



# A collaborative multiyear, multimodel assessment of seasonal influenza forecasting in the United States

Nicholas G. Reich<sup>a,1</sup>, Logan C. Brooks<sup>b</sup>, Spencer J. Fox<sup>c</sup>, Sasikiran Kandula<sup>d</sup>, Craig J. McGowan<sup>e</sup>, Evan Moore<sup>a</sup>, Dave Osthus<sup>f</sup>, Evan L. Ray<sup>g</sup>, Abhinav Tushar<sup>a</sup>, Teresa K. Yamana<sup>d</sup>, Matthew Biggerstaff<sup>e</sup>, Michael A. Johansson<sup>h</sup>, Roni Rosenfeld<sup>i</sup>, and Jeffrey Shaman<sup>d</sup>

<sup>a</sup>Department of Biostatistics and Epidemiology, University of Massachusetts-Amherst, Amherst, MA 01003; <sup>b</sup>Computer Science Department, Carnegie Mellon University, Pittsburgh, PA, 15213; <sup>c</sup>Department of Integrative Biology, University of Texas at Austin, Austin, TX 78712; <sup>d</sup>Department of Environmental Health Sciences, Columbia University, New York, NY 10032; <sup>e</sup>Influenza Division, Centers for Disease Control and Prevention, Atlanta, GA 30333; <sup>f</sup>Statistical Sciences Group, Los Alamos National Laboratory, Los Alamos, NM 87545; <sup>g</sup>Department of Mathematics and Statistics, Mount Holyoke College, South Hadley, MA 01075; <sup>h</sup>Division of Vector-Borne Diseases, Centers for Disease Control and Prevention, San Juan, PR 00920; and <sup>i</sup>Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA 15213

Edited by Sebastian Funk, London School of Hygiene & Tropical Medicine, London, United Kingdom, and accepted by Editorial Board Member Diane E. Griffin December 10, 2018 (received for review July 24, 2018)

**Influenza infects an estimated 9–35 million individuals each year in the United States and is a contributing cause for between 12,000 and 56,000 deaths annually. Seasonal outbreaks of influenza are common in temperate regions of the world, with highest incidence typically occurring in colder and drier months of the year. Real-time forecasts of influenza transmission can inform public health response to outbreaks. We present the results of a multiinstitution collaborative effort to standardize the collection and evaluation of forecasting models for influenza in the United States for the 2010/2011 through 2016/2017 influenza seasons. For these seven seasons, we assembled weekly real-time forecasts of seven targets of public health interest from 22 different models. We compared forecast accuracy of each model relative to a historical baseline seasonal average. Across all regions of the United States, over half of the models showed consistently better performance than the historical baseline when forecasting incidence of influenza-like illness 1 wk, 2 wk, and 3 wk ahead of available data and when forecasting the timing and magnitude of the seasonal peak. In some regions, delays in data reporting were strongly and negatively associated with forecast accuracy. More timely reporting and an improved overall accessibility to novel and traditional data sources are needed to improve forecasting accuracy and its integration with real-time public health decision making.**

influenza | forecasting | statistics | infectious disease | public health

Over the past 15 y, the number of published research articles on forecasting infectious diseases has tripled (Web of Science, [www.webofknowledge.com/](http://www.webofknowledge.com/)). This increased interest has been fueled in part by the promise of “big data,” that near real-time data streams of information ranging from 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 controlled (3). With the spectre of a global pandemic looming, improving infectious disease forecasting continues to be a central priority of global health preparedness efforts (4–6).

Forecasts of infectious disease transmission can inform public health response to outbreaks. Accurate forecasts of the timing and spatial spread of infectious disease incidence can provide valuable information about where public health interventions can be targeted (7). Decisions about hospital staffing, resource allocation, the timing of public health communication campaigns, and the implementation of interventions designed to disrupt disease transmission, such as vector control measures, can be informed by forecasts. 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

(8–11). 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.

While multimodel comparisons exist in the literature for single-outbreak performance (8, 10, 11), here we compare a consistent set of models over seven influenza seasons. This is a documented comparison of multiple real-time infectious-disease forecasting models from different teams across multiple seasons. Since each season has a unique dynamical structure, multiseason comparisons like this have great potential to improve our understanding of how models perform over the long term and which models may be reliable in the future.

Influenza is a respiratory viral infection that can cause mild or severe symptoms. In the United States each year, influenza viruses infect an estimated 9–35 million individuals and cause between 12,000 and 56,000 deaths (12). Influenza incidence typically exhibits a strong annual periodicity in the United States

## Significance

**Accurate prediction of the size and timing of infectious disease outbreaks could help public health officials in planning an appropriate response. This paper compares approaches developed by five different research groups to forecast seasonal influenza outbreaks in real time in the United States. Many of the models show more accurate forecasts than a historical baseline. A major impediment to predictive ability was the real-time accuracy of available data. The field of infectious disease forecasting is in its infancy and we expect that innovation will spur improvements in forecasting in the coming years.**

Author contributions: N.G.R., L.C.B., S.J.F., S.K., C.J.M., E.M., D.O., E.L.R., A.T., T.K.Y., M.B., M.A.J., R.R., and J.S. designed research; N.G.R., L.C.B., S.J.F., S.K., C.J.M., E.M., D.O., E.L.R., A.T., and T.K.Y. performed research; N.G.R., C.J.M., and E.M. analyzed data; and N.G.R. wrote the paper.

Conflict of interest statement: J.S. and Columbia University disclose partial ownership of SK Analytics.

This article is a PNAS Direct Submission. S.F. is a guest editor invited by the Editorial Board.

This open access article is distributed under [Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 \(CC BY-NC-ND\)](https://creativecommons.org/licenses/by-nc-nd/4.0/).

Data deposition: The data and code for this analysis have been deposited in GitHub, <https://github.com/FluSightNetwork/cdc-flusight-ensemble/tree/first-papers>. Also, a permanent repository has been established at <https://doi.org/10.5281/zenodo.1255023> and a public interactive data visualization of data presented in this paper can be found at <https://reichlab.shinyapps.io/FSN-Model-Comparison/>.

See Commentary on page 2802.

<sup>1</sup>To whom correspondence should be addressed. Email: [nick@schoolph.umass.edu](mailto:nick@schoolph.umass.edu).

Published online January 15, 2019.

(and in other global regions with temperate climates), often circulating widely during colder months (i.e., November through April). The social, biological, environmental, and demographic features that contribute to higher-than-usual incidence in a particular season are not fully understood, although contributing factors may include severity of the dominant influenza subtype (13), preexisting population immunity due to prior infections or vaccination (14, 15), temperature and humidity (16), vaccine effectiveness (12), or timing of school vacations (17).

