging

Motivated by the observation that certain models perform better during certain *phases*, we propose a method to supply the phase information during the BMA ensemble training. In the BMA framework, we train independently, one model per location. Considering K models per location, the BMA assumes that the conditional density of observing case count y given the forecasts $f_1, \dots, f_K$ generated from models $M_1, M_2, \dots, M_K$ is given by

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

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 estimate using the standard expectation-maximization (EM) algorithm [2], 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)

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

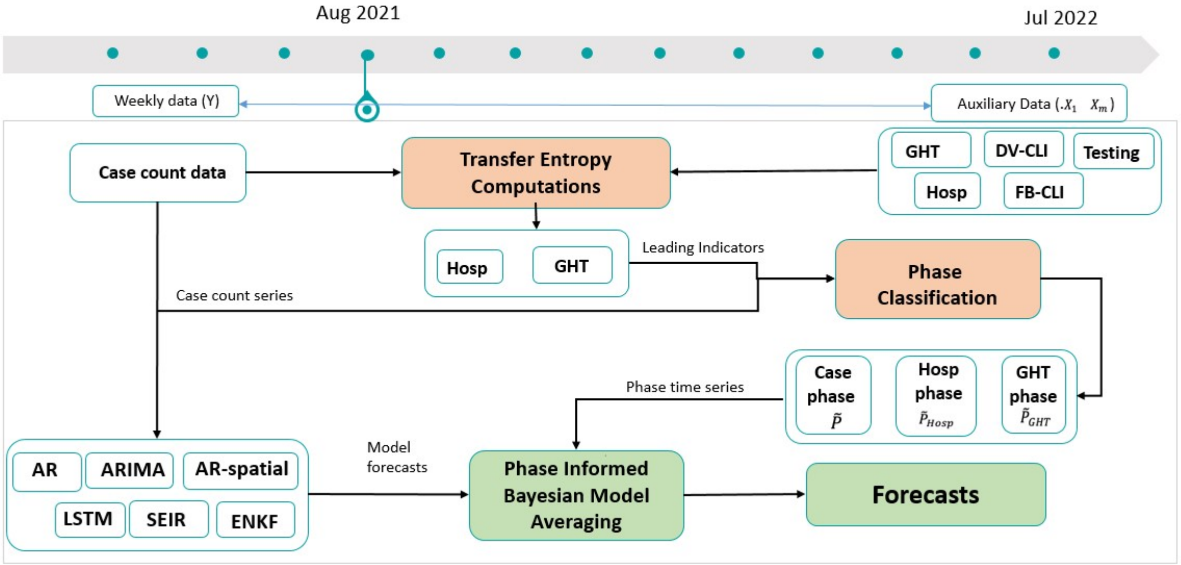


Figure 1: Workflow of the proposed phase-informed BMA

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 [8], $\mathcal{T}$ 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, in order to ensure that the most recent trend is captured, we consider only the most recent N weeks of performance in the training and not the entire set of historical forecasts.

Since the weights are estimated based on the individual model performances over the past N weeks, this introduces a latency in picking the best model during a phase change. The model has to observe sufficient forecasts from the best performing model over the next few weeks to put higher weights on it. This latency leads to the BMA model to produce under-performing forecasts.

We identify and address this issue by designing a BMA ensemble that uses the knowledge of the relevant phase to get improved forecasts. 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 prediction technique that enables us to determine $\mathcal{T}_r$.

2.1. Transfer Entropy for Identifying Leading Indicators

2.1.1. Data Sources. Since early 2020, [24] has collaborated with data partners to collect, curate, and make publicly available numerous real-time spatio-temporal COVID-19 indicators. These indicators have been aggregated to provide multiple views of pandemic activity in the United States [19]. Some of the data sources have been shown to be leading indicators of the case time series [20]. Additionally, these indicators are available at multiple resolutions (example at state level, county level, etc.). In this paper we consider signals only at the state-level as the reporting has been relatively consistent and less noisy compared to county-level signals.

The set of sources or indicator signals are mostly obtained through the COVIDcast application programming interface [19]. Several signals are available but we only consider signals that are categorized as *early indicators*. The individual signals are as follows:

Doctors visits COVID-like illness (DV-CLI): Estimated percentage of outpatient doctor visits primarily about COVID-related symptoms, based on data from health system partners.

Facebook-survey-based COVID-like illness (FB-CLI): Percentage of people with COVID-like illness symptoms estimated from Facebook survey responses (average of ≈ 40,000 surveys per day).

Anti-gen COVID-19 tests (Testing): Percentage of antigen tests that were positive for COVID-19 as provided by Quidel.

