Received 2017-01-01; | Revised 2017-02-01; | Accepted 2017-03-01

DOI: xxx/xxxx

RESEARCH ARTICLE

Improving influenza forecasts by accounting for reporting revisions

Graham C. Gibson*[1,2] | Evan L. Ray[1] | Tom McAndrew[1] | Dave Osthus[2] | Nicholas G. Reich[1]

[1]University of Massachusetts, Amherst
[2]Los Alamos National Laboratory

Correspondence
Graham C Gibson, Email: gcgibson@umass.edu

With an estimated $10.4 billion in medical costs and 31.4 million outpatient visits each year, influenza poses a serious burden of disease in the United States. To provide insights and advance warning into the spread of influenza, the U.S. Centers for Disease Control and Prevention (CDC) has run a challenge for forecasting influenza-like-illness (ILI) and weighted influenza-like-illness (wILI) at the national, regional, and state level. Targets of interest include 1-4 week ahead wILI percentages, as well as seasonal targets such as peak week of incidence, peak week percentage, and season onset. However, because the challenge requires forecasts in real-time, the data used to forecast are subject to revisions. Almost all initial reports of wILI or ILI at a given time are revised later on. Initial reports of wILI (ILI) can be revised either upwards or downwards as additional data from reporting facilities are processed. In order to accurately forecast wILI (ILI) percentages, accounting for these revisions is critical. We present a framework that relies solely on historical revisions which shows improvements in average log score of seasonal and 1-4 step ahead targets at the national, regional, and state level.

1 | INTRODUCTION

1.1 | Importance of Influenza Forecasting

Seasonal influenza hospitalizes over half a million people in the world every year[1]. The United States alone reported approximately 80,000 Influenza related mortalities in the 2017/2018 influenza season, with most serious consequences for vulnerable populations such as children or the elderly. The annual toll of influenza outbreaks in the US provide a frequent reminder of the importance of interventions that could help mitigate the impact of influenza outbreaks.[2]

The main tool in the fight against influenza is vaccination. The CDC recommends that everyone, including children, get vaccinated at the beginning of the season. However, there are only a finite number of vaccines produced each season, begging the question of how to best allocate the limited number of vaccines to protect the largest number of at risk people. Studies have shown that an optimal allocation of influenza vaccine requires an accurate estimate of risk to the population.[3] To this end, accurate probabilistic forecasting models may help with optimal risk assessment and therefore optimal allocation.

As part of their forecasting initiative, the CDC releases forecasts for weighted influenza-like illness (wILI), which measures the proportion of outpatient doctor visits at reporting health care facilities where the patient had influenza-like illness, weighted by state population. Forecasts are made for up to four weeks into the future, as well as seasonal targets including week of peak incidence, peak week wILI value and season onset. Participants in the FluSight challenge have harnessed a variety of models and

methods to forecast the targets under consideration. These efforts have included time series models, mechanistic transmission models, and machine learning techniques.[4] Five teams submitted forecasts from seven models in the 2017/2018 season.[5] Some teams have also incorporated external data to improve forecasts.[6 7 8]

However, many submitted models fall prey to reporting revisions. Each week, the CDC releases updated data that include an initial report of wILI for a particular week as well as revisions to previously reported wILI. These revisions occur for a variety of reasons, and initial reports of wILI may be revised upwards or downwards. For example, initial reports of wILI may be revised upwards if additional cases of ILI are reported or revised downwards if additional outpatient doctor visits without ILI are reported. These revisions have consequences for forecast accuracy since the way the CDC assesses model performance by evaluating forecasts generated using unrevised data are able to predict the revised final data at the end of the season.[9] Revisions may occur up to 10 weeks after the final week of the season, after which the CDC fixes the observed data at that time as the "truth", for purposes of scoring models. Accounting for revisions in real-time may improve log-score on CDC defined targets by recognizing patterns in historical revisions and applying them to currently reported data.

Methods for accounting for reporting revisions have commonly been referred to as "nowcasting" in the literature. This is because we are providing an estimate of a desired signal at the current time (commonly called time "now") based on a partially observed signal.[10 11 12] Early attempts at correcting for reporting revisions focused solely on the under-reporting aspect.[13] The work of Lawless et al. used a non-parametric method to scale up observed incidence levels of human immunodeficiency virus (HIV) based on historical revisions in order to gain a more accurate estimate of the emerging HIV epidemic. Hohle extended this effort to account for arbitrary transmission models in an under-reported setting during a Shiga toxin-producing E. coli epidemic.[14] Recently, Nunes et al. employed a hidden Markov model to estimate the reporting revisions to wILI data from Portugal with success.[15] Stone et al. also investigated the application of state space models to the reporting delay problem with count data in a hierarchical setting.[16]

With respect to influenza-like-illness in the U.S., much of the current efforts are focused on using external data to improve the estimates of the unrevised data. External data show significant improvements to 1-4 week ahead forecasts but not to seasonal forecasts.[17] Reporting delay has previously been modeled without using external data at the national level with mixed results for the 1-4 step ahead targets.[18] We propose a framework that lends itself to modeling delay at both the national and state level and suggest models that improve both seasonal and short term forecast log scores at both the national and state level.

