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

Nicholas G Reich[1], Craig McGowan[2], Logan Brooks[3], Sasikiran Kandula[4], Dave Osthus[5], Evan Ray[6], Abhinav Tushar[1], Teresa Yamana[4], Willow Crawford-Crudell[7], Graham Casey Gibson[1], Evan Moore[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

Seasonal influenza results in substantial annual morbidity and mortality in the United States and worldwide. Accurate forecasts of key features of influenza epidemics, such as the timing and severity of the peak incidence in a given season, can inform public health response to outbreaks. As part of ongoing efforts to incorporate data and advanced analytical methods into public health decision-making, the United States Centers for Disease Control and Prevention (CDC) has organized seasonal influenza forecasting challenges. In the 2017/2018 season, 22 teams contributed a total of 30 models. A subset of four teams participating in the challenge created a research consortium called the FluSight Network in early 2017. During the 2017/2018 season they worked together to produce a "collaborative ensemble" model that combined 21 separate component models into a single model using a machine learning technique called stacking. This approach creates a weighted average of predictive densities where the weight for each component is based on that component's forecast accuracy in past seasons. In the 2017/2018 influenza season, one of the largest seasonal outbreaks in the last 15 years, this ensemble model performed on average better than all individual component models and placed second overall in the CDC challenge. It also outperformed the baseline ensemble model created by the CDC that took a simple average of all models submitted to the forecasting challenge. This project shows that collaborative efforts between research teams to develop ensemble forecasting approaches can bring measurable improvements in forecast accuracy and important reductions in the variability of performance from year to year. Efforts such as this, that emphasize real-time testing and evaluation of forecasting models and facilitate the close collaboration between public health officials and modeling researchers, are critical to improving our understanding of how best to use forecasts to improve public health response to seasonal and emerging

epidemic threats.

1 Introduction

Seasonal influenza results in a substantial annual public health burden in the United States and worldwide. In the influenza season running from October 2017 through May 2018, one of the largest seasonal outbreaks on record, the United States Centers for Disease Control and Prevention (CDC) estimates there were 48.8 million cases of influenza in the US, along with 959,000 hospitalizations and nearly 80,000 deaths.[1] The 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.[3] Forecasts of an influenza season offer the possibility of providing actionable information to improve public health responses, and recent years have seen a large amount of peer-reviewed research describing efforts to predict seasonal influenza.[4, 5, 6, 7, 8, 9, 10, 11]

Ensemble models, i.e. methods that bring together predictions from multiple different component models, have long been seen as a valuable method for improving predictions over any single model.[12, 13, 14] 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[15, 16, 17], and recent work has extended ensemble forecasting to infectious diseases, including influenza, dengue fever, lymphatic filariasis, and Ebola hemorrhagic fever.[18, 19, 20, 21]

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. Among other government-sponsored infectious disease forecasting competitions in recent years,[22, 23] this challenge been unique in its prospective orientation over multiple outbreak seasons. Also, it has provided a venue for close interaction and collaboration between government public health officials and academic and private-sector researchers.

The FluSight challenge has been designed and retooled over the years with an eye towards maximizing the public health utility and integration of forecasts with real-time public health decision making. All forecast targets are derived from the trajectories of U.S. region-level weighted influenza-like illness (wILI), an estimate of the percentage of outpatient visits due to ILI weighted by state populations. ILI is perhaps the most frequently used measure of the burden of influenza-like respiratory illness in epidemiological surveillance. Weekly submissions to the FluSight challenge 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 week, outbreak peak week, and outbreak peak intensity) that do not change for a region within

a season (see Figure 1B and Methods).

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 an ensemble with performance-based model weights. During the 2015/2016 and 2016/2017 FluSight challenges, analysts at the CDC built a simple ensemble model for all targets by taking the arithmetic mean of all submitted models. This model was one of the top performing models each season.[24]

A central goal of the FluSight Network was to demonstrate the benefit of performance-based weights in a real-time, multi-team ensemble setting by outperforming the "simple average" ensemble that CDC uses to inform decision making and situational awareness during the annual influenza season. In this paper, we describe the development of this collaborative ensemble model and present results from seven seasons retrospectively (2010/2011 through 2016/2017) and one season prospectively (2017/2018). The FluSight Network assembled 21 ensemble components to build ensemble models for seasonal influenza outbreaks (Table 1). These components 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. We show that using ensemble models informed by past component performance consistently improved forecast accuracy. Given the fortuitous timing of this experiment, during the most severe seasonal influenza season on record, this work provides the first evidence from a real-time forecasting study that performance-based weights can improve ensemble forecast accuracy during high severity infectious disease outbreaks. This research is an important example of collaboration between government and academic public health experts, setting an important precedent and prototype for real-time collaboration in more severe outbreaks, such as a global influenza pandemic.

2 Results

2.1 Summary of ensemble components

Individual component model 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[26], however we summarize a few key insights from model performance here. 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-, 2-, and 3-week ahead incidence as well as season peak percentage and season peak week. However, season-to-season variability in forecast performance was large, as 10 models had, in at least one season, better overall accuracy than the model with the best average performance across all seasons. 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.

