Collaborative Multi-Season Evaluation of Annual Influenza Forecasting Models 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 22, 2018

Contents

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 large-scale population behavior [**?**] to microscopic changes in a pathogen [**?**] could lead to measurable improvements in how

disease transmission is measured, forecasted, and prevented [**?**]. With the spectre of a global pandemic looming, improving infectious disease forecasting continues to be a central priority of global health preparedness efforts.[**?**, **?**]

Forecasts of infectious disease transmission 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 helath 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.[1] 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 decsision-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[2], but we summarize the key features of the challenge here.

The FluSight challenge focuses 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 public data is available for the US back to the 1997/1998 influenza season. During each influenza season, this data is updated each week by the CDC (Figure **??**). When the most recent data is released, the prior weeks' reported wILI data may also be revised. The unrevised data, available at a particular moment in time, is available via the DELPHI

real-time epidemiological data API beginning in the 2013/2014 season.[3] This API enables researchers to "turn back the clock" to a particular moment in time and use the data available at that time. This enables more accurate assessment of how models would have performed in real-time.

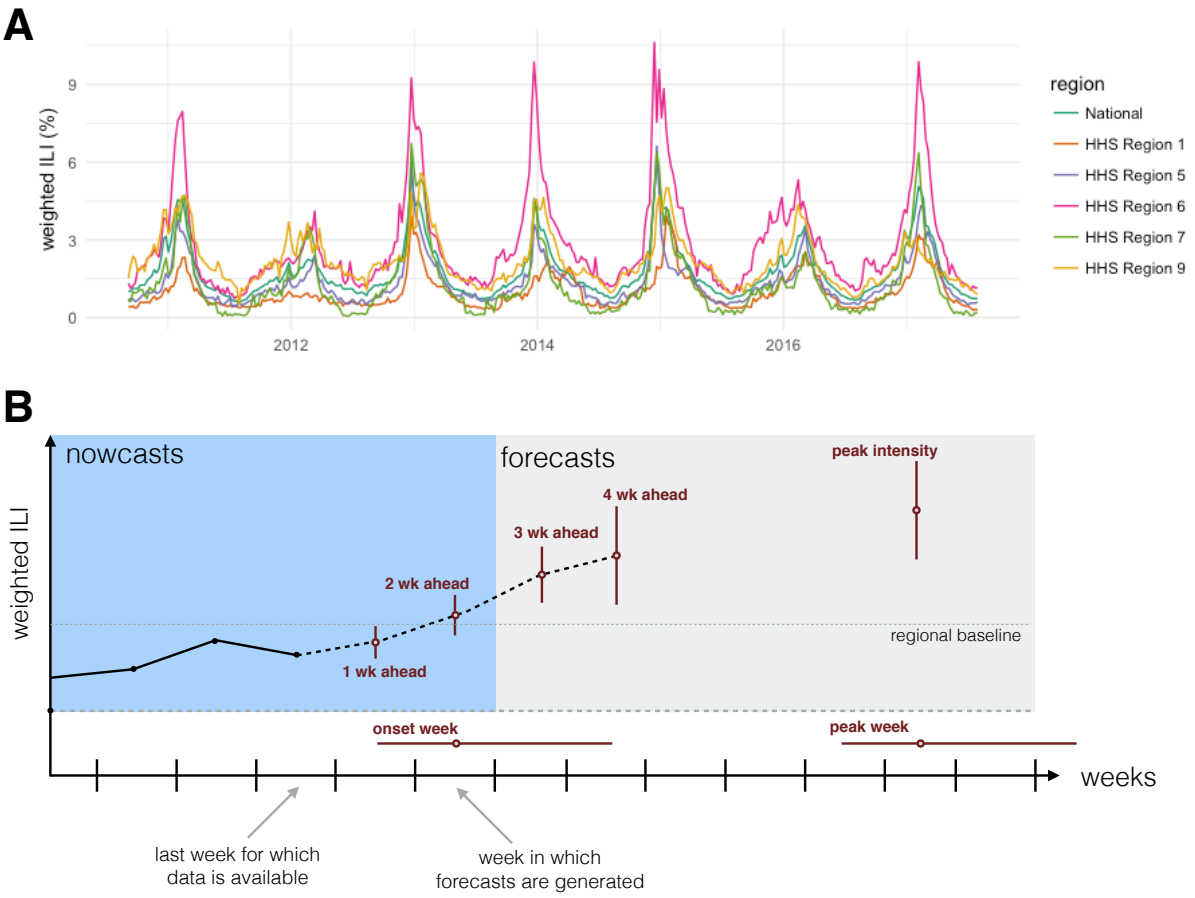

Figure 1: (A) Raw data weighted influenza-like illness data downloaded from the CDC website. The y-axis shows the weighted percentage of doctor's office visits for influenza-like illness for each week between September 2010 through July 2017, which is the time period for which the models presented in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

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

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

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

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

2.2 Summary of Models

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

2.3 Metric Used for Evaluation and Comparison

Influenza forecasts have been evaluated by the CDC primarily using the log-score, a measure that enables evaluation of both the precision and accuracy of a forecast.[4] The log-score is defined as $\log f(\hat{z}|\mathbf{x})$ where $f(z|\mathbf{x})$ is the predicted density function for some target Z, conditional on some data $\mathbf{x}$ and $\hat{z}$ is the observed value of the target Z. The log-score is a "proper" scoring rule, which has the practical implication that linear combinations (i.e. arithmetic means) of log scores will [[preserve overall rankings of forecasts]].

Consistent with the primary evaluation performed by the CDC, we used a modified form of the log-score to evaluate forecasts. The modified log-scores are computed for the targets on the wILI percentage scale such that predictions within +/- 0.5 percentage points are considered accurate, i.e. log score $= \log \int_{\hat{z}-.5}^{\hat{z}+.5} f^{(m)}(z|\mathbf{x})dz$. For the targets on the scale of epidemic weeks, predictions within +/- 1 week are considered accurate, i.e. log score $= \log \int_{\hat{z}-1}^{\hat{z}+1} f^{(m)}(z|\mathbf{x})dz$. 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

