Forecasting Influenza in the U.S. with a Collaborative Ensemble from the FluSight Network: 2019/2020 edition

Nicholas Reich, Nutcha Wattanachit, Tom McAndrew, Evan Ray

November 2019

Abstract

Problem: Our aim is to combine forecasting models for seasonal influenza in the US to create a single ensemble forecast. The central question is **can we provide better information to decision makers by combining forecasting component models**, and specifically by using past performance of the components to inform the ensemble approach.

Materials and Methods: The FluSight Network is a multi-institution and multi-disciplinary consortium of teams that have participated in past CDC FluSight challenges. In the 2017/2018 and 2018/2019 seasons the FluSight Network provided one of the most accurate influenza forecasting models to the CDC. Prior to the start of the 2019/2020 influenza season in the US, we assembled 27 distinct forecasting component models for influenza, each with forecasts from the last nine influenza seasons in the US. Subsequently, we conducted a cross-validation study to compare five different methods for combining these components into a single ensemble forecast.

Conclusions: Across the past nine seasons, three of our collaborative ensemble methods had higher average scores than any of the individual forecasting models. In addition, last year our team's prospective forecasts were the third most accurate forecasts among 33 submitted to and scored by the CDC. Based on updated models for the 2019/2020 season, we chose the best performing ensemble model and are submitting forecasts from this model each week to the CDC 2019/2020 FluSight Challenge.

Executive Summary

In the 2017/2018 and 2018/2019 influenza seasons, the CDC ran the 5th and 6th annual FluSight competition. In each year, they received over 25 model submissions from over 20 teams. The FluSight Network collaborative ensemble model, a weighted combination of models based on past performance, was one of the top three most accurate models during each of the last two seasons. In both years it outperformed an ensemble model built by analysts at the CDC that combined all of the submitted models by taking the average forecast for each influenza target.

In the 2019/2020 influenza season, we have 8 teams submitting 27 unique component models using a diverse array of methodologies. This document provides a high-level overview of the effort to choose a single ensemble model to represent the FluSight Network in real-time during the 2019/2020 U.S. influenza season.

FluSight Network Participation and Results

Season	# of teams	# of components	final rank
2017/2018	4	21	2
2018/2019	5	21	3
2019/2020	8	27	?

FluSight Network Participants for 2019/2020 season

Institution	No. of components	Team members
Delphi team at Carnegie Mellon	6	Logan C. Brooks, Aaron Rumack, David C. Farrow, Sangwon Hyun, Shannon, Gallagher, Ryan J. Tibshirani, Roni Rosenfeld, Rob Tibshirani
Columbia University	7	J Shaman, T Yamana, S Kandula, S Pei, W Yang, H Morita
FluOutlook (Northeastern and Harvard)	2	Alessandro Vespignani, Qian Zhang, Xinyue Xiong, Mauricio Santillana, Fred Lu
FluX (UVA)	2	Aniruddha Adiga, Lijing Wang, Srini Venkatramanan, Bryan Lewis, Jiangzhuo, Chen, Anil Vullikanti, Madhav Marathe
Los Alamos National Laboratory	1	Dave Osthus, Reid Priedhorsky
Protea Analytics	3	Craig J. McGowan, Alysse J. Kowalski
Reich Lab at UMass-Amherst	5	Nicholas G Reich, Evan Ray, Graham C. Gibson
U Arizona	1	Hannah Biegel, Joceline Lega

Choosing an Ensemble Model for Real-time Influenza Forecasting

In late October 2019, the FluSight Network ensemble team (Nick Reich, Nutcha Wattanachit, Tom McAndrew and Evan Ray) chose a single model to submit to the CDC throughout the 2019/2020 influenza season. This model had the highest out-of-sample overall score among all individual component models and ensemble models examined (Figure 1). This model is called the "target-type weight" (TTW) model because it assigns each model an individual weight for each of the "seasonal" and "week-ahead" target types (see next section for details). Forecasts from this model have been, and continue to be, submitted to the CDC in real-time, starting on October 28, 2019. They may be viewed at a public website by visiting: https://flusightnetwork.io.

We used a version of the EM algorithm (for details, see McAndrew and Reich (2019)) to estimate an optimal distribution of weights for the 27 different component models (Figure 2). For example, the `DBMplus` model from the LANL team is given 52% of the weight when creating ensemble forecasts for each of the seasonal targets and 23% of the weight for forecasts of week-ahead ILI.

