

Comparing Mechanistic and Statistical Models to Forecast Influenza in the U.S.

Logan Brooks, Spencer Fox, Craig McGowan, Sasikiran Kandula,
Dave Osthus, Evan Ray, Nicholas G Reich, Roni Rosenfeld, Jeffrey Shaman,
Abhinav Tushar, Teresa Yamana [authorship list to be finalized]

February 3, 2018

Contents

1	Introduction	1
2	Methods	1
2.1	FluSight Challenge Overview	1
2.2	Description of this Forecasting Experiment	3
2.3	Summary of Models	3
2.4	Metrics Used for Evaluation and Comparison	3
2.5	Formal comparisons of model performance	4
3	Results	4
3.1	Performance in forecasting week-ahead incidence	4
3.2	Performance in forecasting seasonal targets	4
3.3	Characterize skill and variability in skill of each model	4
3.4	Performance of models by location	4
3.5	Performance of models by target	7
3.6	Performance of models by time-of-season	7
3.7	Comparison between statistical and mechanistic models	7
4	Discussion	7
4.1	Overview of key results and importance	7
4.2	Overview of statistical vs. mechanistic model comparison	7
4.3	Limitations	7

1 Introduction

Forecasts of infectious disease outbreaks can inform public health response to outbreaks. Close collaboration between public health policy-makers and quantitative modelers is necessary to ensure the forecasts have maximum

impact and are appropriately communicated to the public and the broader public health community.

Infectious disease modeling has proven to be fertile ground for statisticians, mathematicians, and quantitative modelers for over a century. Yet there is not a consensus on a single best modeling approach or method for forecasting the dynamic patterns of infectious disease outbreaks, in both endemic and emergent settings. Mechanistic models consider the biological underpinnings of disease transmission, and are in practice typically implemented as variations on the Susceptible-Infectious-Recovered (SIR) model. Phenomenological models largely ignore the biological underpinnings and theory of disease transmission and focus instead on using data-driven, empirical and statistical approaches to make the best forecasts possible of a given dataset, or phenomenon. Both approaches are commonly used and both have advantages and disadvantages in different settings.

2 Methods

2.1 FluSight Challenge Overview

Starting in the 2013-2014 influenza season, the CDC has run the "Forecast the Influenza Season Collaborative Challenge" (a.k.a. FluSight) each influenza season, soliciting weekly forecasts for specific influenza season metrics from teams across the world. These forecasts are displayed together on a website during the season and are evaluated for accuracy after the season is over.[?] Detailed methodology and results from this challenge have been published[?], but summarize the key features of the challenge here.

The FluSight challenge has been focused on forecasts of the weighted percentage of doctor's office visits for influenza-like-illness (wILI) in a particular region. This is a standard measure of seasonal flu activity, for which public data is available back to the 1997/1998 influenza season. During each influenza season, this data is updated each week by the CDC (Figure 1). When the most recent data is released, the prior weeks' reported wILI data may also be revised. The unrevised data, available at a particular moment in time, is available via the DELPHI real-time epidemiological data API beginning in the 2013/2014 season.[?] This API enables researchers to "turn back the clock" to a particular moment in time and use the data available at that time. This enables more accurate assessment of how models would have performed in real-time.

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

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

To be included in the model comparison presented here, previous participants in the CDC FluSight challenge were invited to provide out-of-sample forecasts for the 2010/2011 through 2016/2017 seasons. For each model, this involved creating 233 separate forecast submission files, one for each of the weeks in the seven training seasons. Each forecast file represented a single submission file, as would be submitted to the CDC challenge. Each team

Forecast pipeline for Epidemic Week "k", or EW(k)

