Forecasting Influenza in the U.S. with a Collaborative Ensemble from the FluSight Network: 2018/2019 edition

Nicholas Reich, Logan Brooks, Katie House, Sasikiran Kandula, Tom McAndrew, Craig McGowan, Dave Osthus, Evan Ray, Nutcha Wattanachit, Teresa Yamana, Jeff Shaman, Roni Rosenfeld

October 2018

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 models**, and specifically by using past performance of the models 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 season the FluSight Network provided one of the most accurate influenza forecasting models to the CDC. Prior to the start of the 2018/2019 influenza season in the US, we assembled 21 distinct forecasting models for influenza, each with forecasts from the last eight influenza seasons in the US. Subsequently, we conducted a cross-validation study to compare five different methods for combining these models into a single ensemble forecast.

Conclusions: Across the past eight seasons, four 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 second most accurate forecasts among 29 submitted to the CDC. Based on updated models for the 2018/2019 season, we chose the best performing ensemble model and are submitting forecasts from this model each week to the CDC 2018/2019 FluSight Challenge.

Executive Summary

In the 2017/2018 influenza season, the CDC ran the 5th annual FluSight competition and received 29 submissions from 20 teams. The FluSight Network collaborative ensemble model, a weighted combination of models based on past performance, was one of the top two most accurate models during this season, outperforming 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.

FluSight Network Participants for 2018/2019 season

Institution	No. of models	Team leaders
Delphi team at Carnegie Mellon	6	Logan Brooks, Roni Rosenfeld
Columbia University	7	Teresa Yamana, Sasikiran Kandula, Jeff Shaman
Los Alamos National Laboratory	1	Dave Osthus, Reid Priedhorsky
Protea Analytics	3	Craig J. McGowan, Alysse J. Kowalski
Reich Lab at UMass-Amherst	4	Nicholas Reich, Evan Ray, Katie House, Tom McAndrew, Nutcha Wattanachit

In March 2017, a group of influenza forecasters who have participated in this challenge in past seasons established the FluSight Network, a multi-institution and multi-disciplinary consortium of forecasting teams. In its inaugural 2017/2018 influenza season, 4 groups participated in the FluSight Network and contributed 21 models using a diverse array of methodologies. This group worked throughout 2017 to create a set of guidelines and an

experimental design that would enable submission of a publicly available, multi-team, real-time submission of a collaborative ensemble model with validated and performance-based weights for each model (i.e. not a simple average of models). In the current 2018/2019 season, we have an updated set of 21 models from 5 teams (some but not all of the models have been updated, removed, or changed from last year) to use in the collaborative ensemble.

This document provides a high-level overview of that effort, highlighting the results and documenting the chosen model that was designated for real-time submission during the 2018/2019 U.S. influenza season.

Choosing an Ensemble Model for Real-time Influenza Forecasting

In late October 2018, the FluSight Network team chose a single model to submit to the CDC throughout the 2018/2019 influenza season. This model had the highest overall score among all individual component models and ensemble models examined (Figure 1). This model is called the "target weight" (TW) model because it assigns each model an individual weight for each target (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 29, 2018. They may be viewed at a public website by visiting: https://flusightnetwork.io.

We used an algorithm to estimate an optimal distribution of weights for the 21 different component models (Figure 2). For example, the `Springbok` model from the Protea Analytics team is given 45% of the weight when creating ensemble forecasts for the week of season onset and 12% of the weight for forecasts of 1-week-ahead incidence.

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.

