

RESEARCH ARTICLE

# Nonmechanistic forecasts of seasonal influenza with iterative one-week-ahead distributions

Logan C. Brooks<sup>1\*</sup>, David C. Farrow<sup>1</sup>, Sangwon Hyun<sup>2</sup>, Ryan J. Tibshirani<sup>1,2</sup>, Roni Rosenfeld<sup>1</sup>

**1** School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania, United States of America, **2** Department of Statistics, Carnegie Mellon University, Pittsburgh, Pennsylvania, United States of America

\* [lcbrooks@cs.cmu.edu](mailto:lcbrooks@cs.cmu.edu)



## Abstract

Accurate and reliable forecasts of seasonal epidemics of infectious disease can assist in the design of countermeasures and increase public awareness and preparedness. This article describes two main contributions we made recently toward this goal: a novel approach to probabilistic modeling of surveillance time series based on “delta densities”, and an optimization scheme for combining output from multiple forecasting methods into an adaptively weighted ensemble. Delta densities describe the probability distribution of the change between one observation and the next, conditioned on available data; chaining together nonparametric estimates of these distributions yields a model for an entire trajectory. Corresponding distributional forecasts cover more observed events than alternatives that treat the whole season as a unit, and improve upon multiple evaluation metrics when extracting key targets of interest to public health officials. Adaptively weighted ensembles integrate the results of multiple forecasting methods, such as delta density, using weights that can change from situation to situation. We treat selection of optimal weightings across forecasting methods as a separate estimation task, and describe an estimation procedure based on optimizing cross-validation performance. We consider some details of the data generation process, including data revisions and holiday effects, both in the construction of these forecasting methods and when performing retrospective evaluation. The delta density method and an adaptively weighted ensemble of other forecasting methods each improve significantly on the next best ensemble component when applied separately, and achieve even better cross-validated performance when used in conjunction. We submitted real-time forecasts based on these contributions as part of CDC’s 2015/2016 FluSight Collaborative Comparison. Among the fourteen submissions that season, this system was ranked by CDC as the most accurate.

## OPEN ACCESS

**Citation:** Brooks LC, Farrow DC, Hyun S, Tibshirani RJ, Rosenfeld R (2018) Nonmechanistic forecasts of seasonal influenza with iterative one-week-ahead distributions. *PLoS Comput Biol* 14(6): e1006134. <https://doi.org/10.1371/journal.pcbi.1006134>

**Editor:** Cecile Viboud, National Institutes of Health, UNITED STATES

**Received:** May 12, 2017

**Accepted:** April 10, 2018

**Published:** June 15, 2018

**Copyright:** © 2018 Brooks et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Data Availability Statement:** The latest ILINet report is available from a Fluview Interactive web module (<https://gis.cdc.gov/grasp/fluview/fluportaldashboard.html>). Past ILINet reports (in addition to the latest one) are available from our delphi-epidata API (<https://github.com/cmu-delphi/delphi-epidata>). (Past and current Delphi-Stat, Delphi-Epicast, and Delphi-Archefilter forecasts, as well as ILI-Nearby nowcasts, are also available from our delphi-epidata API (<https://github.com/cmu-delphi/delphi-epidata>).)

**Funding:** LCB, DCF, and RR were supported by the National Institute Of General Medical Sciences of the National Institutes of Health under Award Number U54 GM088491. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant Nos. 0946825, DGE-1252522, and DGE-1745016. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing interests:** The authors have declared that no competing interests exist.

## Author summary

Seasonal influenza is associated with 250 000 to 500 000 deaths worldwide each year (WHO estimates). In the United States and other temperate regions, seasonal influenza epidemics occur annually, but their timing and intensity varies significantly; accurate and reliable forecasts that quantify their uncertainty can assist policymakers when planning countermeasures such as vaccination campaigns, and increase awareness and preparedness of hospitals and the general public. Starting with the 2013/2014 flu season, CDC has solicited, collected, evaluated, and compared weekly forecasts from external research groups. We developed a new method for forecasting flu surveillance data, which stitches together models of changes that happen each week, and a way of combining its output with other forecasts. The resulting forecasting system produced the most accurate forecasts in CDC's 2015/2016 FluSight comparison of fourteen forecasting systems. We describe our new forecasting methods, analyze their performance in the 2015/2016 comparison and on data from previous seasons, and describe idiosyncrasies of epidemiological data that should be considered when constructing and evaluating forecasting systems.

## Introduction

Seasonal influenza epidemics cause widespread illness which is associated each year with an estimated 250 000 to 500 000 deaths worldwide [1] and 3000 to 56 000 deaths in the United States alone [2–4]. In contrast to influenza “pandemics”, which are rare global outbreaks of especially novel influenza A viruses [5, 6], seasonal epidemics (i.e., non-pandemics), while still having worldwide reach, occur annually in the United States and other countries with (generally) temperate climates. Time series of influenza prevalence in these areas are typically low and flat for the majority of the season, but trace a single, sharp peak sometime during winter, with significant variability in timing and intensity. Accurate and reliable forecasts of seasonal epidemics can help policymakers plan countermeasures such as vaccination campaigns, and increase awareness and preparedness of hospitals and the general public. The Centers for Disease Control and Prevention (CDC) monitors influenza prevalence with several well-established surveillance systems [7]; the recurring nature of seasonal epidemics and availability of historical data provide promising opportunities for the formation, evaluation, and application of statistical models. Starting with the 2013/2014 “Predict the Influenza Season Challenge” [8] and continuing each season thereafter as the Epidemic Prediction Initiative’s FluSight project [9], CDC has solicited and compiled forecasts of influenza-like illness (ILI) prevalence from external research groups and worked with them to develop standardized forecast formats and quantitative evaluation metrics.

Various approaches to influenza epidemic forecasting are summarized in literature reviews [10–12] and descriptions of the CDC comparisons [8, 9]. Some common approaches are described below, with references to work applicable to the current FluSight project and related seasonal dengue forecasting tasks, emphasizing more recent work that may not be listed in the above three literature reviews:

- **Mechanistic models:** describe the disease state and interaction between individuals with causal models, as well as the surveillance data generation process.
- **Compartmental models** (e.g., [13–17]): break down the population into a number of discrete “compartments” describing their characteristics (e.g., age, location) and state (e.g., susceptible to, infectious with, or recovered from a particular disease), and describe how the occupancy of these compartments changes over time, either deterministically or

probabilistically. In many of these models, this division describes solely the state with respect to a single disease, ignoring details regarding age, spatial dynamics, and mixtures of ILI diseases, but keeping the number of parameters to infer low.

- **Agent-based models** (e.g., [11, 18]): also known as individual-based models, these approaches use more detailed descriptions of disease state and/or individual characteristics and behavior, which are not easily simplified into a compartmental form, typically studied using computation-heavy simulations. These approaches usually include many more parameters than compartmental models, which may be set or inferred by heuristics, additional data sources and studies, or Monte Carlo procedures.
- **Phenomenological models**: also referred to as statistical models, these approaches describe the surveillance data without directly incorporating the epidemiological underpinnings.
- **Direct regression models** (e.g., [19–22]): attempt to estimate future prevalence or targets of interest using various types of regression, including nonparametric statistical approaches and alternatives from machine learning literature.
- **Time series models** (e.g., [23–31]): represent the expected value of (transformations of) observations and/or underlying latent state at a particular time as (typically linear) functions of these quantities at previous times and additional covariates, paired with Gaussian, Poisson, negative binomial, or other noise distributions. This category includes linear dynamical systems and frameworks such as SARIMAX.

We present a novel phenomenological approach to epidemiological forecasting using “delta densities”, which assumes an autoregressive dependency structure similar to those of some time series models, but uses a kernel density estimation approach to model these dependencies rather than the common choice of linear relationships plus Gaussian noise. This technique is similar to the method of analogues [19] in that it uses an instance-based, nonparametric estimation procedure, but provides distributional forecasts of entire trajectories rather than point predictions of individual observations. The kernel conditional density estimation (KCDE) forecasting method [22] attacks many of the same issues encountered when applying kernel density estimation methods to seasonal epidemic data, but models the dependency structure of future weeks with a copula, while delta density chains together 1-week-ahead simulations. Compared to approaches that treat the entire season as a unit, such as deterministic, single-strain, fully-mixed compartmental models [13] or our previous empirical Bayes approach based on modifying past seasons’ data [21, 32], this method forms a larger library of possible trajectories by piecing together local models, which appears to help forecast performance, even though the trajectories considered may seem less reasonable on average.

Our second contribution is an adaptively weighted ensemble approach to combining the output of different forecasting methods given their historical and/or cross-validation forecasts. We first implemented this method in preparation for the 2014/2015 FluSight comparison, mixing together our empirical Bayes forecasting method with two baselines (a uniform distribution and an empirical distribution for each target), and later applied it while participating in the Dengue Forecasting project [33] and following FluSight comparisons (adding up to 9 additional components including delta density based methods), and found it improved our forecasts in all cases. Other epidemic forecasting teams have also reported success with concurrently or subsequently developed stacking generalization [34, 35] ensemble approaches to the FluSight forecasting tasks using Bayesian model averaging [36], the fixed weighting scheme that we examine below [37], and alternative adaptive weighting schemes based on gradient tree boosting [37], as well as with earlier ensemble approaches to short-term point predictions

[20]. Methodologically, our adaptively weighted ensemble framework differs from these alternatives in that it selects a weighting over components for a particular forecast using “plug-in” statistical estimators for the optimal weights given the context of the forecast being prepared. Like the adaptive approaches presented in [37], component weights for each forecast are selected using regression, but the type of regression used and the manner of incorporating additional information, such as the forecast week, are distinct.

## Materials and methods

### Surveillance data

Recording every case of influenza is not practicable; infections are often asymptomatic [38] or symptomatic but not clinically attended [39], laboratory testing may not be performed for clinically attended cases or give false negative results, and reporting of lab-confirmed cases is not mandatory in most instances. Forecast comparisons are instead based on syndromic clinical surveillance data from the U.S. Outpatient Influenza-like Illness Surveillance Network (ILINet) [7, 40], a group of health care providers that voluntarily report statistics regarding ILI, where ILI is defined as a 100°F (37.8°C) fever with a cough and/or sore throat without a known cause other than influenza. CDC aggregates these reports and estimates the weekly percentage of patients seen that have ILI, %ILI, across all health care providers using a measure called weighted %ILI (wILI).

- **Geographical resolution:** CDC reports wILI for each of the 10 U.S. Department of Health & Human Services (HHS) regions, as well as for the nation as a whole; the wILI for each of these locations is a weighted average of the ILINet %ILI for state-level units based on population.
- **Temporal resolution:** wILI is available on a weekly basis; weeks begin on Sunday, end on Saturday, and are numbered according to the epidemiological week (epi week) convention in the United States.
- **Timeliness:** Initial wILI estimates for a given week are typically released on Friday of the following week; additional reports and revisions from participating health care providers are incorporated in updates throughout the season and after it ends.
- **Specificity:** Influenza is just one of many potential causes of ILI. Laboratory testing data [7] suggest that influenza is responsible for a significant portion of ILI cases during the flu season, especially for weeks when wILI is high, but only for a very small fraction of cases in the typical flu off-season. Much of the variance and “peakiness” in wILI can be associated with influenza epidemics, but wILI trajectories do not taper off to near-zero values as one might expect in a direct measurement of influenza prevalence.
- **Influence of non-ILI cases:** Since wILI depends on records of both ILI cases and total cases, patterns in non-ILI cases can impact wILI trajectories. We discuss one such pattern in the Holiday effects section.

CDC hosts the latest ILINet report and other types of surveillance data through Fluview Interactive, a collection of web modules [41]; we provide current and historical ILINet reports and some other data sources through our `delphi-epidata` API [42] and `epiviz` visualizer [43].

### Forecasting targets

The FluSight project focuses on in-season distributional forecasts and point predictions of key targets of interest to public health officials:

- **Short-term wILI:** the four wILI values following the last available observation (incorporating all data revisions through some week well after the season's end)
- **Season onset:** the first week in the first run of at least three consecutive weeks with wILI values above a location- and season- specific baseline wILI level set by CDC [7], or "none" if no such runs exist; describes whether and when an influenza epidemic started in a given season
- **Season peak percentage:** the maximum of all wILI values for a given season
- **Season peak week:** the week or weeks in which wILI takes on its maximum value, or "none" if there was no onset in the 2015/2016 comparison

When making distributional forecasts, wILI values are discretized into CDC-specified bins and a probability assigned to each bin, forming a histogram over possible observations. The width of the bins was set at 0.5 %wILI for the 2015/2016 comparison and 0.1 %wILI for the 2016/2017 comparison; we use a width of 0.5 %wILI for analysis of the 2015/2016 comparison prospective forecasts, and a width of 0.1 %wILI for retrospective evaluation. CDC typically presents wILI values rounded to a resolution of 0.1 %wILI; some targets and evaluations are based on these rounded values.

