Forecasting Seasonal Influenza in the U.S.: A collaborative multi-year, multi-model assessment of forecast performance

Logan Brooks, Spencer Fox, Sasikiran Kandula, Craig McGowan, Evan Moore, Dave Osthus, Evan Ray, Nicholas G Reich, Roni Rosenfeld, Jeffrey Shaman, Abhinav Tushar, Teresa Yamana [authorship list to be finalized]

May 11, 2018

Abstract

TBD.

1 Introduction

Over the past 15 years, the number of published research articles on forecasting infectious diseases has tripled (Web of Science). This increased interest has been fueled in part by the promise of 'big data', that near real-time data streams of large-scale population behavior [1] to microscopic changes in a pathogen [2] could lead to measurable improvements in how disease transmission is measured, forecasted, and prevented [3]. With the spectre of a global pandemic looming, improving infectious disease forecasting continues to be a central priority of global health preparedness efforts.[4, 5, 6]

Forecasts of infectious disease transmission can inform public health response to outbreaks. Accurate forecasts of the timing and spatial spread of outbreaks of infectious diseases can provide valuable information about where public health interventions can be targeted.[7] Decisions about hospital staffing, resource allocation, and the timing of public health communication campaigns can be assisted by forecasts. Implementation of interventions designed to disrupt disease transmission, such as vector control measures or mandatory infection prevention protocols at hospitals or health clinics, can be targeted based on forecasted incidence.

In part due to the growing recognition of the importance of systematically integrating forecasting into public health outbreak response, large-scale collaborations have been used in forecasting applications to develop common data standards and facilitate comparisons across multiple models.[8, 9, 10, 11] By enabling a standardized comparison in a single application, these studies greatly improve our understanding of which models perform best in certain settings, of how results can best be disseminated and used by decision-makers, and of what the bottlenecks are in terms of improving forecasts.

The aim of this study is to present a standardized comparison of a range of different forecasting models for influenza in the US, over multiple seasons. Our work brings together models from five different institutions: Carnegie Mellon, Columbia University, Los Alamos National Laboratory, University of Massachusetts-Amherst, and University of Texas-Austin. While most groups developed more than one model for consideration, the models developed within

a single group through the use of common data sources and/or methodologies often bear similarities to each other. Having 22 models from five different teams enhances the diversity of the models presented.

While such multi-model comparisons exist in the literature for single-season performance, a unique aspect of this work is its focus on performance of the models over a longer period of time, i.e. seven flu seasons. To our knowledge, this is the first documented comparison of multiple models (from different teams), in a "real-time" setting, across multiple seasons for any infectious disease application. Since each season has unique dynamical structure, multi-season comparisons like this have great potential to improve our understanding of how models perform over the longer term and which models may be reliable in the future.

This work relies on the forecasting structure developed by existing public forecasting challenges. Starting in the 2013/2014 influenza season, the U.S. Centers for Disease Control and Prevention (CDC) has run the "Forecast the Influenza Season Collaborative Challenge" (a.k.a. FluSight) each influenza season, soliciting weekly forecasts for specific influenza season metrics from teams across the world.[8, 10] These forecasts are displayed together on a website during the season and are evaluated for accuracy after the season is over.[12] This effort has galvanized a community of scientists interested in forecasting, creating an organic testbed for improving both our technical understanding of how different forecast models perform and also how to integrate these models into decision-making.

Building on the structure of the FluSight challenges (and those of other collaborative forecasting efforts[9, 11]), a subset of FluSight participants founded a consortium in early 2017 to facilitate direct comparison and fusion of modeling approaches. In this paper, we provide a detailed analysis of the performance of 22 different models from five different teams over the course of seven influenza seasons. Drawing on the different expertise of the five teams allows us to make fine-grained and standardized comparisons of distinct approaches to disease incidence forecasting that use different data sources and statistical models.

In addition to analyzing comparative model performance over seasons, this work identifies key bottlenecks that limit the accuracy and generalizability of current forecasting efforts. Specifically, we present quantitative analyses of the impact that incomplete or partial case reporting has on forecast accuracy. Additionally, we assess whether purely statistical models show similar performance to models that consider explicit mechanistic models of disease transmission.

2 Methods

2.1 FluSight Challenge Overview

Detailed methodology and results from previous FluSight challenges have been published[8, 10], and we summarize the key features of the challenge here.

The FluSight challenge focuses on forecasts of the weighted percentage of doctor's office visits where the patient showed symptoms of an influenza-like illness in a particular region. Weighting is done by state population as the data are aggregated to the regional level. This is a standard measure of seasonal flu activity, for which public data is available for the US back to the 1997/1998 influenza season (Figure 1A). During each influenza season, these data are updated each week by the CDC. When the most recent data are released, the prior weeks' reported wILI data may also be revised. The unrevised data, available at a particular moment in time, is available via the

DELPHI real-time epidemiological data API beginning in the 2013/2014 season.[13] This API enables researchers to "turn back the clock" to a particular moment in time and use the data available at that time. This tool facilitates more accurate assessment of how models would have performed in real-time.