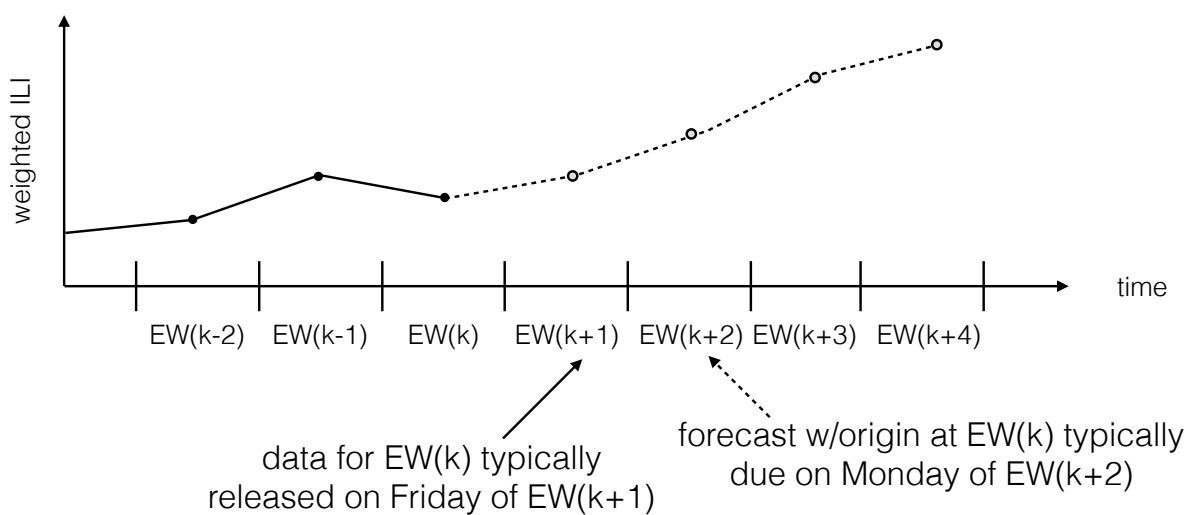


Figure 1: A schematic showing when data arrives in realtime relative to when the forecasts are made available.

created their submitted forecasts in a prospective, out-of-sample fashion, i.e. fitting or training the model only on data available before the time of the forecast (see Figure 1).

2.2 Description of this Forecasting Experiment

2.3 Summary of Models

Particular reference to the 2-3 models that can be considered “baseline” models: - Delphi-Uniform - ReichLab-KDE

2.4 Metrics Used for Evaluation and Comparison

Forecasts have historically been evaluated by the CDC using two metrics, the log-score and the mean absolute error. These two metrics capture different desirable features of performance. The log-score enables evaluation of both the precision and accuracy of a forecast, using the predicted density function.[?] The absolute error provides an interpretable summary of the amount of error the point estimates had on average.[?]

We used a modified form of the log-score to evaluate forecasts, in line with the evaluation performed by the CDC. The log-score is defined as $\log f(\hat{z}|\mathbf{x})$ where $f(z|\mathbf{x})$ is a predictive density function for some target z , conditional on some data \mathbf{x} and \hat{z} is the observed value of the target z . In practice, each model m has a set of log scores associated with it are region-, target-, season-, and week- specific, notated as $\log f_{r,t,s,w}^{(m)}(\hat{z}|\mathbf{x})$. We evaluated model performance based on the exponentiated average log scores, which has been called “forecast skill” and is equivalent to the geometric mean of the probabilities assigned to the eventually observed outcome. For example, the forecast skill for model m and target t would be calculated as

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

$$= \left(\prod_{r,s,w} \hat{f}_{r,t,s,w}^{(m)}(\hat{z}|\mathbf{x}) \right)^{1/N} \quad (2)$$

where N is the total number of log-scores for target t and model m , across all combinations of region, season, and week. Further, within a given region-season-target combination, the weeks included in the calculation of the average forecast skill depend on when the onset and peak occur. Specifically, [[...]]. All weeks are included for the forecast skill calculations for the k -step ahead forecasts of wILI.

The log-scores are computed for the targets on the wILI percentage scale such that predictions within ± 0.5 percentage points are considered accurate, i.e. $\log \text{score} = \log \int_{\hat{z}-.5}^{\hat{z}+.5} f^{(m)}(z|\mathbf{x}) dz$. For the targets on the scale of epidemic weeks, predictions within ± 1 week are considered accurate, i.e. $\log \text{score} = \log \int_{\hat{z}-1}^{\hat{z}+1} f^{(m)}(z|\mathbf{x}) dz$.

- log-score for predictive distribution, aggregated by (model), (model x season), (model x season x location), (model x season x target-type), (model x season x target), , (model x season x week)
- MAE for point predictions

92 2.5 Formal comparisons of model performance

- 93 • permutation test for pairwise statistical comparison between two models
- 94 • beta regression or permutation test for comparison between groups of models (i.e. mechanistic vs. statisti-
95 cal)

96 3 Results

97 Things to confirm: removed weeks that CDC does not score, no onset seasons and multi peak years are handled
98 appropriately

99 3.1 Performance in forecasting week-ahead incidence

- 100 • describe historical model average performance and how it varies (min/max?) across seasons/regions and
101 what that says about inter-season and inter-region variability
- 102 • what fraction of models did better than historical average/uniform.
- 103 • describe skill degradation has horizon increases
- 104 •