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)		
	Stat	Statistical Ensemble (using the eight submitted components, with no backcasting or nowcasting)		
	Uniform	Uniform Distribution		
LANL	DBM	Dynamic Bayesian SIR Model with a hierarchical discrepancy		x
ReichLab	KCDE	Kernel Conditional Density Estimation using recent observations and seasonality		
	KDE	Kernel Density Estimation (seasonal targets) and cyclic penalized splines (week-ahead targets)		
	SARIMA1	SARIMA model without seasonal differencing		
	SARIMA2	SARIMA model with seasonal differencing		
UTAustin	EDM	Empirical Dynamic Model, or topological method of analogues		

Table 1: List of models, with key characteristics. *Comp. model stands for compartmental model.

size, have negligible differences in practice.[4] Additionally, this modified log score has the advantage of having a clear interpretation motivated and designed by public health officials. Hereafter, we will refer to these modified log scores as simply log scores.

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

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

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

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

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

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

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

Following the convention of the CDC challenges, we only included certain weeks in the calculation of the average log scores for each target. Forecasts of season onset are evaluated based on the forecasts that are received up to six weeks after the observed onset week within a given region. Forecasts of season peak and intensity are evaluated through the first forecast received after the weighted ILI goes below the regional baseline for the final time during a given region-season. Week-ahead forecasts are evaluated using forecasts received four weeks prior to the onset week through forecasts received three weeks after the weighted ILI goes below the regional baseline for the final time. In a region-season without an onset, all weeks are scored. To ensure all calculated summary measures would be finite, all modified log scores with values of less than -10 were assigned the value -10, following CDC scoring conventions.

2.4 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 $-\infty$.

2.5 Analysis of delays

[GLM regression model with ...]

3 Results

3.1 Comparing models' forecasting performance 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 0.20, meaning that in an typical season, across all targets and locations, this model assigned on average 0.20 probability to the eventually observed value. The model with the highest average seasonal forecast skill (`Delphi-Stat`) and lowest (`Delphi-EmpiricalBayes2`) had scores of 0.37 and 0.07, respectively. Of the 22 models, 16 models (73%) showed higher average seasonal forecast skill than the historical average. Season-to-season variation was substantial, with 10 models having at least one season with greater average forecast skill than the best model did.