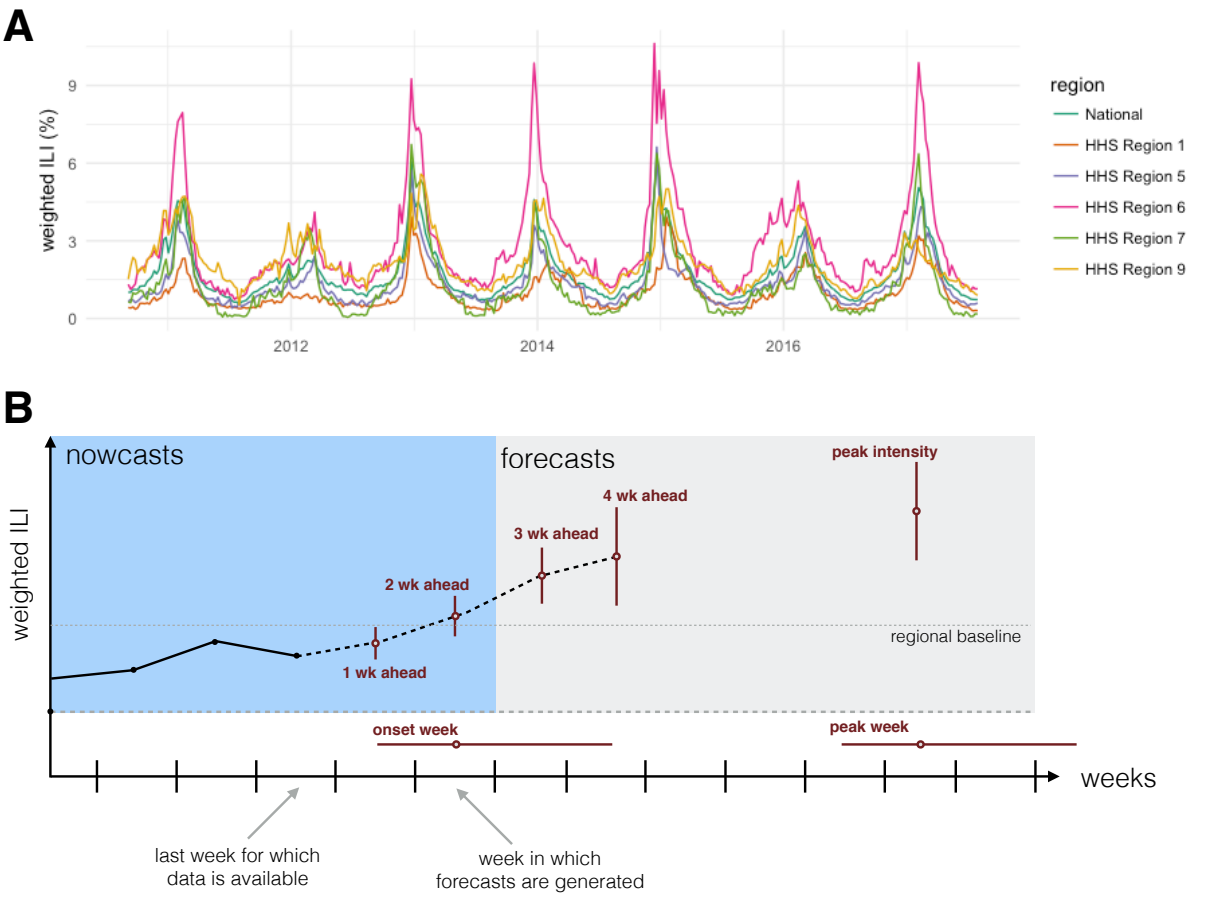

Figure 1: (A) Weighted influenza-like illness data downloaded from the CDC website. The y-axis shows the weighted percentage of doctor's office visits where the patient had influenza-like illness for each week between September 2010 through July 2017, which is the time period for which the models presented in this paper made seasonal forecasts. (B) A diagram showing the anatomy of a single forecast. The seven forecasting targets are illustrated with a point estimate (dot) and interval (uncertainty bars). The five targets on the wILI scale are shown with uncertainty bars spanning the vertical wILI axis, while the two targets for a time-of-year are illustrated with horizontal uncertainty bars along the temporal axis. The onset is defined relative to a region- and season-specific baseline wILI percentage defined by the CDC. Arrows illustrate the timeline for a typical forecast for the CDC FluSight challenge, assuming that forecasts are generated or submitted to the CDC using the most recent reported data. This data includes the first reported observations of wILI% from two weeks prior. Therefore, 1 and 2 week-ahead forecasts are referred to as nowcasts, i.e. at or before the current time. Similarly, 3 and 4 week-ahead forecasts are forecasts, or estimates about events in the future.

The FluSight challenges have defined seven forecasting targets of particular public health relevance. Three of these targets are fixed scalar values for a particular season: onset week, peak week, and peak intensity (i.e. the maximum observed wILI percentage). The remaining four targets are the observed wILI percentages in each of the subsequent four weeks (Figure 1B).

The FluSight challenges have also required that all forecast submissions follow a particular format. A single

submission file (a comma-separated text file) contains the forecast made for a particular epidemic week (EW) of a season. Standard CDC definitions of epidemic week are used. Each file contains binned predictive distributions for seven specific targets across the 10 HHS regions of the US plus the national level. Each file contains over 8000 rows and typically is about 400KB in size.

To be included in the model comparison presented here, previous participants in the CDC FluSight challenge were invited to provide out-of-sample forecasts for the 2010/2011 through 2016/2017 seasons. For each season, files were submitted for EW40 (using standard CDC defined epidemic weeks [14, 15, 16]) of the first calendar year of the season through EW20 of the following calendar year. (For seasons that contained an EW53, an additional file labeled EW53 was included.) For each model, this involved creating 233 separate forecast submission files, one for each of the weeks in the seven training seasons. Each forecast file represented a single submission file, as would be submitted to the CDC challenge. Each team created their submitted forecasts in a prospective, out-of-sample fashion, i.e. fitting or training the model only on data available before the time of the forecast (see Figure 1). Some data sources (e.g. wILI data prior to the 2014/2015 season) were not archived in a way that made data reliably retrievable in this "real-time" manner. In these situations, teams were still allowed to use these data sources with best efforts made to ensure forecasts were made using only data available at the time forecasts would have been made.

2.2 Summary of Models

Five teams each submitted between 1 and 9 separate models for evaluation (Table 1). A wide range of methodological approaches and modeling paradigms are included in the set of forecast models. For example, seven of the models utilize a compartmental structure (e.g. Susceptible-Infectious-Recovered), a model framework that directly encodes both the transmission and the susceptible-limiting dynamics of infectious disease outbreaks. Other less directly mechanistic models use statistical approaches to model the outbreak phenomenon directly by incorporating recent incidence and seasonal trends. Six models directly incorporate external data (i.e. not just the wILI measurements from the CDC ILINet dataset), including historical humidity data and Google search data. Two models stand out as being clear baseline models, that never change based on recent data. The `Delphi-Uniform` model always provides a forecast that assigns equal probability to all possible outcomes. The `ReichLab-KDE` model yields predictive distributions based entirely on data from other seasons using kernel density estimation (KDE) for seasonal targets and a generalized additive model with cyclic penalized splines for weekly incidence. Throughout the manuscript when we refer to the 'historical baseline' model we mean the `ReichLab-KDE` model. Once submitted to the central repository, the models were not updated or modified except in four cases to fix explicit bugs in the code that yielded numerical problems with the forecasts. (In all cases, the updates did not substantially change the performance of the updated models.) Re-fitting of models or tuning of model parameters was explicitly discouraged to avoid unintentional overfitting of models.

2.3 Metric Used for Evaluation and Comparison

Influenza forecasts have been evaluated by the CDC primarily using a variation of the log-score, a measure that enables evaluation of both the precision and accuracy of a forecast.[17] The log-score for a model m is defined as $\log f_m(z^*|\mathbf{x})$ where $f_m(z|\mathbf{x})$ is the predicted density function from moel m for some target Z, conditional on