105 3.2 Performance in forecasting seasonal targets

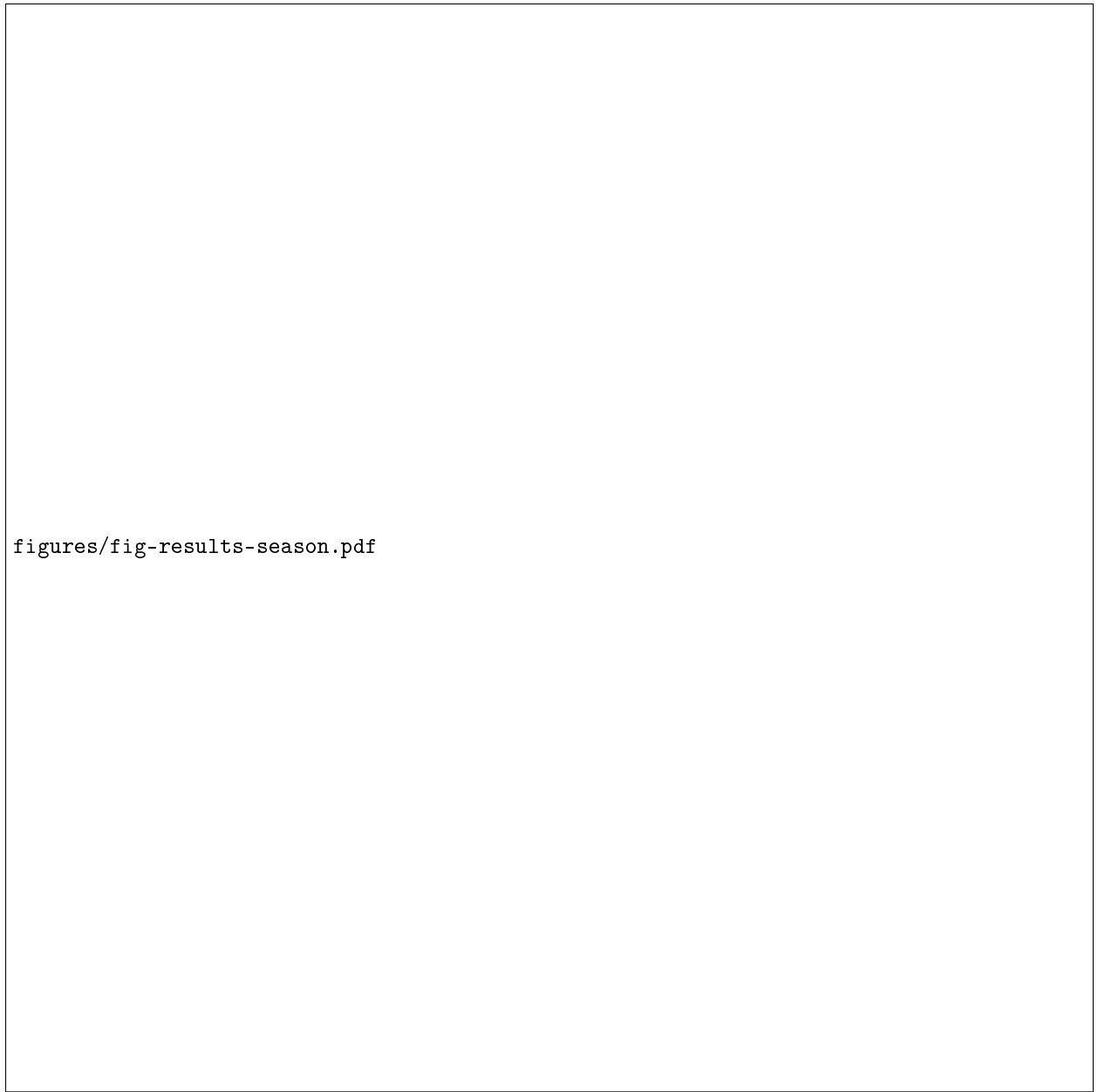
- 106 • describe historical model average performance and how it varies (min/max?) across seasons/regions/targets
- 107 • what fraction of models did better than historical average/uniform.
- 108 • What is hard about Region 2?

