

# Comparing Mechanistic and Statistical Models to Forecast Influenza in the U.S.

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

March 1, 2018

## Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Methods</b>	<b>2</b>
2.1	FluSight Challenge Overview	2
2.2	Description of this Forecasting Experiment	4
2.3	Summary of Models	4
2.4	Metrics Used for Evaluation and Comparison	4
2.5	Formal comparisons of model performance	6
<b>3</b>	<b>Results</b>	<b>6</b>
3.1	Summary of forecast skill by season	6
3.2	Performance in forecasting week-ahead incidence	8
3.3	Performance in forecasting seasonal targets	8
3.4	Performance of models by location	9
3.5	Comparison between statistical and compartmental models	9
3.6	Where do these models fail?	11
<b>4</b>	<b>Discussion</b>	<b>11</b>
4.1	Overview of key results and importance	11
4.2	Overview of statistical vs. mechanistic model comparison	11
4.3	Limitations	11

## 1 Introduction

In recent years, the quantity of research on forecasting infectious diseases has increased XX fold. This increased interest has been fueled in part by the promise of 'big data', that near real-time data streams of everything from

large-scale population behavior to microscopic patterns in disease transmission could lead to measurable improvements in how disease transmission is detected and prevented. With the spectre of a serious pandemic looming over global public health preparedness efforts and national security planning, efforts to improve infectious disease forecasting efforts continue to be a central concern of global health, national security, and broader geopolitical stability.

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

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. Understanding what targets should be forecasted for maximum public health impact is hard to assess without real-time implementation and testing.

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. These forecasts are displayed together on a website during the season and are evaluated for accuracy after the season is over.[?] 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 but also how to integrate these models into decision-making.

Building on the structure of the FluSight challenges (and those of other collaborative forecasting efforts[?]), a subset of participants founded a consortium 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 5 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 modeling approaches that using different data sources. Additionally, it allows us to identify gaps and continued challenges that should be addressed in future modeling efforts.

## 2 Methods

### 2.1 FluSight Challenge Overview

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

The FluSight challenge has been focused on forecasts of the weighted percentage of doctor's office visits for influenza-like-illness (wILI) in a particular region. This is a standard measure of seasonal flu activity, for which

64 public data is available back to the 1997/1998 influenza season. During each influenza season, this data is  
 65 updated each week by the CDC (Figure 1). When the most recent data is released, the prior weeks' reported  
 66 wILI data may also be revised. The unrevised data, available at a particular moment in time, is available via the  
 67 DELPHI real-time epidemiological data API beginning in the 2013/2014 season.[?] This API enables researchers  
 68 to "turn back the clock" to a particular moment in time and use the data available at that time. This enables  
 69 more accurate assessment of how models would have performed in real-time.