Starting in the 2013/2014 influenza season, the US Centers for Disease Control and Prevention (CDC) has run the “Forecast the Influenza Season Collaborative Challenge” (a.k.a. FluSight) each influenza season, soliciting prospective, real-time weekly forecasts of regional-level weighted influenza-like illness (wILI) measures from teams across the world (Fig. 1) (8, 10). The FluSight challenge focuses on forecasts of the weighted percentage of doctor’s office visits where the patient showed symptoms of an ILI in a particular US Health and Human Services (HHS) region. Weighting is done by state population as the data are aggregated to the regional and the national level. This wILI metric is a standard measure of seasonal flu activity, for which public data are available for the United States back to the 1997/1998 influenza season (Fig. 14). The FluSight challenge forecasts are displayed together on a website in real time and are evaluated for accuracy at the end of the season (18). This effort has galvanized a community of scientists interested in forecasting, creating a testbed for improving both the technical understanding of how different forecast models perform and the integration of these models into decision making.

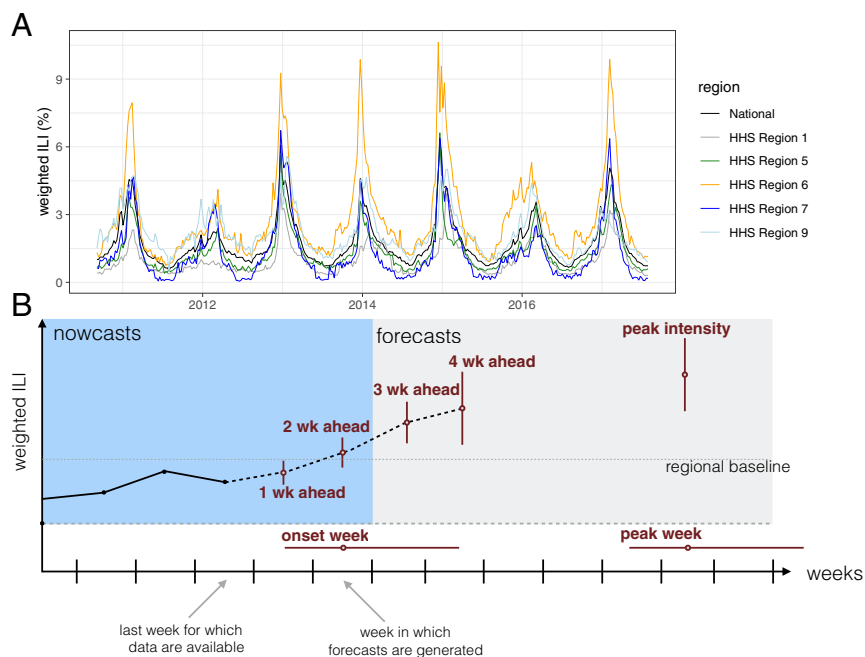
Building on the structure of the FluSight challenges [and those of other collaborative forecasting efforts (9, 11)], a subset of

FluSight participants formed a consortium in early 2017 to facilitate direct comparison and fusion of modeling approaches. Our work brings together 22 models from five different institutions: Carnegie Mellon University, Columbia University, Los Alamos National Laboratory, University of Massachusetts-Amherst, and University of Texas at Austin (Table 1). In this paper, we provide a detailed analysis of the performance of these different models in forecasting the targets defined by the CDC FluSight challenge organizers (Fig. 1*B*). 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 modeling frameworks.

In addition to analyzing comparative model performance over multiple 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 that of models that consider explicit mechanistic models of disease transmission. Overall, this work shows strong evidence that carefully crafted forecasting models for region-level influenza in the United States consistently outperformed a historical baseline model for targets of particular public health interest.

## Results

**Performance in Forecasting Week-Ahead Incidence.** Influenza forecasts have been evaluated by the CDC primarily using a variation of the log score, a measure that evaluates both the precision and accuracy of a forecast (30). Consistent with the primary evaluation performed by the CDC, we used a modified form



**Fig. 1.** (A) wILI data downloaded from the CDC website from selected regions. The y axis shows the weighted percentage of doctor’s office visits in which a patient presents with ILI for each week from 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 (circle) and an 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 outcome 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 (19). 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. These data include the first reported observations of wILI% from 2 wk prior. Therefore, 1- and 2-wk-ahead forecasts are referred to as nowcasts, i.e., at or before the current time. Similarly, 3- and 4-wk-ahead forecasts are forecasts or estimates about events in the future.

**Table 1. List of models, with key characteristics**

Team	Model abbreviation	Model description	Ref.	Ext. data	Mech. model	Ens. model
CU	EAKFC_SEIRS	Ensemble adjustment Kalman filter SEIRS	(20)	x	x	
	EAKFC_SIRS	Ensemble adjustment Kalman filter SIRS	(20)	x	x	
	EKF_SEIRS	Ensemble Kalman filter SEIRS	(21)	x	x	
	EKF_SIRS	Ensemble Kalman filter SIRS	(21)	x	x	
	RHF_SEIRS	Rank histogram filter SEIRS	(21)	x	x	
	RHF_SIRS	Rank histogram filter SIRS	(21)	x	x	
	BMA	Bayesian model averaging	(22)			
Delphi	BasisRegression*	Basis regression, epiforecast defaults	(23)			
	DeltaDensity1*	Delta density, epiforecast defaults	(24)			
	EmpiricalBayes1*	Empirical Bayes, conditioning on past 4 wk	(23, 25)			
	EmpiricalBayes2*	Empirical Bayes, epiforecast defaults	(23, 25)			
	EmpiricalFuture*	Empirical futures, epiforecast defaults	(23)			
	EmpiricalTraj*	Empirical trajectories, epiforecast defaults	(23)			
	DeltaDensity2*	Markovian Delta density, epiforecast defaults	(24)			
	Uniform*	Uniform distribution				
Stat	Ensemble, combination of 8 Delphi models	(24)				x
LANL	DBM	Dynamic Bayesian SIR model with discrepancy	(26)		x	
ReichLab	KCDE	Kernel conditional density estimation	(27)			
	KDE	Kernel density estimation and penalized splines	(28)			
	SARIMA1	SARIMA model without seasonal differencing	(28)			
	SARIMA2	SARIMA model with seasonal differencing	(28)			
UTAustin	EDM	Empirical dynamic model or method of analogues	(29)			

Team abbreviations: CU, Columbia University; Delphi, Carnegie Mellon; LANL, Los Alamos National Laboratories; ReichLab, University of Massachusetts-Amherst; SEIRS, Suceptible-Exposed-Infectious-Recovered-Susceptible, and SIRS, Suceptible-Infectious-Recovered-Susceptible, compartmental models of infectious disease transmission; UTAustin, University of Texas at Austin. The "Ext. data" column notes models that use data external to the ILINet data from CDC. The "Mech. model" column notes models that rely to some extent on a mechanistic or compartmental model formulation. The "Ens. model" column notes models that are ensemble models.

