A Collaborative Ensemble Approach to Real-Time Influenza Forecasting in the U.S.: Results from the 2017/2018 Season

(working author list subject to re-ordering pending final contributions)
Nicholas G Reich[1], Craig McGowan[2], Logan Brooks[3], Sasikiran Kandula[4],
Evan Moore[1], Dave Osthus[5], Evan Ray[6], Abhinav Tushar[1], Teresa Yamana[4],
Willow Crawford-Crudell[7], Graham Casey Gibson[1], Rebecca Silva[8]
Matthew Biggerstaff[2], Michael A Johansson[9], Roni Rosenfeld[3], Jeffrey Shaman[4]

[1]University of Massachusetts-Amherst, Amherst, USA
[2]Influenza Division, Centers for Disease Control and Prevention, Atlanta, USA
[3]Carnegie Mellon University, Pittsburgh, USA
[4]Columbia University, New York, USA
[5]Los Alamos National Laboratory, Los Alamos, USA
[6]Mount Holyoke College, South Hadley, USA
[7]Smith College, Northampton, USA
[8]Amherst College, Amherst, USA
[9]Division of Vector-Borne Diseases, Centers for Disease Control and Prevention, Atlanta, USA

Abstract

TBD.

1   Introduction

Seasonal influenza results in a substantial, annual public health burden in the United States with an estimated 30 million cases and 600,000 hospitalizations during the 2016-2017 influenza season.[1] The United States Centers for Disease Control and Prevention (CDC) utilizes a variety of surveillance methods to assess the severity of an influenza season, including monitoring outpatient visits for influenza-like illness (ILI), influenza-related hospitalizations, and virologic characteristics[2]. However, like all surveillance systems, these are constrained to describing events that have already taken place, and the total burden, along with the timing of the epidemic, can vary substantially from season to season.[1] Forecasts of an influenza season offer the possibility of providing actionable information to improve public health responses, and recent years have seen a large number of efforts in the literature describing efforts to predict season influenza.[3, 4, 5, 6, 7, 8, 9, 10]

Ensemble models, or systematic methods that bring together predictions from multiple different models, have long been seen as a valuable method for improving predictions over any single model. This "wisdom of the crowd" approach has both theoretical and practical advantages. First, it allows for an ensemble forecast to incorporate signals from different data sources and models that may highlight different features of a system. Second, combining signals from models with different biases may allow those biases to offset and result in an ensemble that is more accurate than the individual ensemble components. Weather and climate models have utilized ensemble systems for these very purposes, and recent work has extended ensemble forecasting to forecasts of infectious diseases, including influenza , dengue fever, and Ebola hemorrhagic fever.[11, 12, 13]

Since the 2013/2014 influenza season, the CDC has run an annual prospective influenza forecasting competition, known as the FluSight challenge, in collaboration with outside researchers. Participating teams submit probabalistic forecasts for various influenza-related targets of interest to public health officials weekly from early November through mid May (Figure 1). During the 2015/2016 and 2016/2017 challenges, analysts at the CDC built a simple ensemble model by taking the arithmetic mean of submitted models. This model was one of the top performing models each season (cite McGowan et al when accepted).

In March 2017, a group of influenza forecasters from different institutions who have worked with the CDC in the past, established the FluSight Network. This research consortium worked collaboratively throughout 2017 and 2018 to build and implement in real-time a multi-institution ensemble, with the goal of outperforming the "simple average" ensemble that CDC uses to inform decision making and situational awareness during the annual influenza season. Given the success in previous years of a simple ensemble that incorporated no prior information about the relative performace of ensemble components, the FluSight Network's main goal was to show that an ensemble that uses validated weights based on models' past performance could outperform this simple averaging approach. This research is an important example of a collaboration between government and academic public health experts to systematically test if information for decision making during a high severity influenza season can be improved..