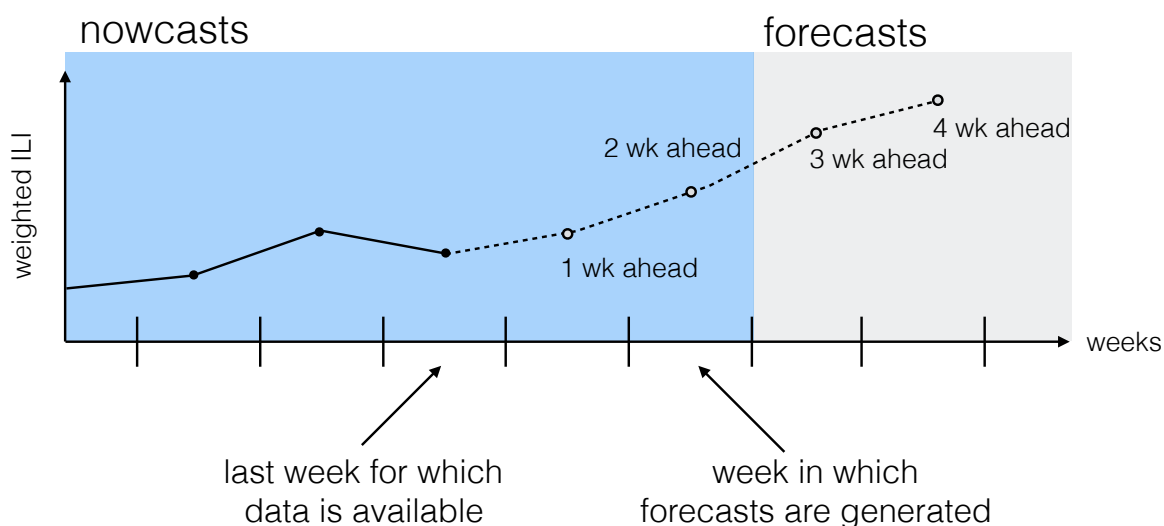


Figure 1: A schematic showing when data arrives in realtime relative to when the forecasts are made available. Based on typical timelines for the CDC FluSight challenge, we assume that forecasts are generated or submitted to the CDC using the most recent reported data. This data includes the first reported observations of wILI% from two weeks prior. Therefore, 1 and 2 week-ahead forecasts are considered nowcasts, i.e. at or before the current time. Similarly, 3 and 4 week-ahead forecasts are considered proper forecasts, or estimates about events in the future.

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

74 The FluSight challenges have also required that all forecast submissions and follow a particular format. A single  
 75 submission file (a comma-separated text file) contains the forecast made for a particular epidemic week (EW) of  
 76 a season. Standard CDC definitions of epidemic week are used. Each file contains binned predictive distributions

77 for seven specific targets across the 10 HHS regions of the US plus the national level. Each file contains over  
78 8000 rows and typically is about 400KB in size.

79 To be included in the model comparison presented here, previous participants in the CDC FluSight challenge were  
80 invited to provide out-of-sample forecasts for the 2010/2011 through 2016/2017 seasons. For each model, this  
81 involved creating 233 separate forecast submission files, one for each of the weeks in the seven training seasons.  
82 Each forecast file represented a single submission file, as would be submitted to the CDC challenge. Each team  
83 created their submitted forecasts in a prospective, out-of-sample fashion, i.e. fitting or training the model only  
84 on data available before the time of the forecast (see Figure 1).

## 85 2.2 Description of this Forecasting Experiment

## 86 2.3 Summary of Models

87 Five teams each submitted between 1 and 9 separate models for evaluation (Table 1). A wide range of method-  
88 ological approaches and modeling paradigms are included in the set of forecast models. For example, seven of  
89 the models utilize a compartmental structure (e.g. Susceptible-Infectious-Recovered), a model framework that  
90 directly encodes both the transmission and the susceptible-limiting dynamics of infectious disease outbreaks.  
91 Other less directly mechanistic models, use statistical approaches to model the outbreak phenomenon directly  
92 by incorporating recent incidence and seasonal trends. XX models directly incorporate external data (i.e. not  
93 just the WILI measurements from the CDC ILINet dataset), including historical humidity data and Google search  
94 data. Two models stand out as being clear naïve baseline models, that never change based on recent data. The  
95 Delphi-Uniform model always provides a forecast that assigns equal probability to all possible outcomes. The  
96 ReichLab-KDE model yields predictive distributions based entirely on data from other seasons using kernel density  
97 estimation (KDE) for seasonal targets and a generalized additive model with cyclic penalized splines for weekly  
98 incidence.

## 99 2.4 Metrics Used for Evaluation and Comparison

100 Forecasts have historically been evaluated by the CDC using two metrics, the log-score and the mean absolute  
101 error. These two metrics capture different desirable features of performance. The log-score enables evaluation of  
102 both the precision and accuracy of a forecast, using the predicted density function.[?] The absolute error provides  
103 an interpretable summary of the amount of error the point estimates had on average.[?]

104 We used a modified form of the log-score to evaluate forecasts, in line with the evaluation performed by the CDC.  
105 The log-score is defined as  $\log f(\hat{z}|\mathbf{x})$  where  $f(z|\mathbf{x})$  is a predictive density function for some target  $z$ , conditional  
106 on some data  $\mathbf{x}$  and  $\hat{z}$  is the observed value of the target  $z$ . In practice, each model  $m$  has a set of log scores  
107 associated with it are region-, target-, season-, and week- specific, notated as  $\log f_{r,t,s,w}^{(m)}(\hat{z}|\mathbf{x})$ . We evaluated  
108 model performance based on the exponentiated average log scores, which has been called “forecast skill” and is  
109 equivalent to the geometric mean of the probabilities assigned to the eventually observed outcome. For example,

Team	Model Abbr	Model Description	External Data	Comp. 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		
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)		
LANL	Stat	Statistical Ensemble (using the eight submitted components, with no backcasting or nowcasting)		
	Uniform	Uniform Distribution		
	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. \*Comp. model stands for compartmental model.

the forecast skill for model  $m$  and target  $t$  would be calculated as

$$FS_t^m = \exp \left( \frac{1}{N} \sum_{r,s,w} \log \hat{f}_{r,t,s,w}^{(m)}(\hat{z}|\mathbf{x}) \right) \quad (1)$$

$$= \left( \prod_{r,s,w} \hat{f}_{r,t,s,w}^{(m)}(\hat{z}|\mathbf{x}) \right)^{1/N} \quad (2)$$

where  $N$  is the total number of log-scores for target  $t$  and model  $m$ , across all combinations of region, season, and week. Further, within a given region-season-target combination, the weeks included in the calculation of the average forecast skill depend on when the onset and peak occur. Specifically,  $[[...]]$ . All weeks are included for the forecast skill calculations for the  $k$ -step ahead forecasts of wILI.