109 3.3 Characterize skill and variability in skill of each model

110 TODO: make point plot of skill by season to show variability and average performance.

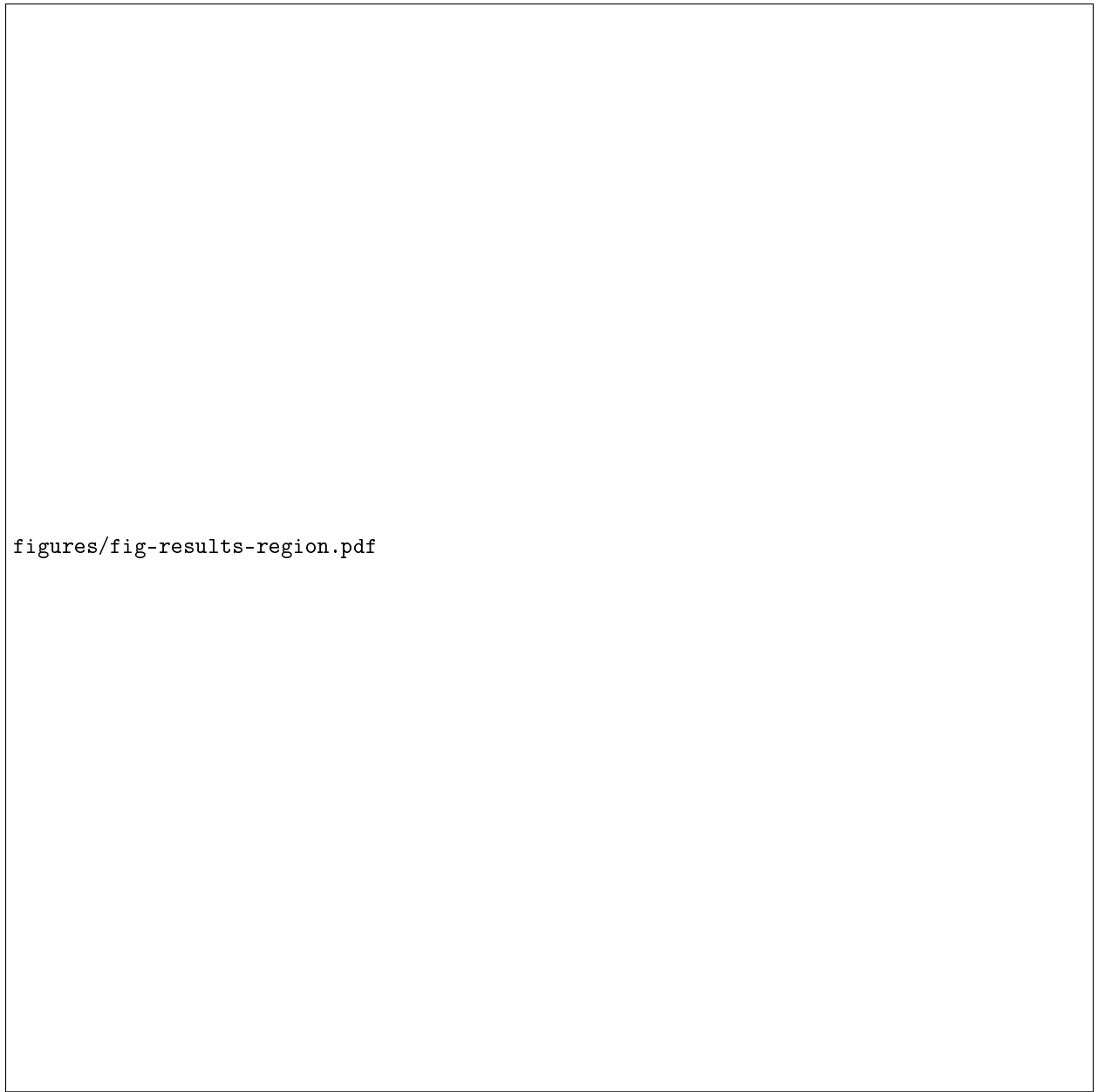
111 3.4 Performance of models by location

112 Consider adding map that averages across all models, or shows max skill per region, as it varies quite a bit.



figures/fig-results-season.pdf

Figure 2: Model results by season.



figures/fig-results-region.pdf

Figure 3: Model results by region and target-type.

113 3.5 Performance of models by target

114 3.6 Performance of models by time-of-season

115 3.7 Comparison between statistical and mechanistic models

116 4 Discussion

117 4.1 Overview of key results and importance

118 The first large-scale comparison of flu forecasting models from different modeling teams/philosophies across
119 multiple years.

120 4.2 Overview of statistical vs. mechanistic model comparison

121 As our knowledge/data about the system mature, we expect mechanistic models to be better, but when true
122 signals of mechanistic model is drowned out by observational noise or spatial aggregation, statistical models may
123 perform better. This comparison serves as a barometer for where the current state of forecast models are.

124 4.3 Limitations

- 125 • relatively few additional data sources incorporated
- 126 • no models that explicitly incorporate strain information
- 127 • no models with spatial information included
- 128 • seven seasons of data is not a lot ($n=7$) to draw strong conclusions about comparative model performance
- 129 • currently limited to models with only recent data...