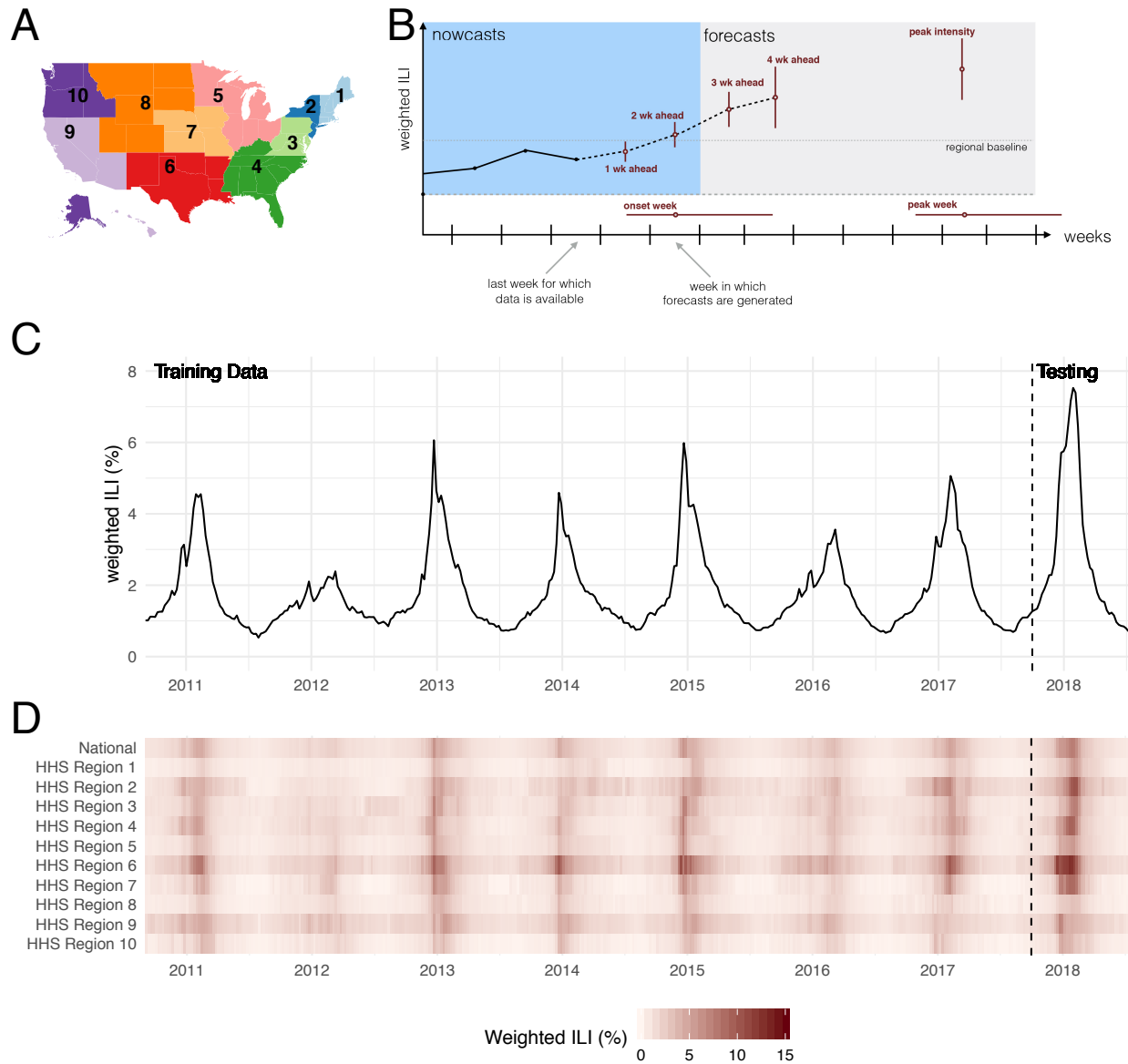[This paper describes the development of ...]

Figure 1: (A) Map of the 10 U.S. Health and Human Services regions. Influenza forecasts are made at this scale. (B) A figure showing the structure of a single forecast. Seven forecasting targets are illustrated with a point estimate (dot) and interval (uncertainty bars). The five targets on the wILI scale are shown with uncertainty bars spanning the vertical wILI axis, while the two targets for a time-of-year 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.[14] 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 two weeks prior. Therefore, 1 and 2 week-ahead forecasts are referred to as nowcasts, i.e., at or before the current time. Similarly, 3 and 4 week-ahead forecasts are forecasts, or estimates about events in the future. This figure has appeared previously in (Reich et al, under review). (C) Publicly available weighted influenza-like illness (wILI) data from the CDC website for the national level. The y-axis shows the weighted percentage of doctor's office visits in which a patient presents with influenza-like illness for each week from September 2010 through July 2018, which is the time period for which the models presented in this paper made seasonal forecasts. (D) Publicly available wILI data for each of the 10 HHS regions. Darker colors indicate higher wILI.

3

2   Results

2.1   Summary of ensemble components

We assembled 21 ensemble components to be evaluated for inclusion in our ensemble models (Table 1). These models encompassed a variety of different modeling philosophies, including Bayesian hierarchical models, mechanistic models of infectious disease transmission, statistical learning methodologies, and classical statistical models for time-series data. As a pre-requisite for inclusion in the ensemble, every model submitted seven seasons of out-of-sample forecasts. As the goal was to submit our ensemble forecast in real-time to the CDC FluSight forecasting challenge, we adhered to guidelines and formats set forth by the challenge in determining forecast format. A season typically consists of forecast files generated weekly for 33 weeks, starting with epidemic week 43 (EW43) of one calendar year and ending with EW18 of the following year. Weekly submissions contain probabilistic and point forecasts for seven targets in each of 11 regions in the U.S. (national-level plus the 10 Health and Human Services (HHS) regions, Figure 1A). There are two classes of targets: "week-ahead" and "seasonal". "Week ahead" targets refer to the four weekly targets (incidence 1, 2, 3 and 4 weeks in the future) that are different for each week of the season. "Seasonal" targets refer to quantities (outbreak onset, outbreak peak week, and outbreak peak intensity) that do not change for a region within a season (see Figure 1B and Methods).