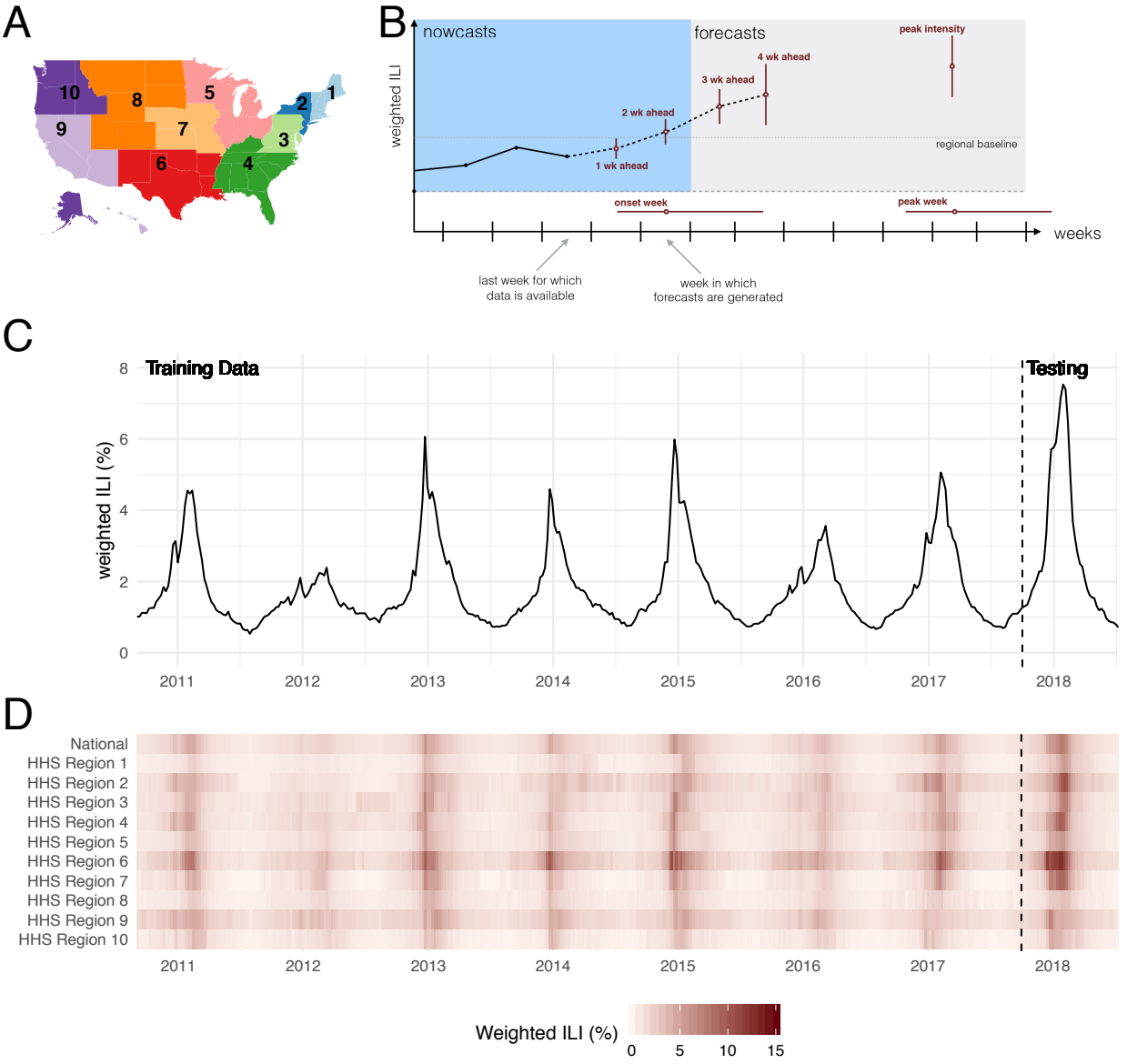

Figure 1: (A) Map of the 10 U.S. Health and Human Services regions. Influenza forecasts are made at this scale. (B) The structure of a single forecast. Seven forecasting targets are illustrated with a point estimate (dot) and interval (uncertainty bars). The five targets on the wILI scale are shown with 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.[25] 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 panel has been adapted from previous work.[26] (C) Publicly available 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. The dashed vertical line indicates the separation of the data used for the training (retrospective) and testing (prospective) phases of analysis. (D) Publicly available wILI data for each of the 10 HHS regions. Darker colors indicate higher wILI.

Team	Model Abbr	Model Description		Ext. Data	Mech. Model	Ens. Model
FSNetwork	EW	Equal Weights (number of estimated weights = 0)			x	
	CW	Constant Weights (20)		x		
	TTW	Target-Type Weights (40)		x		
	TW	Target Weights (140)		x		
	TRW	Target-Region Weights (1,540)		x		
CU	EAKFC_SEIRS	Ensemble Adjustment Kalman Filter SEIRS	[27]	x	x	
	EAKFC_SIRS	Ensemble Adjustment Kalman Filter SIRS	[27]	x	x	
	EKF_SEIRS	Ensemble Kalman Filter SEIRS	[5]	x	x	
	EKF_SIRS	Ensemble Kalman Filter SIRS	[5]	x	x	
	RHF_SEIRS	Rank Histogram Filter SEIRS	[5]	x	x	
	RHF_SIRS	Rank Histogram Filter SIRS	[5]	x	x	
	BMA	Bayesian Model Averaging	[18]			
Delphi	BasisRegression*	Basis Regression (`epiforecast` defaults)	[28]			
	DeltaDensity1*	Delta Density (`epiforecast` defaults)	[10]			
	EmpiricalBayes1*	Empirical Bayes (conditioning on past four weeks)	[29, 28]			
	EmpiricalBayes2*	Empirical Bayes (`epiforecast` defaults)	[29, 28]			
	EmpiricalFuture*	Empirical Futures (`epiforecast` defaults)	[28]			
	EmpiricalTraj*	Empirical Trajectories (`epiforecast` defaults)	[28]			
	DeltaDensity2*	Markovian Delta Density (`epiforecast` defaults)	[10]			
	Uniform*	Uniform Distribution				
	Stat	Ensemble (combination of 8 Delphi models)	[10]			x
LANL	DBM	Dynamic Bayesian SIR Model with discrepancy	[30]		x	
ReichLab	KCDE	Kernel Conditional Density Estimation	[31]			
	KDE	Kernel Density Estimation and penalized splines	[19]			
	SARIMA1	SARIMA model without seasonal differencing	[19]			
	SARIMA2	SARIMA model with seasonal differencing	[19]			
FluSight	unweighted_avg	Average of all models submitted to the CDC	[24]			x

Table 1: List of models, with key characteristics. Team abbreviations are translated as: CU = Columbia University, Delphi = Carnegie Mellon, LANL = Los Alamos National Laboratories, ReichLab = University of Massachusetts Amherst, UTAustin = University of Texas Austin. The FluSight model was not included in the collaborative ensemble, but is used as a reference ensemble model in the analysis. 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. Note that some of these components (marked with *) were not designed as standalone models, so their performance may not reflect the full potential of the method's accuracy (see Methods and Materials).