Google health trends (GHT): This aggregated, anonymized dataset shows trends in search patterns for symptoms. This data reflects the volume of Google searches for a broad set of symptoms, signs and health conditions. We considered keywords *'COVID-19 Symptoms', 'COVID-19 test', 'Sore throat', 'loss of smell', 'COVID-19 Home test'* and considered the median of the relative frequency of searches of these keywords.

COVID-19 Hospitalizations (hosp): The health and human services department provides multiple data concerning COVID-19 hospitalizations. Here we consider the sum of adult and pediatric confirmed COVID-19 hospital admissions occurring each day.

It is to be noted that these data streams undergo considerable amount of revisions or *backfills* across days/weeks. As an example, DV-CLI data observed over the most recent 5-7 days changes substantially over the next few days or even weeks. Estimating the backfill patterns is a non-trivial problem and is referred to as nowcasting [20]. We do not attempt nowcasting in this paper and consider the unrevised data as observed on the date of forecasting in all the subsequent analysis.

Several methods exist in literature for measuring information flows. The most popular is the Granger Causality [21], a statistical test, which looks at the linear dependence of the target time series on the lagged source time series. However, many of the signals of interest do not exhibit linear dependence with the target time series, which is often the case, then Granger causality test fail to capture the information flow. Under such circumstances, information theoretic approach of transfer entropy (TE) [11], [25] is shown to capture the source-target information flows and has been employed in different applications [26], [27]. TE computes the conditional mutual information between a target and lagged version of the source series. For a k order Markov process Y the Shannon transfer entropy measures the information flow from a process X to process Y and is defined as

$$\mathsf{TE}_{X\to Y}(k,l) \quad = \quad \sum_{y\in\mathcal{Y},x\in\mathcal{X}} \left(p\left(y(t+1), y^{k}(t), x^{(l)}(t)\right)\right. \\ \left. \times \log \frac{p(y(t+1)|y^{(k)}(t), x^{(l)}t)}{p\left(y(t+1)|y^{(k)}(t)\right)}\right), \tag{4}$$

where $y^{(k)}(t) = (y(t), \dots, y(t-k+1))$ and $x^{(l)}(t) = (x(t), \dots, x(t-l+1))$ are the observed sequences of Y and X, respectively.

Given m data auxiliary data sources $X_1, \dots, X_m$, we compute the transfer entropy for each pair (X_i, Y). To estimate the joint and conditional densities required in (4), we consider the recent 42 weeks of data for each bi-variate distribution (X_i, Y). We denote $\Lambda = \{1, \dots, m\}$ to be the index set of m auxiliary data source. Note that the set of indicators can change over time and we denote let $\Lambda(t) \subseteq \Lambda$ be the set of leading indicators at time t with non-zero TE values. In Figure 2a and Figure 2b, we observe that the number of leading indicators vary over the observed time period.

2.1.2. Phase inference and Prediction.

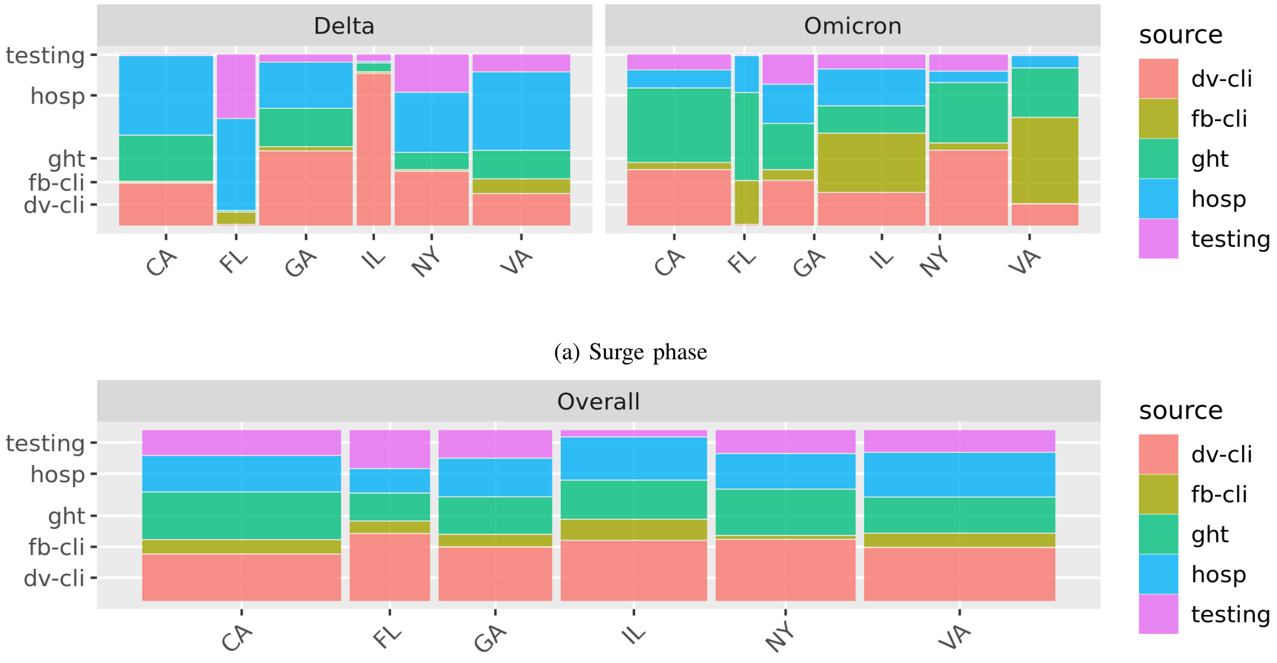
Once the indicators are obtained, we segment the cases and the indicator time series into different phases. Despite heterogeneity in the COVID-19 time series, we broadly observe three phases and classify the observed time period based on the rate of change of reported cases: Surge (period of steep growth in cases), Decline, and Plateau. We would like to note that the definitions of phases are subjective (several exist[1]) and can be user annotated or obtained through standard time-series change point detection algorithms [28]. The main purpose of phase classification is to capture distinct trends in the time series and leverage that information to better train the BMA model.