\*Note that some of these components were not designed as standalone models, so their performance may not reflect the full potential of the method's accuracy (*Materials and Methods*).

of the log score to evaluate forecasts (*Materials and Methods*). The reported scores are aggregated into an average on the log scale and then exponentiated so the reported scores can be interpreted as the (geometric) average probability assigned to the eventually observed value of each target by a particular model. Therefore, higher scores reflect more accurate forecasts. As a common reference point, we compare all models to a historical baseline model, ReichLab-KDE (Table 1), which forecasts the same historical average every week within a season and does not update based on recently observed data.

Average scores for all of the short-term forecasts (1- through 4-wk-ahead targets) varied substantially across models and regions (Fig. 2). The model with the highest average score for the week-ahead targets across all regions and seasons was CU-EKF\_SIRS (Table 1). 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 ReichLab-KDE achieved between 0.12 and 0.37 average scores for all week-ahead targets.

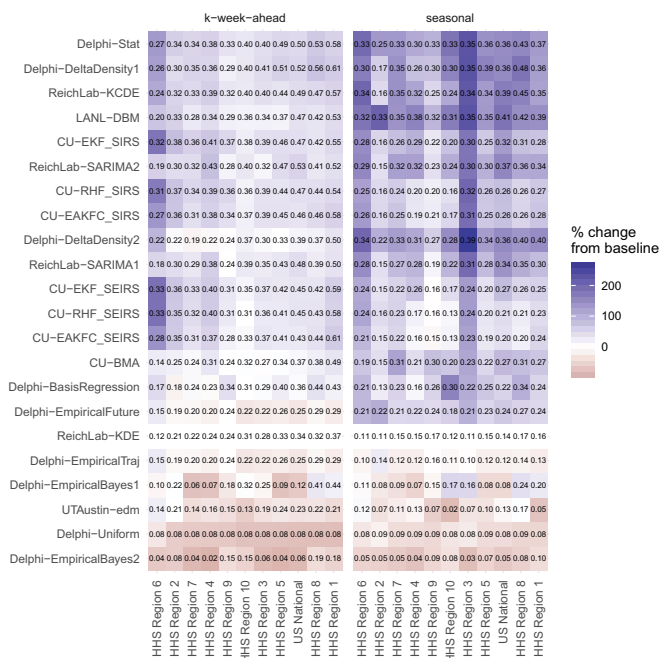
Models were more consistently able to forecast week-ahead wILI in some regions than in others. Predictability for a target can be broken down into two components. First, What is the baseline score that a model derived solely from historical averages can achieve? Second, by using alternate modeling approaches, How much more accuracy can be achieved beyond this historical baseline? Looking at results across all models, HHS region 1 was the most predictable and HHS region 6 was the least predictable (Fig. 2).

The models presented show substantial improvements in accuracy compared with forecasts from the historical baseline model in all regions of the United States. Results that follow are based on summaries from those models that on average showed higher forecast score than the historical baseline model. HHS region 1

showed the best overall week-ahead predictability of any region. Here, the models showed an average forecast score of 0.54 for week-ahead targets (Fig. 3A). This means that in a typical season these models assigned an average of 0.54 probability to the accurate wILI percentages. This resulted from having the highest score from the baseline model (0.37) and having the largest improvement upon baseline predictions (0.17) from the other models (Fig. 3B). In HHS region 6 the average week-ahead score was 0.24. While HHS region 6 showed the lowest baseline model score of any region (0.12), it also showed the second-highest improvement (0.12) upon baseline predictions (Fig. 3B).

Forecast score declined as the target moved farther into the future relative to the most recent observation. Over half of the models outperformed the historical baseline model in making 1-wk-ahead forecasts, as 15 of 22 models outperformed the historical baseline in at least six of the seven seasons. However, only 7 of 22 models outperformed the historical baseline in at least six seasons when making 4-wk-ahead forecasts. For the model with highest forecast score across all 4-wk-ahead targets (CU-EKF\_SIRS), the average scores across regions and seasons for 1- through 4-wk-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 wk ahead) above 0.5.

**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 with the baseline model (Fig. 2). The historical baseline model achieved an overall forecast score of 0.14. The best single model across all seasonal targets was LANL-DBM (Table 1) with



**Fig. 2.** 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- and region-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. The models are sorted in descending order from most accurate (top) to least accurate (bottom) and regions are sorted from high scores (right) to low scores (left).

an overall forecast score of 0.36, more than a twofold 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 (*Materials and Methods*). Of the 77 region–seasons evaluated, 9 had no onset; i.e., the wILI did not remain above a fixed region-specific threshold of influenza activity for 3 or more weeks (see *Materials and Methods* for details). The best model for onset was LANL-DBM, with an 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 an average score of 0.11 in forecasting onset. Overall, 8 of 22 models (36%) had better overall score for onset in at least six of the seven seasons evaluated (Fig. 3E).