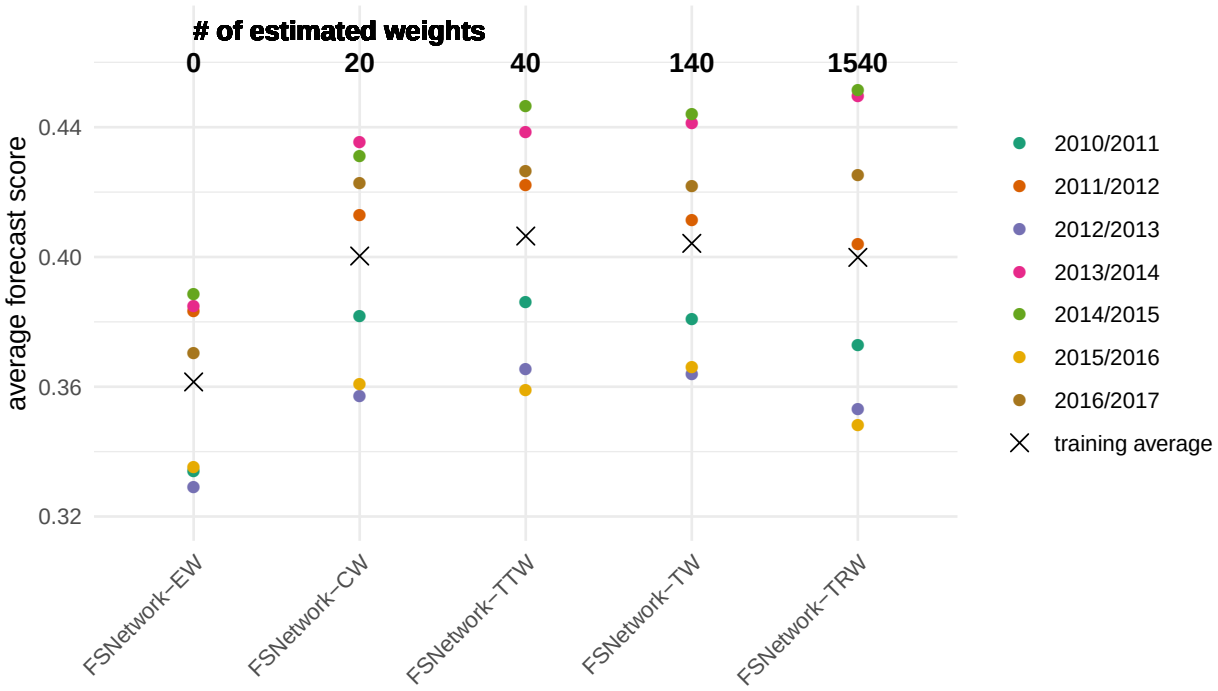

Figure 2: Training phase performance of the five pre-specified ensemble models. The models are sorted from simplest (left) to most complex (right), with the number of estimated weights (see Methods) for each model shown at the top. Each point represents the average forecast score for a particular season, with overall average across all seasons shown by the X.

2.2 Choice of ensemble model based on cross-validation

The FSNetwork Target-Type Weights (`FSNetwork-TTW`) ensemble model outperformed all other ensemble models 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.[32] Using 42 weights, one for each model and target-type (week-ahead 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 0.406, compared with the FSNetwork Target Weights (`FSNetwork-TW`) model with a score of 0.404, the FSNetwork Constant Weights (`FSNetwork-CW`) model with a score of 0.400, and the FSNetwork Target-Region Weights (`FSNetwork-TRW`) model with a score of 0.400 (Figure 4). 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 the pre-specified criteria of it having the highest score of any approach in the cross-validated training phase.[32]

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, six models shared over 99.9% of the

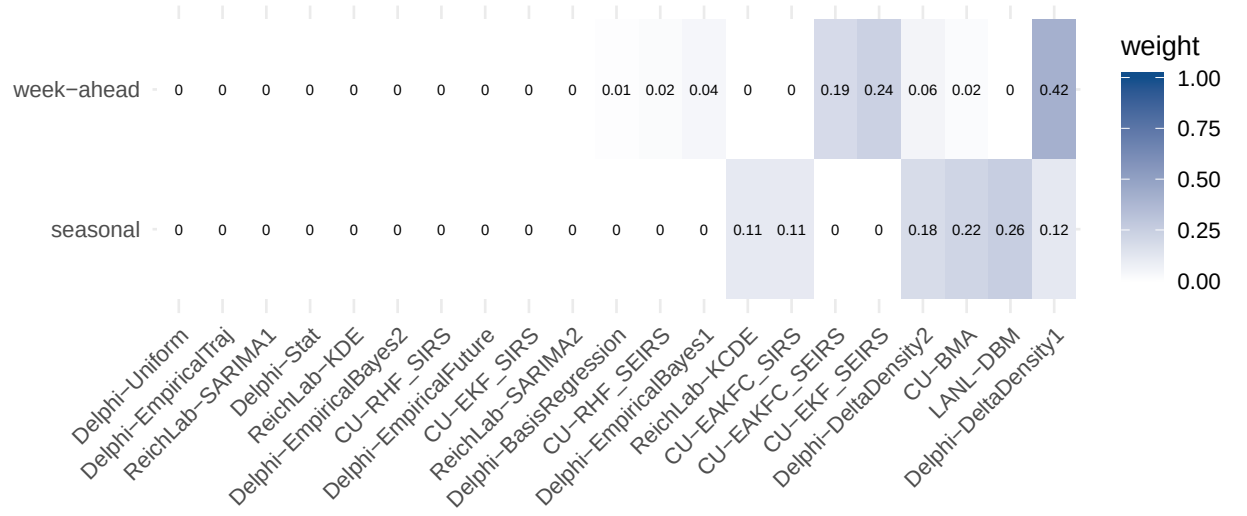

Figure 3: Model weights for the FluSight Network Target-Type Weights (`FSNetwork-TTW`) model in the 2017/2018 season. Weights were estimated using cross-validated forecast performance in the 2010/2011 through the 2016/2017 seasons.

weight, with none of the six having less than 0.11 weight. All four research 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 since 1997: 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% higher than previously observed peaks. Because all forecasting models rely, 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 chosen `FSNetwork-TTW` ensemble model showed the best performance among all models in the 2017/2018 season. In particular, we selected the single best model from each team in training phase stage and the `FluSight-unweighted_avg` model to compare with the `FSNetwork-TTW` model (Figure 4). The results from 2017/2018 were consistent with and confirmed conclusions drawn from the training period, where the `FSNetwork-TTW` model outperformed all other ensemble models and components. The `FSNetwork-TTW` model had the highest average score in the training period (0.337) as well as the highest average score in the 2017/2018 test season (0.406). This strong and consistent performance by the chosen ensemble model is a particularly