We first approximate the nonlinear time series with a piece-wise linear function. We use a standard R package `segmented` [29] to estimate multiple break-points. Note that, in real-time forecasting, since we obtain a new data point each week, the phase segments have to be re-estimated each week. Given the new data point, we would want to refine our estimates of phases but also ensure that they do not change significantly. Hence, each week, we apply the segmentation only on data starting from the recent two break points. The algorithm is described in Algorithm 1.

Algorithm 1 Piece-wise linear fit

```
Input: Ground truth $y(1), \dots, y(T)$
Output: Piece-wise linear version of $y(1), \dots, y(T)$ and set of break points $\{b_1, \dots, b_m\}$
Start with $t_0 = 15$
Get a piece-wise fit for $y(1), y(2), \dots, y(t_0)$ with break points $1 \leq b_1^{(t_0)} \leq \dots \leq b_{k_{t_0}}^{(t_0)}$
$\mathcal{B}(t_0) \leftarrow \{1, b_1^{(t_0)}, \dots, b_{k_{t_0}}^{(t_0)}\}$
while $t_0 + 1 \leq t \leq T$ do
    Get a piece-wise fit for $\{y_s : b_{k_{t-1}-2}^{(t-1)} \leq s \leq t\}$ with break points $1 \leq b_1^{(t)} \leq b_2^{(t)} \leq \dots \leq b_{k_t}^{(t)}$
    $b_1^\star \leftarrow \max \mathcal{B}(t-1)$, $b_2^\star \leftarrow \max \mathcal{B}(t-1) \setminus b_1^\star$
    $\mathcal{B}(t) \leftarrow (\mathcal{B}(t-1) \setminus \{b_1^\star, b_2^\star\}) \cup \{1, b_1^{(t)}, \dots, b_{k_t}^{(t)}\}$
end while
Let $\mathcal{B}(t) = \{b_1, \dots, b_m\}$ and $b_0 = 1$
for $0 \leq i < m$ do
    $\{\tilde{y}(t); b_i \leq t \leq b_{i+1}\}$ is the linearly interpolation between $y(b_i)$ and $y(b_{i+1})$
end for
return $\tilde{y}(1), \dots \tilde{y}(T)$ and breakpoints set $\{b_1, \dots, b_m\}$
```

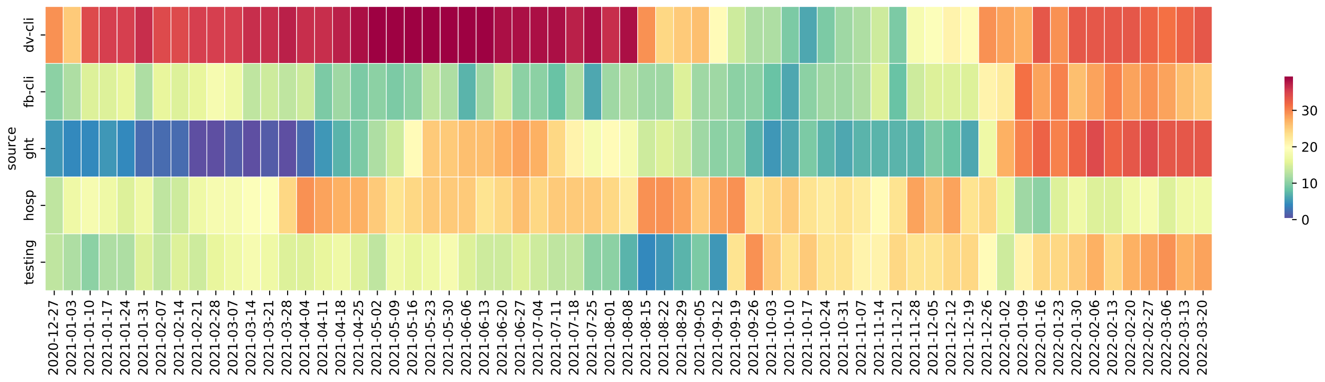
[1] https://www.cdc.gov/flu/pandemic-resources/planning-preparedness/global-planning-508.html