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 a distribution-based variation on nearest-neighbors regression. These models used different implementations and different input variables, but showed similarly strong performance across all seasons. The `UTAustin-edm` and `Delphi-DeltaDensity2` models also used variants of nearest-neighbors regression, although overall skill for these models was not as consistent, indicating that implementation details and/or input variables can impact the performance of this approach. The `LANL-DBM` and `CU-EKF_SIRS` models both rely on a compartmental model of influenza transmission, however the methodologies used to fit and forecast were different for these approaches. The CU model used an ensemble-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 all four week-ahead targets varied substantially across models and regions (Figure 3). The model with the highest average week-ahead forecast skill across all regions and seasons was `CU-EKF_SIRS`. Within regions and across all seasons, this model achieved average forecast skill between 0.32 and 0.55. As a comparison, the historical baseline model achieved between 0.12 and 0.37 average skill for all week-ahead targets.

Even within given models, week-ahead forecast skill showed large region-to-region and year-to-year variation. The forecast skill for specific region-seasons shown by the high-accuracy `CU-EKF_SIRS` model varied from 0.21 to 0.80. The model with the lowest variation in combined week-ahead forecast skill across region-seasons (excluding the uniform model) was Delphi-EmpiricalTraj, with skills ranging between 0.10 and 0.38. The model with the highest