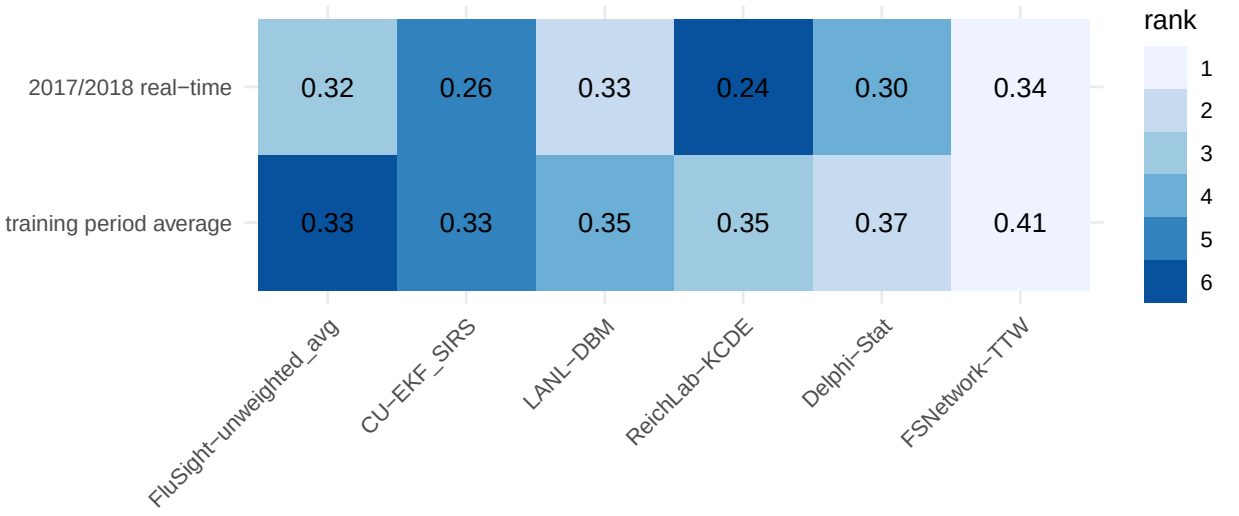

Figure 4: Overall test and training phase performance scores for selected models. Displayed scores are averaged across targets, regions, and weeks, and plotted separately for selected models. Models shown include the `FSNetwork-TTW` model, the top performing model from each team during the training phase and, for the last two training seasons and the test season, the unweighted average of all FluSight models received by CDC. Model ranks within each row are indicated by color of each cell (lighter colors indicates higher rank and more accurate forecasts) and the forecast score (rounded to two decimal places) is printed in each cell. Note that a component's standalone accuracy does not necessarily correlate to its contribution to the overall ensemble accuracy. See discussion at end of Section efsubsec:comp-models.

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 pre-selected model to.

Of particular note is that the `FSNetwork-TTW` model showed a higher performance in both training and testing phase than the `FluSight-unweighted_avg` model (Figure (Figure 4). The `FluSight-unweighted_avg` model contained forecast data from over 20 models submitted to the FluSight competition in 2017/2018 that were not part of the FluSight Network. In 2017/2018, the `FSNetwork-TTW` model earned an average forecast score of 0.337 while the `FluSight-unweighted_avg` model earned an average forecast score of 0.321.

While there there was considerable variation in target- and region-specific performance among the selected models in 2017/2018, the `FSNetwork-TTW` model showed lower variability in performance than other models. We compared the average forecast score for each target-region pair for the top model from each research team, the `FSNetwork-TTW` model and the `FluSight-unweighted_avg` (see Appendix). The `FSNetwork-TTW` model was the only of the six models, across all 77 pairs of targets and regions, that never had the lowest forecast score. Additionally, it only had the 5th lowest score twice. While our ensemble model did not always have the best score in each target-region pair, its consistency and low variability across all combinations secured it the top average score.

Despite being optimized for high log-score values, the `FSNetwork-TTW` showed robust performance in the 2017/2018 season across other performance metrics that measure forecast calibration and accuracy. Overall, the `FSNetwork-TTW` model ranked second among selected models in both RMSE and average bias, behind the `LANL-DBM` model (see Appendix), suggesting that using separate weighting schemes for point estimates and predictive distribution may be valuable. According to the probability integral transform metric[33, 34], the `FSNetwork-TTW` model was well-

calibrated for all four week-ahead targets (see Appendix). It was slightly less well-calibrated for peak performance, and showed indications of having too narrow predictive distributions over the 2017/2018 season. Over the entire training period prior to the 2017/2018 season, the `FSNetwork-TTW` model calibration results suggested that in general the model was a bit conservative, with often a too wide predictive distribution (see Appendix).

When taking into account the performance of ensemble components in the 2017/2018 season, the weights for a subsequent hypothetical ensemble using the same components would be different. Components that received lots of weight in the original ensemble but did particularly poorly in the 2017/2018 season saw the largest drop in weight (see Appendix). Overall, three components were added to the list of six existing components that received more than 0.001 weight for seasonal targets: `CU-EAKFC_SEIRS`, `CU-EKF_SEIRS`, and `ReichLab-SARIMA2`. One component (`ReichLab-SARIMA2`) was added to the list of eight existing components that received more than 0.001 weight for week-ahead targets.

2.4 Ensemble accuracy for peak forecasts

Forecast accuracy around the time of peak incidence is an important indicator of how useful a given model can be in real-time for public health decision-makers. To this end, we evaluated the scores of the `FSNetwork-TTW` ensemble model in each region in the 6 weeks prior to and six weeks after the peak week (Figure 5). Prior to the peak week, forecast scores of the peak percentage were on average lower than in past seasons, assigning on average 0.05, 0.06, and 0.05 probability to the eventually observed value 6, 5, and 4 weeks before the peak, respectively. However, at and after the peak week occured, this probability was over 0.70, quite a bit higher than average levels.

Similarly for peak week, the average forecast scores improved as the peak week approached. With the exception of a large dip in accuracy in HHS Region 7 just after peak occured (due to revisions to observed wILI data in the weeks surrounding peak), the forecast scores for peak week tended to be high (over 50% with a majority over 90%) in the weeks following peak.

3 Discussion

Ensembles hold promise for giving decision makers the ability to use "one answer" that combines the strengths of many different modeling approaches while mitigating their weaknesses. This work presents the first attempt to systematically combine infectious disease forecasts from multiple research groups in real-time using an approach that factors in past performance of each component method. Of the 29 models submitted to the CDC in 2017/2018 as part of their annual FluSight forecasting challenge, this ensemble was the second-highest scoring model overall. (The top scoring model was an ensemble of human judgement forecasts.[35]) In the 2018/2019 influenza season, based on results from this study, the CDC used forecasts from the FluSight Network ensemble model in internal and external communication and planning reports.