(a) Surge phase

(b) Overall

Figure 2: Mosaic plot of the leading indicators for selected states during: 2a Delta wave (15 July 2021 to 15 August 2021) and Omicron wave (15 December 2021 to 15 January 2022) and 2b overall observed time period i.e. from May 2021 to July 2022. Here the width of a bar denotes the frequency for the overall trends for a state and height of each rectangle within each bar represents the frequency proportion of a specific source.



(a) Temporal evolution of sources. Heatmap shows the total number of states for which an individual signal was identified as a sources (signal with significant TE values) for each week. DV-CLI is a source for a significant number of states for a majority of the forecasting weeks.

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 $(b_k, b_{k+1}]$ as 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. Let $P(t)$ and $P_1(t), \ldots, P_m(t)$, be the phase time series for the piece-wise constant versions (as defined in Algorithm 1) of $Y(t)$ and $X_1(t), \ldots, X_m(t)$ respectively. The algorithm to obtain the phase time series for a given time series is described in Algorithm 3.

Let $P^{(k)}(t) = (P(t), \ldots, P(t-k+1))$ and $P_i^{(k)}(t) = (P_i(t), \ldots, P_i(t-k+1))$, then we assume that $P(t)$ is a Markov process that depends only on $P_{\text{aux}}(t-1) = \{P_i(t-1); i \in \Lambda(t)\}$ and $P^{(\tau)}(t-1)$, where $\tau = \max_i \tau_i$. That is, the phase at time t for the case count time series i.e. $P(t)$ can be written as a function of past τ values of itself and past τ_i phases of the X_i when i is restricted to the leading indicators set $\Lambda(t)$. Thus we have

$$P(t) = f\big(P^{(\tau)}(t-1), P_{\text{aux}}(t-1)\big), \tag{5}$$

where f is a $\{S, D, P\}$ valued random function such that for a given phase sequence at time $t-1$, $P^{(\tau)}(t-1) = w$ and $P_i^{(\tau_i)}(t-1) = w_i$

$$f(w, w_1, \dots, w_m) = \begin{cases} S & \text{w.p. } p_1(t) \\ D & \text{w.p. } 1 - p_0(t) - p_1(t) \\ P & \text{w.p. } p_0(t) \end{cases}$$

where the probabilities $p_0(t)$ and $p_1(t)$ are estimated empirically at every time t as

$$\hat{p}_1(t) = \frac{|\{P(t) = S\} \cap \{P^{(\tau)}(t-1) = w, P_i^{(\tau_i)}(t-1) = w_i; i \in \Lambda(t)\}|}{|\{P^{(\tau)}(t-1) = w, P_i^{(\tau_i)}(t-1) = w_i; i \in \Lambda(t)\}|}$$

and

$$\hat{p}_0(t) = \frac{|\{P(t) = P\} \cap \{P^{(\tau)}(t-1) = w, P_i^{(\tau_i)}(t-1) = w_i; i \in \Lambda(t)\}|}{|\{P^{(\tau)}(t-1) = w, P_i^{(\tau_i)}(t-1) = w_i; i \in \Lambda(t)\}|}.$$

As an example, suppose we obtain a subset of signals $\{1, 2, 5\}$ as the leading indicators with lags $\{\tau_1 = 2, \tau_2 = 1, \tau_5 = 1\}$. Since the maximum lag is 2, $\tau = 2$. Thus $\tilde{P}(t) = (P(t-1), P(t-2))$, $P_{\text{aux}}(t) = (P_1(t-1), P_1(t-2), P_2(t-1), P_5(t-1))$. Now, the vector $\mathbf{w} = (w, w_1, w_2, w_5) = [P(t-1), P(t-2), P_1(t-1), P_1(t-2), P_2(t-1), P_5(t-1)]$. In Figure 4, we illustrate the process of predicting the phase.

2.2. Transfer Entropy for Various Sources

The TEs have been computed using the IDTxL toolbox [30]. Figure 3a and Figure **??** describe the distributions of TE (across all states) during two surge phases and the overall observed time period. We observe from these two figures that the distribution of TE for various sources change over time. In-particular, we observe that *ght* and *dv-cli* are both heavy tailed compare to other sources in the overall observed time period whereas during the Delta wave *hosp* and *dv-cli* have higher TE values and during the Omicron wave, only *ght* has slightly higher TE value compared to other sources while all other sources get similar TE values. With this observation, we conclude that it is essential to re-identify the leading indicators with significant TE with every new set of observed data points.