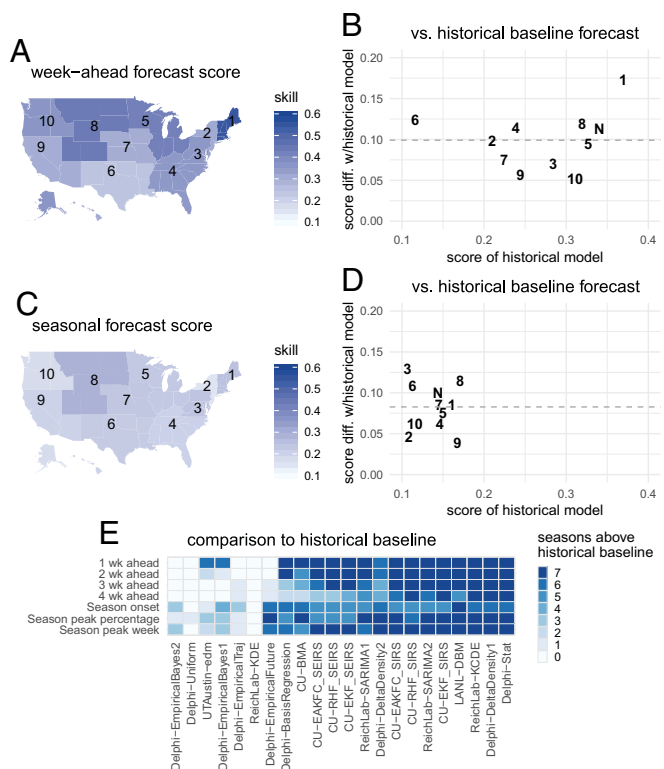
Accuracy in forecasting season onset was also impacted by revisions to wILI data. In some region–seasons current data led models to be highly confident that onset had occurred in one week, only to have revised data later in the season change the week that was considered to be the onset. One good example of this is HHS region 2 in 2015/2016. Here, data in early 2016 showed season onset to be epidemic week 2 (EW2) of 2016. Revisions to the data around EW12 led the models to identify EW51 as the onset. A further revision, occurring in EW21 of 2016, showed the onset actually occurred on EW4 of 2016. Networked metapopulation models that take advantage of observed activity in one location to inform forecasts of other locations have shown promise for improving forecasts of season onset (31).

Models showed an overall average score of 0.23 in forecasting peak week. The best model for peak week was ReichLab-KCDE (Table 1), with an overall average score of 0.35. Region- and

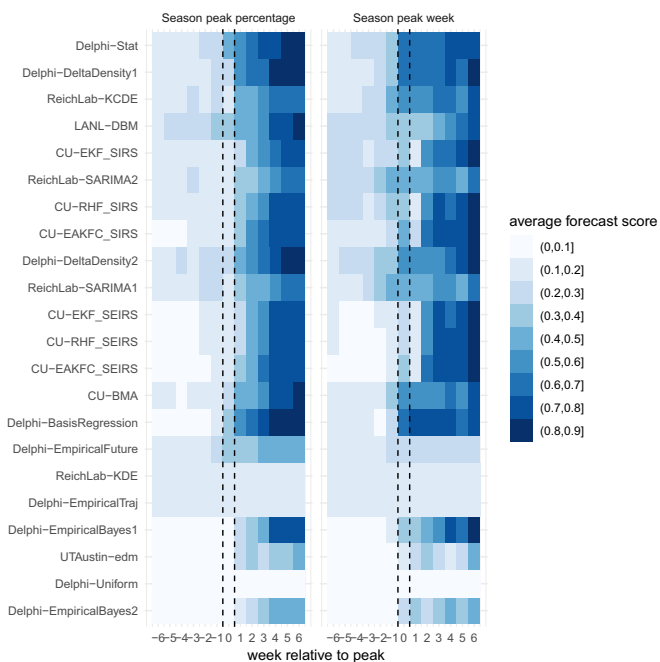
season-specific forecast scores from this model for peak week ranged from 0.01 to 0.67. The historical baseline model showed 0.17 score in forecasting peak week. Overall, 15 of 22 models (68%) had better overall score for peak week in at least six of the seven seasons evaluated (Fig. 3E).

Models showed 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 scores 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, 12 of 22 models (55%) had better overall score in at least six of the seven seasons evaluated (Fig. 3E).

While models for peak week and peak percentage converged on the observed values after the peak occurred, before the peak occurring all models showed substantial uncertainty (Fig. 4). For peak percentage, only one model (LANL-DBM) assigned on average more than 0.3 probability to within 0.5 wILI units of the eventual value (the criteria used by the CDC for evaluating model accuracy) before the peak occurring. At the peak week of the season, four models assigned on average 0.3 or more probability to the eventually observed values. In forecasting peak week, the models were able to forecast the eventual observed value with slightly more certainty earlier than for peak percentage. One week before the peak, 3 models assigned 0.3 or more



**Fig. 3.** Absolute and relative forecast performance for week-ahead (A and B) and seasonal (C and D) targets, summarized across all models that on average performed better than the historical baseline. A and C show maps of the United States that illustrate spatial patterns of average forecast accuracy for week-ahead (A) and seasonal (C) targets. Color shading indicates average forecast score for this model subset. B and D compare historical baseline model score (x axis) with the average score (y axis, horizontal dashed line at average across regions) with one point for each region. For example, a y value of 0.1 indicates that 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, with N indicating the US national region. E shows the number of seasons each model had average performance above the historical baseline.



**Fig. 4.** Average forecast score by model and week relative to peak. Scores for each location–season were aligned to summarize average performance relative to the peak week on the x axis, zero indicates the peak week and positive values represent weeks after the peak week. In general, models that were updating forecasts based on current data showed improved accuracy for peak targets once the peak had passed. Only several of the models consistently assigned probabilities greater than 0.2 to the eventually observed values before the peak week.

probability to within 1 wk of the observed peak week while at the peak week, 14 models assigned on average 0.3 or more probability to the eventually observed peak week.

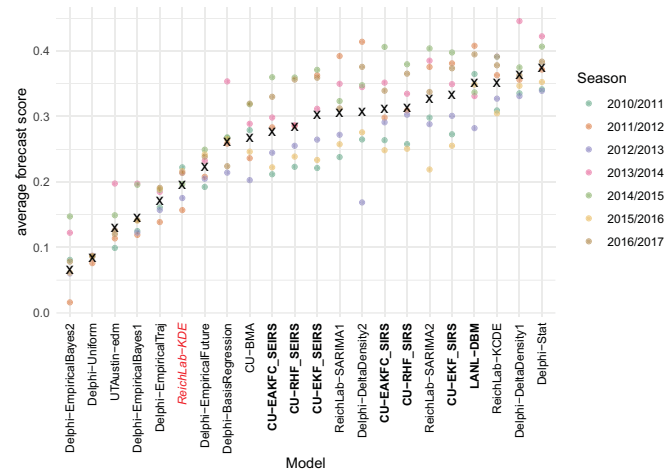
**Comparing Models' Forecasting Performance by Season.** Averaging across all targets and locations, forecast scores varied widely by model and season (Fig. 5). The historical baseline model (ReichLab-KDE) showed an average seasonal score of 0.20, meaning that in a typical season, across all targets and locations, this model assigned on average 0.20 probability to the eventually observed value. The models with the highest average seasonal forecast score (Delphi-Stat) (Table 1) and the lowest one (Delphi-EmpiricalBayes2) (Table 1) 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 used 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 (Table 1) models used kernel conditional density estimation, a non-parametric statistical 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 (Table 1) and Delphi-DeltaDensity2 models also used variants of nearest-neighbors regression, although overall scores for these models were not 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 ReichLab-SARIMA2 (Table 1) model used a classical statistical time-series model, the seasonal autoregressive integrated moving average (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.