Even in a very unusual influenza season, the ensemble approach presented here was a steady contributor and did not see a large reduction in overall performance compared to performance during the training seasons. This bodes well for the long-term robustness of models such as this one, compared to single components that show higher variability in performance across specific years, regions, and targets. During the training and test phases, the weighted ensemble approaches outperformed two equal weight ensembles: one constructed based off of FluSight

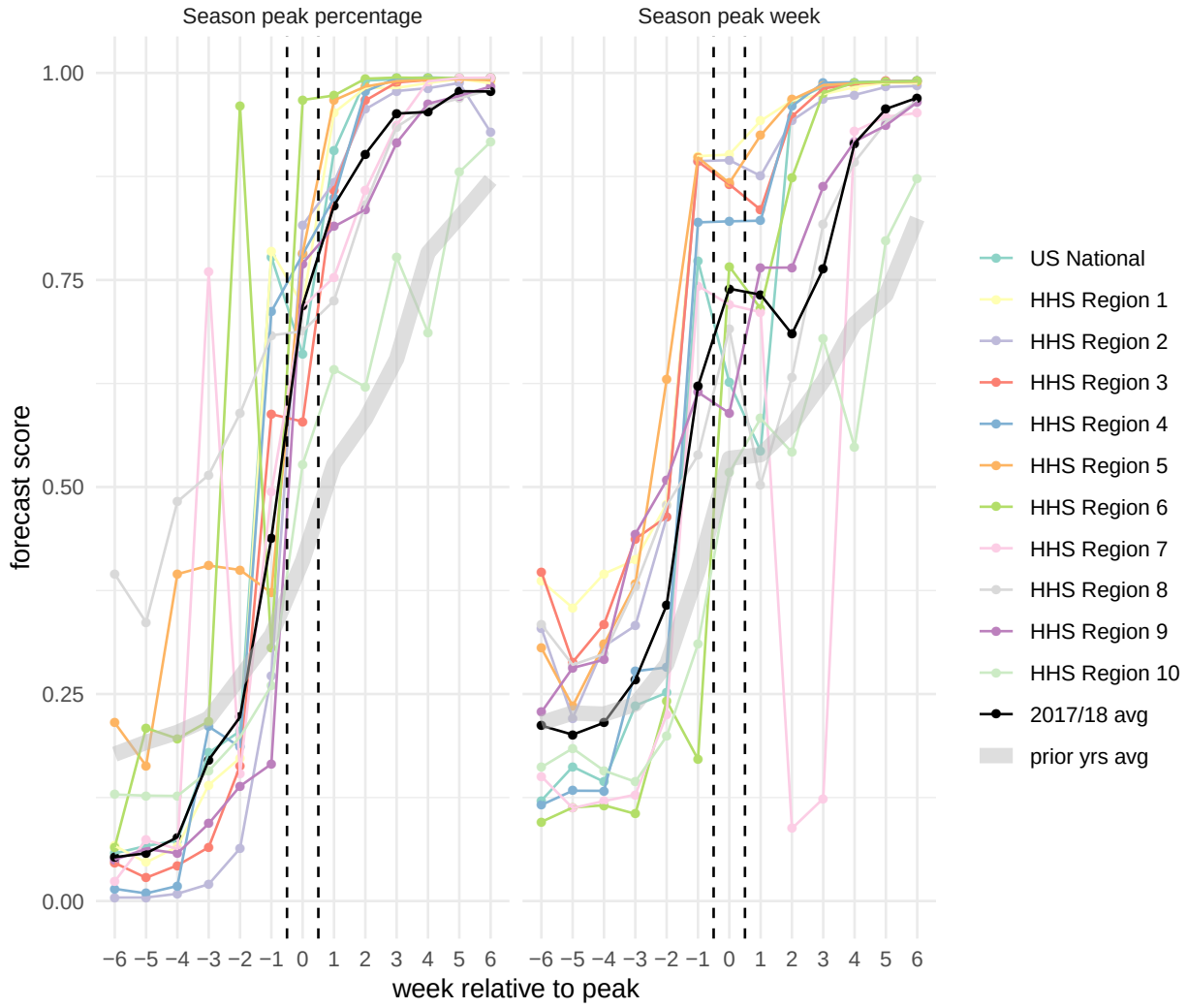

Figure 5: Forecast score for the `FSNetwork-TTW` model in 2017/2018 by week relative to peak. Scores for the two peak targets in each region were aligned to summarize performance relative to the peak week. On the x-axis, zero indicates the peak week and positive values represent weeks after the peak week. The black line indicates the overall geometric average across all regions. The grey band represents the geometric average across all regions and all seasons prior to 2017/2018.

Network models presented here (Table 1) and one constructed by the CDC using a wider array of models.[24]. This clearly illustrates the value for ensemble accuracy in incorporating information on prior component performance.

Overall, the ensemble methods outperformed the components. As shown by the FluSight Network Target-Type Weights component weighting structure presented above (Figure 3), no one model was ever used by the ensemble as the best answer and it instead relied on a combination of components to optimize performance. While there were some specific situations where the ensemble components did better in our training seasons, that is unsurprising given the number of models used. An important consideration in this study was that we only passed along one ensemble model into the testing phase and did not pre-specify any comparisons with ensemble components.

An ongoing challenge to efforts for modeling influenza in the US is that the outcome of interest to many public health officials, the percentage of doctor's office visits due to "influenza-like illness" (wILI), is an imperfect measure of actual influenza activity. It is a measure with a mixture of signals from outbreaks of different pathogens that cause influenza-like illness, e.g. common cold viruses. Additionally, the final measurements of wILI are estimates and not 'true' values. This suggests that forecast evaluation approaches such as the modified log score (which allow for some uncertainty around the eventually observed value to be scored as "correct") may play an important role in ensuring the models are not over-trained on a noisy outcome.

A critical limitation to our approach is that, as currently implemented, it relies on having multiple years of past performance for each component to be able to construct a reliable ensemble. As currently implemented, this may not a viable method for use in emerging pandemics, where there may not be any historical data on how models have performed nor reliable real-time data to train on. However, adaptive weighting approaches that dynamically update the weights over the course of a season could remove the requirement that all models have a substantial track-record of performance. Preliminary work on adaptive weighting has shown some promise, though such approaches still rely on accurately reported real-time data. Furthermore, a simple average of forecasts remains available in such situations and, as illustrated by the relatively strong performance of the FluSight Network Equal Weights model, can still offer advantages over individual models.