## Evaluation metrics

We focus on three metrics for evaluating performance of a forecast for a given target:

- **Unibin log score:**  $\log \hat{p}_i$ , where  $\hat{p}_i$  is the probability assigned to  $i$ , the bin containing the observed value. We use this score for ensemble weight selection and most internal evaluation as it has ties to maximum likelihood estimation, and is "proper score" [44]. A score for a (reported) distributional prediction  $\hat{p}$  is called "proper" if its expected value according to any (internal) distributional prediction  $\hat{q}$  is maximized when reporting  $\hat{p} = \hat{q}$ , i.e., forecasters can maximize their expected scores by reporting their true beliefs. We refer to the "unibin log score" simply as the "log score" except for when comparing it with the multibin log score, which is defined next. The exponentiated mean log score is the (geometric) average probability assigned to events that were actually observed. The exponentiated difference in the mean log scores of method A and method B is an estimate of the (geometric) expected winnings of unit-sized bets of the form "this bin will hold the true value" when bets are placed optimally according to the forecasts of A, and (relative) prices are set optimally according to the forecasts of B.
- **Multibin log score:**  $\log \sum_{i \text{ near observed value}} \hat{p}_i$ , where the  $i$ 's considered are typically bins within 0.5 %wILI of observed values for a wILI target, or within 1 week for a timing target. The multibin log score was designed by FluSight hosts in consultation with participants, and the judgment "near observed value" was selected as a level of error that would not significantly impact policymakers' decisions. The exponentiated mean multibin log score is the (geometric) average amount of mass a forecaster placed within this margin for error of observed target values.
- **Absolute error:**  $|\hat{y} - y|$ , where  $\hat{y}$  is the point prediction and  $y$  is an observed value. (In the case of onset, we consider point predictions for the value of onset conditioned on the fact that an onset actually occurs. We do not consider absolute error for onset in instances where no onset occurred. Some methods considered would sometimes fail to produce such conditional onset point predictions when they were confident that there was no onset, but these methods are not included in any of the figures containing absolute errors.)

The FluSight 2015/2016 forecast comparison evaluations were based solely on the multibin log score [45].

### Terminology and notation

The “flu season” is typically defined as epi week 40 of one year through 20 of the next; we also include data from the rest of the year as part of the season for the purpose of fitting models. In all mathematical notation, we will number the first week of the season as 1 rather than using the corresponding epi week. Let

- $W_{1..t}^t$  denote the  $t$ -th CDC report of the current season, containing wILI values for weeks 1 through  $t$ , inclusive, which is normally published on Friday of week  $t + 1$ ;
- $T$  be the number of weeks in the current season (either 52 or 53); we omit all details regarding differing season lengths, presenting forecasting methods and labeling epi week plot axes as if all seasons were of length  $T$ ;
- $Y_{1..T}$  be the ground truth wILI for the current season: the wILI values used for forecast evaluation, specifically the epi week 28 report for the FluSight comparison, or later revisions as they are available for cross-validation analysis;
- $Y_{1..T}^s$  be the ground truth wILI for past season  $s$ ; and
- $Z^t$  be a vector containing the forecasting targets of interest at the  $t$ -th wILI report of the current season:  $Y_{t+1..t+4}$  and the seasonal onset, peak week, and peak percentage; for the FluSight comparison, forecasts for these targets were typically due on Monday of week  $t + 2$ , and allowed to use ILINet and any other data released before the deadline.

Our goal is to forecast  $Z^t$  given  $W_{1..t}^t$  and previous reports. This can be broken down into multiple steps, such as:

1. “Backcast” updates to the data through time  $t$ , producing a distribution over  $Y_{1..t}$  based on the value of  $W_{1..t}^t$  and previous reports.
2. Connect the backcast for  $Y_{1..t}$  with corresponding forecasts for  $Y_{t+1..T}$ , yielding a distribution for the entire trajectory  $Y_{1..T}$ .
3. Calculate the distribution for  $Z^t$  corresponding to this distribution over  $Y_{1..T}$ .

We first introduce the delta density method, which forecasts  $Y_{t+1..T}$  given  $Y_{1..t}$  (step 2). We then discuss a separate procedure for combining multiple forecasts into an adaptively weighted ensemble, forecasting  $Z^t$  given either  $Y_{1..t}$  or  $W_{1..t}^t$  (steps 2–3 or 1–3). We also outline a method for estimating the distribution of  $Y_{1..t}$  given  $W_{1..t}^t$  (step 1), and analyze its performance when used in conjunction with the delta density method.

### Delta density method

Consider the task of estimating the density function  $f_{Y_{t+1..T}|Y_{1..t}}$  using an instance-based approach. Kernel density estimation and kernel regression use smoothing kernels to produce flexible estimates of the density of a random variable (e.g.,  $f_{Y_{t+1..T}}$ ) and the conditional expectation of one random variable given the value of another (e.g.,  $\mathbb{E}[Y_{t+1..T} | Y_{1..t}]$ ), respectively; we can combine these two methods to obtain estimates of the conditional density of one random variable given another. One possible approach would be to use the straightforward estimate

$$\hat{f}_{Y_{t+1..T}|Y_{1..t}}(y_{t+1..T} | y_{1..t}) = \frac{\sum_{s=1}^S I^{1..t}(y_{1..t}, Y_{1..t}^s) O^{t+1..T}(y_{t+1..T}, Y_{t+1..T}^s)}{\sum_{s=1}^S I^{1..t}(y_{1..t}, Y_{1..t}^s)}$$

where  $\{1..S\}$  is the set of fully observed historical training seasons, and  $I^{1..t}$  and  $O^{t+1..T}$  are smoothing kernels describing similarity between “input” trajectories and between “output” trajectories, respectively. However, while basic kernel smoothing methods can excel in low-dimensional settings, their performance scales very poorly with growing dimensionality. During most of the season, neither  $Y_{1..t}$  nor  $Y_{t+1..T}$  is low-dimensional, and the current season’s observations are extremely unlikely to closely match any past  $Y_{1..t}^s$  or  $Y_{t+1..T}^s$ . This, in turn, can lead to kernel density estimates for  $Y_{t+1..T}$  based almost entirely on the single season  $s$  with the closest  $Y_{1..t}^s$  when conditioning on  $Y_{1..t}$ , and excessively narrow density estimates for  $Y_{t+1..T}$  even without conditioning on  $Y_{1..t}$ . So, instead of applying kernel density estimation directly, we first break the task down into a sequence of low-dimensional sub-tasks. We avoid the high-dimensional output problem by chaining together estimates of  $f_{\Delta Y_u | Y_{1..u-1}}$  for each  $u$  from  $t + 1$  to  $T$ , where  $\Delta Y_u = Y_u - Y_{u-1}$ ; estimating these single-dimensional densities requires relatively little data. However, this reformulation exacerbates the high-dimensional input problem since we are conditioning on  $Y_{1..u-1}$ , which can be considerably longer than  $Y_{1..t}$ . We address the high-dimensional input problem by approximating  $f_{\Delta Y_u | Y_{1..u-1}}$  with  $f_{\Delta Y_u | \mathbf{R}^u}$ , where  $\mathbf{R}^u$  is some low-dimensional vector of features derived from  $Y_{1..u-1}$ . Smoothing kernel methods are used to approximate the conditional density functions using data from past seasons.

We use two sets of choices for the approximate conditional density function and summary features to form two versions of the method.

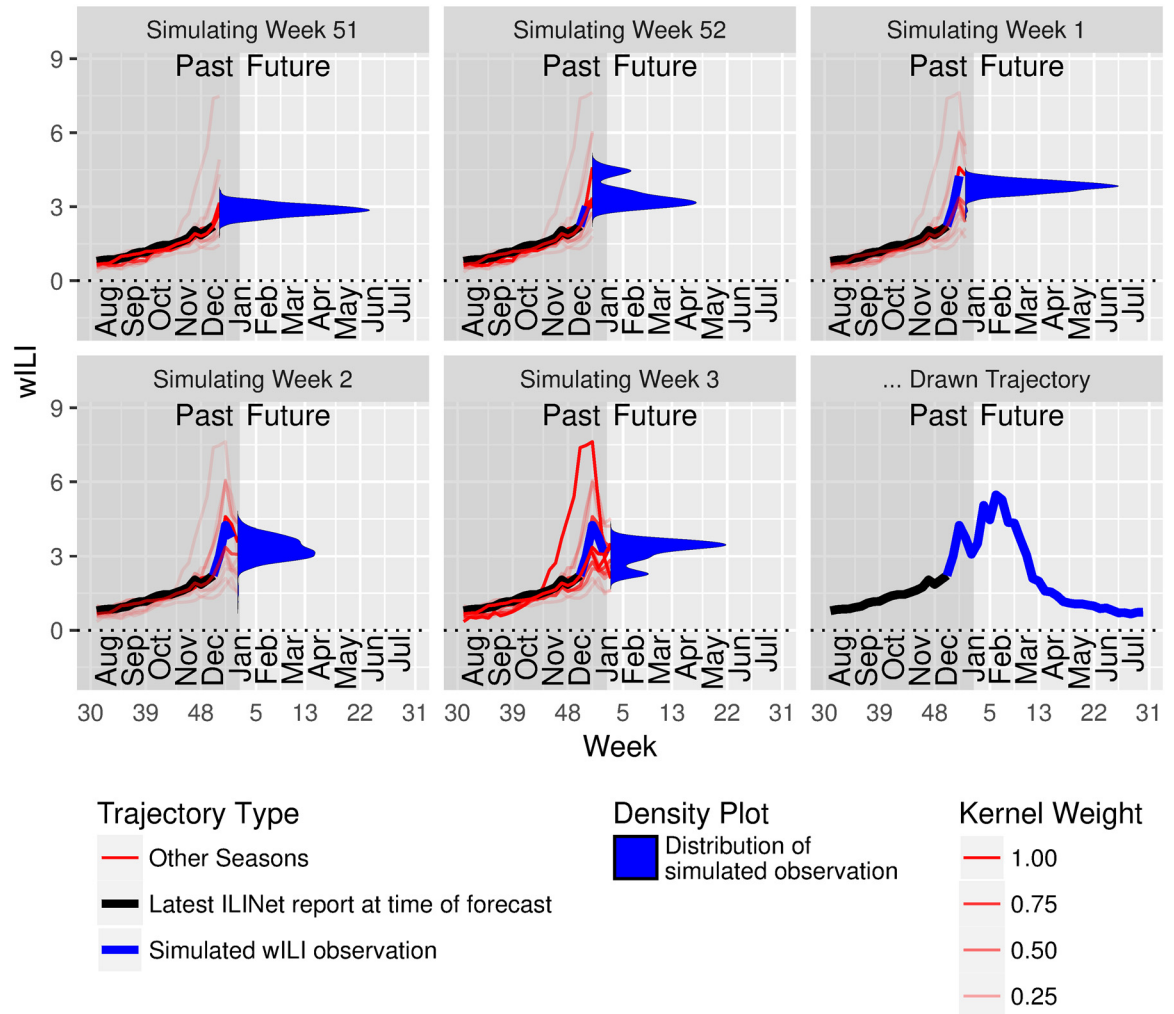
- **Markovian delta density:** approximates the conditional density of  $\Delta Y_u$  given  $Y_{1..u-1}$  with its conditional density given just the previous (real or simulated) observation,  $Y_u$ :

$$\begin{aligned} \hat{f}_{Y_{t+1..T} | Y_{1..t}}(y_{t+1..T} | y_{1..t}) &= \prod_{u=t+1}^T \hat{f}_{\Delta Y_u | Y_{1..u-1}}(\Delta y_u | y_{1..u-1}) \\ &= \prod_{u=t+1}^T \hat{f}_{\Delta Y_u | Y_{u-1}}(\Delta y_u | y_{u-1}) \\ &= \prod_{u=t+1}^T \frac{\sum_s I^u(y_{u-1}, Y_{u-1}^s) \cdot O^u(\Delta y_u, \Delta Y_u^s)}{\sum_s I^u(y_{u-1}, Y_{u-1}^s)}, \end{aligned}$$

where  $I^u$  and  $O^u$  are Gaussian smoothing kernels. The first equality corresponds to the chain rule of probability on the actual (not estimated) densities; the second incorporates the Markov assumption (i.e., selects  $\mathbf{R}^u = [Y_{u-1}]$ ); and the third gives our choice of estimators for the conditional densities  $\hat{f}_{\Delta Y_u | Y_{u-1}}$  for each  $u$ . The bandwidth of each  $I^u$  and  $O^u$  is chosen separately using bandwidth selection procedures for regular kernel density estimation of  $Y_{u-1}$  and  $\Delta Y_u$ , respectively. (Specifically, we use the `bw.SJ` function from the R [46] built-in `stats` package, with `bw.nrd0` as a fallback in the case of errors. These functions do not accept weights for the inputs; it may be possible to improve forecast performance by incorporating these weights or by using other approaches to select the bandwidths.) Note that density estimates for  $\Delta Y_u$  are based on data from past seasons on week  $u$  only, allowing the method to incorporate seasonality and holiday effects (for holidays that consistently occur at the same time of year).

Forecasts are based on Monte Carlo simulations of  $Y_{t+1..T} | Y_{1..t}$ . The following procedure draws a single possibility for  $Y_{t+1..T}$ :

1. Draw  $\Delta Y_{t+1}^{\text{sim}}$  from its conditional density estimate given  $Y_{1..t}$ .
2. Draw  $\Delta Y_{t+2}^{\text{sim}}$  from its conditional density estimate given  $Y_{1..t}$  and  $Y_{t+1} = Y_t + \Delta Y_{t+1}^{\text{sim}}$ .