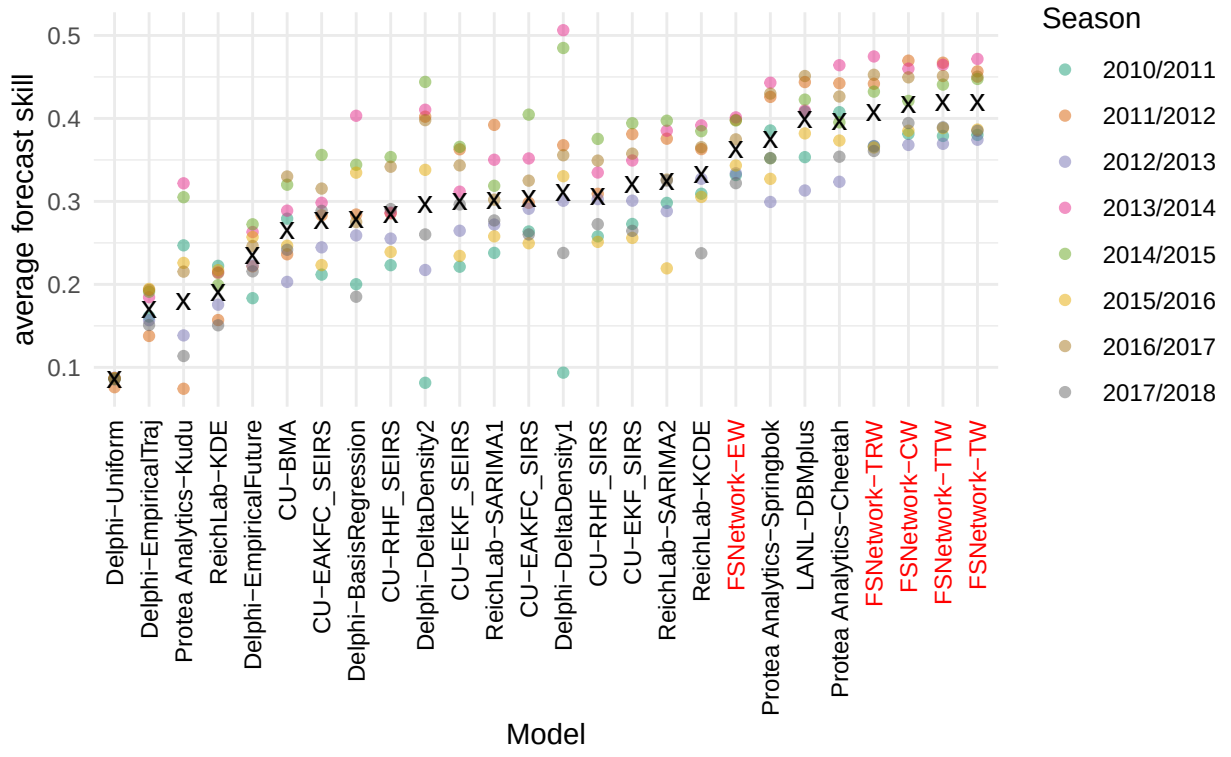

Figure 1: Average performance for all models, by season. The average forecast skill for each model within a season is plotted with a colored dot. The average across all seasons is shown with a black 'x'. Higher values indicate more accurate predictive performance, as they are a measure of how much probability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

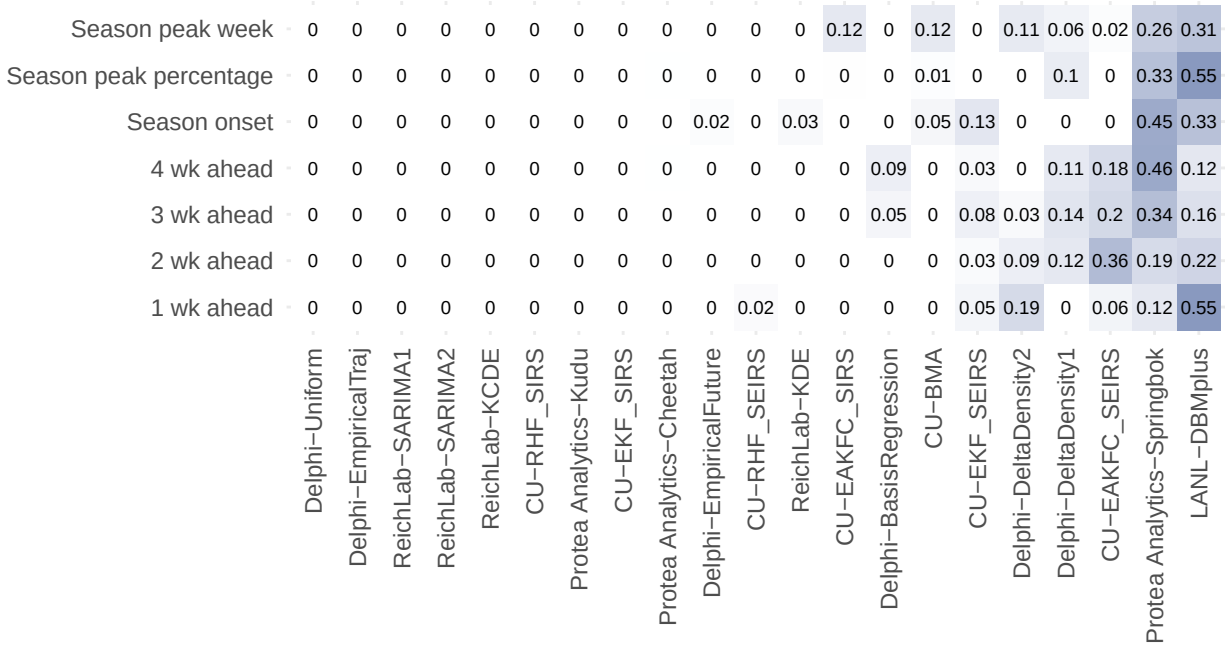

Figure 2: Estimated model weights by target. The number in each cell corresponds to the weight assigned to the model in each column and the target in each row. The weights in each row sum to 1. These weights are used to create the weighted average ensemble model for the 2017/2018 season.

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. Forecasts within 0.5 percentage points of the target wILI and within 1 week of the weekly targets are given full credit for having been "correct".

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.

Model	No. of weights	description
Equal weights (EW)	1	Every model gets same weight.
Constant weights (CW)	21	Every model gets a single weight, not necessarily the same.
Target-type-based weights (TTW)	42	Two sets of weights, one for seasonal targets and one for weekly wILI targets.
Target-based weights (TW)	147	Seven sets of weights, one for each target separately.
Target-and-region-based weights (TRW)	1,617	Target-based weights estimated separately for each region.

Forecast Evaluation: We measured performance by (1) comparing the average score across all targets and all relevant weeks in the last seven seasons and (2) comparing the variability in average score. The variability is important because a model can achieve good average performance by having a model that captures typical trends fairly well, but in a season showing unusual timing or dynamics it might fail. We want to ensure that we choose a model that shows good average performance but also is consistently good, especially in unusual seasons.

For submitting in real-time in 2018/2019, we selected the ensemble model that achieved the best overall score in a cross-validation experiment over the last seven seasons. This was the target-based model that assigned one set of weights to each component model for each target separately..