Individual component forecast performance in the seven training seasons (2010/2011 - 2016/2017) varied widely across region, season, and target. A detailed comparative analysis of component forecast performance can be found elsewhere (Reich et al, under review), however we summarize a few key insights from model performance here. To evaluate model accuracy, we followed CDC convention and used a metric that takes the geometric average of the probabilities assigned to the eventually observed value. This measure, which we refer to as "forecast score", can be interpreted as the average probability a given forecast model assigned to the eventually observed value. As such, higher values, on a scale of 0 to 1, indicate more accurate models. A seasonal baseline model, whose forecasts for a particular target are based on data from previous seasons and do not update based on data from the current season, was used as a reference point for other models. Over 50% of the ensemble components out-performed the seasonal baseline model in forecasting 1- and 2-week ahead incidence as well as season peak percentage and season peak week.

Each ensemble component's weekly forecasts were submitted in real-time to a central data repository throughout the 2017/2018 influenza season.[19] The season began with submissions for epidemic week 43, 2017 (EW43-2017) on Monday, November 6, 2017 and ended with submissions for EW18-2018 on Monday, May 18, 2018. For each of the 33 weeks of the 2017/2018 season, each of 21 models submitted real-time weekly probabilistic forecasts for 7 targets in each of 11 regions. In all, the seven seasons of training data consisted of about 95MB and over 1.5m rows of data per model and about 2GB of forecast data for all models combined. The real-time forecasts for the 2017/2018 season added about 300MB of data.

2.2   Choice of ensemble model based on cross-validation

The FSNetwork Target-Type Weights (FSNetwork-TTW) ensemble model outperformed all other ensemble and ensemble components in the training phase by a slim margin. This model was one of five pre-specified ensemble approaches defined prior to any systematic evaluation of ensemble component performance in previous seasons and prior to the 2017/2018 season. Using 42 estimated weights, one for each model and target-type (week-ahead

Team	Model Abbr	Model Description	Ens.	Model
FSNetwork	EW	Equal Weights (number of unique weights = 1)	x	
	CW	Constant Weights (21)	x	
	TTW	Target-Type Weights (42)	x	
	TW	Target Weights (147)	x	
	TRW	Target-Region Weights (1,617)	x	
CU	EAKFC_SEIRS	Ensemble Adjustment Kalman Filter SEIRS		[15]
	EAKFC_SIRS	Ensemble Adjustment Kalman Filter SIRS		[15]
	EKF_SEIRS	Ensemble Kalman Filter SEIRS		[4]
	EKF_SIRS	Ensemble Kalman Filter SIRS		[4]
	RHF_SEIRS	Rank Histogram Filter SEIRS		[4]
	RHF_SIRS	Rank Histogram Filter SIRS		[4]
	BMA	Bayesian Model Averaging		[11]
Delphi	BasisRegression	Basis Regression (epiforecast defaults)		[16]
	DeltaDensity1	Delta Density (epiforecast defaults)		[9]
	EmpiricalBayes1	Empirical Bayes (conditioning on past four weeks)		[17, 16]
	EmpiricalBayes2	Empirical Bayes (epiforecast defaults)		[17, 16]
	EmpiricalFuture	Empirical Futures (epiforecast defaults)		[16]
	EmpiricalTraj	Empirical Trajectories (epiforecast defaults)		[16]
	DeltaDensity2	Markovian Delta Density (epiforecast defaults)		[9]
	Uniform	Uniform Distribution		
	Stat	Ensemble (combination of 8 Delphi models)	x	[9]
LANL	DBM	Dynamic Bayesian SIR Model with discrepancy		[8]
ReichLab	KCDE	Kernel Conditional Density Estimation		[18]
	KDE	Kernel Density Estimation and penalized splines		[12]
	SARIMA1	SARIMA model without seasonal differencing		[12]
	SARIMA2	SARIMA model with seasonal differencing		[12]