**Fig 1. The delta density method conditions on real and simulated observations up to week  $u - 1$  when building a probability distribution over the observation at week  $u$ .** This figure demonstrates the process for drawing a single trajectory from the Markovian delta density estimate, ignoring the data revision process. The latest ILINet report,  $W_{1..t}^i$ , which incorporates observations through week 48, is shown in black. Kernel smoothing estimates for future values at times  $u$  from  $t + 1$  to  $T$  are shown in blue, as are simulated observations drawn from these estimates. Past seasons' trajectories are shown in red, with alpha values proportional to the weight they are assigned by the kernel  $P^u$ .

<https://doi.org/10.1371/journal.pcbi.1006134.g001>

3. Continue in the same manner for all remaining weeks.

This process is illustrated in Fig 1. Repeating this procedure many times yields a sample from the model for  $Y_{t+1..T} | Y_{1..t}$ ; stopping at 2000 draws seems sufficient for use in our ensemble forecasts, while at least 7000 are needed to smooth out noise when displaying distributional target forecasts for the delta density method in isolation. Any negative simulated wILI values in these trajectories are clipped off and replaced with zeroes.

- **Extended delta density:** approximates the conditional density of  $\Delta Y_u$  given  $Y_{1..u-1}$  with its conditional density given four features:
  - the previous wILI value,  $Y_{u-1}$ ;
  - the sum of the previous  $k^u$  wILI values, roughly corresponding to the sum of wILI values for the current season;



- an exponentially weighted sum of the previous  $k^u$  wILI; values, where the weight assigned to time  $v$  is  $0.5^{u-v}$ ; and
- the previous change in wILI value,  $\Delta Y_{u-1}$ .

The approximate conditional density assigns each of these features a weight (0.5, 0.25, 0.25, and 0.5, respectively) in order to reduce overfitting and emphasize some relative to the others, and incorporates data from other weeks close to  $u$  (specifically, within  $l^u$  weeks; the choice of  $l^u$  is discussed in a later section) with a truncated Laplacian kernel. We selected these weights and other settings, such as kernel bandwidth selection rules, somewhat arbitrarily based on intuition and experimentation on out-of-sample data; a cross-validation subroutine could be used to make the selection as well, but would multiply the amount of computation required. In case the resulting product of Gaussian and Laplacian kernels is too narrow, we mix its results with a wide boxcar kernel which evenly weights all data from time  $u - l^u$  to  $u + l^u$ :

$$\begin{aligned} & \hat{f}_{\Delta Y_u | Y_{1..u-1}}(\Delta y_u | y_{1..u-1}) \\ &= 0.9 \cdot \frac{\sum_s \sum_{v=u-l^u}^{u+l^u} 0.7^{|v-u|} [I_1^u(y_{u-1}, Y_{v-1}^s)]^{0.5} \cdots O^u(\Delta y_u, \Delta Y_v^s)}{\sum_s \sum_{v=u-l^u}^{u+l^u} 0.7^{|v-u|} [I_1^u(y_{u-1}, Y_{v-1}^s)]^{0.5} \cdots [I_4^u(\Delta y_{u-1}, \Delta Y_{v-1}^s)]^{0.5}} \\ &+ 0.1 \cdot \frac{\sum_s \sum_{v=u-l^u}^{u+l^u} O^u(\Delta y_u, \Delta Y_v^s)}{\sum_s \sum_{v=u-l^u}^{u+l^u} 1} \end{aligned}$$

Using data from  $v \neq u$  incorporates additional reasonable outcomes for  $\Delta Y_u$  by incorporating past wILI patterns with different timing, but risks including some very unreasonable possibilities produced by repeatedly drawing from the same  $v$  rather than following seasonal trends with increasing  $v$ 's. For example, when a portion of a past season that is more similar to itself with a slight time shift than to any other past season, it may be selected for multiple consecutive  $u$ 's and produce an unreasonable trajectory. This could potentially occur when drawing data from the relatively flat regions of wILI trajectories of many seasons, or when incorporating observations around an unusually early, late, high, or low peak. To prevent this possibility, we combine the natural estimate for  $Y_u$  arising from the density estimate for  $\Delta Y_u$  with a random draw  $Y_u^{\text{uncond}}$  from the unconditional density estimate for  $Y_u$  (using a Gaussian kernel and only data from week  $u$ ):

$$Y_u^{\text{sim}} = 0.9 \cdot (Y_{u-1} + \Delta Y_u^{\text{sim}}) + 0.1 \cdot Y_u^{\text{uncond}}.$$

### Residual density method

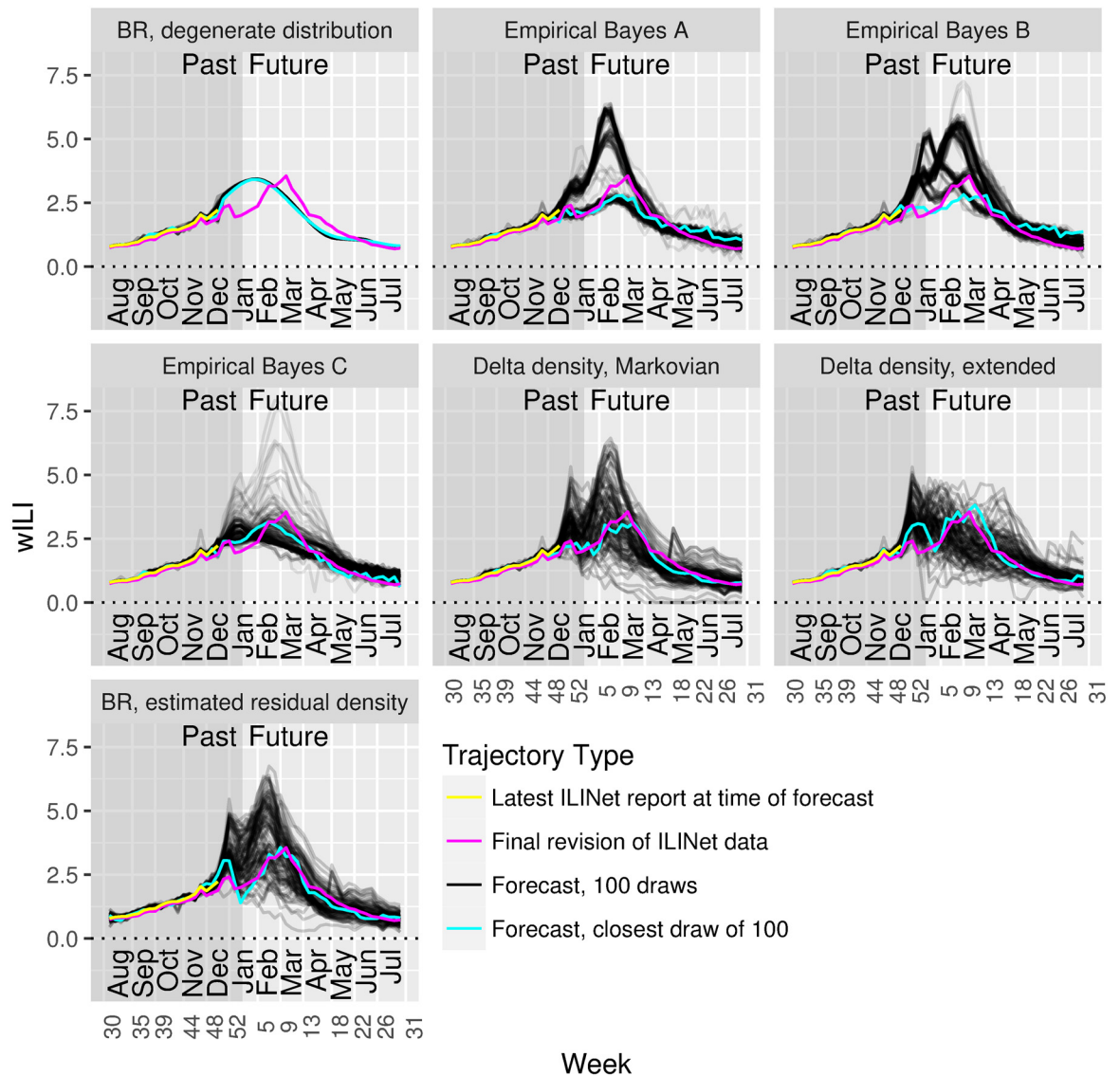
This same approach can be applied to estimate the distribution of residuals of a wILI point predictor. Suppose that we have observed

- $Y_{1..t_1}$ , the partial wILI trajectory up to time  $t_1$ , and
- $X_{1..t_2}$ , point estimates of the trajectory up to some later time  $t_2$ ;

our goal is to estimate the conditional distribution of

- $(Y - X)_{t_1+1..t_2}$ , the unknown residuals,

given  $Y_{1..t_1}$  and  $X_{1..t_2}$ , using data from past seasons. This can be achieved by chaining together draws from conditional density estimates of  $(Y - X)_u | \mathbf{R}^u$  for  $u$  from  $t_1 + 1$  to  $t_2$ , where  $\mathbf{R}^u$  is a



**Fig 2. Delta and residual density methods generate wider distributions over trajectories than methods that treat entire seasons as units.** These plots show sample forecasts of wILI trajectories generated from models that treat seasons as units (BR, Empirical Bayes) and from models incorporating delta and residual density methods. Yellow, the latest wILI report available for these forecasts; magenta, the ground truth wILI available at the beginning of the following season; black, a sample of 100 trajectories drawn from each model; cyan, the closest trajectory to the ground truth wILI from each sample of 100.

<https://doi.org/10.1371/journal.pcbi.1006134.g002>

function of  $Y_{1..u-1}$  and  $X_{1..t_2}$ . The delta density method can be seen as a special case where  $t_1 = t$ ;  $t_2 = T$ ;  $X_{1..t} = Y_{1..t}$ , past values of  $Y$  which are treated as known and are simply duplicated in the simulated trajectories; and  $X_{t+1..T} = Y_{t..T-1}$ , values of  $Y$  which begin as unknown but are filled in as needed by previous simulation steps, giving  $(Y - X)_{t+1..T} = \Delta Y_{t+1..T}$ . We use the residual density method to backcast  $Y_{1..t}$  from  $W_{1..t}^t$  and as the basis for another forecaster in the ensemble.

Fig 2 shows sample forecasts over wILI trajectories generated by each of these approaches and compares them to some alternatives described in S1 Appendix.

## Combining multiple methods: Stacking approach to model averaging

Forecasting systems that select effective combinations of predictions from multiple models can improve on the performance of the individual components, as demonstrated by their successful application in many domains. For each probability distribution and point prediction in a forecast, we treat the choice of an effective combination as a statistical estimation problem, and base each decision on the models' behavior in leave-one-out cross-validation forecasts. Additional cross-validation analysis indicates that this approach achieves performance comparable to or better than the best individual component.

**Background, motivation for combining forecasts.** Methods that combine the output of different models, called “ensembles”, “multi-model ensembles”, “super-ensembles”, “model averages”, or various other names based on the domain and type of approach, have been applied successfully in many problem settings, improving upon the results of the best individual model. An ensemble approach is motivated in the context of seasonal epidemic forecasting by factors such as:

- **Model misspecification and overconfidence in distributional forecasts:** Many methods overlook the possibility of a significant proportion of observed outcomes, or assign otherwise inappropriate probabilities. These omissions and other mistakes are not identical across models; the gaps left by one component can be filled in by another.
- **Leveraging partially correlated errors in point predictions:** The point prediction errors of individual methods can vary in magnitude and are often only partially correlated with each other, allowing ensemble methods to improve performance, e.g., by highly weighting more accurate predictors, or by reducing the variance when combining multiple unbiased estimators.
- **Strengths and weaknesses in different targets:** Some methods may work well for certain forecasting targets, but have poor performance or fail to produce predictions for others; model averages can be smoothly adjusted to account for different behaviors for different targets.
- **Changes in performance within seasons:** Making predictions at the beginning, middle, and end of a season can be seen as different tasks, and the relative performance characteristics of the components may change based on the time of season (or whether it is around a holiday). Just as ensemble methods can account for distinct patterns based on forecasting target, they can be tailored to account for changes in behavior within a season.

We developed an adaptively weighted model average that consistently outperforms the best individual component. Other teams submitting forecasts to the FluSight comparison have concurrently developed other ensemble systems and found similar success [36, 37]. Our approach is distinguished from these other methods in that it very directly estimates the best model average weights for a given location, time, target, and evaluation metric.

**A stacking approach to model averaging.** For each location  $l$ , week  $t$ , target  $i$ , and evaluation metric  $e$ , we choose a (weighted) model average as the final prediction: an ensemble forecast of the form  $\mathbf{X}\mathbf{w}$ , where

- $\mathbf{X}$  is the output of the  $m$  ensemble components—either (a) a row vector of point predictions with  $m$  entries, or (b) a matrix of distributional predictions with  $m$  columns—and
- $\mathbf{w} \in [0, 1]^m$  is a (column) vector of weights, one per component, with  $\sum_{j=1}^m w_j = 1$ .