While the weights are optimized to choose the best combination, they should not be interpreted as a ranking of models. For example, if two models make very similar forecasts but one is always a little bit better, it is possible that the slightly worse model will receive very little weight since most of the information it has to contribute is already contained in the better model. Typically, the ensemble will choose a set of models that contribute different information to the forecast, as this "diversity of opinion" will improve the forecast.

A few observations

1. Five different teams are represented in the top five most heavily-weighted components in the ensemble. However, all components from three different teams receive negligible weight.
2. For some components, the distribution of log scores is quite skewed (Figure 3), such that the "average" score is quite low however the component still receives quite a bit of weight. This highlights a point we have made in previous work: that weights should not be interpreted as a "ranking" of the component models.

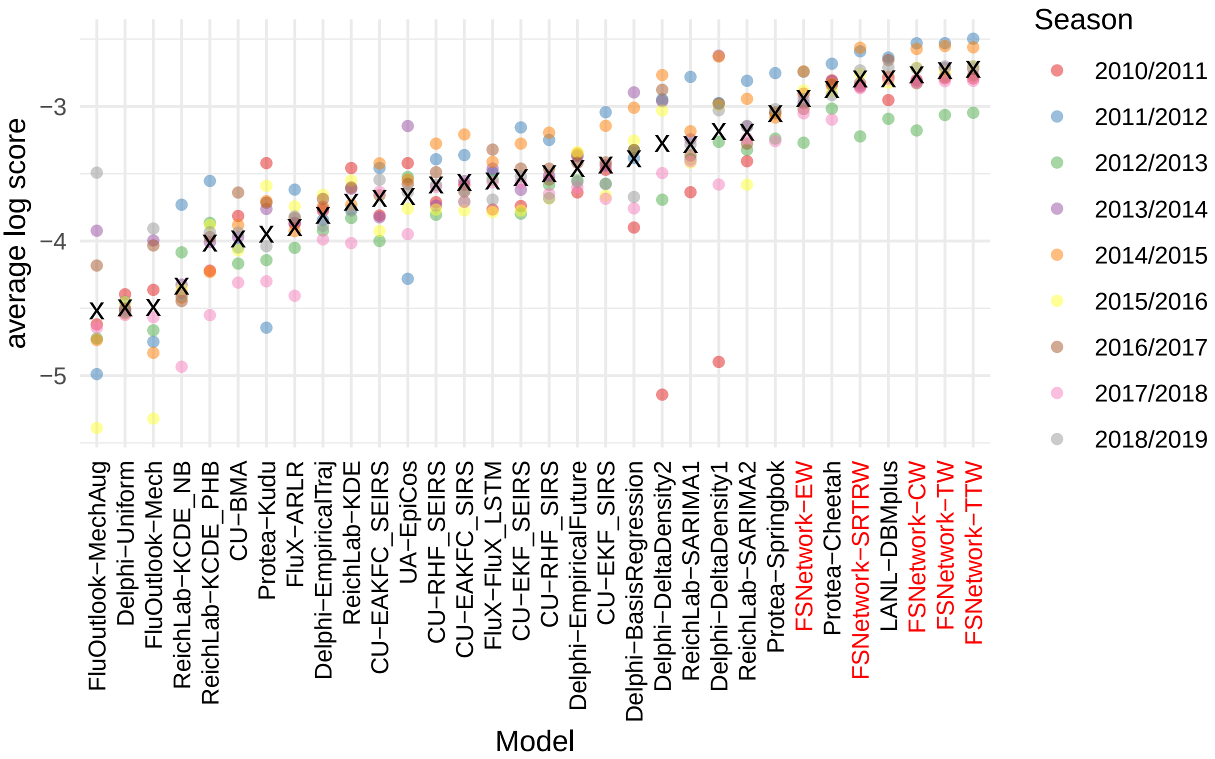

Figure 1: Average performance for all models, by season. The average log score for each component model within a season is plotted with a colored dot. The average across all seasons is shown with a black 'x'. Higher values indicate more accurate predictive performance, as they are a measure of how much probability on average a forecast from the given model assigned to the eventually observed value. The FluSightNetwork ensemble models are highlighted in red text. Components are sorted left to right in order of increasing accuracy.