The log-scores are computed for the targets on the wILI percentage scale such that predictions within  $\pm 0.5$  percentage points are considered accurate, i.e.  $\log \text{ score} = \log \int_{\hat{z}-0.5}^{\hat{z}+0.5} f^{(m)}(z|\mathbf{x}) dz$ . For the targets on the scale of epidemic weeks, predictions within  $\pm 1$  week are considered accurate, i.e.  $\log \text{ score} = \log \int_{\hat{z}-1}^{\hat{z}+1} f^{(m)}(z|\mathbf{x}) dz$ .

- log-score for predictive distribution, aggregated by (model), (model x season), (model x season x location), (model x season x target-type), (model x season x target), , (model x season x week)
- MAE for point predictions

## 2.5 Formal comparisons of model performance

Model-based comparisons of forecast accuracy are hindered by the high correlation of sequential forecasts and by outlying observations. When observations assign no probability to the eventually observed outcome they have a log-score of  $-\infty$ .

## 3 Results

### 3.1 Summary of forecast skill by season

Averaging across targets and locations, forecast skill varied widely by model and season (Figure 2). The historical baseline model showed an average seasonal skill of 0.2, meaning that in an typical season, across all targets and locations, this model assigned on average 0.2 probability to the eventually observed value. The model with the highest average seasonal forecast skill (Delphi-Stat) and lowest (Delphi-EmpiricalBayes2) had skills of 0.37 and 0.07, respectively. Of the 22 models, 16 models (73%) showed higher average seasonal forecast skill than the historical average. Season-to-season variation was substantial, with 10 models having at least one season with greater forecast skill than the best model did on average.

The top six 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 methodology that is distribution-based variation on nearest-neighbors regression. These models used different implementations and

figures/fig-results-model-season.pdf

Figure 2: Model forecast season, overall and by season. Models are sorted from least skill (left) to most skill (right). Dots show average skill across all targets and regions for a given season. The x marks the average of the seven seasons. The names of compartmental models are shown in bold face. The ReichLab-KDE model can be thought of as the historical baseline model.

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 skill for these models was not as consistent, indicating that implementation and or input variables can impact the performance of this approach. The LANL-DBM and CU-EKS\_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-adjustment Kalman filter approach to generate forecasts, the LANL model used particle filtering. The ReichLab-SARIMA2 model used a classical statistical time-series model, the seasonal auto-regressive integrated moving average model, 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 skill for the four week-ahead targets varied substantially across models and regions. The best model (XX) achieved between XX and XX average skill in forecasting week-ahead targets across all seasons and regions. The historical baseline model achieved between XX and XX average skill. Of the XX model-region-seasons available for comparison to the historical baseline, XX% showed greater skill than the historical baseline for a given region-season.

Even within given models, forecast skill showed large region-to-region and year-to-year variation. The model with the lowest variation in forecast skill across region-seasons was XX, with skills ranging between XX and XX. The model with the highest variation in forecast skill across region-seasons was XX, with skills ranging between XX and XX. In general, Region XX was the easiest to forecast and XX was the hardest, with models showing an average forecast skill of XX and XX across all seasons, respectively.

Forecast skill declined as the target moved further into the future. For the model with highest forecast skill across all four week-ahead targets (XX), the average skill across region and season for 1 through 4 week-ahead forecasts were XX, XX, XX and XX. This mirrored an overall decline in skill observed across most models. The historical baseline model showed average forecast skill of XX for all week-ahead targets. (Performance does not decline for the historical model, since it always forecasts the same thing for every week, without updating based on recent data.) For 1 week-ahead forecasts, XX of 21 models showed more skill than a historical baseline. For the 4 week-ahead forecasts, only XX of 21 models showed more skill than the historical baseline. In Regions XX, XX, and XX, the average forecast skill for the “nowcast” targets (1 and 2 weeks ahead) were both above 0.5.