Algorithm 2 Transfer Entropy

```
Input: Auxiliary sources $\Lambda = \{1, \dots, m\}$.
Observed target case count time series $y(t)$.
Observed source time series $x_i(t)$, for each $i \in \Lambda$.
Output: Probabilistic estimate of $P(t+1)$

Obtain phase time series $P(t), P_1(t), \dots P_m(t)$ from Algorithm 3 for $y, x_1, \dots, x_m$.
Compute $h_i = \mathsf{TE}_{X_i \to Y}$ for each pair $(X_i, Y)$ with appropriate order $\tau_i$.
Obtain $\Lambda(t)$ using Python package IDTxL.
$\tau \leftarrow \max_{i \in \Lambda(t)} \tau_i$
$\tilde{P}(t) := (P(t), P(t-1), P(t-\tau+1))$
for $i \in \Lambda(t)$ do
    $\tilde{P}_i(t) := (P_i(t), P_i(t-1), \dots, P_i(t-\tau_i+1))$
end for
Given $\tilde{P}_i(t) = w_i$
$n_o \leftarrow |\{\tilde{P}_i(s) = w_i;\ i \in \Lambda(t), 1 \le s \le t-1\}|$
if $\Lambda(t) = \emptyset$ or $n_o = 0$ then
    $P(t+1) \leftarrow P(t)$
else
    while $n_o/t \le 0.1$ do
        update $\Lambda(t) \leftarrow \Lambda(t) \setminus \{j : h_j = \min_{i \in \Lambda(t)} h_i\}$
    end while
    $P(t+1) \leftarrow f(\tilde{P}(t), \tilde{P}_i(t))$
end if
return $(\hat{p}_1(t), \hat{p}_0(t))$
```

Algorithm 3 Phase classification

```
Input: Time series $y(t)$.
Output: Phase time series $P(t)$.
Parameters: $\delta$
Approximate $y(t)$ with a piece-wise linear time series $\tilde{y}(t)$ with breakpoints $b_1, \dots b_n$
if $y_{b_{i+1}} > (1+\delta) y_{b_i}$ then
    $P(t) \leftarrow S$ for $t \in (b_i, b_{i+1}]$
else
    if $y_{b_{i+1}} < (1-\delta) y_{b_i}$ then
        $P(t) \leftarrow D$ for $t \in (b_i, b_{i+1}]$
    else
        $P(t) \leftarrow P$
    end if
end if
```

3. Results

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 [10].

3.1. Retrospective Evaluation: A Comparison with *The Hub* Models

Since early 2020, over 100 models from dozens of teams have submitted forecasts to *The Hub* with the numbers varying each week. The model details are available in [31].