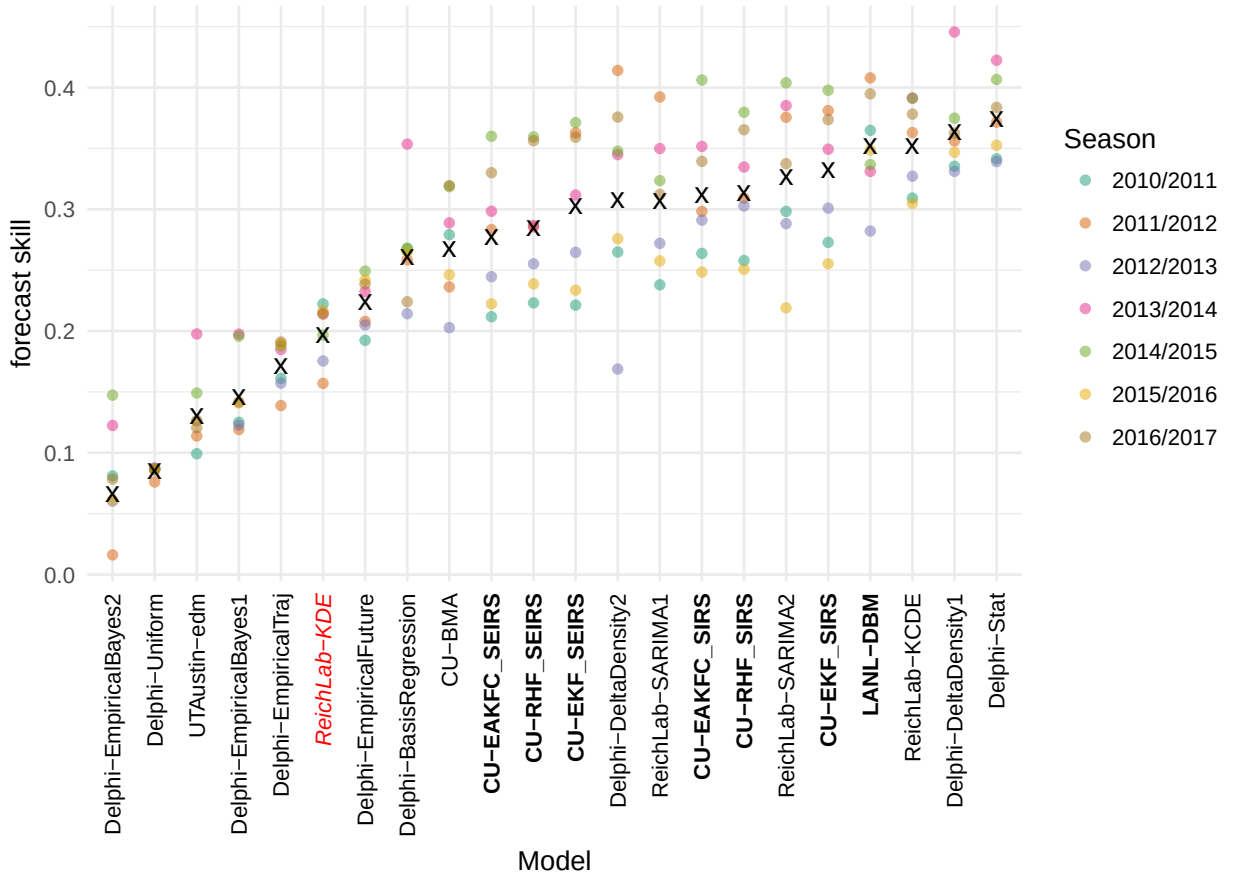

Figure 2: Average forecast skill, aggregated across targets and plotted separately for each model and season. Models are sorted from least skill (left) to most skill (right). Dots show average skill across all targets and regions for a given season. The x marks the geometric mean of the seven seasons.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