Variants of the same models, or methods based on related approaches or assumptions, may at times produce similar forecasts that commit the same errors while producing a misleading impression of consensus; a successful ensemble may need to consider not only the performance of each individual component, but also the relationships between the raw output of the components. To this end, we use a “stacking generalization” approach [34, 35], treating the selection of weights  $w$  for the current season,  $S + 1$ , as the task of frequentist estimation of the risk-optimal weight vector,

$$w^* = \arg \max_{\substack{w \in [0,1]^m \\ \sum_{j=1}^m w_j = 1}} \mathbb{E}[\text{Score}(w, S + 1, l, t, i, e)],$$

based on leave-one-season-out cross-validation:

$$\hat{w} = \mu e_{\text{uniform}} + (1 - \mu) \arg \max_{\substack{w \in [0,1]^m \\ \sum_{j=1}^m w_j = 1}} \sum_{\substack{s' \in \{1..S\}, \\ l', t', i', e'}} \text{RelevanceWeight}(s', l', t', i', e'; S + 1, l, t, i, e) \cdot \text{CrossValidationScore}(w, s', l', t', i', e'),$$

where  $\mu$  is an inflation factor that gives addition weight to the uniform component ( $e_{\text{uniform}}$  is a vector containing a 1 in the position corresponding to the uniform distribution component, and 0 in every other position). We changed the RelevanceWeight function used for real-time forecasts throughout the 2015/2016 season, but study only the following RelevanceWeight function in the cross-validation analysis of the adaptively weighted ensemble:

$$\text{RelevanceWeight}(s, l, t, i, e; s', l', t', i', e') = \begin{cases} 1, & |t - t'| \leq 4, i = i', e = e' \\ 0, & \text{otherwise.} \end{cases}$$

A larger collection of cross-validation data can be considered by assigning relevance weights of 1 to additional training instances; relevance weights can also be gradually decreased for less similar data rather than jumping down to zero.

When  $e$  is the unibin or multibin log score:

- Using the rule of three [47] to estimate the frequency of events that we haven't seen before, we chose  $\mu = \frac{3}{s \cdot L}$  for most submissions. (Prior to the submission for 2015 EW43, we used a constant  $\mu = 0.01$  to guarantee a certain minimum log score.)
- The optimization problem is equivalent to fitting a mixture of distributions, and we can use the degenerate EM algorithm [48] to efficiently find the weights; convex optimization techniques such as the logarithmic barrier method are also appropriate.

When  $e$  is mean absolute error:

- We choose  $\mu = 0$  (and further, exclude the uniform distribution method from the ensemble entirely).
- This optimization problem is referred to as least absolute deviation regression or median regression, with linear inequality and equality constraints on the coefficients; we reformulate the problem as a linear program and use the `lpSolve` package [49] to find a solution.

We compare the “adaptive” weighting scheme above to two alternatives:

- **Fixed-weightset-based stacking:** the same approach as above, with the same  $\mu$  selections but a different RelevanceWeight function:

$$\text{RelevanceWeight}(s, l, t, i, e; s', l', t', i', e') = \begin{cases} 1, & e = e' \\ 0, & \text{otherwise;} \end{cases}$$

and

- **Uniform weights:** does not use the above stacking scheme; instead, for every prediction, assigns each component the same weight in the ensemble,  $\frac{1}{m}$  (replacing  $\hat{\mathbf{w}}$  with  $\frac{1}{m} \mathbf{1}$ ).

The ensemble and each of its components forecast the targets  $\mathbf{Z}^t$  given a point or distributional estimate for  $Y_{1..T}$ .

### Considering details of the data generation process

Two important features of ILINet data to consider in models and forecast evaluation are 1. timeliness and accuracy of initial wILI values for each week and subsequent updates to these values, and 2. changes in behavior on and around major holidays. We examine these details of the data generation process, describe how they are addressed in the delta density model, and demonstrate the importance of considering the update procedure when performing retrospective evaluation and prospective forecasting.

**Modeling backfill updates to past wILI.** When a wILI value is reported for a given week, it is not set in stone; as ILINet members provide “backfill” reports or revisions for past weeks and data is cleaned, wILI observations are updated accordingly. When generating a retrospective forecast, it is important to use the version of the data that would have been available at the time rather than the final revision in order to get a more appropriate estimate of the future performance of a forecasting system. Furthermore, forecasting performance can be improved by modeling and “backcasting” these backfill updates, accounting for the following sources of error:

- **Biased early reports:** earlier wILI versions are generally biased downwards early in the in-season, and upwards towards the end of the in-season, which may lead to forecasts of lower, later peaks early in the season, and of longer epidemic duration later in the season;
- **Overconfident short-term distributional forecasts:** since updates in wILI can cause “observed” data, e.g., of the wILI at the presumed peak week, to shift, ignoring backfill may lead to “thin”, overconfident forecast distributions;
- **Revisions of “observed” seasonal targets:** wILI updates sometimes cause large changes in the apparent onset week or peak week when there are bumps or multiple peaks in the trajectory: wILI updates can cause a measurement to change from above the CDC baseline to below (or vice versa), or for an earlier, lower peak to rise above a later peak (or vice versa); ignoring backfill updates can cause models to completely miss some possibilities when these targets appear to be determined. A similar type of error can arise from revisions to the peak height value (regardless of whether the peak week changes); even small updates can result in large unibin log score penalties.

We estimate the distribution of backfill updates using the residual density method with  $t_1 = 0$ ,  $t_2 = t$ ,  $X_{1..t} = W_{1..t}^t$  the latest version of wILI available,  $Y_{1..t}$  the corresponding final revisions, and  $\mathbf{R}^u = [Y_{u-1}]$ . The weight given to a historical nonfinal-to-final residual is based on three factors:

- **Lag amount:** later revisions of wILI values tend to be closer to the final revision than earlier revisions are; thus, when estimating the distribution of  $n$ -week-old wILI to finalized wILI residuals, only  $n$ -week-old wILI to finalized wILI data is considered; backfill data for other lags is ignored (i.e., has zero weight);
- **The current season's nonfinal wILI value:** historical backfill updates with nonfinal wILI values closer to the nonfinal wILI value from the current season are given greater weights according to a Gaussian kernel (with bandwidth based on a rule for kernel density estimation of the historical nonfinal wILI values);
- **Epi week of observation:** since the backfill pattern changes throughout a season, historical backfill updates corresponding to nearby epi weeks are weighted more highly than those from a different time of the season, using a Laplacian kernel (with an arbitrarily selected bandwidth).

The bandwidth of the density estimate is based on a kernel density estimate of the nonfinal-to-final residuals.

The backcasting method is modular and can combine with any forecaster expecting ground truth wILI as input. The straightforward approach is to sample a few thousand trajectories from the backfill simulator, feed each of these into the forecaster to obtain a trajectory or a distribution over targets, and aggregate the results. Some forecasting methods in the Delphi-Stat ensemble do not have a simple way to quickly generate single-trajectory forecasts, so we also use alternative approaches to reduce computation, such as randomly pairing backcasts and trajectory forecasts, where the trajectory forecasts are efficiently generated in batch, based on the pointwise mean of the backcasts.

**Latency of initial wILI value and “nowcasting”.** The initial ILINet wILI value for a given “target” week (from Sunday to Saturday) is typically released on Friday of the following week. Data sources with lower latency and higher temporal resolution can be used to prepare wILI estimates (“nowcasts”) earlier in the following week or even during the target week itself. More generally, auxiliary data for past and current weeks can improve not only models of disease activity in these weeks but also forecasts of future disease activity. Given a backcaster that simulates finalized data for past weeks  $Y_{1..t}$  given observed ILINet and auxiliary data, a nowcaster that simulates  $Y_{t+1}$  given these observations and (a simulated)  $Y_{1..t}$ , and a forecaster that simulates  $Y_{t+2..T}$  given these observations and (a simulated)  $Y_{1..t+1}$ , we can sample from an enhanced model of  $Y_{1..T}$  (given the latest wILI observations  $W_{1..t}^t$ , previous versions of wILI, and auxiliary data) using the following procedure:

1. Repeatedly draw a random value  $Y_{1..t}^{\text{sim}}$  for  $Y_{1..T}$  by:
  1. drawing a random value  $Y_{1..t}^{\text{sim}}$  for  $Y_{1..t}$  conditioned on the observed data, using the backcaster, then
  2. drawing a random value  $Y_{t+1}^{\text{sim}}$  for  $Y_{t+1}$  conditioned on the observed data and  $Y_{1..t} = y_{1..t}^{\text{sim}}$ , using the nowcaster, then
  3. drawing a random value  $Y_{t+2..T}^{\text{sim}}$  for  $Y_{t+2..T}$  conditioned on the observed data and  $Y_{1..t+1} = Y_{1..t+1}^{\text{sim}}$ , using the forecaster, then
  4. combine  $Y_{1..t}^{\text{sim}}$ ,  $Y_{t+1}^{\text{sim}}$ , and  $Y_{t+2..T}^{\text{sim}}$  into a single (random) trajectory  $Y_{1..T}^{\text{sim}}$ , and
2. Collect these individual, randomly drawn trajectories into a list (i.e., a random sample).

As with the earlier method of combining backcasts and forecasts without a nowcaster, this procedure may be too computationally expensive for some implementations of some

forecasters; we use these steps exactly with the delta density methods, for example, but consider modifications and approximations for some other forecasters.

This methodology can be applied in conjunction with one of many available nowcasters. We focus on ILI-Nearby [50, 51], which produces nowcasts for wILI by fusing together several “sensors” using another type of stacked generalization, where each sensor is also a nowcast of wILI data; we reproduce a list of references from [50] on other methodologies for nowcasting and incorporating auxiliary data here [15, 28, 31, 52–71] along with some more recent work [27, 29, 30], with special note of other work using multiple auxiliary data sources [29] or nowcasters [71]. We consider four distributional nowcasters:

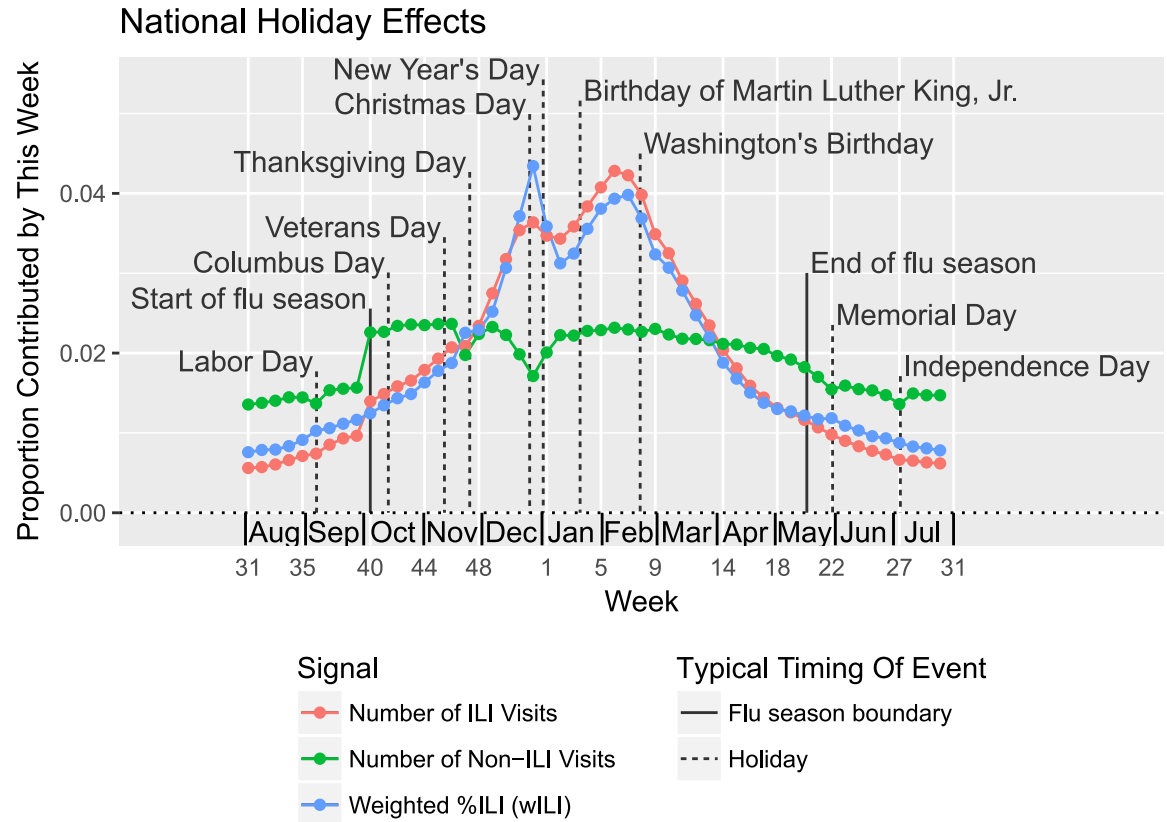
- $Y_{t+1}^{\text{sim}}$  produced by the forecaster, i.e., not using separate nowcasts at all—the basis for all performance estimates unless otherwise noted, as no nowcasts were incorporated into the Delphi-Stat forecasts for the 2015/2016 season;
- $Y_{t+1}^{\text{sim}}$  following a normal distribution with mean and standard deviation given by the ILI-Nearby nowcasting system (ignoring the backcaster’s output);
- $Y_{t+1}^{\text{sim}}$  following a Student’s  $t$  distribution with two degrees of freedom, centrality parameter set to the ILI-Nearby point estimate, and scale parameter set to the ILI-Nearby standard deviation estimate, intended to be a wide-tailed variant of the above (ignoring the backcaster’s output);
- an ensemble of first and third approaches, with associated weights (probabilities) of 15% and 85% respectively. (The choice of weights was inherited from a similar approach that mixed “1 wk ahead” delta density forecasts with nowcasts, rather than ensemble forecasts (including a uniform component) based on these two approaches; a nowcast weight of 85% was selected on a limited amount of out-of-sample (preseason) forecasts to maximize log score.)