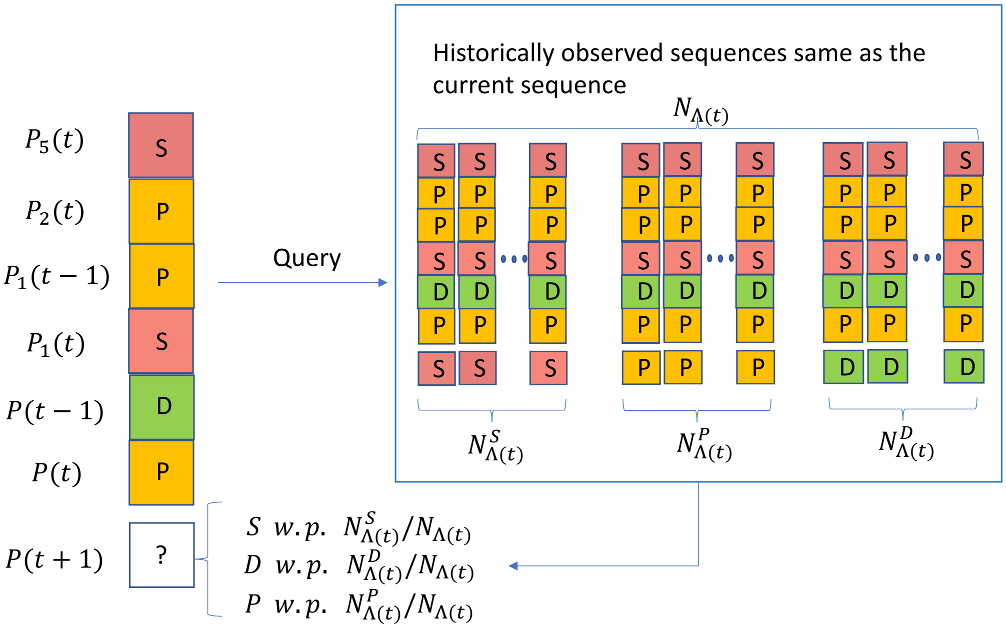
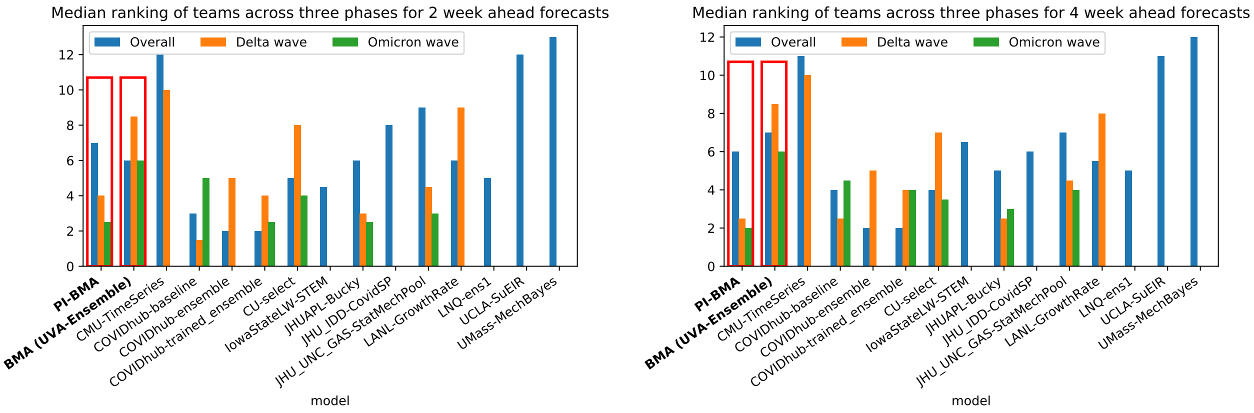

$P_5(t)$
S
$P_2(t)$
P
$P_1(t-1)$
P
$P_1(t)$
S
$P(t-1)$
D
$P(t)$
P
$P(t+1)$
?
Query
Historically observed sequences same as the current sequence
$N_{\Lambda(t)}$
S P P S D P S
S P P S D P P
S P P S D P D
$N^S_{\Lambda(t)}$
$N^P_{\Lambda(t)}$
$N^D_{\Lambda(t)}$
$S \ w.p. \ N^S_{\Lambda(t)}/N_{\Lambda(t)}$
$D \ w.p. \ N^D_{\Lambda(t)}/N_{\Lambda(t)}$
$P \ w.p. \ N^P_{\Lambda(t)}/N_{\Lambda(t)}$

Figure 4: An example of phase prediction.

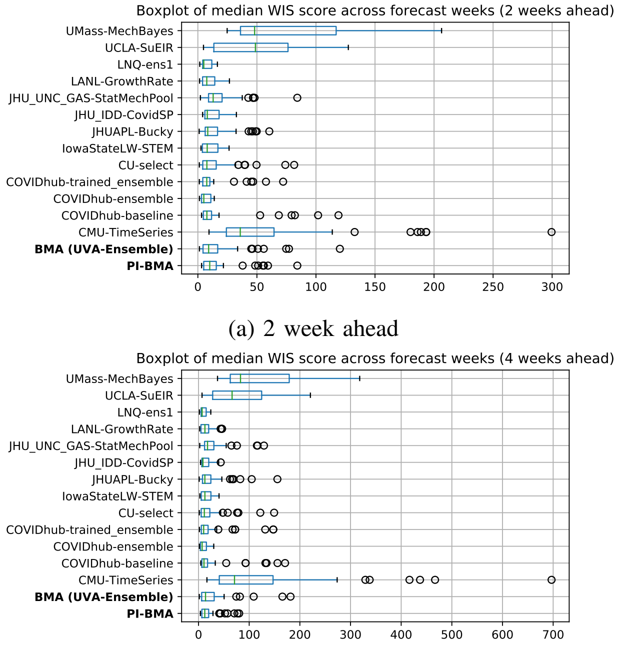


(a) 2 week ahead

(b) 4 week ahead

Figure 5: A comparison of several *The Hub* models performance. The median ranking of models for 2 and 4 week ahead forecasts in (a) and (d) computed across different regimes, respectively. Blue bars shows 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. **The rankings across different phases indicate that the PI-BMA (red box) is able to provide significantly better forecasts that our UVA-Ensemble model, especially 4 weeks ahead, for critical surge phases corresponding to the *Delta wave* (median ranking of 2) and the *Omicron wave* (median ranking of 2)**.