One risk of complex ensemble approaches is that they may be "overfit" to the data, resulting in models that place too much emphasis on one approach in a particular scenario or setting. This is a particular concern in applications such as this one, where the number of observations is fairly limited (hundreds to thousands of observations instead of hundreds of thousands). Against this backdrop, the relative simplicity of the FluSight Network Target-Type weights model is a strength, as there is less danger of these models being overfit to the data. Additionally, approaches that use regularization or penalization to reduce the number of effective parameters estimated by a particular model have been shown to have some practical utility in similar settings and may also have a role to play in future ensembles for infectious disease forecasting.[19]

As the success of this collaborative effort shows, there are significant gains to be made by working across disciplines and research groups, incorporating experts from government, academia and industry. However, at this point, collaborative efforts that consist of pooling results together (in many scientific applications, not just infectious disease forecasting) largely rely on bespoke technological solutions. We built a highly customized solution that relied on GitHub, Travis Continuous Integration server, unix scripts, and model code in R, python, and MatLab. 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. To move ensemble infectious disease forecasting into a more generalizable, operational phase, technological advancements are necessary to both standardize data sources, model structures, and forecast

formats as well as develop modeling tools that can facilitate the development and implementation of component and ensemble models.

This effort shows that collaborative efforts between research teams to develop ensemble forecasting approaches bring measurable improvements in accuracy and reductions in variability. We therefore are moving substantially closer to forecasts that can and should be used to inform routine, ongoing public health surveillance of infectious diseases. With the promise of new, real-time data sources and continued methodological innovation for both component models and ensemble approaches, there is good reason to believe that infectious disease forecasting will continue to mature and improve in upcoming years. As modeling efforts become more commonplace in the support of public health decision-making worldwide, it will be critical to develop infrastructure so that multiple models can more easily be brought online to both create ensemble forecasts as well as to develop our understanding of how best to communicate the forecasts and their uncertainty to decision-makers and the general public. Efforts such as this, that emphasize real-time testing and evaluation of forecasting models and facilitate the close collaboration between public health officials and modeling researchers, are critical to improving our understanding of how best to use forecasts to improve public health response to seasonal and emerging epidemic threats.

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) regions. Reporting of 'current' wILI is typically delayed by approximately one to two weeks from the calendar date of a doctor's office visit as data are collected and processed, and each weekly publication can also include revisions of prior reported values if new data become available. Larger revisions have been shown to be associated with decreased forecast accuracy.[26] For the U.S. and each HHS Region, CDC publishes an annual baseline level of ILI activity based on off-season ILI levels.[2]

4.2 Forecast Targets and Structure

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. Forecasts for the CDC FluSight challenge consist of seven targets: three seasonal targets and four short-term or 'week-ahead' targets (Figure 1B). The seasonal targets consist of season onset (defined as the first MMWR week where wILI is at or above baseline and remains above for three consecutive weeks), season peak week (defined as the MMWR week of maximum wILI), and season peak percentage (defined as the maximum wILI value for the season). The short-term targets consist of forecasts for wILI values 1, 2, 3, and 4 weeks ahead of the most recently published data. With the two-week reporting delay in

the publication of ILINet, these forecasts are for the level of wILI occurring 1 week prior to the week the forecast is made, the current week, and the two weeks after the forecast is made (Figure 1B). Forecasts are created for all targets for the US as a whole and for each of the 10 HHS Regions (Figure 1A,C,D).

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

4.3 Forecast Evaluation

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

While a true log score only evaluates the probability assigned to the exact observed value z^*, the CDC uses a modified log score that classifies additional values as "accurate". For predictions of season onset and peak week, probabilities assigned to the week before and after the observed week are included as correct, so the modified 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 assigned to wILI values within 0.5 units of the observed values are included as correct, so the modified log score 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.

Individual log scores can be averaged across different combinations of forecast regions, target, weeks, or seasons. Each model m has an associated predictive density for each combination of region (r), target (t), season (s), and week (w). Each of these densities has an accompanying scalar log score, which could be represented as $\log f_{m,r,t,s,w}(z^*_{r,t,s,w}|\mathbf{x})$. These individual log scores can be averaged across combinations of regions, targets, seasons, and weeks to compare model performance.

To enhance interpretability, we report exponentiated average log scores which are the geometric mean of probability a model assigned to the value(s) eventually deemed to be accurate. In this manuscript, we refer to these as "average forecast scores". As an example, to determine the average forecast score for model m in season s (as shown in Figure 2), they are computed as

$$\begin{aligned} S_{m,\cdot,\cdot,s,\cdot} &= \exp\left(\frac{1}{N}\sum_{r,t,w}\log f_{m,r,t,s,w}(z^*_{r,t,s,w}|\mathbf{x})\right) \\ &= \left(\prod_{r,t,w} f_{m,r,t,s,w}(z^*_{r,t,s,w}|\mathbf{x})\right)^{1/N} \end{aligned} \tag{1}$$

As other forecasting efforts have used mean square error (MSE) or root mean square error (RMSE) as an evaluation method, we additionally evaluated the prospective forecasts received during the 2017-2018 season using RMSE. The submitted point forecast was used to score each component, and a point forecast was generated for each FSNetwork model by taking the median of the predicted distribution. For each model m, we calculated $RMSE_{m,t}$ for target t, averaging over all weeks w in the s =2017/2018 season and all regions r, as

$RMSE_{m,t} = \sqrt{\sum_{r,w}(\hat{z}_{m,r,t,w} - z^*_{r,t,w})^2}$, where $\hat{z}_{m,r,t,w}$ is the point prediction of model m for observed value $z^*_{r,t,w}$.

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. Teams submitted out-of-sample forecasts for the 2010/2011 through 2016/2017 influenza seasons. These models and their performance is evaluated in separate work.[26] Teams constructed their forecasts in a prospective fashion, 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 best 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.[36, 37, 38] 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 where 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 (Table 1). Seven of the models used a compartmental structure (i.e. Susceptible-Infectious-Recovered) to model the disease transmission process, while other models used more statistical approaches to directly model the observed wILI curve. Six of the models explicitly incorporated additional data sources beyond previous wILI data, including weather data and Google search data.

Additionally, we obtained the predictive distributions from the CDC-created "unweighted average" model. This ensemble combined all forecast models received by the CDC in real-time in the 2015/2016 (14 models), 2016/2017 (28 models), and 2017/2018 (33 models) seasons.[24] These included models that are not part of the collaborative ensemble effort described in this manuscript, although some variations on the components presented here were also submitted to the CDC. Including this model allowed us to compare our ensemble accuracy to the model used by the CDC in real-time during these three seasons.