Team	Model Abbr	Model Description	Ext. Data	Comp. Model*	Ens. Model
CU	EAKFC_SEIRS	Ensemble Adjustment Kalman Filter SEIRS	x	x	
	EAKFC_SIRS	Ensemble Adjustment Kalman Filter SIRS	x	x	
	EKF_SEIRS	Ensemble Kalman Filter SEIRS	x	x	
	EKF_SIRS	Ensemble Kalman Filter SIRS	x	x	
	RHF_SEIRS	Rank Histogram Filter SEIRS	x	x	
	RHF_SIRS	Rank Histogram Filter SIRS	x	x	
	BMA	Bayesian Model Averaging			x
Delphi	BasisRegression	Basis Regression (epiforecast package defaults)			
	DeltaDensity1	Delta Density (epiforecast package defaults)			
	EmpiricalBayes1	Empirical Bayes (conditioning on past four weeks only)			
	EmpiricalBayes2	Empirical Bayes (epiforecast package defaults)			
	EmpiricalFuture	Empirical Futures (epiforecast package defaults)			
	EmpiricalTraj	Empirical Trajectories (epiforecast package defaults)			
	DeltaDensity2	Markovian Delta Density (epiforecast package defaults)			
	Uniform	Uniform Distribution			
	Stat	Statistical Ensemble (combining predictions from the other 8 models submitted by Delphi)			x
LANL	DBM	Dynamic Bayesian SIR Model with a hierarchical discrepancy		x	
ReichLab	KCDE	Kernel Conditional Density Estimation using recent observations and seasonality			
	KDE	Kernel Density Estimation (seasonal targets) and cyclic penalized splines (week-ahead targets)			
	SARIMA1	SARIMA model without seasonal differencing			
	SARIMA2	SARIMA model with seasonal differencing			
UTAustin	EDM	Empirical Dynamic Model, or topological method of analogues			

Table 1: List of models, with key characteristics. Team abbreviations are translated as: CU = Columbia University (PI: Shaman), Delphi = Carnegie Mellon (PI: Rosenfeld), LANL = Los Alamos National Laboratories (PI: Osthus), ReichLab = University of Massachusetts Amherst (PI: Reich), UTAustin = University of Texas Austin (PI: Lauren Meyers). The 'Ext data' column notes models that use data external to the ILI Net data from CDC. The 'Comp. model' column notes models that rely to some extent on an compartmental model formulation. The 'Ens. model' column notes models that are ensemble models.

some data $\mathbf{x}$ and z^* is the observed value of the target Z. The log-score is a "proper" scoring rule, which has the practical implication that linear combinations (i.e. arithmetic means) of log scores will also be proper.

Consistent with the primary evaluation performed by the CDC, we used a modified form of the log-score to evaluate forecasts. The modified log-scores are computed for the targets on the wILI percentage scale such that predictions within +/- 0.5 percentage points are considered accurate, i.e. modified log score $= \log \int_{z^*-.5}^{z^*+.5} f_m(z|\mathbf{x})dz$. For the targets on the scale of epidemic weeks, predictions within +/- 1 week are considered accurate, i.e. modified log score $= \log \int_{z^*-1}^{z^*+1} f_m(z|\mathbf{x})dz$. While this modification means that the resulting score is not formally a proper scoring rule, some have suggested that improper scores derived from proper scoring rules may, with large enough sample size, have negligible differences in practice.[17] Additionally, this modified log score has the advantage of having a clear interpretation and was motivated and designed by public health officials. Hereafter, we will refer to these modified log scores as simply log scores.

Average log scores can be used to compare models' performance in forecasting for different locations, seasons, targets, or times of season. In practice, each model m has a set of log scores associated with it are region-, target-, season-, and week-specific. We represent one specific scalar log score value as $\log f_{m,r,t,s,w}(z^*|\mathbf{x})$. These values can be averaged across any of the indices to create a summary measure of performance. For example,

$$LS_{m,\cdot,t,\cdot,\cdot} = \frac{1}{N}\sum_{r,s,w} \log f_{m,r,t,s,w}(z^*|\mathbf{x}) \quad (1)$$

represents a log score for model m and target t averaged across all regions, seasons and weeks.

While log scores are not on a particularly interpretable scale, a simple transformation enhances interpretability substantially. Exponentiating an average log score yields a forecast score equivalent to the geometric mean of the probabilities assigned to the eventually observed outcome. The geometric mean is an alternative measure of central tendency to an arithmetic mean, representing the N^{th} root of a product of N numbers. Using the example above, we then have that

$$S_{m,\cdot,t,\cdot,\cdot} = \exp\left(LS_{m,\cdot,t,\cdot,\cdot}\right) = \exp\left(\frac{1}{N}\sum_{r,s,w} \log f_{m,r,t,s,w}(z^*|\mathbf{x})\right) \quad (2)$$

$$= \left(\prod_{r,s,w} f_{m,r,t,s,w}(z^*|\mathbf{x})\right)^{1/N} \quad (3)$$

In this setting, this score has the intuitive interpretation of being the average probability assigned to the true outcome (where average is considered to be a geometric average). Hereafter, we will refer to an average score as an exponentiated average log score. In all cases, we compute the averages arithmetically on the log scale and only exponentiate before reporting and interpreting a final number Therefore, all reported average scores can be interpreted as the corresponding geometric means, or as the correponding average probabilities assigned to the true outcome.

Following the convention of the CDC challenges, we only included certain weeks in the calculation of the average log scores for each target. This focuses model evaluation on periods of time that are more relevant for public health decision making. Forecasts of season onset are evaluated based on the forecasts that are received up to six weeks after the observed onset week within a given region. Peak week and peak intensity forecasts were scored for all weeks in a specific region-season up until the wILI measure drops below the regional baseline level for the

final time. (All weeks are scored if wILI never goes above the baseline.) Week-ahead forecasts are evaluated using forecasts received four weeks prior to the onset week through forecasts received three weeks after the weighted ILI goes below the regional baseline for the final time. In a region-season without an onset, all weeks are scored. To ensure all calculated summary measures would be finite, all modified log scores with values of less than -10 were assigned the value -10, following CDC scoring conventions. All scores were based on "ground truth" values of wILI data obtained as of September 27, 2017.

2.4 Specific model comparisons

Analysis of data revisions

The CDC publicly releases data on doctor's office visits due to ILI each week. These data for previous weeks (especially the most recent ones) are occasionally revised, due to new or updated data being reported to the CDC since their last report. While often these revisions are fairly minor or non-existent, at other times, these revisions can be substantial, changing the reported wILI value by over 50% of the originally reported value. Since these data are used by forecasters to generate current forecasts, these forecasts can be contaminated by the initially reported, biased data.

We used a regression model to analyze the impact of these unrevised reports on forecasting. Specifically, for each region and epidemic week we calculated the difference between the first and the last reported ILI values for each epidemic week for which forecasts were generated in the seven seasons under consideration. We then created a categorical variable (X) with a binned representation of these differences using the following six categories covering the entire range of obesrved values: (-3.5,-2.5], (-2.5,-1.5], ..., (1.5,2.5]. Using the forecasting results from the four most accurate individual non-ensemble models, (`ReichLab-KCDE`, `LANL-DBM`, `Delphi-DeltaDensity1`, `CU-EKF_SIRS`), we then fit the following linear regression model

$$S_i = \beta + \alpha_{m(i)} + \gamma_{t(i)} + \lambda_{w(i)} + \theta \cdot X_i + \epsilon_i \tag{4}$$

where the index i indexes a specific set of subscripts $\{m, r, t, s, w\}$, and the $\alpha_{m(i)}$, $\gamma_{t(i)}$, and $\lambda_{w(i)}$ are model-, target-, and week-specific fixed effects, respectively. (The notation $m(i)$ refers to the model contained in the ith observation of the dataset.) The error term is assumed to follow a Gaussian distribution with mean zero and an estimated variance parameter. The parameter of interest in the model is the vector θ, which represent the expected changes in average score from the reference category (defined as the bin representing changes of between +/- 0.5 from the original reported value) adjusting for model, target and week-of-season.