**Holiday effects.** On average, wILI tends to be higher on holidays and during the winter holiday season than would be expected based on interpolating observations from nearby weeks [21, 50], as seen in Fig 3. Sharp rises and drops in wILI are common from early or mid-December to early January (roughly coinciding with a four week period beginning with epi week 50), with either the season’s peak or a lower, secondary peak commonly occurring on epi week 52. This pattern appears to arise from two factors:

- spikes downward in the number of non-ILI visits during the holiday season (corresponding to increases in wILI), perhaps caused by patients choosing not to visit the doctor for less serious issues on holidays, and
- decreases in the average number of ILI visits at the end of the holidays, perhaps due to decreased transmission of ILI during holidays, which make the preceding increases in wILI appear even sharper.

Similarly, there are spikes or minor blips downward in the average number of non-ILI visits (which can result in small increases in wILI) associated with Thanksgiving Day; Labor Day; Independence Day; Memorial Day; Birthday of Martin Luther King, Jr.; Washington’s Birthday; Columbus Day; and perhaps other holidays. The spike upward in wILI at Thanksgiving can push wILI unexpectedly over the onset threshold, and holiday effects may help explain the surprising frequency at which peaks occur on epi week 7 but not neighboring weeks. Additional age-specific patterns may be obscured by this analysis of aggregate ILI and non-ILI visit counts.

**Impact of holiday effects on choice of kernels.** Each holiday above occurs at roughly the same time of year every year, falling on one of two possible epi weeks. Thus, models that



**Fig 3. On average, wILI is higher on holidays than expected based on neighboring weeks.** Weekly trends in wILI values, as expressed by the contribution of a each week to a sum of wILI values from seasons 2003/2004 to 2015/2016, excluding 2008/2009 and 2009/2010 (which include portions of the 2009 influenza pandemic), show spikes and bumps upward on and around major holidays. (U.S. federal holidays are indicated with event lines.) The number of non-ILI visits to ILINet health care providers spikes downwards on holidays (disproportionately with any drops in the number of ILI visits), contributing to higher wILI. The number of ILI visits generally declines in the second half of the winter holiday season, causing winter holiday peaks to appear even higher relative to nearby weeks. In addition to holiday effects, we see that average ILINet participation jumps upward on epi week 40, and gradually tapers off later in the season and in the off-season.

<https://doi.org/10.1371/journal.pcbi.1006134.g003>

predict behavior at a given epi week by prioritizing or focusing solely on past behavior at that given epi week will automatically perform a rough adjustment for holiday effects. This factor informs our decision to use historical data only from corresponding weeks in the Markovian delta density method, and a truncated Laplacian kernel with narrower width near winter holidays in the extended delta density method. Specifically, for the extended delta density method, we choose the half-width of the kernel to be  $l^u = \min\{10, \max\{0, |u - 22| - 1\}\}$ , which assigns  $l^u = 0$  for  $u$  within one week of epi week 52, and larger  $l^u$ s the farther  $u$  is from this time period, up to a maximum value of 10.

## Results

### 2015/2016 FluSight comparison

During the 2015/2016 FluSight comparison, we submitted weekly, prospective forecasts from three forecasting systems:

- **Delphi-Stat:** an adaptively weighted ensemble of instance-based statistical forecasting methods, and the topic of this paper;





**Fig 4. The three Delphi systems had similar overall scores; Delphi-Stat gave the best distributional forecasts overall, while Delphi-Epicast gave the best point predictions overall.** These bar plots contain evaluations for the 2015/2016 season, averaged across 11 locations and 29 forecast weeks, for each target and evaluation metrics. Shorter bars indicate better performance. Each entry for a specific target is an average of 319 evaluations, giving a total of 2233 evaluations overall for each system. This figure’s data is shown in tabular form in [S2 Appendix](#).

<https://doi.org/10.1371/journal.pcbi.1006134.g004>

- **Delphi-Archefilter:** forms an empirical (rather than mechanistic) process model describing wILI trajectories, and incorporates both wILI and multiple forms of digital surveillance data using statistical filtering techniques [50]; and
- **Delphi-Epicast:** wisdom-of-crowds approach based on combining predictions submitted by several human participants [72].

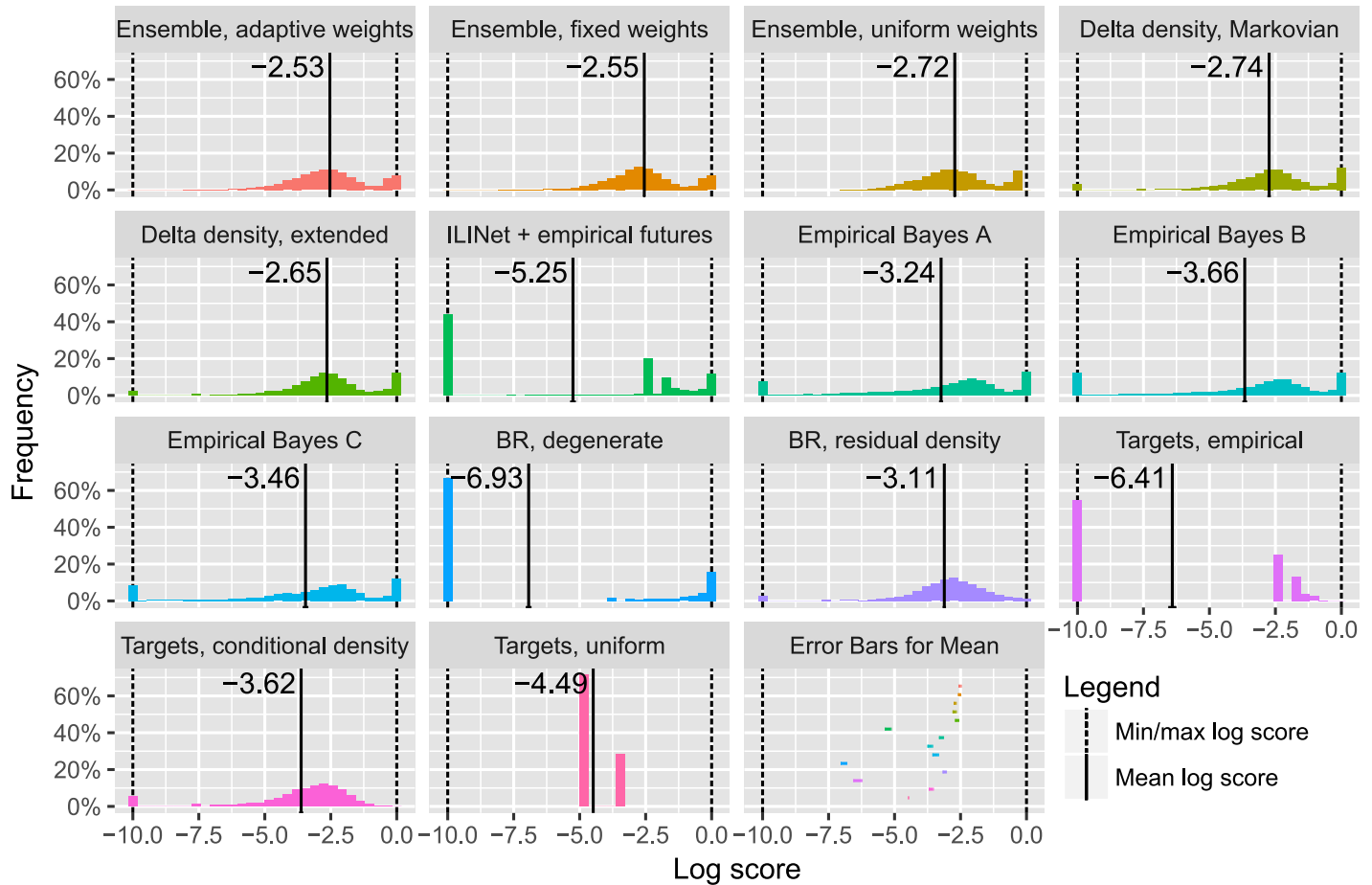
Our past and ongoing forecasts, as well as Python [73] and R [46] code for components of the systems used to generate them, are publicly available online [74–76]. Changes made to Delphi-Stat throughout the 2015/2016 season are described in [S5 Appendix](#). These three forecasting systems were ranked as the top three in the 2015/2016 comparison in terms of overall multibin score, with Delphi-Stat at the top. [Fig 4](#) shows the performance of the three Delphi forecasting systems, broken down by evaluation metric and forecasting target. [S1](#) and [S2](#) Figs. show the multibin scores broken down by location and by forecasting week. Delphi-Stat had consistently strong aggregate multibin scores across different targets, locations, and forecasting weeks, and the best overall multibin log score of all FluSight 2015/2016 submissions. Delphi-Stat’s unibin log score evaluations relative to the other two Delphi systems seem similar to or better than the corresponding multibin log score evaluations, as Delphi-Stat has the best unibin score of the three for each target rather than just overall; this observation seems natural since Delphi-Stat was developed to optimize unibin log score, and may suggest that optimizing for multibin log score rather than unibin log score when selecting ensemble weights or as a post-processing step could produce multibin log score improvements. However, the system’s point predictions, while optimized for the mean absolute error metric, were less accurate than (but still competitive with) the other two when averaged across all predictions.

## Cross-validation analysis

Prospective forecast evaluation ensures that performance estimates are truly out-of-sample, not inflated by design decisions or model fits that are influenced by the evaluation data; however, such evaluation data is not readily generated, as it is expensive in terms of physical time: new wILI observations arrive once per week, and performance can vary significantly from season to season and from week to week. The evaluations from the 2015/2016 comparison may be noisy due to these season-to-season fluctuations. To address this issue, we use pseudo-out-of-sample retrospective analysis to provide more stable estimates of performance. Specifically, we use leave-one-season-out cross-validation: for each evaluation season  $s$ , we form and evaluate retrospective forecasts for  $s$  at every evaluation week using all training seasons except for  $s$  as inputs to the forecasting methods as if they were past seasons. (We exclude seasons prior to 2010/2011 from the evaluation set because records of HHS region ILINet data revisions are only available beginning in late 2009. We exclude seasons prior to 2003/2004 from the training set because year-round ILINet observations, which are required by some of the ensemble components, started in 2003. The 2009/2010 season—containing the peak of the 2009 pandemic according to our adjusted definition of “season”—is also removed from the training set. Finally, we do not include the season currently underway ( $S + 1$ ) in evaluation or training as it has not been completely observed.) Using cross-validation prevents most direct model fitting to evaluation data, and basing design decisions on motivations other than the effects on cross-validation evaluation helps limit fitting through iterative design.

[Fig 5](#) shows the distribution of log scores for several forecasting methods, described earlier in the text and in [S1 Appendix](#), and the three ensemble approaches specified earlier in the text. Except for the uniform distribution and ensembles, all forecasting methods miss some possibilities completely, reporting unreasonable probabilities less than  $\exp(-10) \approx 0.0000454$  for events that actually occurred. In these situations, the log score has been increased to the cap of  $-10$  (as CDC does for multibin log scores). Delta and residual density forecasting methods (Delta density, Markovian; Delta density, extended; and BR, residual density) are less likely to commit these errors than other non-ensemble, non-uniform approaches, and have higher average log scores. Ensemble approaches combine forecasts of multiple components, missing fewer possibilities, and ensuring that a reasonable log score is obtained by incorporating the uniform distribution as a component. For the full Delphi-Stat ensemble, the main advantage of the ensemble over its best component appears to be successfully filling in possibilities missed by the best component with other models to avoid  $-10$  and other low log scores appears, while for ensembles of subsets of the forecasting methods, there are other benefits; [S3 Appendix](#) shows the impact of these missed possibilities and the log score cap.

[Fig 5](#) also includes estimates of the mean log score for each method and rough error bars for these estimates. We expect there to be strong statistical dependence across evaluations for the same season and location, and weaker dependencies between different seasons and locations; thus, the most common approaches to calculating standard errors, confidence intervals, and hypothesis test results will be inappropriate. Properly accounting for such dependencies and calibrating intervals and tests is an important but difficult task and is left for future investigation. We use “rough standard error bars” on estimates of mean evaluations: first, the relevant data (e.g., all cross-validation evaluations for a particular method and evaluation metric) is summarized into one value for each season-location pair by taking the mean of all evaluations for that season-location pair; we then calculate the mean and standard error of the mean of these season-location values using standard calculations as if these values were independent. Under some additional assumptions which posit the existence of a single underlying true mean log score for each method, these individual error bars—or rough error bars for the mean



**Fig 5. Delta and residual density methods cover more observed events and attain higher average log scores than alternatives operating on seasons as a unit; ensemble approaches can eliminate missed possibilities while retaining high confidence when justified.** This figure contains histograms of cross-validation log scores for a variety of forecasting methods, averaged across seasons 2010/2011 to 2015/2016, all locations, forecast weeks 40 to 20, and all forecasting targets. A solid black vertical line indicates the mean of the scores in each histogram, which we use as the primary figure of merit when comparing forecasting methods; a rough error bar for each of these mean scores is shown as a colored horizontal bar in the last panel, and as a black horizontal line at the bottom of the corresponding histogram if the error bar is wider than the thickness of the black vertical line.