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 were built using a method that combines component predictive distributions or densities using weighted averages. In the literature, this approach has been called stacking[14] or weighted density ensembles[19], and is similar to methods used in Bayesian model averaging[16]. Let $f_c(z_{t,r,w})$ represent the predictive density of ensemble component c for the value of the target $Z_{t,r,w}$, where t indexes the particular target, r indexes the region, and w indexes the week. We combine these components together into an ensemble model $f(z_{t,r,w})$ as follows:

$$f(z_{t,r,w}) = \sum_{c=1}^{C} \pi_{c,t,r} f_c(z_{t,r,w}) \quad (2)$$

where $\pi_{c,t,r}$ is the weight assigned to component c for predictions of target t in region r. 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 ensemble weighting schemes were considered, with varying complexity and number of estimated weights (Table 1).

- Equal Weight (`FSNetwork-EW`): This model consisted of assigning all components the same weight regardless of performance and is equivalent to an equally weighted probability density mixture of the components: $\pi_{c,t,r} = 1/C$.
- Constant Weight model (`FSNetwork-CW`): The weights vary across components but have the same value for all targets and regions, for a total of 21 weights: $\pi_{c,t,r} = \pi_c$. For purposes of statistical estimation, we say that the degrees of freedom (df) is $(21 - 1) = 20$. For each set of weights, once 20 weights are estimated the 21st is determined since they must add up to 1.
- Target Type Weight model (`FSNetwork-TTW`): Weights are estimated separately for our two target-types (tt), short-term and seasonal targets, with no variation across regions. This results in a total of 42 weights (df=40): $\pi_{c,t,r} = \pi_{c,tt}$.
- Target Weight model (`FSNetwork-TW`): The weights are estimated separately for each of the seven targets for each component with no variation across regions, resulting in 147 weights (df=140): $\pi_{c,t,r} = \pi_{c,t}$.
- Target-Region Weight model (`FSNetwork-TRW`): The most complex model considered, this model estiamted weights separately for each component-target-region combination, resulting in 1617 unique weights (df=1540): $\pi_{c,t,r} = \pi_{c,t,r}$.

Weights were estimated using the EM algorithm (see Supplement). Weights for components 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 available for cross-validation, 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.

4.7 Ensemble evaluation

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. This criteria for this choice were pre-specified in September of 2017, prior to conducting the cross-validation experiments.[32] Component weights for

the FSNetwork-TTW model were estimated using all seven seasons of training data. In real-time 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 remained 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 an 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. In fact, we saw this in our results since the Delphi-Stat model was the top-performing component model but was a linear combination of other Delphi models. It received zero weight in all of our ensemble specifications. Additionally, components can be assigned small weights but have a large impact on ensemble accuracy compared to an ensemble excluding it.

Reproducibility and data availability

To maximize the reproducibility and data availability for this project, the data and code for the entire project are publicly available. The project is available on GitHub[39], with a permanent repository stored on Zenodo[40]. Code for specific models are either publicly available or available upon request from the modeling teams, with more model-specific details available at the related citations (see Table 1). Retrospective and real-time forecasts from the FluSight Network may be interactively browsed on the website http://flusightnetwork.io. Additionally, this manuscript was dynamically generated using R version 3.5.1 (2018-07-02), Sweave, knitr, and make. These tools enable the intermingling of manuscript text with R code that run the central analyses, automatically regenerate parts of the analysis that have changed, and minimize the chance for errors in transcribing or translating results.[41, 42].

References

[1] Centers for Disease Control and Prevention. Estimated Influenza Illnesses, Medical visits, Hospitalizations, and Deaths in the United States – 2017-2018 influenza season | Seasonal Influenza (Flu) | CDC; 2018. Available from: https://www.cdc.gov/flu/about/burden/estimates.htm.

[2] Overview of Influenza Surveillance in the United States; 2017. https://www.cdc.gov/flu/weekly/overview.htm. Available from: https://www.cdc.gov/flu/weekly/overview.htm.

[3] Estimated Influenza Illnesses, Medical Visits, Hospitalizations, and Deaths Averted by Vaccination in the United States; 2018. https://www.cdc.gov/flu/about/disease/2016-17.htm. Available from: https://www.cdc.gov/flu/about/disease/2016-17.htm.

[4] Shaman J, Karspeck A, Yang W, Tamerius J, Lipsitch M. Real-time influenza forecasts during the 2012-2013 season. Nature Communications. 2013;4.

[5] Yang W, Karspeck A, Shaman J. Comparison of Filtering Methods for the Modeling and Retrospective Forecasting of Influenza Epidemics. PLoS Computational Biology. 2014;10(4):e1003583. doi:10.1371/journal.pcbi.1003583.

[6] Yang S, Santillana M, Kou SC. Accurate estimation of influenza epidemics using Google search data via ARGO. Proceedings of the National Academy of Sciences of the United States of America. 2015;112(47):14473–8. doi:10.1073/pnas.1515373112.

[7] Chretien JP, George D, Shaman J, Chitale RA, McKenzie FE. Influenza Forecasting in Human Populations: A Scoping Review. PLOS ONE. 2014;9(4):1–8. doi:10.1371/journal.pone.0094130.

[8] Kandula S, Hsu D, Shaman J. Subregional Nowcasts of Seasonal Influenza Using Search Trends. Journal of medical Internet research. 2017;19(11):e370. doi:10.2196/jmir.7486.

[9] Osthus D, Gattiker J, Priedhorsky R, Del Valle SY. Dynamic Bayesian Influenza Forecasting in the United States with Hierarchical Discrepancy. arXiv. 2017;.

[10] Brooks LC, Farrow DC, Hyun S, Tibshirani RJ, Rosenfeld R. Nonmechanistic forecasts of seasonal influenza with iterative one-week-ahead distributions. PLOS Computational Biology. 2018;14(6):e1006134. doi:10.1371/journal.pcbi.1006134.

[11] Pei S, Kandula S, Yang W, Shaman J. Forecasting the spatial transmission of influenza in the United States. Proceedings of the National Academy of Sciences of the United States of America. 2018;115(11):2752–2757. doi:10.1073/pnas.1708856115.

[12] Polikar R. Ensemble based systems in decision making. IEEE Circuits and Systems Magazine. 2006;doi:10.1109/MCAS.2006.1688199.