## 3.3 Performance in forecasting seasonal targets

Of the three seasonal targets, models showed the lowest average skill in forecasting season onset, with an overall average skill of XX. Due to the variable timing of season onset, different numbers of weeks were included in the final scoring for each region-season, varying from XX to XX weeks per region-season (see methods for details). Of the XX region-seasons evaluated, XX had no onset. The best model for onset was XX, with overall average skill of XX and minimum skill for a region-season of XX. Overall, XX of XX models had more forecast skill than the historical baseline model in the scoring period of interest.

Models showed less overall skill in forecasting the peak week and peak intensity when compared to the week-ahead



forecasts. Model-specific average skill for peak week, across region and season, was above 0.25 for XX models. This means that XX models assigned on average at least 25% probability to a week within +/- 1 week of the eventually observed peak week during the scoring period of interest. Similarly, for peak intensity, XX models assigned on average at least 25% probability to a wILI percentage within +/- 0.5% of the eventually observed value during the scoring period of interest. During the same time-periods, the historical model forecasts assigned XX% probability to the eventually observed peak week and XX% probability to the eventually observed peak intensity.

Average forecast skill showed substantial variation by region and by season. [[Forecasts of peak intensity and peak week showed lower skill in seasons that experienced higher than average peak incidence.]] [[Forecasts of peak intensity and peak week showed lower skill in regions with larger variability in peak incidence.]]

### 3.4 Performance of models by location

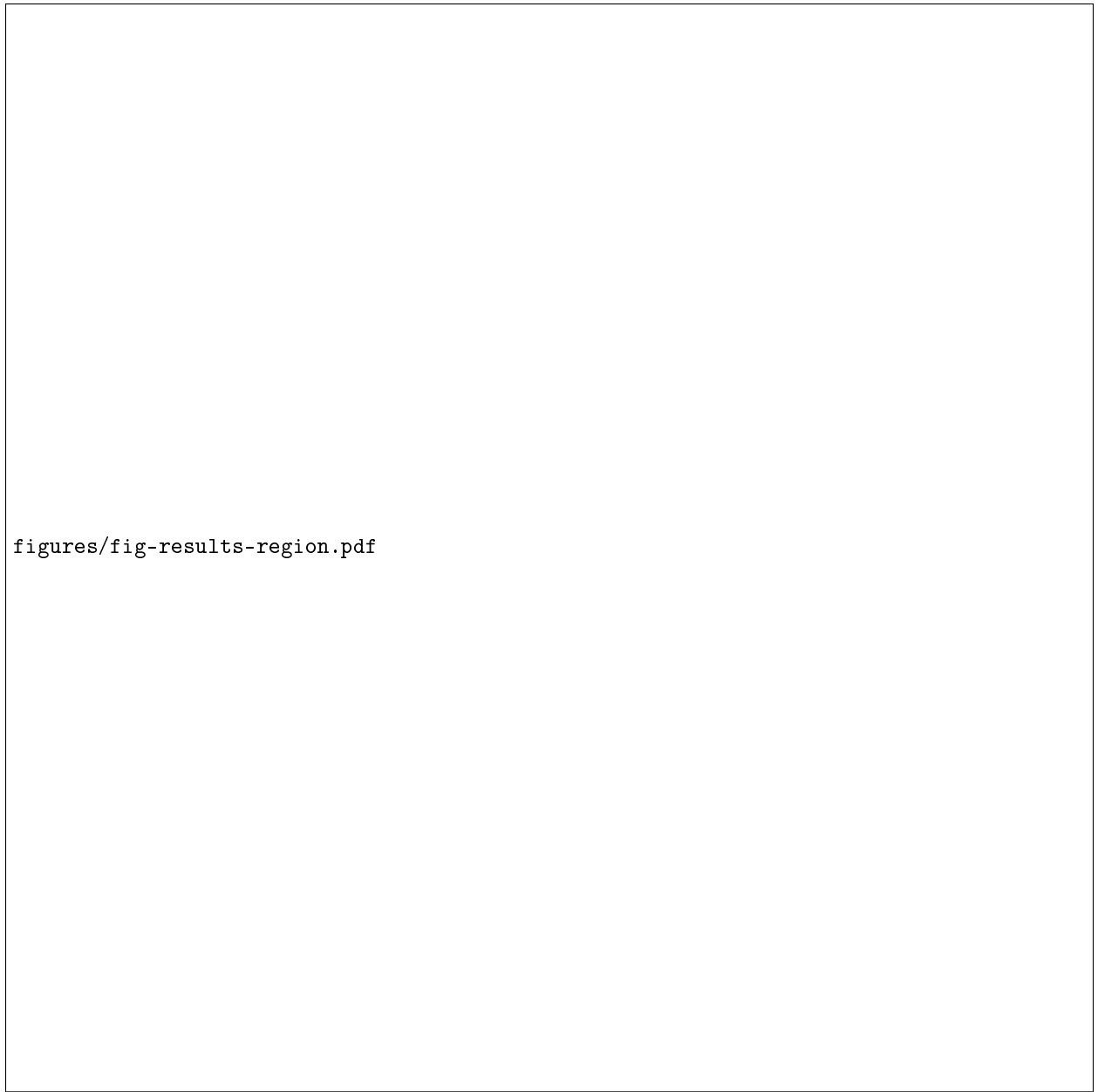
Consider adding map that averages across all models, or shows max skill per region, as it varies quite a bit. two columns: by target type mapa: average performance across all models(-KDE)/targets by region mapb: performance of historical model by region figc: scatter plot of region-level historical model skill on x and avg model skill on y, colored by season