Table 1: List of models, with key characteristics. Team abbreviations are translated as: FSNetwork = FluSight Network, CU = Columbia University, Delphi = Carnegie Mellon, LANL = Los Alamos National Laboratories, ReichLab = University of Massachusetts Amherst. 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 an mechanistic or compartmental model formulation. The 'Ens. model' column notes models that are ensemble models.
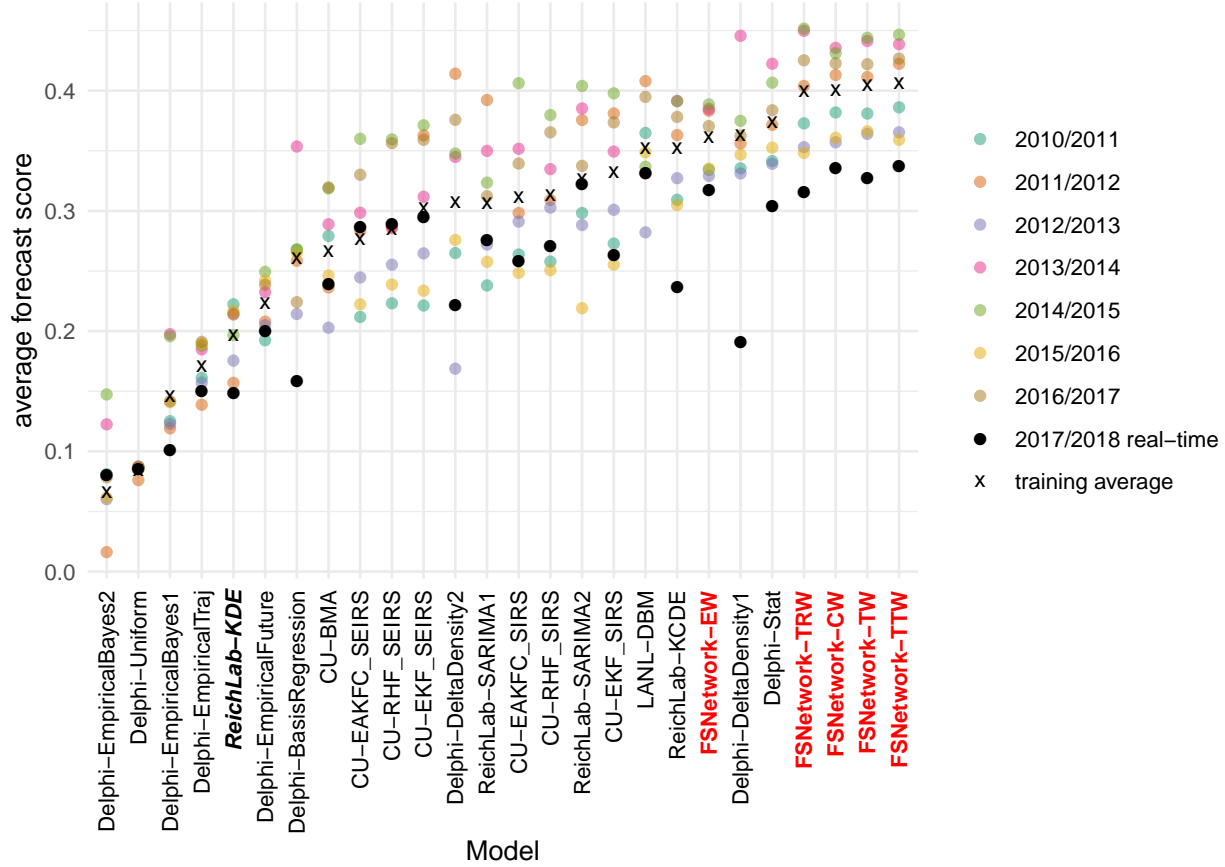
Figure 2: Average forecast score, aggregated across targets, regions, and weeks, plotted separately for each model and season. Models are sorted from lowest scores (left) to highest scores (right). Higher scores indicate better performance. Dots show average scores across all targets, regions, and weeks within a given season. The 'x' marks the geometric mean of the seven seasons. The black dot represents the average score for the prospective, real-time forecasts from the 2017/2018 season. The ReichLab-KDE model (bold, italicized font) is considered the historical baseline model. The models highlighted in red, boldface font are the FluSight Network ensemble models. Note that a componentâĂŹs standalone accuracy does not necessarily correlate to its contribution to the overall ensemble accuracy. See discussion at end of Section efsubsec:comp-models.
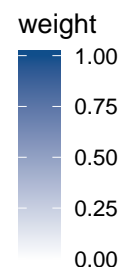
Figure 3: Model weights for the FluSight Network Target-Type Weights (`FSNetwork-TTW`) model. Weights were estimated using cross-validated forecast performance in the 2010/2011 through the 2016/2017 seasons.

and seasonal) combination, the `FSNetwork-TTW` model built a weighted model average using a predictve density stacking approach (see Figure 3 and Methods). In the training period consisting of the seven influenza seasons prior to 2017/2018, this model achieved a leave-one-season-out cross-validated average forecast score of 0.41, compared with the FSNetwork Target Weights (`FSNetwork-TW`) model with a score of 0.40, the FSNetwork Constant Weights (`FSNetwork-CW`) model with a score of 0.40, and the FSNetwork Target-Region Weights (`FSNetwork-TRW`) model with a score of 0.40 (Figure 2). We chose the target-type weights model as the model that would be submitted in real-time to the CDC during the 2017/2018 season, based on it having the highest score of any approach in the cross-validated training phase.

Using out-of-sample cross-validated performance of all ensemble components across the seven training seasons, we estimated weights for the chosen `FSNetwork-TTW` ensemble model that would be used for the 2017/2018 real-time forecasting. The `FSNetwork-TTW` model assigned non-negligible weight (greater than 0.001) to 8 models for week-ahead targets and 6 models for seasonal targets (Figure 3). For week-ahead targets, the highest non-zero weight (0.42) was given to the `Delphi-DeltaDensity1` model. For seasonal targets, the highest weight (0.26) was given to the `LANL-DBM` model. In the weights for the seasonal targets, 6 models shared fairly equal weights, with none of the 6 having less than 0.11 weight. All four teams had at least one model with non-negligible weight in the chosen model. Ensemble weights themselves are not a measure of a component's standalone accuracy nor its contribution to the overall accuracy of the ensemble.