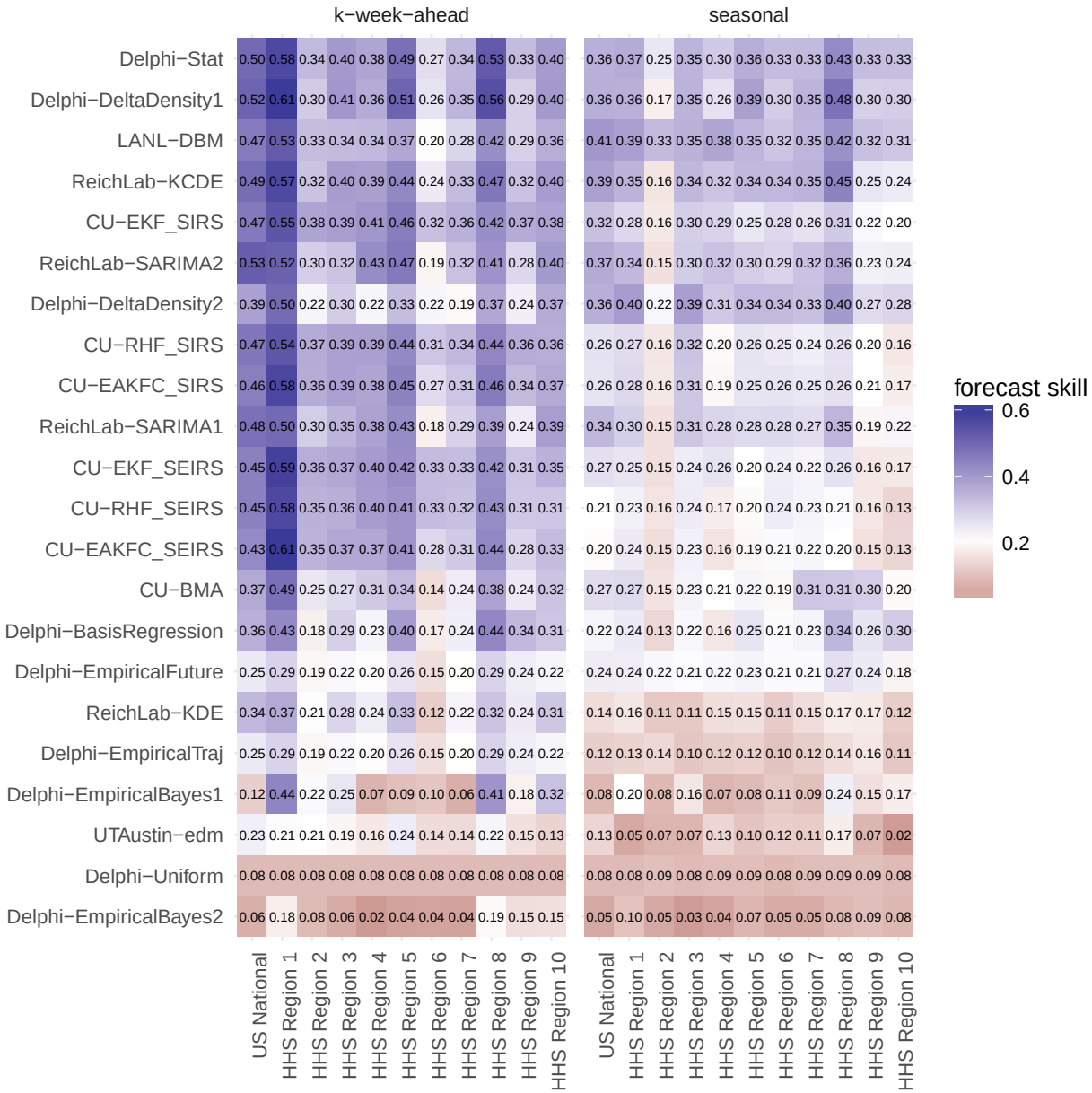

Figure 3: Average forecast skill by model region and target-type, averaged over weeks and seasons. The white midpoint of the color scale is set to be the overall average of the historical baseline model, `ReichLab-KDE`.

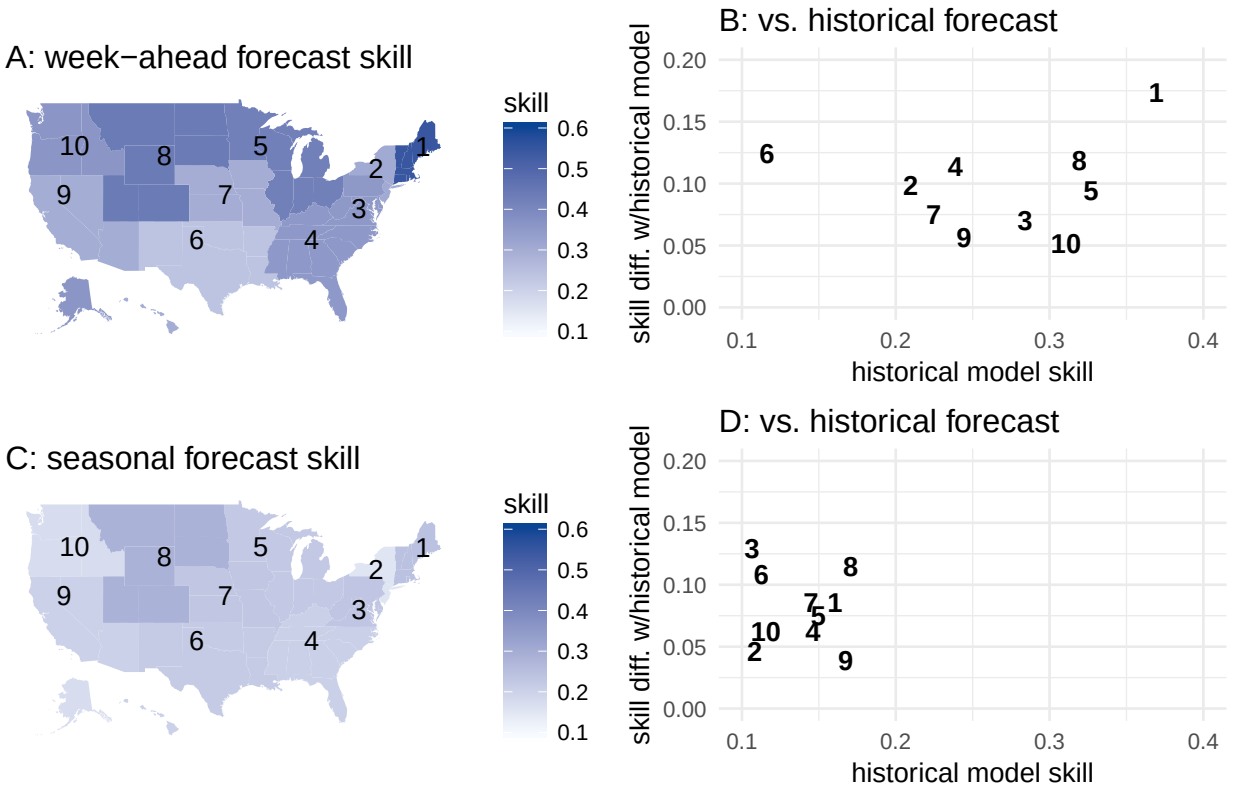

Figure 4: Absolute and relative forecast performance for week-ahead and seasonal targets, summarized for all models. Panels A & C show maps of the U.S. that illustrate spatial patterns of average forecast accuracy for all models for week-ahead (A) and seasonal (C) targets. Color shading indicates average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

variation in forecast skill across region-seasons was Delphi-EmpiricalBayes2, with skills ranging between 0.00 and 0.69.