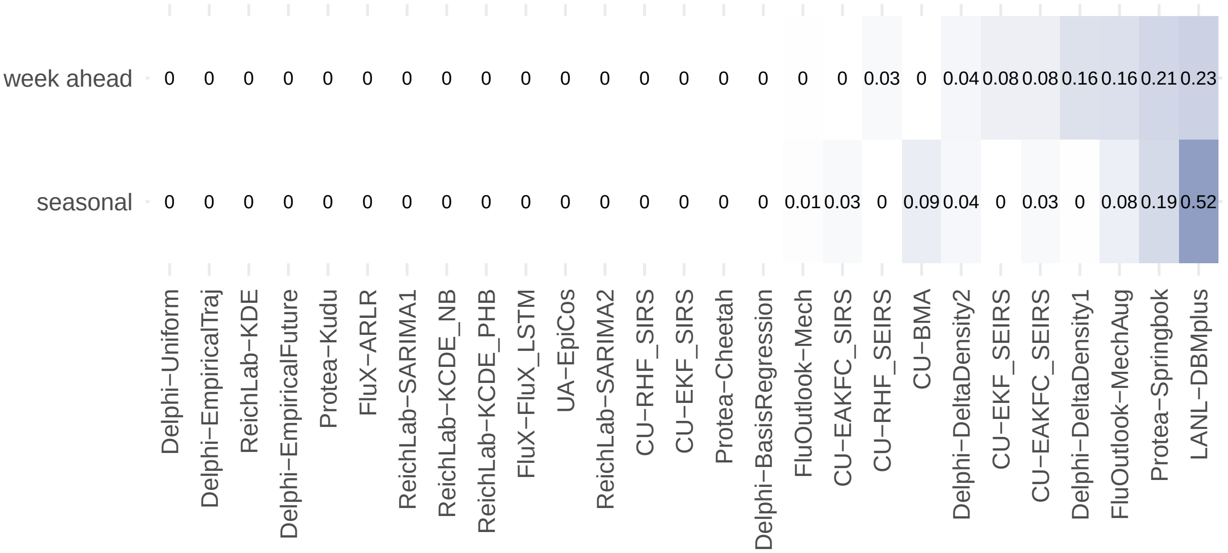

Figure 2: Estimated component weights by target-type. The number in each cell corresponds to the weight assigned to the component in each column and the target-type in each row. The weights in each row sum to 1. These weights are used to create the weighted average ensemble model for the 2019/2020 season.

Technical details

Targets: For every week in a season, each component model submission contains forecasts for seven targets of public health interest specified by the CDC for each of the 11 HHS regions. The region-level targets are: weighted influenza-like-illness (wILI) in each of the next four weeks of the season, the week of season onset, the week in which the peak wILI occurs, and the level of the peak wILI.

Ensemble specifications: All of our ensemble models are built by taking weighted averages of the component models. We examined the performance of five different possible ensemble specifications (see table below). The "equal weights" model takes a simple average of all of the models, with no consideration of past performance. The other four approaches estimated weights for models based on past performance. The specification with regularized weights (SRTRW) used methods for shrinking estimated weights towards equal weights as described in McAndrew and Reich (2019).

Model	No. of weights	description
Equal weights (EW)	1	Every model gets same weight.
Constant weights (CW)	27	Every model gets a single weight, not necessarily the same.
Target-type-based weights (TTW)	54	Two sets of weights, one for seasonal targets and one for weekly wILI targets.
Target-based weights (TW)	189	Seven sets of weights, one for each target separately.
Static regularized target-region-based weights (SRTRW)	2,079	Target-based weights estimated separately for each region with shrinkage towards equal weights.

Forecast Evaluation: We measured performance by comparing the average log-score across all targets and all relevant weeks in the last nine seasons. The log-score is taken as the "single-bin" log score, that is the natural log of the probability assigned to the outcome bin (as defined by the FluSight submission template and guidelines) in which the eventually observed true values lies. Scores from past seasons were included only if they lie in the "region of interest" according to CDC definitions (e.g., scores 6 weeks after onset occurred were excluded). Log-scores lower than -10 were truncated to be -10. Log scores were averaged on the log-scale and then exponentiated to represent a "forecast skill" measure that is on the probability scale.