Mechanistic vs. statistical models

There is not a consensus on a single best modeling approach or method for forecasting the dynamic patterns of infectious disease outbreaks, in both endemic and emergent settings. Semantically, modelers and forecasters often use a dichotomy of mechanistic vs. statistical (or 'phenomenological') models to represent two different philosophical approaches to modeling. Mechanistic models for infectious disease consider the biological underpinnings of disease transmission, and are in practice are implemented as variations on the Susceptible-Infectious-Recovered (SIR) model. Statistical models largely ignore the biological underpinnings and theory of disease transmission and

focus instead on using data-driven, empirical and statistical approahces to make the best forecasts possible of a given dataset, or phenomenon.

However, in practice, this dichotomized distinction is less clear than it is in theory. For example, statistical models for infectious disease counts may encode an autoregressive term for incidence (i.e. as done by the `ReichLab-KCDE` model). This could be interpreted as representing a transmission process from one time period to another. In another example, the `LANL-DBM` model has an explicit SIR compartmental model component but also leverages a hierarchical discrepancy which is purely statistical. The models from Columbia University used a statistical 'now-casting' approach for their 1-week ahead forecasts, but after that relied on different variations of an SIR model.

We categorized models according to whether or not they had any explicit compartmental framework (Table 1). We then took the top three performing compartmental models (i.e. models with some kind of an underlying compartmental structure) and compared their performance with the top three individual component models without compartmental structure. We excluded multi-model ensemble models (i.e. `Delphi-Stat`) from this comparison. Separately for each target, we computed the average score for the top three compartmental models and compared this to the average score for the top three non-compartmental models.

2.5 Reproducibility and data availability

To maximize the reproducibility and data availability for this project, the data and code for the entire project (excluding specific model code) are publicly available. The project is available on GitHub[18], with a permanent repository [[stored on Zotero]]. All of the forecasts may be interactively browsed on the website http://flusightnetwork.io. A web applet with interactive visualizations of the model evaluations is available at https://ermoore.shinyapps.io/FSN_Model_Comp [[link to change]] Additionally, this manuscript was dynamically generated using R version 3.4.1 (2017-06-30), Sweave, and knitr, tools for intermingling manuscript text with R code that run the central analyses, minimizing the chances for errors transcribing or translating results.[19, 20].

3 Results

3.1 Comparing models' forecasting performance by season

Averaging across targets and locations, forecast score varied widely by model and season (Figure 2). The historical baseline model (`ReichLab-KDE`) showed an average seasonal score of 0.20, meaning that in a typical season, across all targets and locations, this model assigned on average 0.20 probability to the eventually observed value. The model with the highest average seasonal forecast score (`Delphi-Stat`) and lowest (`Delphi-EmpiricalBayes2`) had scores of 0.37 and 0.07, respectively. Of the 22 models, 16 models (73%) showed higher average seasonal forecast score than the historical average. Season-to-season variation was substantial, with 10 models having at least one season with greater average forecast score than the `Delphi-Stat` model did.

The six top-performing models utilized a range of methodologies, highlighting that very different approaches can result in very similar overall performance. The overall best model was an ensemble model (`Delphi-Stat`) that used a weighted combination of other models from the Delphi group. Both the `ReichLab-KCDE` and the `Delphi-DeltaDensity1` model utilized kernel conditional density estimation, a non-parametric statistical

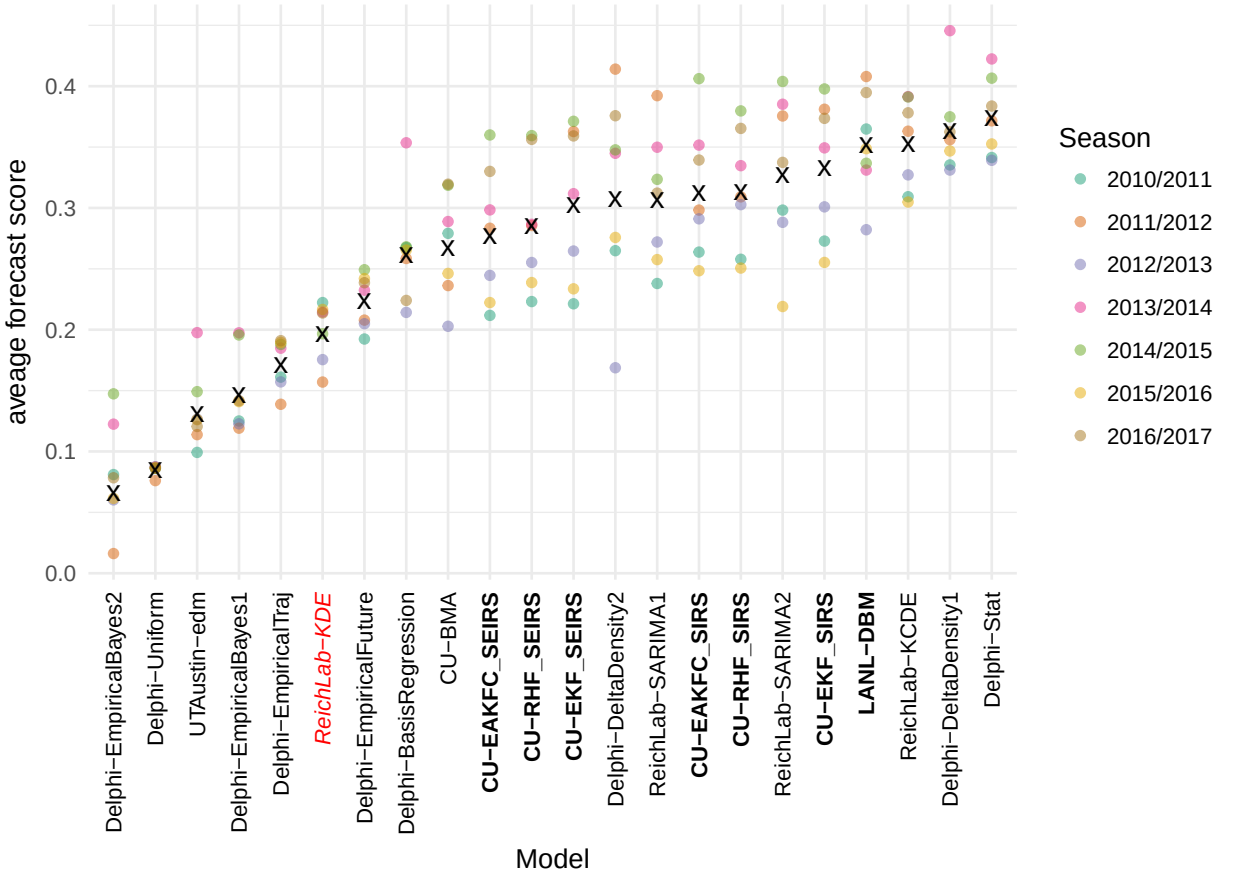

Figure 2: Average forecast score, aggregated across targets, regions, and weeks, plotted separately for each model and season. Models are sorted from lowest scores (left) to highest scores (right). Higher scores indicate better performance. Dots show average score across all targets, regions, and weeks within a given season. The x marks the geometric mean of the seven seasons. The names of compartmental models are shown in bold face. The `ReichLab-KDE` model (italicized red font) can be thought of as the historical baseline model.