2.3 Summary of ensemble real-time performance in 2017/2018 season

The 2017/2018 influenza season in the U.S. exhibited features that were unlike that of any season in the past 15 years. As measured by wILI percent at the national level, the 2017/2018 season was on par with the other two highest peaks on record: the 2003/2004 season and the 2009 H1N1 pandemic. In some regions, for example HHS Region 2 (New York and New Jersey) and HHS Region 4 (southeastern states), the highest reported wILI percent for the 2017/2018 season was more than 20% above previously observed peaks. Because all forecasting models rely,
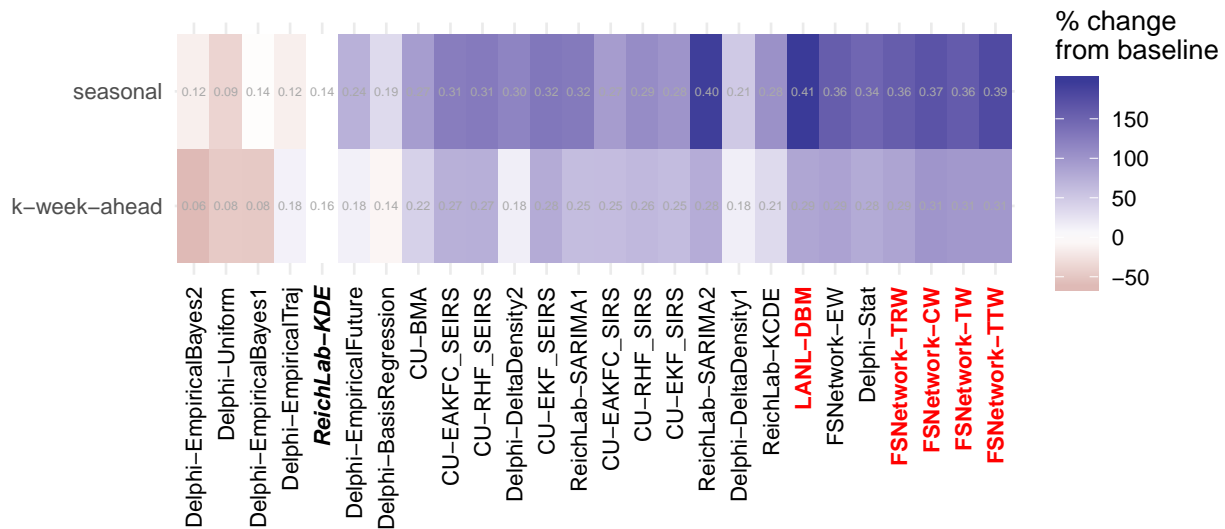
Figure 4: Average forecast score in 2017/2018 by model target-type. The text within the grid shows the score itself. The white midpoint of the color scale is set to be the target-specific average of the historical baseline model, `ReichLab-KDE`, with darker blue colors representing models that have better scores than the baseline and darker red scores representing models that have worse scores than the baseline. The models are sorted in descending order from most accurate (right) to least accurate (left).

to some extent, on future trends mimicking observed patterns in the past, the anomalous dynamics in 2017/2018 posed a challenging "test season" for all models, including the new ensembles. Indeed, some of the models that saw the largest drop in performance (e.g., `Delphi-DeltaDensity1`, `ReichLab-KCDE`, and `Delphi-DeltaDensity2`) are ones that explicity rely on using kernel conditional density estimation to find previously observed regions of the time-series that bear a similarity to current trends.

In spite of these unusual dynamics negatively impacting the forecast accuracy of the top-performing ensemble components, the `FSNetwork-TTW` ensemble model that varied weights by model and target-type was tied for the best performance among all component and ensemble models in 2017/2018. The `FSNetwork-TTW` model tied with the `FSNetwork-CW` model with a score of 0.34. These results from 2017/2018 were consistent with and confirmed conclusions drawn from the training period, where these same models outperformed all other ensemble and ensemble components. This strong and consistent performance by the ensemble models is a particularly noteworthy achievement given that just matching the single model that happens to do best in a particular season (without being a pre-specified comparison) is a high standard to hold a model to.

[Describe additional performance metrics: bias and PIT results from Rebecca.]

[Score breakdown by target-types and by region?]

The model accuracy patterns observed in the cross-validation phase was largely preserved in the real-time phase. [[add model ranks to table (avg for CV years, single for 2017/2018)]] In the cross-validation phase, ensemble models had slightly higher average performance than all ensemble components, and [[ranked as the top model in each season]].

[Could also present/describe updated weights that incorporate 2017/2018 performance.]