For submitting in real-time in 2019/2020, we selected the ensemble model that achieved the best overall score in a cross-validation experiment over the last nine seasons. This was the target-type-based model (TTW) that assigned one set of weights to each component model for each target type (seasonal and week-ahead) separately.

References

McAndrew, Thomas, and Nicholas G. Reich. 2019. "Adaptively Stacking Ensembles for Influenza Forecasting with Incomplete Data." *arXiv*.

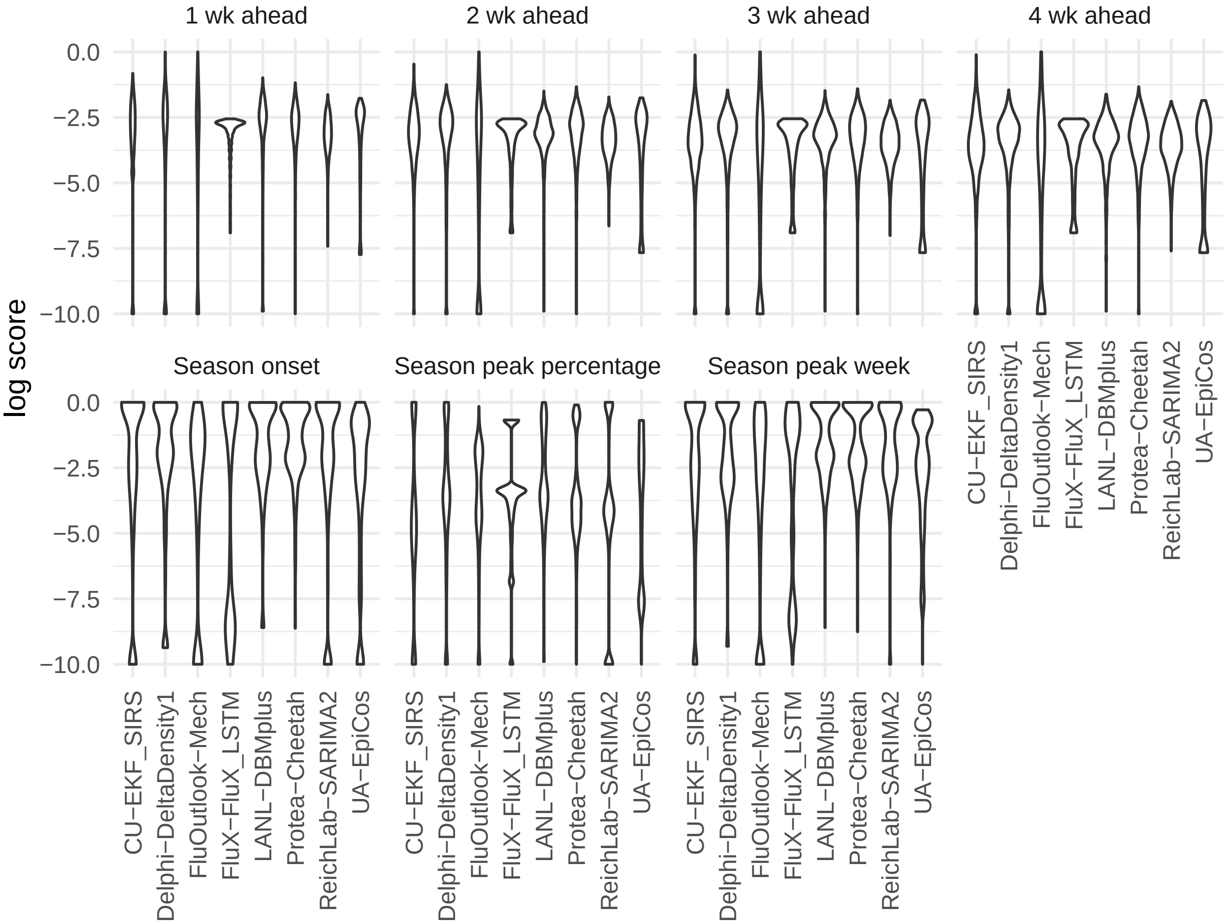

Figure 3: Densities of raw log-scores by model and target. The top model for each team is shown.