methodology that is a distribution-based variation on nearest-neighbors regression. These models used different implementations and different input variables, but showed similarly strong performance across all seasons. The `UTAustin-edm` and `Delphi-DeltaDensity2` models also used variants of nearest-neighbors regression, although overall score for these models was not as consistent, indicating that implementation details and/or input variables can impact the performance of this approach. The `LANL-DBM` and `CU-EKF_SIRS` models both rely on a compartmental model of influenza transmission, however the methodologies used to fit and forecast were different for these approaches. The CU model used an ensemble Kalman filter approach to generate forecasts, while the LANL model sampled from the posterior predictive distribution using Markov chain Monte Carlo (MCMC). The `ReichLab-SARIMA2` model used a classical statistical time-series model, the seasonal auto-regressive integrated moving average model (SARIMA), to fit and generate forecasts. Interestingly, several pairs of models, although having strongly contrasting methodological approaches, showed similar overall performance; e.g., `CU-EKF_SIRS` and `ReichLab-SARIMA2`, `LANL-DBM` and `ReichLab-KCDE`.

3.2 Performance in forecasting week-ahead incidence

Average forecast score for all four week-ahead targets varied substantially across models and regions (Figure 3). The model with the highest average score for the week-ahead targets across all regions and seasons was `CU-EKF_SIRS`. This model achieved a region-specific average forecast score for week-ahead targets between 0.32 and 0.55. As a comparison, the historical baseline model achieved between 0.12 and 0.37 average score for all week-ahead targets.

Even within given models, week-ahead forecast score showed large region-to-region and year-to-year variation. The forecast score for specific region-seasons shown by the high-accuracy `CU-EKF_SIRS` model varied from 0.21 to 0.80.

Models were more consistently able to forecast week-ahead wILI in some regions than in others. Predictability for a target can be broken down into two components. First, what is the basline score that a model based on historical averages can achieve? Second, how much value do other models add beyond the historical baseline? Looking at results across all models, HHS Region 1 was the most predictable and HHS Region 6 was the least predictable.

In HHS Region 1, the 22 models showed an average forecast score of 0.42 for k-week-ahead targets (Figure 4). This means that in a typical season these models assigned an average of 0.42 probability to the eventually observed wILI percentages. HHS Region 1 showed the best overall week-ahead predictability of any region. This resulted from having the highest score from the baseline baseline model and having the largest improvement upon baseline predictions from the other models (Figure 4B).

In HHS Region 6 the average week-ahead score was 0.18. While HHS Region 6 showed the lowest baseline model score of any region, it also showed the second highest improvement upon baseline predictions (Figure 4B).

Forecast score declined as the target moved further into the future relative to the last observed data. For the model with highest forecast score across all four week-ahead targets (`CU-EKF_SIRS`), the average scores across region and season for 1 through 4 week-ahead forecasts were 0.55, 0.44, 0.36, and 0.31. This mirrored an overall decline in score observed across most models. Only in HHS Region 1 were the forecast scores from the `CU-EKF_SIRS` model for both the "nowcast" targets (1 and 2 weeks ahead) above 0.5. The historical baseline model showed average forecast score of 0.26, for all week-ahead targets. (Performance does not decline at longer horizons for