2 | DATA & FORECAST TARGETS

2.1 | U.S. Influenza Surveillance Data

For the national challenge, the CDC wILI data are provided at both the national level and broken down into 10 Health and Human Services (HHS) regions, mostly organized by geographical proximity. The national level data extend from 1997 to the present and the HHS regional data are available starting from 2013. The revised wILI data are highly seasonal and vary by region (Figure 1 A). For the state level challenge, only data from 2016 to the present is available. Data is reported for each state at the ILI level, except for Florida, which does not participate.

These data are reported by the ILINet system, a consortium of over 3,500 outpatient healthcare facilities across all states and territories in the US. Each week, around 2,200 of these providers report both total number of patient visits and total number of patients presenting with influenza-like symptoms. These two numbers are combined to report the percentage of cases reporting with influenza-like symptoms and are weighted by population size of the state to generate the final regional or national wILI level.[19]

The CDC releases updated wILI data on a weekly basis for all states and regions. These updates include new wILI estimates for the most recent week, in addition to revisions for all prior weeks of the season. As noted above, revisions can be made either upwards or downwards due to updates to both the total number of visits and total of number of ILI visits. An example of historical revisions is shown in Figure 1 B). Revision data extends back to 1997 for the national level data, but only includes the 2017/2018 season for state level data.

2.2 | Targets

The CDC FluSight challenge identifies three key seasonal targets of interest: season onset, season peak week percentage, and season peak week. Season onset is defined as the first week of the season which exceeds a pre-specified baseline for at least 3

consecutive weeks. That is, the week at which we have observed three wILI values above a set baseline in a row. This is helpful for public health officials in their preparation and planning for the upcoming season. This target is restricted to national and regional level data, where such baselines are defined. This definition makes season onset particularly susceptible to reporting revisions. Revising the wILI just slightly can flip the wILI value above or below the pre-specified season onset, and therefore change the identified season onset week. Season peak week percentage is defined as the maximum wILI or ILI value observed for the season. This target is also sensitive to reporting revisions if a season peak percentage value that has been observed is revised downwards. This is because forecasting models can only place probability of a season peak percentage larger than the observed peak percentage. Finally, season peak week is defined as the week in which the maximum observed wILI value occurs. This target is less sensitive to reporting revisions since revisions usually respect the relative ordering of wILI or ILI values within a season. This is also in part because of the relatively consistent direction of reporting revisions throughout a season.
In addition to seasonal targets the CDC is interested in 1-4 week ahead forecasts for short term projections of ILI. These are helpful for real-time public health decision making and resource allocation. In practice, the data is delivered with a two week lag, so the 1 step head forecast is actually a "hindcast", the 2 step ahead forecast is a "nowcast", and the 3-4 step ahead forecast is a true forecast.

2.3 | Notation

Due to the complicated nature of reporting revisions, we first introduce notation to describe the data that is available at given time and the final data used to score the models.

We begin by examining the traditional forecasting models used to create predictive distributions of the form:

$$f(z_w|y_1, ..., y_w, \theta) \tag{1}$$

for some observed data $y_1, ..., y_w$ and parameters θ. For example, y_w may be a measure of disease incidence at epiweek w. We use Z_w to indicate an arbitrary forecast target relative to time w. For example, $Z_w = Y_{w+1}$ would be a 1-step ahead prediction and $Z_w = argmax_{w \in S}(Y_1, ..., Y_w)$ would be a season peak target for some season S.

In order to capture the inherent variability of the observed data due to revisions, we instead consider $(Y_1, ..., Y_w)$ as random, not fixed. We denote the revised wILI (or ILI) for epiweek w at time $w + l$ as $Y_{w,l}$. Borrowing from the notation of HÃűhle [14] denote the final reported data as $Y_{w,\infty}$ where $l = \infty$ denotes the revision at time ∞, that is, the final revision. We also notate the most up to date set of data for a given season s in a given region r at epiweek w as

$$\vec{Y}_{w,l} = \{Y_{1,w}, Y_{2,w-1}, ..., Y_{w,0}\} \tag{2}$$

and similarly we define the vector of initially reported data and finally reported data as follows:

$$\vec{Y}_{w,0} = \{Y_{1,0}, Y_{2,0}, ..., Y_{w,0}\} \tag{3}$$

$$\vec{Y}_{w,\infty} = \{Y_{1,\infty}, Y_{2,\infty}, ..., Y_{w,\infty}\} \tag{4}$$

This notation is further illustrated in Figure 3 B.