<https://doi.org/10.1371/journal.pcbi.1006134.g005>

difference in log scores between pairs of methods—suggest that the observed data is unlikely to have been recorded if the true mean log score of the extended delta density method were greater than that of the adaptively weighted ensemble, or if the true mean log score of the “Empirical Bayes A” method were greater than the extended delta density method. The mean and rough standard error estimates in Fig 5 also appear in tabular form in S4 Appendix.

Methods that model wILI trajectories and “pin” past wILI to its observed values have a large number of log scores near 0 because they are often able to confidently “forecast” many onsets and peaks that have already occurred; ensemble methods also have a large number of log scores near 0. Note that these scores are closer to 0 for ensembles that optimize weighting of different methods than for the ensemble with uniform weights. For this particular set of forecasting methods, targets, and evaluation seasons:

- the uniformly weighted ensemble has lower average log score than the best individual component (extended delta density),

- using the stacking approach to assign weights to ensemble components improves ensemble performance significantly and gives higher average log score than the best individual component,
- the adaptive weighting scheme does not provide a major benefit over a fixed-weight scheme using a single set of weights for each evaluation metric.

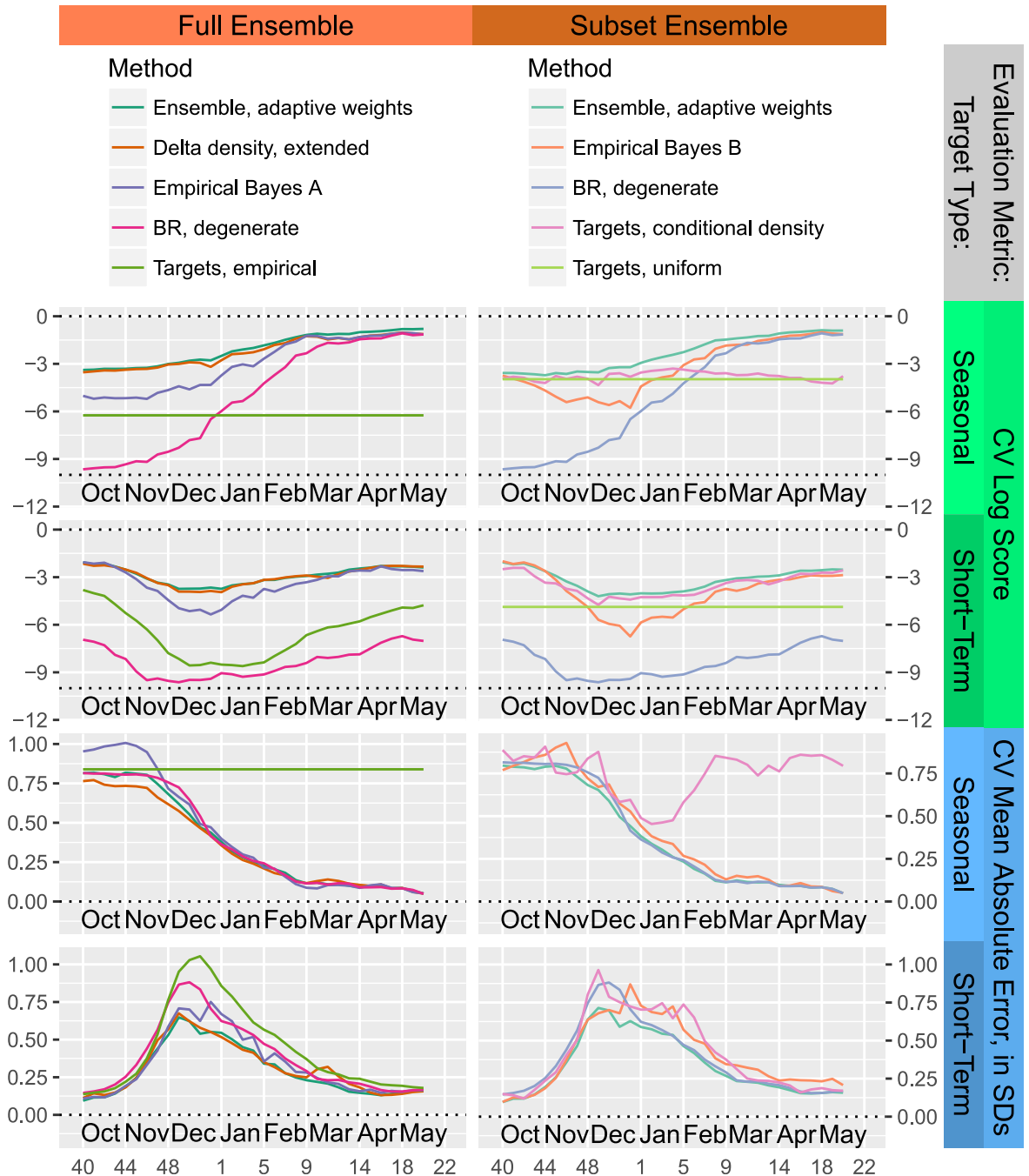
When given subsets of these forecasting methods as input, with regard to average performance:

- the uniformly weighted ensemble often outperforms the best individual, but is sometimes slightly ( $\approx 0.1$  log score) worse;
- the stacking approach improves upon the performance of the uniformly weighted ensemble; and
- the adaptive weighting scheme's performance is equal to or better than that of the fixed-weight scheme, sometimes improving on the log score by  $\approx 0.1$ . The adaptive weighting scheme's relative performance appears to improve with more input seasons, fewer ensemble components, and increased variety in underlying methodologies and component performance. These trends suggest that using wider RelevanceWeight kernels, regularizing the component weights, or considering additional data from 2003/2004 to 2009/2010, for which ground truth wILI but not weekly ILINet reports are available, may improve the performance of the adaptive weighting scheme. In addition to these avenues for possible improvement in ensemble weights for the components presented in Fig 5, the adaptive weighting scheme provides a natural way of incorporating forecasting methods that generate predictions for only a subset of all targets, forecast weeks, or forecast types (distributional forecast or point prediction). For example, in the 2015/2016 season, we incorporated a generalized additive model that provided point predictions (and later, distributional forecasts) for peak week and peak height given at least three weeks of observations from the current season.

Fig 6 shows a subset of the cross-validation data used to form the ensemble and evaluate the effectiveness of the ensemble method, for two sets of components: one using all the components of Delphi-Stat, and the other incorporating three of the lower-performance components and a uniform distribution for distributional forecasts. The Delphi-Stat ensemble near-uniformly dominates the best component, extended delta density, in terms of log score, and has comparable mean absolute error overall. The ensemble approach produces greater gains for the smaller subset of methods, surpassing not only its best components, but all forecasting methods in the wider Delphi-Stat ensemble except for the delta density approaches.

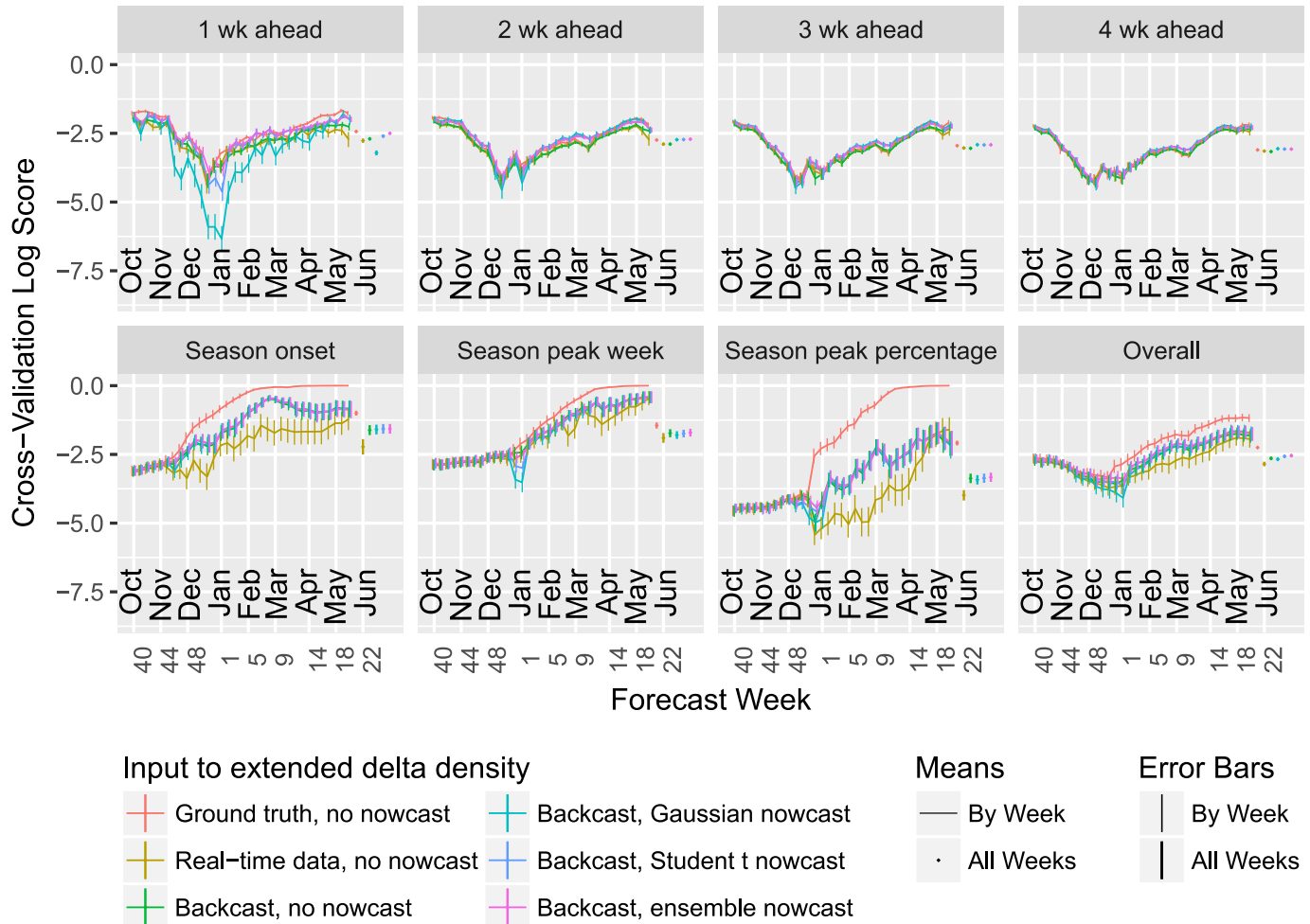
Fig 7 shows cross-validation performance estimates for the extended delta density method based on three evaluation schemes:

- **Ground truth, no nowcast:** the ground truth wILI for the left-out season up to the forecast week is provided as input, resulting in an optimistic performance estimate;
- **Real-time data, no nowcast:** the appropriate wILI report is used for data from the left-out season, but no adjustment is made for possible updates; this performance estimate is valid, but we can improve upon the underlying method;
- **Backcast, no nowcast:** the appropriate wILI report is used for data from the left-out season, but we use a residual density method to “backcast” updates to this report; this performance estimate is valid, and the backcasting procedure significantly improves the log score;



**Fig 6. The ensemble method matches or beats the best component overall, consistently improves log score across all times, and, for some sets of components, can provide significant improvements in both log score and mean absolute error.** These plots display cross-validation performance for two ensembles and some components broken down by evaluation metric, target type, and forecast week; each point is an average of cross-validation evaluations for all 11 locations, seasons 2010/2011 to 2015/2016, and all targets of the given target type; data from the appropriate ILINet reports is used as input for the left-out seasons, while finalized wILI is used for the training seasons. Top half: log score evaluations (higher is better); bottom half: mean absolute error, normalized by the standard deviation of each target (lower is better). Left side: full Delphi-Stat ensemble, which includes additional methods not listed in the legend; right side: ensemble of the three methods listed in the legend, plus a uniform distribution component for distributional forecasts. Many components of the full ensemble are not displayed. The “Targets, uniform” method is excluded from any mean absolute error plots as it was not incorporated into the point prediction ensembles.

<https://doi.org/10.1371/journal.pcbi.1006134.g006>



**Fig 7. Using finalized data for evaluation leads to optimistic estimates of performance, particularly for seasonal targets, “backcasting” improves predictions for seasonal targets, and nowcasting can improve predictions for short-term targets.** Mean log score of the extended delta density method, averaged across seasons 2010/2011 to 2015/2016, all locations, all targets, and forecast weeks 40 to 20, both broken down by target and averaged across all targets (“Overall”). Rough standard error bars for the mean score for each target (or overall) appear on the right, in addition to the error bars at each epi week.

<https://doi.org/10.1371/journal.pcbi.1006134.g007>

- **Backcast, Gaussian nowcast:** same as “Backcast, no nowcast” but with another week of simulated data added to the forecast, based on a Gaussian-distributed nowcast; and
- **Backcast, Student *t* nowcast:** same as “Backcast, Gaussian nowcast” but using a Student *t*-distributed nowcast in place of the Gaussian nowcast.
- **Backcast, ensemble nowcast:** same as the previous two but using the ensemble nowcast (which combines “no nowcast” with “Student *t* nowcast”).

For every combination of target and forecast week, using ground truth as input rather than the appropriate version of these wILI observations produces either comparable or inflated performance estimates.