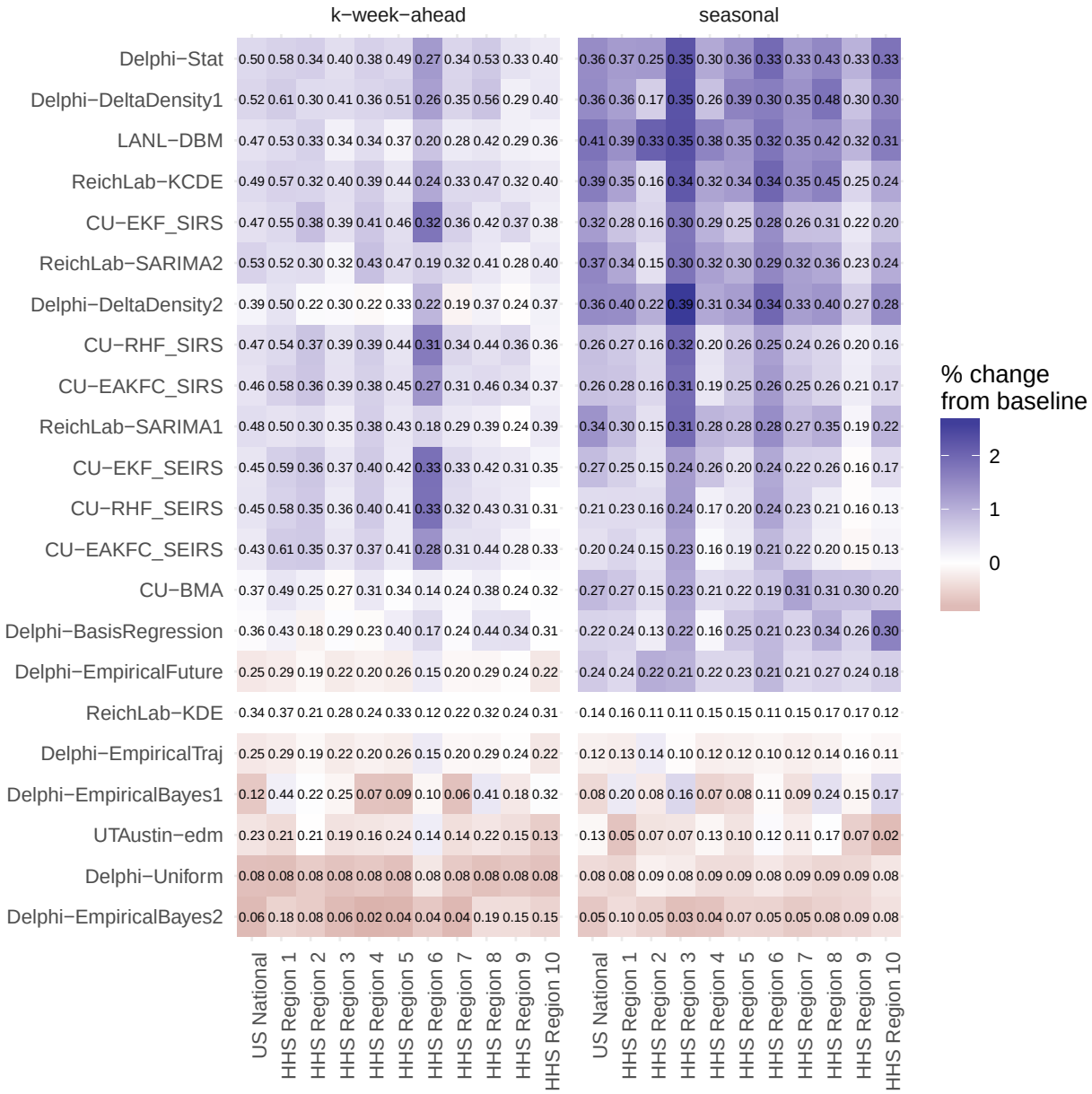

Figure 3: Average forecast score by model region and target-type, averaged over weeks and seasons. The text within the grid shows the score itself. The white midpoint of the color scale is set to be the target-specific average of the historical baseline model, `ReichLab-KDE`, with darker blue colors representing models that have better scores than the baseline and darker red scores representing models that have worse scores than the baseline.

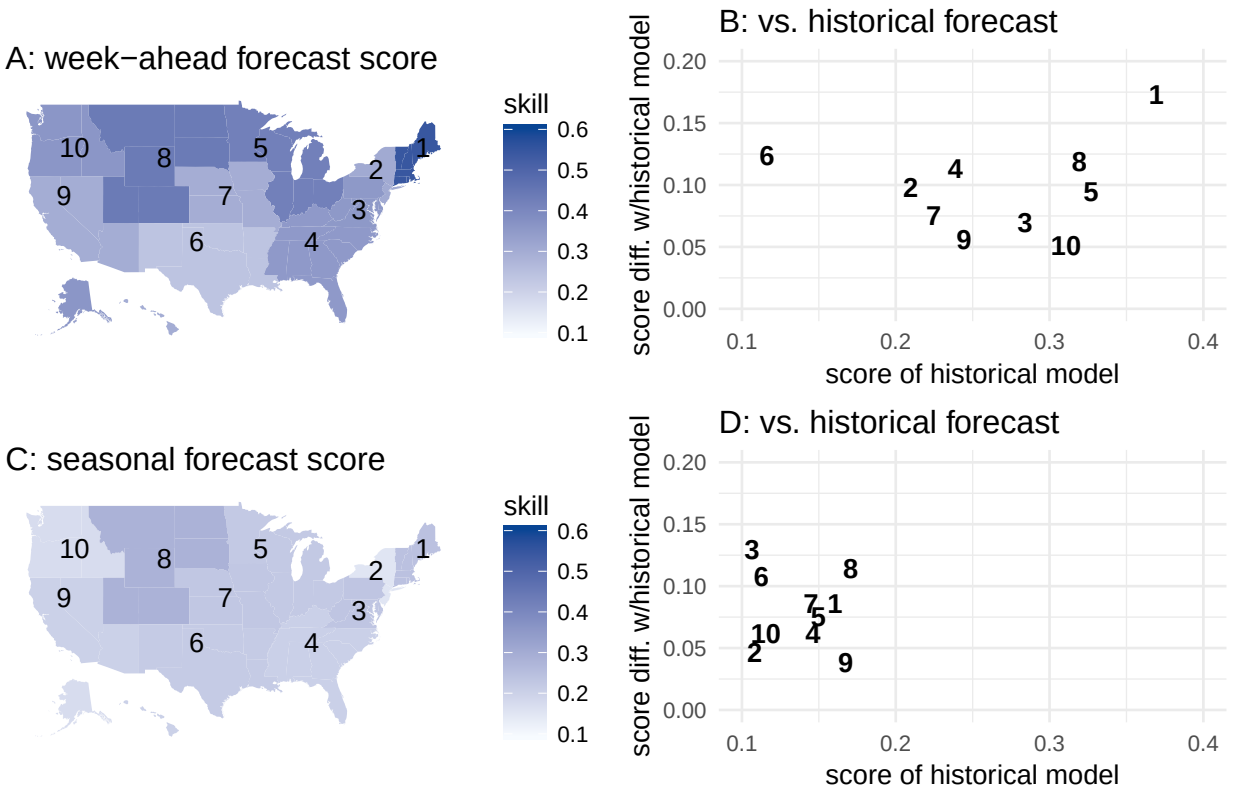

Figure 4: Absolute and relative forecast performance for week-ahead and seasonal targets, summarized for all models. Panels A & C show maps of the U.S. that illustrate spatial patterns of average forecast accuracy for all models for week-ahead (A) and seasonal (C) targets. Color shading indicates average forecast score for all models. Panels B & D compare historical model score (x-axis) with the average score of all models (y-axis) with one point for each region. For example, a y-value of 0.1 indicates that in the models on average assigned 10% more probability to the eventually observed value than the historical baseline model. The digits in the plot refer to the corresponding HHS Region number.

the historical model, since its forecasts are always the same for a given week.) For 1 week-ahead forecasts, 17 of 22 models (77%) showed higher scores than a historical baseline. For the 4 week-ahead forecasts, only 12 of 22 models (55%) showed higher scores than the historical baseline.

3.3 Performance in forecasting seasonal targets

Overall, forecast score was lower for seasonal targets than for week-ahead targets, although the models showed greater relative improvement compared to the baseline model (Figure 3). The historical average model achieved an overall forecast score of 0.14. The best single model across all seasonal targets was `LANL-DBM` with an overall forecast score of 0.36, more than a two-fold increase in score over the baseline.

Of the three seasonal targets, models showed the lowest average score in forecasting season onset, with an overall average score of 0.15. Due to the variable timing of season onset, different numbers of weeks were included in the final scoring for each region-season(see methods for details).Of the 77 region-seasons evaluated, 9 had no onset. The best model for onset was `LANL-DBM`, with overall average score of 0.33 and region-season-specific scores for onset that ranged from 0.03 to 0.81. The historical baseline model showed 0.11 average score in forecasting onset. Overall, 16 of 22 models (73%) had more forecast score than the historical baseline model in the scoring period of interest.

Models showed an overall average score of 0.23 in forecasting peak week. The best model for peak week was `ReichLab-KCDE`, with overall average score of 0.35. Region- and season-specific forecast score from this model for peak week ranged from 0.01 to 0.67. The historical baseline model showed 0.17 score in forecasting peak week. Overall, 16 of 22 models (73%) had more forecast score than the historical baseline model in the scoring period of interest.

Models showed the an overall average score of 0.20 in forecasting peak intensity. The best model for peak intensity was `LANL-DBM`, with overall average score of 0.38. Region- and season-specific forecast score from this model for peak intensity ranged from 0.13 to 0.61. The historical baseline model showed 0.13 score in forecasting peak intensity. Overall, 16 of 22 models (73%) had more forecast score than the historical baseline model in the scoring period of interest for peak intensity.