Among the several teams, only a handful have provided forecasts consistently, especially at the county-level. As 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 significant number of forecasts. As the pandemic progressed, we do observe that the number of models start to drop after July 2021. 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 epidemiological forecasting



(a) 2 week ahead

(b) 4 week ahead

Figure 6: A comparison variation in median WIS scores across counties for several *The Hub* models performance.

community for probabilistic forecast evaluation [32]:

$$WIS_{\alpha_{0:k}}(F, y) = \frac{1}{K + 0.5} \sum_{k=0}^{K} \frac{\alpha_k}{2}(u_k - l_k) + \frac{2}{\alpha_k}(l_k - y)\mathbb{1}(y < l_k) + \frac{2}{\alpha_k}(y - u_k)\mathbb{1}(y > u_k) \tag{6}$$

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.

The model performances are first ranked for each forecast week and target horizon by considering its median WIS score across all the counties (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 6a and 6b for 2 week ahead and 4 week ahead forecast horizons, respectively. The blue bars, which corresponds 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 harder 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 2 out of 9) and Omicron wave surge (median ranking of 2 out of 6). The PI-BMA's performance indicates that the model is able to effectively incorporate the phase information and provide considerably better forecasts during important 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**.

4. Conclusion

This paper, based on the observations made during the COVID-19 forecasting efforts, proposed a novel phase-based Bayesian model averaging, a modification of the current model. The paper provides three critical contributions, (i) a phase-based sampling approach for training the ensemble, (ii) a novel transfer-entropy-based leading indicator identification method, and (iii) a phase prediction method that uses the phases from the leading indicators to make prediction for the future phase. The performance of the PI-BMA model validates the utility if the phase-based training methodology. The proposed method is fairly generic and can be incorporated in most ensemble model.

In future work, we plan on exploring other variants of transfer entropy. In addition, several of the auxiliary data sources undergo revision and developing a model to correct for it might improve phase forecasting.

References

[1] O. Sagi and L. Rokach, "Ensemble learning: A survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 4, p. e1249, 2018.

[2] A. E. Raftery, T. Gneiting, F. Balabdaoui, and M. Polakowski, "Using bayesian model averaging to calibrate forecast ensembles," *Monthly weather review*, vol. 133, no. 5, pp. 1155–1174, 2005.

[3] T. K. Yamana, S. Kandula, and J. Shaman, "Superensemble forecasts of dengue outbreaks," *Journal of The Royal Society Interface*, vol. 13, no. 123, p. 20160410, 2016.

[4] 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 *et al.*, "A collaborative multiyear, multimodel assessment of seasonal influenza forecasting in the united states," *Proceedings of the National Academy of Sciences*, vol. 116, no. 8, pp. 3146–3154, 2019.

[5] E. Y. Cramer, E. L. Ray, V. K. Lopez, J. Bracher, A. Brennen, A. J. Castro Rivadeneira, A. Gerding, T. Gneiting, K. H. House, Y. Huang *et al.*, "Evaluation of individual and ensemble probabilistic forecasts of covid-19 mortality in the united states," *Proceedings of the National Academy of Sciences*, vol. 119, no. 15, p. e2113561119, 2022.

[6] Q. Duan, N. K. Ajami, X. Gao, and S. Sorooshian, "Multi-model ensemble hydrologic prediction using bayesian model averaging," *Advances in water Resources*, vol. 30, no. 5, pp. 1371–1386, 2007.

[7] J. M. Montgomery, F. M. Hollenbach, and M. D. Ward, "Improving predictions using ensemble bayesian model averaging," *Political Analysis*, vol. 20, no. 3, pp. 271–291, 2012.

[8] A. Adiga, L. Wang, B. Hurt, A. Peddireddy, P. Porebski, S. Venkatramanan, B. L. Lewis, and M. Marathe, "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*, 2021, pp. 2505–2513.

[9] E. Y. Cramer, Y. Huang, Y. Wang, E. L. Ray, M. Cornell, J. Bracher, A. Brennen, A. J. Castro Rivadeneira, A. Gerding, K. House, D. Jayawardena, A. H. Kanji, A. Khandelwal, K. Le, J. Niemi, A. Stark, A. Shah, N. Wattanachit, M. W. Zorn, N. G. Reich, and U. C.-. F. H. Consortium, "The united states covid-19 forecast hub dataset," *medRxiv*, 2021. [Online]. Available: https://www.medrxiv.org/content/10.1101/2021.11.04.21265886v1

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

[11] T. Schreiber, "Measuring information transfer," *Physical review letters*, vol. 85, no. 2, p. 461, 2000.

[12] S. Yang, M. Santillana, and S. C. Kou, "Accurate estimation of influenza epidemics using google search data via argo," *Proceedings of the National Academy of Sciences*, vol. 112, no. 47, pp. 14 473–14 478, 2015.