Using the “backcasting” method to model the difference between the ground truth and the available report helps close the gap between the update-ignorant method. The magnitude of the performance differences depends on the target and forecast week. Differences in mean scores for the short-term targets are small and may be reasonably explained by random chance

alone; the largest potential difference appears to be an improvement in the “1 wk ahead” target by using backcasting. More significant differences appear in each of the seasonal targets following typical times for the corresponding onset or peak events; most of the improvement can be attributed to preventing the method from assigning inappropriately high probabilities (often 1) to events that look like they must or almost certainly will occur based on available wILI observations for past weeks, but which are ultimately not observed due to revisions of these observations. The magnitude of the mean log score improvement depends in part on the resolution of the log score bins; for example, wider bins for “Season peak percentage” may reduce the improvement in mean log score (but would also shrink the scale of all mean log scores). Similarly, the differences in scores may be reduced but not eliminated by use of multi-bin scores for evaluation or ensembles incorporating uniform components for forecasting.

Using the heavy-tailed Student *t* nowcasts or nowcast ensemble appears to improve on short-term forecasts without damaging performance on seasonal targets. The performance of the nowcast ensemble is further explored in [S5](#), [S6](#), [S7](#), [S8](#), [S9](#), [S10](#), [S11](#) and [S12](#) Figs. The Gaussian nowcast has a similar effect as the other nowcasters except on the “1 wk ahead” target that it directly predicts: its distribution is too thin-tailed, resulting in lower mean log scores than using the forecaster by itself on this target.

## Discussion

Delphi-Stat forecasts submitted to the 2015/2016 comparison were based solely on wILI observations from the 2015/2016 “pre-season” (EW21–EW39) and season (EW40–EW20) and nonmechanistic models (with a majority of the ensemble weight assigned to the delta and residual density based methods). Additional data and mechanistic models dealing with categories of ILI or type and subtype of influenza, climate, digital surveillance, season-to-season patterns, spatial interaction, etc. were not incorporated. We do think that these types of data are useful, but analyzing their dynamics and effects on wILI is complicated by the fact that the smallest geographical units for which real-time wILI data is readily available for the entire US are HHS regions (on a weekly time scale). We expect that mechanistic components incorporating climate data and separating diseases, types, and subtypes will be more useful when we are able to model, forecast, and validate data at a finer geographical resolution, ideally at the metro area level. Similarly, we believe that digital surveillance data is useful; in fact, we currently use a sensor fusion framework to combine several such data sources and short-term forecasters to produce “nowcasts” for the current week [[50](#), [51](#)], and improve the performance of forecasting methods by incorporating these nowcasts in a manner similar to the ILINet-based backcasts.

During development and throughout this manuscript, we have focused on (thresholded) unibin log score as a (near-)proper, simple-to-implement metric for distributional forecasts. CDC FluSight organizers, on the other hand, selected exponentiated mean thresholded multi-bin log scores over the entire influenza season as the evaluation metric for forecast comparisons to 1. encourage high-quality distributional predictions rather than point predictions, for better understanding of the risk of certain scenarios, 2. make the scoring metric more accessible to policymakers than unibin and non-exponentiated variants, and 3. avoid  $-\infty$  scores due to a single forecaster mistake or unmodeled data revisions. We believe that it is up to policymakers to decide whether these forecasts are ready for use in decision support at the current level of accuracy. For other potential users and forecast comparisons, we provide absolute error evaluations for all targets in 2015/2016 in [Fig 4](#) and [S2 Appendix](#), as well as absolute error and percent absolute error for short-term targets from cross-validation in [S5](#), [S6](#), [S7](#), [S8](#), [S9](#), [S10](#), [S11](#) and [S12](#) Figs.

## Conclusion

The delta density forecasting method, stacking-based adaptively weighted ensemble, distributional “backcasts” of wILI updates, and nowcasts from ILI-Nearby provide significant improvements upon other individual forecasting approaches that we considered. Promising avenues for further improvements include refining the methodology to rely less on arbitrary and heuristic feature, kernel, bandwidth, and parameter selections, regularization of ensemble weights, incorporating conditional density estimators from statistical literature, and using additional data sources and finer-resolution data models.

## Supporting information

**S1 Fig. Multibin log scores for the 2015/2016 season—Averaged across 7 targets and 29 forecast weeks, broken down by location.** Smaller bars indicate better performance. Each bar is an average of 203 evaluations. This figure’s data is included in tabular form in [S2 Appendix](#). (PDF)

**S2 Fig. Multibin log scores for the 2015/2016 season—Averaged across 7 targets and 11 locations, broken down by forecast week.** Higher log scores indicated better performance. Each point is an average of 77 evaluations. This figure’s data is included in tabular form in [S2 Appendix](#). (PDF)

**S3 Fig. Unibin log score of the delta density method with backcasting—Averaged across targets and locations, broken down by season and epi week.** (PDF)

**S4 Fig. Unibin log score of the delta density method with backcasting—Averaged across targets and seasons, broken down by location and epi week.** (PDF)

**S5 Fig. Mean absolute error of short-term wILI point predictions of the extended delta density method with backcasting and ensemble nowcasting, and of wILI reports—Averaged across seasons, locations, and epi weeks, broken down by timeliness of the estimates.** (PDF)

**S6 Fig. Empirical cumulative distribution function of the absolute error of short-term wILI point predictions of the extended delta density method with backcasting and ensemble nowcasting, and of wILI reports—Combined across seasons, locations, and epi weeks, broken down by timeliness of the estimates.** (PDF)

**S7 Fig. Mean absolute error of short-term wILI point predictions of the extended delta density method with backcasting and ensemble nowcasting, and of wILI reports—Averaged across seasons and epi weeks, broken down by location and timeliness of the estimates.** (PDF)

**S8 Fig. Empirical cumulative distribution function of the absolute error of short-term wILI point predictions of the extended delta density method with backcasting and ensemble nowcasting, and of wILI reports—Combined across seasons and epi weeks, broken down by location and timeliness of the estimates.** (PDF)



**S9 Fig. Mean absolute percent error of short-term wILI point predictions of the extended delta density method with backcasting and ensemble nowcasting, and of wILI reports—Averaged across seasons, locations, and epi weeks, broken down by timeliness of the estimates.**

(PDF)

**S10 Fig. Empirical cumulative distribution function of the absolute percent error of short-term wILI point predictions of the extended delta density method with backcasting and ensemble nowcasting, and of wILI reports—Combined across seasons, locations, and epi weeks, broken down by timeliness of the estimates.**

(PDF)

**S11 Fig. Mean absolute percent error of short-term wILI point predictions of the extended delta density method with backcasting and ensemble nowcasting, and of wILI reports—Averaged across seasons and epi weeks, broken down by location and timeliness of the estimates.**

(PDF)

**S12 Fig. Empirical cumulative distribution function of the absolute percent error of short-term wILI point predictions of the extended delta density method with backcasting and ensemble nowcasting, and of wILI reports—Combined across seasons and epi weeks, broken down by location and timeliness of the estimates.**

(PDF)

**S1 Appendix. Description of all ensemble components in the 2015/2016 Delphi-Stat forecasting system.**

(PDF)

**S2 Appendix. Tabular evaluations of 2015/2016 forecasts.**

(PDF)

**S3 Appendix. “Missed possibilities” and -10 log score threshold.**

(PDF)

**S4 Appendix. Tabular evaluations of CV forecasts.**

(PDF)

**S5 Appendix. Log of changes to Delphi-Stat throughout the 2015/2016 season and for cross-validation analysis.**

(PDF)

## Acknowledgments

We thank CDC for organizing the infectious disease forecasting challenges and comparisons, and for their helpful advice, feedback, and encouragement.

Robert Tibshirani devised and provided an implementation of a “pinned spline” forecasting method, which evolved into the basis regression approaches presented. Shannon Gallagher constructed the generalized additive model for direct prediction of forecasting targets that was included in the 2015/2016 Delphi-Stat ensemble.

Robert Tibshirani, Shannon Gallagher, Daren Wang, and Zirui Wang provided helpful input and feedback during the process of devising the Delphi-Stat ensemble and nowcast analysis.

## Author Contributions

**Investigation:** Logan C. Brooks, David C. Farrow, Sangwon Hyun, Ryan J. Tibshirani, Roni Rosenfeld.

**Methodology:** Logan C. Brooks, David C. Farrow, Sangwon Hyun, Ryan J. Tibshirani, Roni Rosenfeld.

**Visualization:** Logan C. Brooks, David C. Farrow, Sangwon Hyun, Ryan J. Tibshirani, Roni Rosenfeld.

**Writing – original draft:** Logan C. Brooks.

**Writing – review & editing:** David C. Farrow, Sangwon Hyun, Ryan J. Tibshirani, Roni Rosenfeld.

## References

1. World Health Organization. WHO | Influenza (Seasonal); 2016. Available from: <http://www.who.int/mediacentre/factsheets/fs211/en/>.
2. Centers for Disease Control and Prevention, National Center for Immunization and Respiratory Diseases (NCIRD). Estimating Seasonal Influenza-Associated Deaths in the United States | Seasonal Influenza (Flu) | CDC; 2016. Available from: [https://www.cdc.gov/flu/about/disease/us\\_flu-related\\_deaths.htm](https://www.cdc.gov/flu/about/disease/us_flu-related_deaths.htm).
3. Thompson M, Shay D, Zhou H, Bridges C, Cheng P, Burns E, et al. Estimates of deaths associated with seasonal influenza—United States, 1976–2007. *Morbidity and Mortality Weekly Report*. 2010; 59(33):1057.
4. Rolfes M, Foppa I, Garg S, Flannery B, Brammer L, Singleton J, et al. Estimated Influenza Illnesses, Medical Visits, Hospitalizations, and Deaths Averted by Vaccination in the United States; 2016. <https://www.cdc.gov/flu/about/disease/2015-16.htm>.
5. Centers for Disease Control and Prevention, National Center for Immunization and Respiratory Diseases (NCIRD). Pandemic Influenza | Pandemic Influenza (Flu) | CDC; 2017. <https://www.cdc.gov/flu/pandemic-resources/>.
6. Centers for Disease Control and Prevention, National Center for Immunization and Respiratory Diseases (NCIRD). Pandemic Basics | Pandemic Influenza (Flu) | CDC; 2016. <https://www.cdc.gov/flu/pandemic-resources/basics/index.html>.
7. Centers for Disease Control and Prevention. Overview of influenza surveillance in the United States; 2013. <https://www.cdc.gov/flu/weekly/overview.htm>.
8. Biggerstaff M, Alper D, Dredze M, Fox S, Fung ICH, Hickmann KS, et al. Results from the Centers for Disease Control and Prevention’s Predict the 2013–2014 Influenza Season Challenge. *BMC Infectious Diseases*. 2016; 16:357. <https://doi.org/10.1186/s12879-016-1669-x> PMID: 27449080
9. Biggerstaff M, Johansson M, et al. Results from the second year of a collaborative effort to forecast influenza seasons in the United States; 2017.
10. Chretien JP, George D, Shaman J, Chitale RA, McKenzie FE. Influenza forecasting in human populations: a scoping review. *PloS one*. 2014; 9(4):e94130. <https://doi.org/10.1371/journal.pone.0094130> PMID: 24714027
11. Nsoesie EO, Brownstein JS, Ramakrishnan N, Marathe MV. A systematic review of studies on forecasting the dynamics of influenza outbreaks. *Influenza and other respiratory viruses*. 2014; 8(3):309–316. <https://doi.org/10.1111/irv.12226> PMID: 24373466
12. Unkel S, Farrington C, Garthwaite PH, Robertson C, Andrews N. Statistical methods for the prospective detection of infectious disease outbreaks: a review. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*. 2012; 175(1):49–82. <https://doi.org/10.1111/j.1467-985X.2011.00714.x>
13. Shaman J, Karspeck A. Forecasting seasonal outbreaks of influenza. *Proceedings of the National Academy of Sciences*. 2012; 109(50):20425–20430. <https://doi.org/10.1073/pnas.1208772109>
14. Shaman J, Karspeck A, Yang W, Tamerius J, Lipsitch M. Real-time influenza forecasts during the 2012–2013 season. *Nature communications*. 2013; 4. <https://doi.org/10.1038/ncomms3837> PMID: 24302074