3.4 Comparison between statistical and compartmental models

On the whole, statistical models showed the same score as compartmental models at forecasting week-ahead targets, and slightly higher score for the seasonal targets, although the differences were small and of minimal practical significance. Using the best three overall models from each category, we computed the average forecast score for each combination of region, season, and target (Table 2). For the week-ahead forecasts, the difference in model score was slight, never greater than 0.02. For the three seasonal targets, the difference in model skill was larger, ranging from 0.01 for peak week to 0.05 for peak intensity. We note that the 1 week-ahead forecasts from the compartmental models from the CU team are driven largely by a statistical "nowcast" model that uses data from the Google Search API to create the ILI+ metric.[21] Therefore, the only compartmental model making 1 week-ahead forecasts is the LANL-DBM model.

target	stat. model score	compartmental model score	difference
1 wk ahead	0.49	0.51	-0.02
2 wk ahead	0.40	0.41	-0.01
3 wk ahead	0.35	0.34	0.00
4 wk ahead	0.32	0.30	0.02
Season onset	0.23	0.22	0.01
Season peak percentage	0.32	0.27	0.05
Season peak week	0.34	0.32	0.02

Table 2: Comparison of the top three statistical models (`Delphi-DeltaDensity1`, `ReichLab-KCDE`, `ReichLab-SARIMA2`) and the top three compartmental models, (`LANL-DBM`, `CU-EKF_SIRS`, `CU-RHF_SIRS`) based on best average region-season forecast score The difference column represents the difference in the average probability assigned to the eventual outcome for the target in each row. Positive values indicate the top statistical models showed more average score than the compartmental models.

3.5 Delayed case reporting impacts forecast score

In the seven years examined in this study, wILI percentages were often revised after first being reported. For example, 22% of all weekly reported wILI percentages ended up being over 20% different than the initally reported value. The frequency and magnitude of revisions varies substantially by region.

When the first report of the wILI measurement for a given region-week was not accurate (due to incomplete or delayed reporting), we observed a corresponding strong negative impact on forecast accuracy. We found that larger biases in the initially reported data were strongly associated with a decrease in the forecast score for the forecasts made using the incomplete data. Specifically, among the four top-performing models we observed an expected change in forecast score of -0.29 when the first observed wILI measurement is between 2.5 and 3.5 percentage points lower than the final observed value, adjusting for model, week-of-year, and target (Figure 5). These results are based on results from four top-performing models: `ReichLab-KCDE`, `LANL-DBM`, `Delphi-DeltaDensity1`, and `CU-EKF_SIRS`. This pattern is symmetric for under- and over-reported values, although there are more extreme under-reported values than there are over-reported values.

4 Discussion

This work presents the first large-scale comparison of real-time forecasting models from different modeling teams across multiple years. With the rapid increase in infectious disease forecasting efforts, it can be difficult to parse the literature and understand the relative importance of different methodological advances when there is not an agreed-upon set of standard evaluations. We have built on the foundational work of the CDC efforts to establish and evaluate models against a set of shared benchmarks which other models can use as comparisons. Our collaborative, team-science approach highlights the ability of multiple research groups working together to expose patterns and trends of model performance that are harder to observe in single-team studies.

Seasonal influenza in the US, given the relative accessibility of historical surveillance data and recent history of coordinated forecasting 'challenges', is an important testbed system for understanding the current state of the art of infectious disease forecasting models. Using models from some of the most experienced influenza forecasting teams in the country, we have obtained several key observations forecasting seasonal influenza in the US.

- On the whole, models improve substantially on baseline forecasts based on historical averages, both in

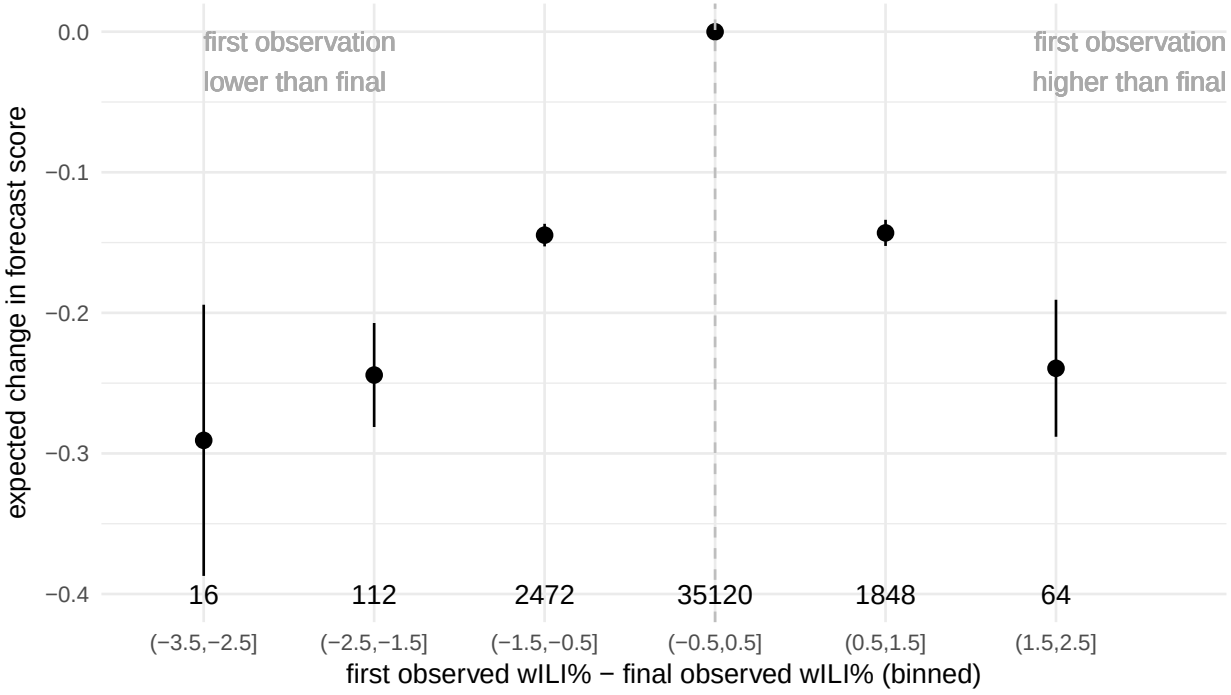

Figure 5: Model-estimated changes in forecast skill due to bias in initial reports of wILI %. The figure shows estimated coefficient values (and 95% confidence intervals) from a multivariable linear regression using model, week-of-year, target, and a categorized version of the bias in the first reported wILI % to predict forecast skill (Section 2.4). The x-axis labels show the range of bias (e.g. "(-0.5,0.5]" represents all observations whose first observations were within +/- 0.5 percentage points of the final reported value). Values to the left of the dashed grey line are observations whose first reported value were lower than the final. Y-axis values of less than zero (the reference category) represent decreases in expected forecast skill. The total number of observations in each category are shown above the x-axis labels.

regions that have more and less consistent seasonal trends (Figure 3B, D);

- At the presented spatial and temporal resolutions for influenza forecasts, we do not a meaningful difference between models that rely on an underlying mechanistic (i.e. compartmental) model and those that are more statistical in nature (Section 3.4);
- A major impediment to improved forecasting is the reporting biases in initially reported real-time data (Section 3.5).