**Comparison Between Statistical and Compartmental Models.** On the whole, statistical models achieved similar or slightly higher scores to those of compartmental models when forecasting both week-ahead and 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 all targets, except 1-wk-ahead forecasts and peak intensity, the difference in model score was slight and never greater than 0.02. For 1-wk-ahead forecasts, the statistical models had slightly higher scores on average than mechanistic models (0.06, on the probability scale). We note that the 1-wk-ahead forecasts from the compartmental models from the CU team are driven by a statistical nowcast model that uses data from the Google Search application programming interface (API) (32). Therefore, the CU models were not counted as mechanistic models for 1-wk-ahead forecasts. For peak percentage forecasts, the statistical models had slightly higher scores on average than mechanistic models (0.05).

**Delayed Case Reporting Impacts Forecast Score.** In the seven seasons examined in this study, wILI percentages were often revised after first being reported. The frequency and magnitude of revisions varied by region, and the majority of initial values (nearly 90%) are within  $\pm 0.5\%$  of the final observed value. For example, in HHS region 9, over 51% of initially reported wILI values ended up being revised by over 0.5 percentage points while in HHS region 5 less than 1% of values were revised that much.



**Fig. 5.** 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. Circles show average scores 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 type. The ReichLab-KDE model (red italics) is considered the historical baseline model.

**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) (Table 1) based on best average region–season forecast score**

Target	Score		
	Statistical model	Compartmental model	Difference
1 wk ahead	0.49	0.43	0.06
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

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 higher average score than the top compartmental models.

Across all regions, 10% of observations were ultimately revised by more than 0.5 percentage points.

When the first report of the wILI measurement for a given region–week was revised in subsequent weeks, we observed a corresponding strong negative impact on forecast accuracy. Larger revisions to the initially reported data were strongly associated with a decrease in the forecast score for the forecasts made using the initial, unrevised data. Specifically, among the four top-performing nonensemble models (ReichLab-KCDE, LANL-DBM, Delphi-DeltaDensity1, and CU-EKF\_SIRS), there was an average change in forecast score of  $-0.29$  (95% CI:  $-0.39, -0.19$ ) when the first observed wILI measurement was between 2.5 and 3.5 percentage points lower than the final observed value, adjusting for model, week of year, and target (Fig. 6; see *Materials and Methods* for details on regression model). Additionally, we observed an expected change in forecast score of  $-0.24$  (95% CI:  $-0.29, -0.19$ ) when the first observed wILI measurement was between 1.5 and 2.5 percentage points higher than the final observed value. This pattern is similar for under- and overreported values, although there were more extreme underreported values than there were overreported values. Some of the variation in region-specific performance could be attributed to the frequency and magnitude of data revisions.

## Discussion

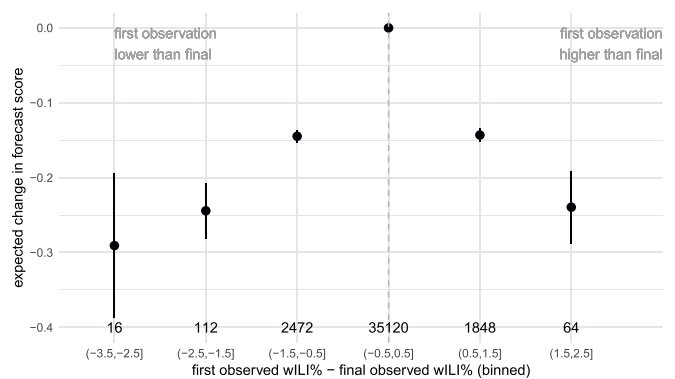
This work presents a large-scale comparison of real-time forecasting models from different modeling teams across multiple years. With the rapid increase in infectious disease forecasting efforts, it can be difficult to understand the relative importance of different methodological advances in the absence of an agreed-upon set of standard evaluations. We have built on the foundational work of CDC efforts to establish and evaluate models against a set of shared benchmarks which other models can use for comparison. Our collaborative, team science approach highlights the ability of multiple research groups working together to uncover patterns and trends of model performance that are harder to observe in single-team studies.

Seasonal influenza in the United States, 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 forecasting teams in the country, this work reveals several key results about forecasting seasonal influenza in the United States: A majority of models consistently showed higher

accuracy than historical baseline forecasts, both in regions with more predictable seasonal trends and in those with less consistent seasonal patterns (Figs. 3 B, D, and E); a majority of the presented models showed consistent improvement over the historical baseline for 1- and 2-wk-ahead forecasts, although fewer models consistently outperformed the baseline model for 3- and 4-wk-ahead forecasts (Fig. 3E); at the presented spatial and temporal resolutions for influenza forecasts, we did not identify substantial or consistent differences between high-performing models that rely on an underlying mechanistic (i.e., compartmental) model of disease transmission and those that are more statistical in nature (Table 2); and forecast accuracy is significantly degraded in some regions due to initial partially reported real-time data (Fig. 6).

As knowledge and data about a given infectious disease system improve and become more granular, a common question among domain-area experts is whether mechanistic models will outperform more statistical approaches. However, the statistical vs. mechanistic model dichotomy is not always a clean distinction in practice. In the case of influenza, mechanistic models simulate a specific disease transmission process governed by the assumed parameters and structure of the model. But observed “influenza-like illness” data are driven by many factors that have little to do with influenza transmission (e.g., clinical visitation behaviors, the symptomatic diagnosis process, the case-reporting process, a data-revision process, etc.). Since ILI data represent an impure measure of actual influenza transmission, purely mechanistic models may be at a disadvantage in comparison with more structurally flexible statistical approaches when attempting to model and forecast ILI. To counteract this potential limitation of mechanistic models in modeling noisy surveillance data, many forecasting models that have a mechanistic core also use statistical approaches that explicitly or implicitly account for unexplained discrepancies from the underlying model (20, 26).