Models were more consistently able to forecast week-ahead wILI in some regions than in others. Looking at results across all models, HHS Region 1 was the most predictable and HHS Region 6 was the least predictable. In HHS Region 1, the 22 models showed an average forecast skill of 0.42 for k-week-ahead targets (Figure 4). This means that in a typical season these models together assigned an average of 0.42 probability to the eventually observed wILI percentages. On the flip side, in HHS Region 6 the average week-ahead skill was 0.18. This was just marginally better than the week-ahead forecast skill for the historical average model in HHS Region 6, which was 0.12.

Forecast skill declined as the target moved further into the future. For the model with highest forecast skill across all four week-ahead targets (`CU-EKF_SIRS`), the average skill across region and season for 1 through 4 week-ahead forecasts were 0.55, 0.44, 0.36, and 0.31. This mirrored an overall decline in skill observed across most models. Only in HHS Region 1 were the forecast skills from the `CU-EKF_SIRS` model for both the "nowcast" targets (1 and 2 weeks ahead) above 0.5. The historical baseline model showed average forecast skill of 0.26, 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, 17 of 22 models (77%) showed more skill than a historical baseline. For the 4 week-ahead forecasts, only 12 of 22 models (55%) showed more skill than the historical baseline.

3.3 Performance in forecasting seasonal targets

Overall, forecast skill was lower for seasonal targets than for week-ahead targets (Figure **??**). While the scale of the log score can depend on the number of possible bins in the predictive distribution (i.e. more bins means less probability on average assigned to each bin), the model that assigned uniform probabilities to all possible outcomes achieved similar forecast skill for both seasonal and week-ahead targets. This shows that as a whole, the models performed worse on predicting seasonal targets relative to this uniform baseline. The best single model across all seasonal targets was `LANL-DBM` with an overall forecast skill of 0.36. The historical average model achieved an overall forecast skill of 0.14.

Of the three seasonal targets, models showed the lowest average skill in forecasting season onset, with an overall average skill 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, varying from XX to XX weeks per region-season (see methods for details). Of the 77 region-seasons evaluated, 9 had no onset. The best model for onset was `LANL-DBM`, with overall average skill of 0.33 and region-season-specific skills for onset that ranged from 0.03 to 0.81. The historical baseline model showed 0.11 average skill in forecasting onset. Overall, 16 of 22 models (73) had more forecast skill than the historical baseline model in the scoring period of interest.

Models showed the an overall average skill of 0.23 in forecasting peak week. Following CDC evaluation rules, peak week and peak intensity forecasts were scored for all weeks in a specific region-season up until the wILI measure drops below the regional baseline level for the final time. All weeks are scored if wILI never goes above the baseline. The best model for peak week was `ReichLab-KCDE`, with overall average skill of 0.35. Region- and season-specific forecast skill from this model for peak week ranged from 0.01 to 0.67. The historical baseline model showed 0.17 skill in forecasting peak week. Overall, 16 of 22 models (73%) had more forecast skill than the historical baseline model in the scoring period of interest.

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

3.4 Comparison between statistical and compartmental models

On the whole, statistical models showed the same amount of skill as compartmental models at forecasting week-ahead targets, and slightly more skill for the seasonal targets, although the differences were small. 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 slight, never greater than 3.0 percentange points of probability. For the three seasonal targets, the difference in model skill was larger, ranging from 0.03 for Season peak week to 0.07 for Season peak percentage . 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 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	0.51	0.00
2 wk ahead	0.41	0.41	-0.00
3 wk ahead	0.36	0.34	0.02
4 wk ahead	0.33	0.30	0.03
Season onset	0.26	0.22	0.04
Season peak percentage	0.34	0.27	0.07
Season peak week	0.35	0.32	0.03

Table 2: Comparison of the top three statistical models (`Delphi-Stat`, `Delphi-DeltaDensity1`, `ReichLab-KCDE`) and the top three compartmental models, (`LANL-DBM`, `CU-EKF_SIRS`, `CU-RHF_SIRS`) 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.