8

3   Discussion

Outline

- something about CDC using these data to decide whether or not to recommend the new ensemble approach for the 2018-19 season and how the weights will be updated going forward.

- something about how this is a collaboration between CDC and outside partners and how we were able to accomplish more by collaboration in real time on a common problem.

- ensembles hold promise for giving decision makers the ability to use "one answer" that combines the strength of many different approaches. highlight that no one model was ever used by the ensemble as the best answer and instead it relied on a combination of models to give the best answer.

- i do a think limitation or a something to consider working on is how to adapt these methods when there aren't historic analogous (e.g. ebola, pandemic flu)

- Overall, the ensemble methods outperform ensemble components. There are some specific situations where the ensemble components do better, but that is to be expected when testing a lot of situations.

- Even in a very unusual season, the ensemble approach holds steady and was high-performing. This bodes well for its long-term robustness, compared to single ensemble components. In this way, the relative simplicity of CW or TTW is a strength, as there is not too much danger of these models being "overfit" to the data.

- Important limitation/opportunity: not an obvious way to incorporate new models that, say, rely on a new data source and don't have 7 years of training. Could be important to include such a model, but how to do it is not trivial, and won't have a good sense about how the long run performance of this model is.

- places where we could improve: dynamic weighting over the course of a season, increasing diversity in models, model calibration, consideration of other ways to combine models.

- We did not explore the use of more complex regularization schemes for estimation in this work, but in future work penalized estimation procedures could be employed for the more complex models, yielding an intermediate number of effective parameters. Opportunity to cite awes.

- This new approach shows terrific promise for future collaborative forecasting efforts. Was a close second place in the 2017/2018 competition. Highlights the value of standard data formats and requiring multiple "training seasons".

4   Methods

4.1   Influenza Data

Forecasting targets for the CDC FluSight challenge are based on the US Outpatient Influenza-like Illness Surveillance Network (ILINet). ILINet is a syndromic surveillance system that measures the weekly percentage of outpatient visits due to influenza-like illness (ILI) from a network of more than 2,800 providers, and publishes a weighted estimate of ILI (wILI) based on state populations. Estimates of wILI are reported weekly by the CDC's Influenza Division for the United States as a whole as well as for each of the 10 Health and Human Services (HHS)

¹⁶⁴ regions. Reporting of 'current' wILI is typically delayed by approximately one to two weeks from the calendar date
¹⁶⁵ of a doctor's office visit as data are collected and processed, and each weekly publication can also include revisions
¹⁶⁶ of prior reported values if new data become available. For the U.S. and each HHS Region, CDC publishes an
¹⁶⁷ annual baseline level of ILI activity based on off-season ILI levels.

4.2 Forecast Targets and Structure

¹⁶⁹ Forecasts for the CDC FluSight challenge consist of seven targets: three seasonal targets and four short-term or
¹⁷⁰ 'week-ahead' targets (Figure 1B). The seasonal targets consist of season onset, defined as the first MMWR week
¹⁷¹ where wILI is at or above baseline and remains above for three consecutive weeks, season peak week, defined as
¹⁷² the MMWR week of maximum wILI, and season peak percentage, defined as the maximum wILI value for the
¹⁷³ season. The short-term targets consist of forecasts for wILI values 1, 2, 3, and 4 weeks ahead of the most recently
¹⁷⁴ published data. With the two-week reporting delay in the publication of ILINet, these forecasts are for the level
¹⁷⁵ of wILI occurring 1 week prior to the week the forecast is made, the current week, and the two weeks after the
¹⁷⁶ forecast is made (Figure 1B). Forecasts are created for all targets for the US as a whole and for each of the 10
¹⁷⁷ HHS Regions (Figure 1A,C,D).

¹⁷⁸ For all targets, forecasts consist of probability distributions within bins of possible values for the target. For season
¹⁷⁹ onset and peak week, forecast bins consist of individual weeks within the influenza season, with an additional bin for
¹⁸⁰ onset week corresponding to a forecast of no onset. For short-term targets and peak intensity, forecast bins consist
¹⁸¹ of levels of observed wILI rounded to the nearest 0.1% up to 13%, which is the level of resolution publicly for ILINet
¹⁸² reported by the CDC. Formally, the bins are defined as $[0.00, 0.05)$, $[0.05, 0.15)$, ..., $[12.85, 12.95)$, $[12.95, 100]$.

4.3 Forecast Evaluation

¹⁸⁴ Submitted forecasts were evaluated using the modified log score used by the CDC in their forecasting challenge,
¹⁸⁵ which provides a simultaneous measure of forecast accuracy and precision. The log score for a probabalistic
¹⁸⁶ forecast 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 some
¹⁸⁷ target Z, conditional on some data $\mathbf{x}$ and z^* is the observed value of the target Z.