As knowledge and data about a given infectious disease system improve and become more granular, a common expectation among domain-area experts is that mechanistic models will outperform more statistical approaches. However, the statistical vs. mechanistic model distinction is not always a clean distinction in practice. In the case of influenza, mechanistic models are models for a disease transmission process, while the data-generating mechanism for wILI captures much more than just disease transmission (e.g., clinical visitation process, symptomatic diagnosis process, reporting process, a backfill process, etc...). That is, a disease transmission model and the wILI data generating model are fundamentally different, suggesting a limitation to purely mechanistic models.

There are several important limitations to this work as presented. While we have presented a range of models from experienced influenza forecasting teams, there are large gaps in the types of data and models represented in our library of models. For example, relatively few additional data sources have been incorporated into these models, no models that explicitly incorporate information about circulating strains of influenza, and no model explicitly includes spatial relationships between regions. Additionally, while seven seasons of data and forecasts is the largest study we know of that compares models from multiple teams, this remains a smaller-than-desired sample size about model performance. Since each season represents a set of highly correlated dynamics across regions, this 'N=7' is not a lot of data from which to draw strong conclusions about comparative model performance. From the perspective of model evaluation, there is no external benchmark defined by the CDC (or others) as to what constitutes a 'good' or 'useful' forecast. While relative comparisons areuseful, it could be beneficial to have public health officials declare a given threshold, a forecast score of, for example, 0.7 or better as 'useful'.

Public health officials are still learning how to best integrate forecasts into real-time decision making. Close collaboration between public health policy-makers and quantitative modelers is necessary to ensure the forecasts have maximum impact and are appropriately communicated to the public and the broader public health community. Real-time implementation and testing of forecasting methods is helpful for planning and assessing what targets should be forecasted for maximum public health impact.

References

[1] Natalie A. Molodecky, Isobel M. Blake, Kathleen M. O'Reilly, Mufti Zubair Wadood, Rana M. Safdar, Amy Wesolowski, Caroline O. Buckee, Ananda S. Bandyopadhyay, Hiromasa Okayasu, and Nicholas C. Grassly. Risk factors and short-term projections for serotype-1 poliomyelitis incidence in Pakistan: A spatiotemporal analysis. *PLOS Medicine*, 14(6):e1002323, jun 2017.

[2] Xiangjun Du, Aaron A King, Robert J Woods, and Mercedes Pascual. Evolution-informed forecasting of seasonal influenza A (H3N2). *Science translational medicine*, 9(413):eaan5325, oct 2017.

[3] Shweta Bansal, Gerardo Chowell, Lone Simonsen, Alessandro Vespignani, and Cécile Viboud. Big Data for Infectious Disease Surveillance and Modeling. *Journal of Infectious Diseases*, 214(suppl 4):S375–S379, dec 2016.

[4] M.F. Myers, D.J. Rogers, J. Cox, A. Flahault, and S.I. Hay. Forecasting disease risk for increased epidemic preparedness in public health. *Advances in Parasitology*, 47:309–330, jan 2000.

[5] World Health Organization. Anticipating Emerging Infectious Disease Epidemics. Technical report, World Health Organization, Geneva, Switzerland, 2016.

[6] Jean-Paul Chretien, David Swedlow, Irene Eckstrand, Dylan George, Michael Johansson, Robert Huffman, and Andrew Hebbeler. Advancing Epidemic Prediction and Forecasting: A New US Government Initiative. *Online Journal of Public Health Informatics*, 7(1), 2015.

[7] Marc Lipsitch, Lyn Finelli, Richard T Heffernan, Gabriel M Leung, Stephen C Redd, and 2009 H1n1 Surveillance Group. Improving the evidence base for decision making during a pandemic: the example of 2009 influenza A/H1N1. *Biosecurity and bioterrorism : biodefense strategy, practice, and science*, 9(2):89–115, jun 2011.

[8] Matthew Biggerstaff, David Alper, Mark Dredze, Spencer Fox, Isaac Chun-Hai Fung, Kyle S. Hickmann, Bryan Lewis, Roni Rosenfeld, Jeffrey Shaman, Ming-Hsiang Tsou, Paola Velardi, Alessandro Vespignani, and Lyn Finelli. Results from the centers for disease control and prevention's predict the 2013âĂŞ2014 Influenza Season Challenge. *BMC Infectious Diseases*, 16(1):357, dec 2016.

[9] Morgan E Smith, Brajendra K Singh, Michael A Irvine, Wilma A Stolk, Swaminathan Subramanian, T Déirdre Hollingsworth, and Edwin Michael. Predicting lymphatic filariasis transmission and elimination dynamics using a multi-model ensemble framework. *Epidemics*, 18:16–28, 2017.

[10] Matthew Biggerstaff, Michael Johansson, David Alper, Logan C. Brooks, Prithwish Chakraborty, David C. Farrow, Sangwon Hyun, Sasikiran Kandula, Craig McGowan, Naren Ramakrishnan, Roni Rosenfeld, Jeffrey Shaman, Rob Tibshirani, Ryan J. Tibshirani, Alessandro Vespignani, Wan Yang, Qian Zhang, and Carrie Reed. Results from the second year of a collaborative effort to forecast influenza seasons in the United States. *Epidemics*, feb 2018.

[11] Cécile Viboud, Kaiyuan Sun, Robert Gaffey, Marco Ajelli, Laura Fumanelli, Stefano Merler, Qian Zhang, Gerardo Chowell, Lone Simonsen, and Alessandro Vespignani. The RAPIDD ebola forecasting challenge: Synthesis and lessons learnt. *Epidemics*, aug 2017.

[12] PhiResearchLab. Epidemic Prediction Initiative.

[13] DELPHI. Real-time Epidemiological Data API.

[14] New Mexico Department of Health. Indicator-Based Information System for Public Health Web.

[15] Jarad Niemi. *MMWRweek: Convert Dates to MMWR Day, Week, and Year*, 2015. R package version 0.1.1.

[16] Abhinav Tushar. *pymmwr: MMWR weeks for Python*, 2018. python library version 0.2.2.

[17] Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

[18] A Tushar, NG Reich, T Yamana, D Osthus, C McGowan, EL Ray, SJ Fox, LC Brooks, and E Moore. FluSightNetwork: cdc-flusight-ensemble repository.

[19] Yihui Xie. *Dynamic Documents with R and knitr*. Chapman and Hall/CRC, Boca Raton, Florida, 2nd edition, 2015. ISBN 978-1498716963.

[20] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017.

[21] Wan Yang, Alicia Karspeck, and Jeffrey Shaman. Comparison of Filtering Methods for the Modeling and Retrospective Forecasting of Influenza Epidemics. *PLoS Computational Biology*, 10(4):e1003583, apr 2014.