There are several important limitations to this work as presented. While we have assembled and analyzed 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 are included that explicitly incorporate information about circulating strains



**Fig. 6. Model-estimated changes in forecast skill due to bias in initial reports of wILI %. Shown are 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 score. The x-axis labels show the range of bias [e.g., “(−0.5,0.5]” represents all observations whose first observations were within  $\pm 0.5$  percentage points of the final reported value]. Values to the left of the dashed gray line are observations whose first reported value was lower than the final value. y-axis values of less than zero (the reference category) represent decreases in expected forecast skill. The total number of observations in each category is shown above each x-axis label.**

of influenza, and no model explicitly includes spatial relationships between regions. Given that several of the models rely on similar modeling frameworks, adding a more diverse set of modeling approaches would be a valuable contribution. Additionally, while seven seasons of forecasts from 22 models is the largest study we know of that compares models from multiple teams, this remains a less-than-ideal sample size to draw strong conclusions about model performance. Since each season represents a set of highly correlated dynamics across regions, few data are available from which to draw strong conclusions about comparative model performance. Finally, these results should not be used to extrapolate hypothetical accuracy in pandemic settings, as these models were optimized specifically to forecast seasonal influenza.

What is the future of influenza forecasting in the United States and globally? While long-run forecast accuracy for influenza will vary based on a variety of factors [including, e.g., data quality, the geographical scale of forecasts, population density of forecasted areas, and consistency of weather patterns over time (33)], we expect to see continued forecast improvement through competition, collaboration, and methodological and technological innovation. Further analyses that help elucidate factors that drive forecast accuracy in specific settings will be particularly instructive. We see particular promise in models that leverage different data sources, such as pathogen-specific and highly localized incidence data. Additionally, building ensemble models that capitalize on the strengths of a diverse set of individual component models will be critical to improving accuracy and consistency of models in all infectious disease forecasting settings. Ensemble forecasting was the motivation behind the creation of the FluSight Network, although it is out of the scope of this paper.

To advance infectious disease forecasting broadly, a complete enumeration and understanding of the challenges facing the field are critical. In this work, we have identified and quantified some of these challenges, specifically focusing on timely reporting of surveillance data. However, other barriers may be of equal or greater importance to continued improvement of forecasts. Often, researchers either lack access to or do not know how best to make use of novel data streams (e.g., Internet data, electronic medical health record data). Increased methodological innovation in models that merge together an understanding of biological drivers of disease transmission (e.g., strain-specific dynamics and vaccination effectiveness) with statistical approaches to combine data hierarchically at different spatial and temporal scales will be critical to moving this field forward. From a technological perspective, additional efforts to standardize data collection, format, storage, and access will increase interoperability between groups with different modeling expertise, improve accessibility of novel data streams, and continue to provide critical benchmarks and standards for the field. Continuing to refine forecasting targets to more closely align with public health activities will improve integration of forecasts with decision making. Recent work from the CDC has developed standardized algorithms to classify the severity of influenza seasons (19), which could be used to inform the development of new forecasting targets.

Public health officials are still learning how to best integrate forecasts into real-time decision making. Close collaboration between public health policymakers and quantitative modelers is necessary to ensure that 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 play a central role in planning and assessing what targets should be forecasted for maximum public health impact.

## Materials and Methods

**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.

During each influenza season, the wILI 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, are available via the DELPHI real-time epidemiological data API beginning in the 2014/2015 season (34). 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.

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 4 wk (Fig. 1B). A season has an onset week when at least 3 consecutive weeks are above a CDC-defined regional baseline for wILI. The first of these weeks is considered to be the onset week.

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 EW of a season. Standard CDC definitions of EW are used (35–37). Each file contains binned predictive distributions for seven specific targets across the 10 HHS regions of the United States plus the national level. Each file contains over 8,000 rows and typically is about 400 kB 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 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. In total, the forecasts represent over 40 million rows and 2.5 GB of data. Each forecast file represented a single submission file, as would be submitted to the CDC challenge. Each team created 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 (Fig. 1). All teams used the Delphi epidata API to retrieve ILINet data (34). Some data sources (e.g., wILI data before 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.

**Summary of Models.** Five teams each submitted between one and nine 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 use a compartmental structure (e.g., susceptible–infectious–recovered), a model framework that explicitly 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 by incorporating recent incidence and seasonal trends. One model, Delphi-Stat, is an ensemble model, a combination of other models from the Delphi team. No team had early access to CDC surveillance data or any other data from sentinel surveillance sites. Every team accessed the data using the same API and baseline datasets. The Columbia University team used data from Google Extended Health Trends API to nowcast for the 1-wk-ahead target in all of its models. In their six mechanistic models, the nowcast was also used as an observation, i.e., as if it were CDC surveillance data. An analysis of results from the 2016/2017 season showed that the forecast quality of these models improved by about 7% (38). Additionally, the Columbia University team used daily specific humidity averaged over 24 y (1979–2002) in their six mechanistic models. These climatological estimates were calculated from the National Land Data Assimilation System (NLDAS) project-2 dataset.

Three models stand out as being reference models. One shared feature of these models is that their forecasts do not depend on observed data from the season being forecasted. 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. The Delphi-EmpiricalTraj model uses KDE for all targets. The "historical baseline" model named throughout this paper refers to the ReichLab-KDE model. Because this model represents a prediction that

essentially summarizes historical data, we consider this model an appropriate baseline model to reflect historical trends.

We note that some of the models presented here were developed as standalone forecasting models whereas others were developed as components of a larger ensemble system. We define a standalone model as one that is rigorously validated to show optimal performance on its own. Component models could also be optimized, although they could also be developed solely to provide a specific or supplemental signal as part of a larger system. All of the Delphi group's models except for Delphi-Stat were developed as components rather than standalone models. Despite this, some of the Delphi models, in particular, Delphi-DeltaDensity1, performed quite well relative to other standalone models. Component models can also provide useful baselines for comparison, e.g., the Delphi-Uniform model, which assigns uniform probability to all possible outcomes, and the Delphi-EmpiricalTraj model, which creates a seasonal average model that is not updated based on current data.

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.) Refitting of models or tuning of model parameters was explicitly discouraged to avoid unintentional overfitting of models.

**Metric Used for Evaluation and Comparison.** 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 model  $m$  for target  $Z$  conditional on some data  $\mathbf{x}$ ,  $z^*$  is the observed value of the target  $Z$ , and  $\log$  is the natural logarithm. 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 (30).