[13] Hastie T, Tibshirani R, Friedman J. Springer Series in Statistics; 2009.

[14] Wolpert DH. Stacked generalization. Neural Networks. 1992;5(2):241–259.

[15] Palmer TN. Predicting uncertainty in numerical weather forecasts. International Geophysics. 2002;doi:10.1016/S0074-6142(02)80152-8.

[16] Raftery AE, Gneiting T, Balabdaoui F, Polakowski M. Using Bayesian Model Averaging to Calibrate Forecast Ensembles. Monthly Weather Review. 2005;133(5):1155–1174.

[17] Leutbecher M, Palmer TN. Ensemble forecasting. Journal of Computational Physics. 2008;doi:10.1016/j.jcp.2007.02.014.

[18] Yamana TK, Kandula S, Shaman J. Individual versus superensemble forecasts of seasonal influenza outbreaks in the United States. PLOS Computational Biology. 2017;13(11):e1005801. doi:10.1371/journal.pcbi.1005801.

[19] Ray EL, Reich NG. Prediction of infectious disease epidemics via weighted density ensembles. PLOS Computational Biology. 2018;14(2):e1005910. doi:10.1371/journal.pcbi.1005910.

[20] Smith ME, Singh BK, Irvine MA, Stolk WA, Subramanian S, Hollingsworth TD, et al. Predicting lymphatic filariasis transmission and elimination dynamics using a multi-model ensemble framework. Epidemics. 2017;18:16–28. doi:10.1016/j.epidem.2017.02.006.

[21] Viboud C, Sun K, Gaffey R, Ajelli M, Fumanelli L, Merler S, et al. The RAPIDD ebola forecasting challenge: Synthesis and lessons learnt. Epidemics. 2017;doi:10.1016/J.EPIDEM.2017.08.002.

[22] DARPA. CHIKV Challenge Announces Winners, Progress toward Forecasting the Spread of Infectious Diseases; 2015. https://www.darpa.mil/news-events/2015-05-27. Available from: https://www.darpa.mil/news-events/2015-05-27.

[23] NOAA, CDC. Dengue Forecasting;. Available from: http://dengueforecasting.noaa.gov/about.php.

[24] McGowan C, Biggerstaff M, Johansson MA, Apfeldorf K, Ben-Nun M, Brooks L, et al. Collaborative efforts to forecast seasonal influenza in the United States, 2015-2016. Nature Scientific Reports. 2018;.

[25] Biggerstaff M, Kniss K, Jernigan DB, Brammer L, Bresee J, Garg S, et al. 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. 2018;187:1040–1050.

[26] Reich NG, Brooks L, Fox S, Kandula S, McGowan C, Moore E, et al. Forecasting seasonal influenza in the U.S.: A collaborative multi-year, multi-model assessment of forecast performance. bioRxiv. 2018;doi:10.1101/397190.

[27] Pei S, Shaman J. Counteracting structural errors in ensemble forecast of influenza outbreaks. Nature Communications. 2017;8(1):925. doi:10.1038/s41467-017-01033-1.

[28] Brooks LC, Farrow DC, Hyun S, Tibshirani RJ, Rosenfeld R. epiforecast: Tools for forecasting semi-regular seasonal epidemic curves and similar time series; 2015. https://github.com/cmu-delphi/epiforecast-R.

[29] Brooks LC, Farrow DC, Hyun S, Tibshirani RJ, Rosenfeld R. Flexible Modeling of Epidemics with an Empirical Bayes Framework. PLOS Computational Biology. 2015;11(8):e1004382. doi:10.1371/journal.pcbi.1004382.

[30] Osthus D, Gattiker J, Priedhorsky R, Del Valle SY. Dynamic Bayesian Influenza Forecasting in the United States with Hierarchical Discrepancy. Bayesian Analysis. 2018;doi:10.1214/18-BA1117.

[31] Ray EL, Sakrejda K, Lauer SA, Johansson MA, Reich NG. Infectious disease prediction with kernel conditional density estimation. Statistics in Medicine. 2017;doi:10.1002/sim.7488.

[32] Reich N. Guidelines for a CDC FluSight ensemble (2017-2018); 2017. https://github.com/FluSightNetwork/cdc-flusight-ensemble/blob/eadf553fcf85d89e16322ef1b44bc9990fc9e0a7/README.md. Available from: https://github.com/FluSightNetwork/cdc-flusight-ensemble/blob/eadf553fcf85d89e16322ef1b44bc9990fc9e0a7/README.md.

[33] Angus JE. The probability integral transform and related results. SIAM review. 1994;36(4):652–654.

[34] Diebold FX, Gunther TA, Tay A. Evaluating density forecasts; 1997.

[35] Farrow DC, Brooks LC, Hyun S, Tibshirani RJ, Burke DS, Rosenfeld R. A human judgment approach to epidemiological forecasting. PLoS computational biology. 2017;13(3):e1005248.

[36] New Mexico Department of Health. Indicator-Based Information System for Public Health Web; 2018. https://ibis.health.state.nm.us/resource/MMWRWeekCalendar.html. Available from: https://ibis.health.state.nm.us/resource/MMWRWeekCalendar.html.

[37] Niemi J. MMWRweek: Convert Dates to MMWR Day, Week, and Year; 2015. https://CRAN.R-project.org/package=MMWRweek. Available from: https://CRAN.R-project.org/package=MMWRweek.

[38] Tushar A. pymmwr: MMWR weeks for Python; 2018. https://pypi.org/project/pymmwr/. Available from: https://pypi.org/project/pymmwr/.

[39] Tushar A, Reich N, Yamana T, Osthus D, McGowan C, Ray E, et al.. FluSightNetwork: cdc-flusight-ensemble repository; 2018. https://github.com/FluSightNetwork/cdc-flusight-ensemble. Available from: https://github.com/FluSightNetwork/cdc-flusight-ensemble.

[40] Tushar A, Reich N, Yamana T, Osthus D, McGowan C, Ray E, et al.. FluSightNetwork/cdc-flusight-ensemble v1.0; 2018. https://doi.org/10.5281/zenodo.1255023. Available from: https://doi.org/10.5281/zenodo.1255023.

[41] Xie Y. Dynamic Documents with R and knitr. 2nd ed. Boca Raton, Florida: Chapman and Hall/CRC; 2015. Available from: https://yihui.name/knitr/.

[42] R Core Team. R: A Language and Environment for Statistical Computing; 2017. https://www.R-project.org/. Available from: https://www.R-project.org/.