[13] P. Rangarajan, S. K. Mody, and M. Marathe, "Forecasting dengue and influenza incidences using a sparse representation of google trends, electronic health records, and time series data," *PLOS Computational Biology*, vol. 15, no. 11, pp. 1–24, 11 2019. [Online]. Available: https://doi.org/10.1371/journal.pcbi.1007518

[14] J. Shaman and A. Karspeck, "Forecasting seasonal outbreaks of influenza," *Proceedings of the National Academy of Sciences*, vol. 109, no. 50, pp. 20 425–20 430, 2012.

[15] A. Kapoor, X. Ben, L. Liu, B. Perozzi, M. Barnes, M. Blais, and S. O'Banion, "Examining covid-19 forecasting using spatio-temporal graph neural networks," *arXiv preprint arXiv:2007.03113*, 2020.

[16] L. Wang, X. Ben, A. Adiga, A. Sadilek, A. Tendulkar, S. Venkatramanan, A. Vullikanti, G. Aggarwal, A. Talekar, J. Chen *et al.*, "Using mobility data to understand and forecast covid19 dynamics," *IJCAI 2021 Workshop on AI for Social Good*, 2021.

[17] C. Fritz, E. Dorigatti, and D. Rügamer, "Combining graph neural networks and spatio-temporal disease models to predict covid-19 cases in germany," *arXiv preprint arXiv:2101.00661*, 2021.

[18] A. Rodriguez, A. Tabassum, J. Cui, J. Xie, J. Ho, P. Agarwal, B. Adhikari, and B. A. Prakash, "Deepcovid: An operational deep learning-driven framework for explainable real-time covid-19 forecasting," *medRxiv*, 2020.

[19] A. Reinhart, L. Brooks, M. Jahja, A. Rumack, J. Tang, S. Agrawal, W. Al Saeed, T. Arnold, A. Basu, J. Bien *et al.*, "An open repository of real-time covid-19 indicators," *Proceedings of the National Academy of Sciences*, vol. 118, no. 51, p. e2111452118, 2021.

[20] D. J. McDonald, J. Bien, A. Green, A. J. Hu, N. DeFries, S. Hyun, N. L. Oliveira, J. Sharpnack, J. Tang, R. Tibshirani *et al.*, "Can auxiliary indicators improve covid-19 forecasting and hotspot prediction?" *Proceedings of the National Academy of Sciences*, vol. 118, no. 51, p. e2111453118, 2021.

[21] C. W. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica: journal of the Econometric Society*, pp. 424–438, 1969.

[22] A. Dionisio, R. Menezes, and D. A. Mendes, "Mutual information: a measure of dependency for nonlinear time series," *Physica A: Statistical Mechanics and its Applications*, vol. 344, no. 1-2, pp. 326–329, 2004.

[23] M. Staniek and K. Lehnertz, "Symbolic transfer entropy," *Physical review letters*, vol. 100, no. 15, p. 158101, 2008.

[24] T. Arnold, J. Bien, L. Brooks, S. Colquhoun, D. Farrow, J. Grabman, P. Maynard-Zhang, A. Reinhart, and R. Tibshirani, *covidcast: Client for Delphi's COVIDcast Epidata API*, 2021, r package version 0.4.2. [Online]. Available: https://cmu-delphi.github.io/covidcast/covidcastR/

[25] T. Bossomaier, L. Barnett, M. Harré, and J. T. Lizier, "Transfer entropy," in *An introduction to transfer entropy*. Springer, 2016, pp. 65–95.

[26] R. Marschinski and H. Kantz, "Analysing the information flow between financial time series," *The European Physical Journal B-Condensed Matter and Complex Systems*, vol. 30, no. 2, pp. 275–281, 2002.

[27] R. Vicente, M. Wibral, M. Lindner, and G. Pipa, "Transfer entropy—a model-free measure of effective connectivity for the neurosciences," *Journal of computational neuroscience*, vol. 30, no. 1, pp. 45–67, 2011.

[28] S. Aminikhanghahi and D. J. Cook, "A survey of methods for time series change point detection," *Knowledge and information systems*, vol. 51, no. 2, pp. 339–367, 2017.

[29] V. Muggeo, "segmented: An r package to fit regression models with broken-line relationships," *R NEWS*, vol. 8/1, pp. 20–25, 2008.

[30] P. Wollstadt, J. T. Lizier, R. Vicente, C. Finn, M. Martinez-Zarzuela, P. Mediano, L. Novelli, and M. Wibral, "Idtxl: The information dynamics toolkit xl: a python package for the efficient analysis of multivariate information dynamics in networks," *arXiv preprint arXiv:1807.10459*, 2018.

[31] "The COVID-19 Forecast Hub community," https://covid19forecasthub.org/, 2021.

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