3.5 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, ...

Delayed case reporting impacts forecast skill

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

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

Intensity of season not reliably correlated with forecast skill

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).

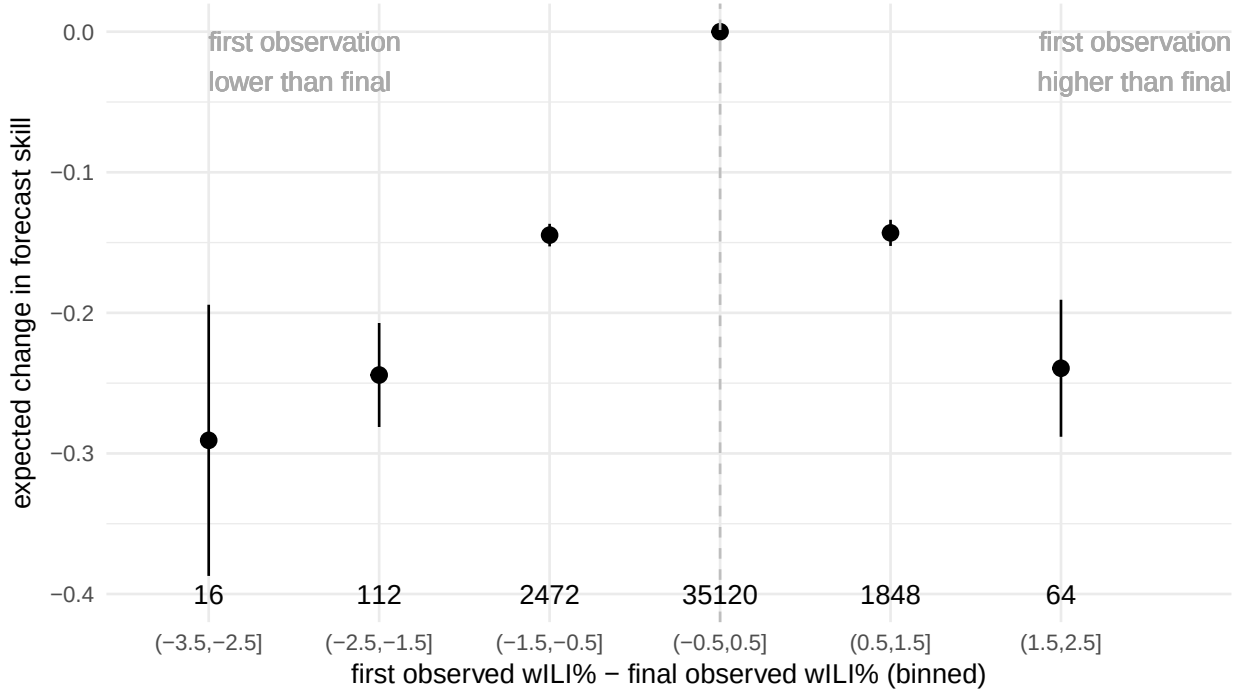

Figure 5: Model-estimated changes in forecast skill due to bias in initial reports of wILI %. The figure shows estimated coefficient values (and 95% confidence intervals) from a multivariable linear regression using model, week-of-year, target, and a categorized version of the bias in the first reported wILI % to predict forecast skill. 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.

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. Importance of this work (and the foundational work of the CDC) in establishing and evaluating against a set of shared benchmarks against which other methods/models can use as comparisons.

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
- seven seasons of data is not a lot (n=7) to draw strong conclusions about comparative model performance
- no standard methods for evaluating repeated forecasts of the same targets with the same models...
- currently limited to models with only recent data...

References

[1] PhiResearchLab. Epidemic Prediction Initiative.

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

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

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