15. Hickmann KS, Fairchild G, Priedhorsky R, Generous N, Hyman JM, Deshpande A, et al. Forecasting the 2013–2014 Influenza Season Using Wikipedia. *PLOS Computational Biology*. 2015; 11(5): e1004239. <https://doi.org/10.1371/journal.pcbi.1004239> PMID: 25974758
16. Zhang Q, Perra N, Perrotta D, Tizzoni M, Paolotti D, Vespignani A. Forecasting seasonal influenza fusing digital indicators and a mechanistic disease model. In: Proceedings of the 26th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee; 2017. p. 311–319.
17. Kandula S, Yang W, Shaman J. Type-and Subtype-Specific Influenza Forecast. *American journal of epidemiology*. 2017; p. 1.
18. Deodhar S, Chen J, Wilson M, Soundarapandian M, Bisset K, Lewis B, et al. Flu Caster: A Pervasive Web Application for High Resolution Situation Assessment and Forecasting of Flu Outbreaks. In: Healthcare Informatics (ICHI), 2015 International Conference on. IEEE; 2015. p. 105–114.
19. Viboud C, Boëlle PY, Carrat F, Valleron AJ, Flahault A. Prediction of the spread of influenza epidemics by the method of analogues. *American Journal of Epidemiology*. 2003; 158(10):996–1006. <https://doi.org/10.1093/aje/kwg239> PMID: 14607808
20. Chakraborty P, Khadivi P, Lewis B, Mahendiran A, Chen J, Butler P, et al. Forecasting a moving target: Ensemble models for ILI case count predictions. In: Proceedings of the 2014 SIAM international conference on data mining. SIAM; 2014. p. 262–270.
21. 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. <https://doi.org/10.1371/journal.pcbi.1004382> PMID: 26317693
22. Ray EL, Sakrejda K, Lauer SA, Johansson MA, Reich NG. Infectious disease prediction with kernel conditional density estimation; Accessed 2017-05-02. <https://github.com/reichlab/article-disease-pred-with-kcde/raw/master/inst/article/infectious-disease-prediction-with-kcde.pdf>.
23. Paul MJ, Dredze M, Broniatowski D. Twitter improves influenza forecasting. *PLOS Currents Outbreaks*. 2014;. <https://doi.org/10.1371/currents.outbreaks.90b9ed0f59bae4ccaa683a39865d9117>
24. Höhle M, Meyer S, Paul M. surveillance: Temporal and Spatio-Temporal Modeling and Monitoring of Epidemic Phenomena; 2017. Available from: <https://CRAN.R-project.org/package=surveillance>.
25. Martinez EZ, Silva EASd, Fabbro ALD. A SARIMA forecasting model to predict the number of cases of dengue in Campinas, State of São Paulo, Brazil. *Revista da Sociedade Brasileira de Medicina Tropical*. 2011; 44(4):436–440. <https://doi.org/10.1590/S0037-86822011000400007> PMID: 21860888
26. Lowe R, Bailey TC, Stephenson DB, Jupp TE, Graham RJ, Barcellos C, et al. The development of an early warning system for climate-sensitive disease risk with a focus on dengue epidemics in Southeast Brazil. *Statistics in medicine*. 2013; 32(5):864–883. <https://doi.org/10.1002/sim.5549> PMID: 22927252
27. Lamos V, Miller AC, Crossan S, Stefansen C. Advances in nowcasting influenza-like illness rates using search query logs. *Scientific Reports*. 2015; 5:12760. <https://doi.org/10.1038/srep12760> PMID: 26234783
28. Yang S, Santillana M, Kou SC. Accurate estimation of influenza epidemics using Google search data via ARGO. *Proceedings of the National Academy of Sciences*. 2015; 112(47):14473–14478. <https://doi.org/10.1073/pnas.1515373112>
29. Yang S, Santillana M, Brownstein JS, Gray J, Richardson S, Kou SC. Using electronic health records and Internet search information for accurate influenza forecasting. *BMC Infectious Diseases*. 2017; 17:332. <https://doi.org/10.1186/s12879-017-2424-7> PMID: 28482810
30. Johansson MA, Reich NG, Hota A, Brownstein JS, Santillana M. Evaluating the performance of infectious disease forecasts: A comparison of climate-driven and seasonal dengue forecasts for Mexico. *Scientific Reports*. 2016; 6:33707. <https://doi.org/10.1038/srep33707> PMID: 27665707
31. Generous N, Fairchild G, Deshpande A, Valle SYD, Priedhorsky R. Global Disease Monitoring and Forecasting with Wikipedia. *PLOS Computational Biology*. 2014; 10(11):e1003892. <https://doi.org/10.1371/journal.pcbi.1003892> PMID: 25392913
32. van Panhuis WG, Hyun S, Blaney K, Marques ET Jr, Coelho GE, Siqueira JB Jr, et al. Risk of dengue for tourists and teams during the World Cup 2014 in Brazil. *PLoS Negl Trop Dis*. 2014; 8(7):e3063. <https://doi.org/10.1371/journal.pntd.0003063> PMID: 25079960
33. US Department of Commerce N. Dengue Forecasting; Accessed 2017-05-02. <http://dengueforecasting.noaa.gov/>.
34. Wolpert DH. Stacked generalization. *Neural networks*. 1992; 5(2):241–259. [https://doi.org/10.1016/S0893-6080\(05\)80023-1](https://doi.org/10.1016/S0893-6080(05)80023-1)
35. Breiman L. Stacked regressions. *Machine learning*. 1996; 24(1):49–64.
36. Yamana TK, Kandula S, Shaman J. Superensemble forecasts of dengue outbreaks. *Journal of The Royal Society Interface*. 2016; 13(123):20160410. <https://doi.org/10.1098/rsif.2016.0410>

37. Ray EL, Reich NG. Prediction of infectious disease epidemics via weighted density ensembles. arXiv preprint arXiv:170310936. 2017;.
38. Leung NH, Xu C, Ip DK, Cowling BJ. Review Article: The Fraction of Influenza Virus Infections That Are Asymptomatic: A Systematic Review and Meta-analysis.; 2015.
39. Hayward AC, Fragaszy EB, Bermingham A, Wang L, Copas A, Edmunds WJ, et al. Comparative community burden and severity of seasonal and pandemic influenza: results of the Flu Watch cohort study. *The Lancet Respiratory Medicine*. 2014; 2(6):445–454. [https://doi.org/10.1016/S2213-2600\(14\)70034-7](https://doi.org/10.1016/S2213-2600(14)70034-7) PMID: 24717637
40. Brammer L, Budd AP, Finelli L. 12. In: Seasonal and pandemic influenza surveillance. John Wiley & Sons Ltd; 2013. p. 200–210. <http://dx.doi.org/10.1002/9781118543504.ch16>.
41. Centers for Disease Control and Prevention, National Center for Immunization and Respiratory Diseases (NCIRD). FluView Interactive | Seasonal Influenza (Flu) | CDC; 2017. <https://www.cdc.gov/flu/weekly/fluviewinteractive.htm>.
42. The Delphi Group at Carnegie Mellon University. The Delphi Epidemiological Data API; Accessed 2017-04-26. <https://github.com/cmu-delphi/delphi-epidata>.
43. Farrow DC. EpiVis; Accessed 2017-04-26. <http://delphi.midas.cs.cmu.edu/epivis/epivis.html>.
44. Hendrickson AD, Buehler RJ. Proper scores for probability forecasters. *The Annals of Mathematical Statistics*. 1971; p. 1916–1921. <https://doi.org/10.1214/aoms/1177693057>
45. Epidemic Prediction Initiative. Forecast Evaluation; 2016. Available from: <https://predict.phiresearchlab.org/legacy/flu/evaluation.html>.
46. R Core Team. R: A Language and Environment for Statistical Computing; 2015. <https://www.R-project.org/>.
47. Jovanovic BD, Levy PS. A look at the rule of three. *The American Statistician*. 1997; 51(2):137–139. <https://doi.org/10.1080/00031305.1997.10473947>
48. Rosenfeld R. The “degenerate EM” algorithm for finding optimal linear interpolation coefficients  $\lambda$ ; Accessed 2017-03-21. <http://www.cs.cmu.edu/~roni/11761/Presentations/degenerateEM.pdf>.
49. Berkelaar M, others. IpSolve: Interface to ‘Lp\_solve’ v. 5.5 to Solve Linear/Integer Programs; 2015. <http://CRAN.R-project.org/package=IpSolve>.
50. Farrow DC. Modeling the Past, Present, and Future of Influenza [PhD Thesis]. Carnegie Mellon University; 2016. Available from: <http://reports-archive.adm.cs.cmu.edu/anon/cbd/CMU-CB-16-101.pdf>.
51. Farrow DC, Rosenfeld R. Multiple Resolution Nowcasting of Influenza through Sensor Fusion; 2018. *Manuscript in preparation*.
52. Soebiyanto RP, Adimi F, Kiang RK. Modeling and Predicting Seasonal Influenza Transmission in Warm Regions Using Climatological Parameters. *PLOS ONE*. 2010; 5(3):e9450. <https://doi.org/10.1371/journal.pone.0009450> PMID: 20209164
53. Shaman J, Karspeck A. Forecasting seasonal outbreaks of influenza. *Proceedings of the National Academy of Sciences*. 2012; 109(50):20425–20430. <https://doi.org/10.1073/pnas.1208772109>
54. Shaman J, Karspeck A, Yang W, Tamerius J, Lipsitch M. Real-time influenza forecasts during the 2012–2013 season. *Nature Communications*. 2013; 4:2837. <https://doi.org/10.1038/ncomms3837> PMID: 24302074
55. Eysenbach G. Infodemiology: Tracking Flu-Related Searches on the Web for Syndromic Surveillance. *AMIA Annual Symposium Proceedings*. 2006; 2006:244–248.
56. Polgreen PM, Chen Y, Pennock DM, Nelson FD, Weinstein RA. Using Internet Searches for Influenza Surveillance. *Clinical Infectious Diseases*. 2008; 47(11):1443–1448. <https://doi.org/10.1086/593098> PMID: 18954267
57. Hulth A, Rydevik G, Linde A. Web Queries as a Source for Syndromic Surveillance. *PLOS ONE*. 2009; 4(2):e4378. <https://doi.org/10.1371/journal.pone.0004378> PMID: 19197389
58. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. *Nature*. 2009; 457(7232):1012–1014. <https://doi.org/10.1038/nature07634> PMID: 19020500
59. Dugas AF, Jalalpour M, Gel Y, Levin S, Torcaso F, Igusa T, et al. Influenza Forecasting with Google Flu Trends. *PLOS ONE*. 2013; 8(2):e56176. <https://doi.org/10.1371/journal.pone.0056176> PMID: 23457520
60. Araz OM, Bentley D, Muellemann RL. Using Google Flu Trends data in forecasting influenza-like-illness related ED visits in Omaha, Nebraska. *The American Journal of Emergency Medicine*. 2014; 32(9):1016–1023. <https://doi.org/10.1016/j.ajem.2014.05.052> PMID: 25037278

61. Santillana M, Zhang DW, Althouse BM, Ayers JW. What Can Digital Disease Detection Learn from (an External Revision to) Google Flu Trends? *American Journal of Preventive Medicine*. 2014; 47(3):341–347. <https://doi.org/10.1016/j.amepre.2014.05.020> PMID: 24997572
62. Preis T, Moat HS. Adaptive nowcasting of influenza outbreaks using Google searches. *Royal Society Open Science*. 2014; 1(2):140095. <https://doi.org/10.1098/rsos.140095> PMID: 26064532
63. Ritterman J, Osborne M, Klein E. Using prediction markets and Twitter to predict a swine flu pandemic. In: *Proceedings of the 1st International Workshop on Mining Social Media*. vol. 9; 2009. p. 9–17.
64. Culotta A. Towards Detecting Influenza Epidemics by Analyzing Twitter Messages. In: *Proceedings of the First Workshop on Social Media Analytics. SOMA'10*. New York, NY, USA: ACM; 2010. p. 115–122. <http://doi.acm.org/10.1145/1964858.1964874>.
65. Signorini A, Segre AM, Polgreen PM. The Use of Twitter to Track Levels of Disease Activity and Public Concern in the U.S. during the Influenza A H1N1 Pandemic. *PLOS ONE*. 2011; 6(5):e19467. <https://doi.org/10.1371/journal.pone.0019467> PMID: 21573238
66. Achrekar H, Gandhe A, Lazarus R, Yu SH, Liu B. Predicting Flu Trends using Twitter data. In: *2011 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*; 2011. p. 702–707.
67. Broniatowski DA, Paul MJ, Dredze M. National and Local Influenza Surveillance through Twitter: An Analysis of the 2012-2013 Influenza Epidemic. *PLOS ONE*. 2013; 8(12):e83672. <https://doi.org/10.1371/journal.pone.0083672> PMID: 24349542
68. Paul MJ, Dredze M, Broniatowski D. Twitter Improves Influenza Forecasting. *PLoS Currents*. 2014; 6.
69. McIver DJ, Brownstein JS. Wikipedia Usage Estimates Prevalence of Influenza-Like Illness in the United States in Near Real-Time. *PLOS Computational Biology*. 2014; 10(4):e1003581. <https://doi.org/10.1371/journal.pcbi.1003581> PMID: 24743682
70. Dredze M, Cheng R, Paul MJ, Broniatowski D. HealthTweets.org: A Platform for Public Health Surveillance Using Twitter. In: *Workshops at the Twenty-Eighth AAAI Conference on Artificial Intelligence*; 2014. p. 593–596. Available from: <https://www.aaai.org/ocs/index.php/WS/AAAIW14/paper/view/8723>.
71. Santillana M, Nguyen AT, Dredze M, Paul MJ, Nsoesie EO, Brownstein JS. Combining Search, Social Media, and Traditional Data Sources to Improve Influenza Surveillance. *PLOS Computational Biology*. 2015; 11(10):e1004513. <https://doi.org/10.1371/journal.pcbi.1004513> PMID: 26513245
72. 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. <https://doi.org/10.1371/journal.pcbi.1005248> PMID: 28282375
73. Van Rossum G, Drake FL. Python language reference manual. *Network Theory*; 2003.
74. Reichek K, Gao L. CMU Delphi forecasts; Accessed 2017-04-26. <http://delphi.midas.cs.cmu.edu/forecast/>.
75. Carnegie Mellon University Delphi group. Delphi; Accessed 2017-04-26. <https://github.com/cmu-delphi>.
76. Brooks LC, Farrow DC, Hyun S, Tibshirani RJ, Rosenfeld R. *epiforecast: Tools for forecasting semi-regular seasonal epidemic curves and similar time series*; 2015.