### 3.5 Comparison between statistical and compartmental models

Statistical models showed the same amount of skill as compartmental models at forecasting week-ahead targets, and slightly more skill for the seasonal targets. Using the best three overall models from each category, we computed the average forecast skill for each combination of region, season, and target (Table 2). For the week-ahead forecasts, the difference in model skill was below XX. For the three seasonal targets, the difference in model skill was larger, ranging from XX for [target X] to XX for [target X]. 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 and influenza laboratory testing data from the CDC to create the ILI+ metric.[?] Therefore, the only truly compartmental model making 1 week-ahead forecasts is the LANL-DBM model.

target	stat. model skill	compartment model skill	difference
1 wk ahead	0.51	0.52	-0.01
2 wk ahead	0.42	0.42	-0.01
3 wk ahead	0.37	0.36	0.01
4 wk ahead	0.35	0.32	0.03
Season onset	0.41	0.38	0.04
Season peak percentage	0.38	0.28	0.11
Season peak week	0.40	0.33	0.07

Table 2: Comparison of the top three statistical models (Delphi-Stat, Delphi-DeltaDensity1, Delphi-DeltaDensity2) and the top three compartmental models, (LANL-DBM, CU-EKF\_SIRS, CU-EKF\_SEIRS) based on best average region-season forecast skill. 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 skill than the compartmental models.



figures/fig-results-region.pdf

Figure 3: Model results by region and target-type.

## 3.6 Where do these models fail?

In addition to examining where our models perform well, we also identified situations in which where current state-of-the-art forecast models still need improvement. We identify and quantify several of these challenges, including revisions to initially reported data, ....

When the first report of the wILI measurement for a given region-week is not accurate (due to incomplete or delayed reporting), this has a strong negative impact on forecast accuracy. In the seven years examined in this study, wILI percentages were often revised after first being reported. For example, 21.8% of all weekly reported wILI percentages ended up being over 20% different than the originally reported value. We found that an increase in the bias of the initially reported data was strongly associated with a decrease in the forecast skill for the forecasts made using the biased data. Specifically, among top-performing models we see an expected change in forecast skill of  $-0.2906703$ , *NA* 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 4). 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.

We anticipated seeing a relationship between the peak intensity of the season and the observed forecast skill for the peak. However, no clear relationship between the peak intensity of the season and the skill at forecasting the peak intensity was observed (data not shown).

[[Add a paragraph about capturing a holiday spike?]]

## 4 Discussion

### 4.1 Overview of key results and importance

The first large-scale comparison of flu forecasting models from different modeling teams/philosophies across multiple years.

### 4.2 Overview of statistical vs. mechanistic model comparison

As our knowledge/data about the system mature, we expect mechanistic models to be better, but when true signals of mechanistic model is drowned out by observational noise or spatial aggregation, statistical models may perform better. This comparison serves as a barometer for where the current state of forecast models are.

### 4.3 Limitations

- relatively few additional data sources incorporated
- no models that explicitly incorporate strain information
- no models with spatial information included

figures/fig-delay-model-coefs.pdf

Figure 4: Model-estimated changes in forecast skill due to bias in initial reports of wLI %. 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 wLI % to predict forecast skill. The model was fit to 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.s

- 231 • seven seasons of data is not a lot ( $n=7$ ) to draw strong conclusions about comparative model performance
- 232 • currently limited to models with only recent data...