¹⁸⁸ While a true log score only evaluates the probability assigned to the exact observed value z^*, the CDC uses a
¹⁸⁹ modified log score that classifies additional values as "accurate". For predictions of season onset and peak week,
¹⁹⁰ probabilities assigned to the week before and after the observed week are included as correct, so the modified
¹⁹¹ log score becomes $\log \int_{z^*-1}^{z^*+1} f_m(z|\mathbf{x})dz$. For season peak percentage and the short-term forecasts, probabilities
¹⁹² assigned to wILI values within 0.5 units of the observed values are included as correct, so the modified log score
¹⁹³ becomes $\log \int_{z^*-.5}^{z^*+.5} f_m(z|\mathbf{x})dz$. We refer to these modified log scores as simply log scores hereafter.

¹⁹⁴ Individual log scores can be averaged across different combinations of forecast regions, target, weeks, or seasons.
¹⁹⁵ Formally, each model m has a large number of region-, target-, season-, and week-specific log scores, and we
¹⁹⁶ represent a specific scalar log score as $\log f_{m,r,t,s,w}(z^*|\mathbf{x})$. These individual log scores can be averaged across
¹⁹⁷ combinations of regions, targets, seasons, and weeks to compare model performance.

4.4 Ensemble components

To provide training data for the ensemble, four teams submitted between 1 and 9 models each, for a total of 21 ensemble components. *(NOTE: Could refer to comparison manuscript here if it's out)* Teams submitted out-of-sample forecasts for the 2010/2011 through 2016/2017 influenza seasons. Teams constructed their forecasts in a prospective fashioon, using only data that were available at the time of the forecast. For some data sources (i.e. wILI prior to the 2014/2015 influenza season), data as they were published at the time were not available. In such cases, teams were still allowed to use those data sources while making efforts to only use data available at the time forecasts would have been made.

For each influenza season, teams submitted weekly forecasts from epidemic week 40 (EW40) of the first year through EW20 of the following year, using standard CDC definitions for epidemic week (citation). If a season contained EW53, forecasts were submitted for that week as well. In total, teams submitted 233 individual forecast files representing forecasts across the seven influenza seasons. Once submitted, the forecast files were not updated except in four instances were explicit programming bugs had resulted in numerical issues in the forecast. Teams were explicitly discouraged from re-tuning or adjusting their models for different prior seasons to avoid issues with over-fitting.

Teams utilized a variety of methods and modeling approaches in the construction of their submissions. Seven of the models used a compartmental structure in their models (i.e. Susceptible-Infectious-Recovered) to model the disease transmission process in some way, while other models used more statistical approaches to directly model the observed wILI curve. Six of the models explicitly incorporate additional data sources beyond previous wILI data, including weather data and Google search data.

4.5 Distinction between standalone models and ensemble components

It is important to distinguish ensemble components from standalone forecasting models. Standalone models are optimized to be as accurate as possible on their own by, among other things, using proper smoothing. Ensemble components might be designed to be accurate on their own, or else they may be included merely to complement weak spots in other components, i.e. to reduce the ensemble's variance. Because we had sufficient cross-validation data to estimate ensemble weights for several dozen components, some groups contributed non-smoothed "complementing" components for that purpose. Such components may perform poorly on their own, yet their contribution to overall ensemble accuracy may still be significant.

4.6 Ensemble Construction

All ensemble models can be represented with the same notation. Let $f_c(y_{t,r,w})$ represent the predictive density of ensemble component c for the value of the target $Y_{t,r,w}$, where t indexes the particular target, r indexes the region, and w indexes the week. We combine these components together into an ensemble model $f(y_t)$ as follows:

$$f(y_{t,r,w}) = \sum_{c=1}^{C} \pi_{c,t,r} f_c(y_{t,r,w}) \tag{1}$$

where $\pi_{c,t,r,w}$ is the weight assigned to component c for predictions of target t in region r on week w. We require $\sum_{c=1}^{C} \pi_{c,t,r} = 1$ and thereby ensure that $f(y_{t,r,w})$ remains a valid probability distribution.

A total of five weighted ensemble models were considered, with varying complexity and number of estimated weights (Table 1).

- Equal Weight (`FSNetwork-EW`): $\pi_{c,t,r} = 1/C$. This model consisted of assigning all components the same weight regardless of performance and is equivalent to the arithmetic mean of the components.

- Constant Weight model (`FSNetwork-CW`): $\pi_{c,t,r}$ varies across components but has the same value for all targets and regions, for a total of 21 weights.

- Target Type Weight model (`FSNetwork-TTW`): $\pi_{c,t,r}$ is estimated separately for short-term targets and seasonal targets with no variation across regions, resulting in a total of 42 weights.

- Target Weight model (`FSNetwork-TW`): $\pi_{c,t,r}$ is estimated separately for each of the seven targets for each component with no variation across regions, resulting in 147 weights

- Target-Region Weight model (`FSNetwork-TRW`): The most complex model considered, which involved estimating $\pi_{c,t,r}$ separately for each component-target-region combination, resulting in 1617 unique weights.

[EM algorithm/weighting/density stacking]

Component weights were trained using a leave-one-season out cross-validation approach on component forecasts from the 2010/2011 through 2016/2017 seasons. Given the limited number of seasons, we used data from all other seasons as training data to estimate weights for a given test season, even if the training season occured chronologically after the test season of interest.