Following CDC FluSight evaluation procedures, we computed modified log scores for the targets on the wLI percentage scale such that predictions within  $\pm 0.5$  percentage points are considered accurate; i.e., modified log score =  $\log \int_{z^*-0.5}^{z^*+0.5} f_m(z|\mathbf{x}) dz$ . For the targets on the scale of EWs, predictions within  $\pm 1$  wk 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 (30). Additionally, this modified log score has the advantage of having a clear interpretation and was motivated and designed by public health officials to reflect an accuracy of practical significance. Hereafter, we 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 that 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 indexes to create a summary measure of performance. For example,

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

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 (or, more specifically for the modified log score, to regions of the distribution eventually considered accurate). 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 from Eq. 1 above, we then have that

$$\begin{aligned} S_{m,r,t,s,w} &= \exp(LS_{m,r,t,s,w}) \\ &= \exp\left(\frac{1}{N} \sum_{r,s,w} \log f_{m,r,t,s,w}(z^*|\mathbf{x})\right) \\ &= \left(\prod_{r,s,w} f_{m,r,t,s,w}(z^*|\mathbf{x})\right)^{1/N}. \end{aligned} \quad [2]$$

In this setting, this score  $S$  has the intuitive interpretation of being the average probability assigned to the true outcome (where average is considered

to be a geometric average). Throughout this paper, we refer to an exponentiated average log score as an average score. In all cases, we compute the averages arithmetically on the log scale and exponentiate only before reporting and interpreting a final number. Therefore, all reported average scores can be interpreted as the corresponding geometric means or as the corresponding average probabilities assigned to the true outcome.

Following the convention of the CDC challenges, we included only 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 6 wk 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 wLI measure drops below the regional baseline level for the final time. Week-ahead forecasts are evaluated using forecasts received 4 wk before the onset week through forecasts received 3 wk after the wLI 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 log scores with values of less than  $-10$  were assigned the value  $-10$ , following CDC scoring conventions. This rule was invoked for 2,648 scores or 0.8% of all scores that fell within the scoring period. All scores were based on "ground truth" values of wLI data obtained as of September 27, 2017.

**Specific Model Comparisons.**

**Analysis of data revisions.** The CDC publicly releases data on doctor's office visits due to ILI each week. These data, especially for the most recent weeks, are occasionally revised, due to new or updated data being reported to the CDC since their last publication. While often these revisions are fairly minor or nonexistent, at other times these revisions can be substantial, changing the reported wLI value by over 50% of the originally reported value. Since the unrevised data are used by forecasters to generate current forecasts, real-time forecasts can be biased by the initially reported, preliminary data.

We used a regression model to analyze the impact of these unrevised reports on forecasting. Specifically, for each region and EW we calculated the difference between the first and the last reported wLI values for each EW 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 observed values:  $(-3.5, -2.5]$ ,  $(-2.5, -1.5]$ ,  $\dots$ ,  $(1.5, 2.5]$ . Using the forecasting results from the four most accurate individual nonensemble models (ReichLab-KCDE, LANL-DBM, Delphi-DeltaDensity1, CU-EKF\_SIRS), we then fitted the linear regression model

$$S_i = \beta + \alpha_{m(i)} + \gamma_{t(i)} + \lambda_{w(i)} + \theta \cdot X_i + \epsilon_i, \quad [3]$$

where  $S_i$  is the score; 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  $i$ th 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  $\theta$ , which represents the average change in score based on the magnitude of the bias in the latest available wLI value, adjusted for differences based on model, target, and week of season. The  $[-0.5, +0.5]$  change category was taken as a baseline category and the corresponding  $\theta$  entry constrained to be 0, so that other  $\theta$  entries represent deviations from this baseline.

**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 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 approaches to make the best forecasts possible of a given dataset or phenomenon.

However, in practice, this dichotomy is less clear than it is in theory. For example, statistical models for infectious disease counts may have an autoregressive term for incidence (e.g., as done by the ReichLab-SARIMA1 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 uses a purely statistical model for the discrepancy of the compartmental model with observed



trends. The models from Columbia University used a statistical nowcasting approach for their 1-wk-ahead forecasts, but after that relied on different variations of a 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 that of the top three individual component models without compartmental structure. We excluded multimodel ensemble models (i.e., Delphi-Stat) from this comparison and also excluded the 1-wk-ahead forecasts of the CU models from the compartmental model category, since they were generated by a statistical nowcast. Separately for each target, we compared the average score of the top three compartmental models to the average score of the top three noncompartmental models.

**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 (39), with a permanent repository stored on Zenodo (40). All of the forecasts may be interactively browsed on the website [flusightnetwork.io](https://flusightnetwork.io) (41). A web applet with interactive visualizations of the model evaluations is available at <https://reichlab.shinyapps.io/FSN-Model-Comparison/> (42). Additionally, this paper was dynamically generated using R version 3.5.1 (2018-07-02), Sweave, and knit, which are tools for intermingling manuscript text with R code that run the central analyses and minimize the chance for errors in transcribing or translating results (43, 44).

**ACKNOWLEDGMENTS.** The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention, Defense Advanced Research Projects Agency, Defense Threat Reduction Agency, the National Institutes of Health, National Science Foundation, or Uptake Technologies.

- Molodecky NA, et al. (2017) Risk factors and short-term projections for serotype-1 poliomyelitis incidence in Pakistan: A spatiotemporal analysis. *PLoS Med* 14: e1002323.
- Du X, King AA, Woods RJ, Pascual M (2017) Evolution-informed forecasting of seasonal influenza A (H3N2). *Sci Transl Med* 9:eaan5325.
- Bansal S, Chowell G, Simonsen L, Vespignani A, Viboud C (2016) Big data for infectious disease surveillance and modeling. *J Infect Dis* 214:S375–S379.
- Myers MF, Rogers DJ, Cox J, Flahault A, Hay SI (2000) Forecasting disease risk for increased epidemic preparedness in public health. *Adv Parasitol* 47: 309–330.
- World Health Organization (2016) Anticipating emerging infectious disease epidemics (World Health Organization, Geneva). Available at <http://apps.who.int/iris/bitstream/handle/10665/252646/WHO-OHE-PED-2016.2-eng.pdf>. Accessed January 25, 2018.
- Chretien J-P, et al. (2015) Advancing epidemic prediction and forecasting: A new US government initiative. *Online J Public Health Inform* 7:e13.
- Lipsitch M, Finelli L, Heffernan RT, Leung GM, Redd SC (2009) H1N1 Surveillance group. Improving the evidence base for decision making during a pandemic: The example of 2009 influenza A/H1N1. *Biosecur Bioterror Biodefense Strategy Pract Sci* 9:89–115.
- Biggerstaff M, Velardi P, Vespignani A, Finelli L (2016) Results from the Centers for Disease Control and Prevention's predict the 2013–2014 influenza season challenge. *BMC Infect Dis* 16:357.
- Smith ME, et al. (2017) Predicting lymphatic filariasis transmission and elimination dynamics using a multi-model ensemble framework. *Epidemics*, 18:16–28.
- Biggerstaff M, et al. (2018) Results from the second year of a collaborative effort to forecast influenza seasons in the United States. *Epidemics* 24:26–33.
- Viboud C, et al. (2018) The RAPIDD Ebola forecasting challenge: Synthesis and lessons learnt. *Epidemics* 22:13–21.
- Rolfes MA, et al. Estimated influenza illnesses, medical visits, hospitalizations, and deaths averted by vaccination in the United States. Available at <https://www.cdc.gov/flu/about/disease/2015-16.htm>. Accessed May 29, 2018.
- Thompson WW, et al. (2003) Mortality associated with influenza and respiratory syncytial virus in the United States. *JAMA* 289:179–186.
- Woolthuis RG, Wallinga J, van Boven M (2017) Variation in loss of immunity shapes influenza epidemics and the impact of vaccination. *BMC Infect Dis* 17:632.
- Zarnitsyna VI, et al. (2018) Intermediate levels of vaccination coverage may minimize seasonal influenza outbreaks. *PLoS One* 13:e0199674.
- Anice CL, Mubareka S, Steel J, Peter P (2007) Influenza virus transmission is dependent on relative humidity and temperature. *PLoS Pathog* 3:e151.
- Simon C, Valleron A-J, Boelle P-Y, Antoine F, Ferguson NM (2008) Estimating the impact of school closure on influenza transmission from sentinel data. *Nature* 452:750–754.
- PhiResearchLab (2018) Epidemic prediction initiative. Available at <https://predict.phiresearchlab.org/>. Accessed January 25, 2018.
- Biggerstaff M (2018) Systematic assessment of multiple routine and near-real time indicators to classify the severity of influenza seasons and pandemics in the United States, 2003–04 through 2015–2016. *Am J Epidemiol* 187:1040–1050.
- Pei S, Shaman J (2017) Counteracting structural errors in ensemble forecast of influenza outbreaks. *Nat Commun* 8:925.
- Yang W, Karspeck A, Shaman J (2014) Comparison of filtering methods for the modeling and retrospective forecasting of influenza epidemics. *PLoS Comput Biol* 10:e1003583.
- Yamana TK, Kandula S, Shaman J (2017) Individual versus superensemble forecasts of seasonal influenza outbreaks in the United States. *PLoS Comput Biol* 13: e1005801.
- Brooks LC, Farrow DC, Hyun S, Tibshirani RJ, Rosenfeld R (2015) *Epiforecast: Tools for Forecasting Semi-Regular Seasonal Epidemic Curves and Similar Time Series*. Available at <https://github.com/cmu-delphi/epiforecast-r>. Accessed May 29, 2018.
- Brooks LC, Farrow DC, Hyun S, Tibshirani RJ, Rosenfeld R (2018) Nonmechanistic forecasts of seasonal influenza with iterative one-week-ahead distributions. *PLoS Comput Biol* 14:e1006134.
- Brooks LC (2015) Flexible modeling of epidemics with an empirical Bayes framework. *PLoS Comput Biol* 11:e1004382.
- Osthus D, Gattiker J, Reid P, Del Valle, SY (2018) Dynamic Bayesian influenza forecasting in the United States with hierarchical discrepancy. *Bayesian Anal*, in press.
- Ray EL, Sakrejda K, Lauer SA, Johansson MA, Reich NG (2017) Infectious disease prediction with kernel conditional density estimation. *Stat Med* 36:4908–4929.
- Ray EL, Reich NG (2018) Prediction of infectious disease epidemics via weighted density ensembles. *PLoS Comput Biol* 14:e1005910.
- George S, May RM (1990) Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series. *Nature* 344:734–741.
- Gneiting T, Raftery AE (2007) Strictly proper scoring rules, prediction, and estimation. *J Am Stat Assoc* 102:359–378.
- Pei S, Kandula S, Yang W, Shaman J (2018) Forecasting the spatial transmission of influenza in the United States. *Proc Natl Acad Sci USA* 115:2752–2757.
- Kandula S, Hsu D, Shaman J (2017) Subregional nowcasts of seasonal influenza using search trends. *J Med Internet Res* 19:e370.
- Dalziel BD, et al. (2018) Urbanization and humidity shape the intensity of influenza epidemics in US cities. *Science* 362:75–79.
- DELPHI (2018) Real-time epidemiological data API. Available at <https://github.com/cmu-delphi/delphi-epidata>. Accessed January 25, 2018.
- New Mexico Department of Health (2018) Indicator-based information system for public health web. Available at <https://ibis.health.state.nm.us/resource/MMWRWeekCalendar.html>. Accessed May 6, 2018.
- Niemi J (2015) MMWRweek: Convert dates to MMWR day, week, and year. R package version 0.1.1. Available at <https://cran.r-project.org/web/packages/MMWRweek/index.html>. Accessed April 6, 2018.
- Tushar A (2018) *Pymmw: MMWR Weeks for Python*. Python library version 0.2.2. Available at <https://pypi.org/project/pymmw/>. Accessed May 6, 2018.
- Kandula S, et al. (Jul 2018) Evaluation of mechanistic and statistical methods in forecasting influenza-like illness. *J R Soc Interf* 15:20180174.
- Tushar A, et al. (2018) FluSightNetwork: CDC-flusight-ensemble repository. Available at <https://github.com/FluSightNetwork/cdc-flusight-ensemble>. Accessed April 6, 2018.
- Tushar A, et al. (2018) FluSightNetwork/CDC-flusight-ensemble v1.0. Available at <https://doi.org/10.5281/zenodo.1255023>. Accessed April 6, 2018.
- Tushar A, House K, Reich NG (2018) FluSight network collaborative ensemble interactive visualization. Available at <http://flusightnetwork.io/>. Accessed May 6, 2018.
- Moore E, Reich NG (2018) FluSight network model comparison. Available at <https://reichlab.shinyapps.io/FSN-Model-Comparison/>. Accessed May 6, 2018.
- Xie Y (2015) *Dynamic Documents with R and Knitr* (Chapman and Hall/CRC, Boca Raton, FL), 2nd Ed. Available at <https://yihui.name/knitr/>. Accessed April 6, 2018.
- R Core Team. (2017) R: A language and environment for statistical computing, Version 3.5.1. Available at <https://www.R-project.org/>. Accessed December 7, 2018.