Based on the results of the cross-validation study, we selected one ensemble model as the official FluSight Network entry to the CDC's 2017/2018 influenza forecasting challenge. Component weights for that model were estimated using all seven seasons of training data. Over the course of the 2017/2018 influenza season, participating teams submitted weekly forecasts from each component, which were combined using the estimated weights into the FluSight Network model and submitted to the CDC. The component weights for the submitted model were unchanged throughout the course of the season.

It should be noted that ensemble weights are not a measure of ensemble components' standalone accuracy nor do they measure the overall contribution of a particular model to the ensemble accuracy. For example, consider a setting where a duplicate of a identical (or highly similar) ensemble component with weight π^* is added to a given ensemble. The accuracy of the original ensemble can be maintained in a number of ways, including (a) assigning each copy a weight of $\pi^*/2$, or (b) assigning the first copy a weight of π^* and the second copy a weight of 0. In both of these weightings, at least one high accuracy ensemble component would be assigned significantly lower weight based on the presence of another identical or similar component. Additionally, components can be assigned small weights but have a large impact on ensemble accuracy compared to an ensemble excluding it.

References

[1] Estimated influenza illnesses, medical visits, hospitalizations, and deaths averted by vaccination in the united states. https://www.cdc.gov/flu/about/disease/2016-17.htm, May 2018.

[2] Overview of influenza surveillance in the united states. https://www.cdc.gov/flu/weekly/overview.htm, 2017.

[3] Jeffrey Shaman, Alicia Karspeck, Wan Yang, James Tamerius, and Marc Lipsitch. Real-time influenza forecasts during the 2012âĂŞ2013 season. *Nature Communications*, 4, 2013.

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

[5] Shihao Yang, Mauricio Santillana, and S C Kou. Accurate estimation of influenza epidemics using Google search data via ARGO. *Proceedings of the National Academy of Sciences of the United States of America*, 112(47):14473–8, nov 2015.

[6] Jean-Paul Chretien, Dylan George, Jeffrey Shaman, Rohit A. Chitale, and F. Ellis McKenzie. Influenza forecasting in human populations: A scoping review. *PLOS ONE*, 9(4):1–8, 04 2014.

[7] Sasikiran Kandula, Daniel Hsu, and Jeffrey Shaman. Subregional Nowcasts of Seasonal Influenza Using Search Trends. *Journal of medical Internet research*, 19(11):e370, nov 2017.

[8] Dave Osthus, James Gattiker, Reid Priedhorsky, and Sara Y. Del Valle. Dynamic Bayesian Influenza Forecasting in the United States with Hierarchical Discrepancy. *arXiv*, aug 2017.

[9] Logan C. Brooks, David C. Farrow, Sangwon Hyun, Ryan J. Tibshirani, and Roni Rosenfeld. Nonmechanistic forecasts of seasonal influenza with iterative one-week-ahead distributions. *PLOS Computational Biology*, 14(6):e1006134, jun 2018.

[10] Sen Pei, Sasikiran Kandula, Wan Yang, and Jeffrey Shaman. Forecasting the spatial transmission of influenza in the United States. *Proceedings of the National Academy of Sciences of the United States of America*, 115(11):2752–2757, mar 2018.

[11] Teresa K. Yamana, Sasikiran Kandula, and Jeffrey Shaman. Individual versus superensemble forecasts of seasonal influenza outbreaks in the United States. *PLOS Computational Biology*, 13(11):e1005801, nov 2017.

[12] Evan L. Ray and Nicholas G. Reich. Prediction of infectious disease epidemics via weighted density ensembles. *PLOS Computational Biology*, 14(2):e1005910, feb 2018.

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

[14] Matthew Biggerstaff, Krista Kniss, Daniel B Jernigan, Lynnette Brammer, Joseph Bresee, Shikha Garg, Erin Burns, and Carrie Reed. 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. *American Journal of Epidemiology*, 187:1040–1050, 2018.

[15] Sen Pei and Jeffrey Shaman. Counteracting structural errors in ensemble forecast of influenza outbreaks. *Nature Communications*, 8(1):925, dec 2017.

[16] Logan C Brooks, David C Farrow, Sangwon Hyun, Ryan J Tibshirani, and Roni Rosenfeld. epiforecast: Tools for forecasting semi-regular seasonal epidemic curves and similar time series. https://github.com/cmu-delphi/epiforecast-R, 2015.

[17] Logan C. Brooks, David C. Farrow, Sangwon Hyun, Ryan J. Tibshirani, and Roni Rosenfeld. Flexible Modeling of Epidemics with an Empirical Bayes Framework. *PLOS Computational Biology*, 11(8):e1004382, aug 2015.

[18] Evan L. Ray, Krzysztof Sakrejda, Stephen A. Lauer, Michael A. Johansson, and Nicholas G. Reich. Infectious disease prediction with kernel conditional density estimation. *Statistics in Medicine*, sep 2017.

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