With the data notation established, we can now consider a joint distribution over the finally reported data and the forecast target, conditional on the currently reported data, thereby treating the observed data as random.

$$f(z_{t,\infty}, y_{1,\infty}, y_{2,\infty}..., y_{t,\infty}|\theta, y_{1,t}, y_{2,t-1}, ..., y_{t,0}) \tag{5}$$

To simplify notation, we condense the set of observed data for a given week w as $\vec{Y}_{w,l}$ (Figure 3). In order to leverage existing process models that historically yield well calibrated forecast distributions, we factor the above distribution into a forecast distribution conditional on some observed data, and an "observed data distribution" that captures our uncertainty over the currently reported data.

$$f(z_{w,\infty}|y_{1,\infty}, ..., y_{w,\infty}|\theta)g(y_{1,\infty}, ..., y_{w,\infty}|\phi, \vec{y}_{w,l}) \tag{6}$$

This factorization also allows us to recover our real goal when forecasting, the marginal distribution over the target $Z_{w,\infty}$:

$$\tilde{f}(z_{w,\infty}) = \int f(z_{w,\infty}|\vec{y}_{w,\infty}, \theta)g(\vec{y}_{w,\infty}|\vec{y}_{w,l}, \phi)\, d\, y_1, .., y_w \tag{7}$$

In cases where the distribution of $Y_1, ..., Y_w$ does not have a pdf g, this integral can be written in terms of the cdf $G_{Y_1,...,Y_w}$ using a Stieltjes integral:

$$\tilde{f}(z_{w,\infty}) = \int f(z_{w,\infty}|\vec{y}_{w,\infty}, \theta)\, dG_{\vec{Y}_{w,\infty}}(\vec{y}_{w,\infty}|\phi, \vec{y}_{w,l}) \tag{8}$$

This integral will often be intractable, especially in a generic setting where we allow arbitrary process model and observed data distributions. In practice, we will approximate it using Monte Carlo techniques.

$$\tilde{f}(z_{w,\infty}) \approx \frac{1}{n}\sum_{i}^{n} f(z_{w,\infty}|\vec{y}_{w,\infty}, \theta) \tag{9}$$

where

$$\vec{y}_{w,\infty} \sim G|\vec{Y}_{w,l}, \phi$$

. With the general framework established, we turn our attention to specific models for g, designed to remedy the reporting revision challenges described above.

3 | MODELS FOR THE AVAILABLE DATA

In what follows we develop two separate models for the available data based on the two types of targets found in the challenge, seasonal and short term. This is motivated by the separate challenges reporting revisions present to each target. We first examine the effects of reporting revisions on seasonal targets followed by the effects of reporting revisions on 1-4 step ahead targets.

3.1 | Reporting Revision Effects on Seasonal Targets

Season onset requires three or more wILI values to be above the region specific season onset baseline. This definition makes the target very sensitive to reporting revisions. As noted in Figure 7 , the season onset can be revised below the baseline, moving the truth more than 1 week away from the currently reported season onset. This results in a log score of negative infinity when predicting season onset after it has been initially observed, since the model places all probability on the available data season onset. Historically, this has occurred in around 10% of seasons across all regions. However, for each season this occurs in, a score of negative infinity is assigned multiple times, since the season onset target is evaluated for all weeks in a season, and only when the versions move the available data to the correct side of the season onset baseline does the model predict season onset correctly. A similar story happens in the peak week percentage. The peak week percentage is calculated by sampling process model trajectories forward in time and choosing the max over the sampled trajectories. Only sampled trajectories that exceed the currently observed peak will be included in the predictive distribution. Therefore, only peak week percentages larger than the currently observed peak percentage receive non-zero probability. If the peak week percentage is revised downwards this also results in a negative infinity log score. Historically, this has occurred in over half the seasons across all regions. Peak week, however, does not suffer from a similar problem as the other two targets defined above. This is again because the reporting revisions usually respect the relative ordering of wILI (or ILI) values throughout a season. Therefore, even though the season maximum may be revised, the week at which the maximum occurs is relatively stable. In fact, historically reporting revisions have only shifted the week at which the maximum occurs twice, across all regions.

These observations highlight the need for uncertainty around the available data. Our model for g must take into account the possibility of future revisions that alter that season targets. In order to add uncertainty to the observed data we non-parametrically sample from historical revisions using the revision difference we call d. More specifically, at a given region r, season s, week w, and lag l as a central concept in our model. We denote the difference between the reported wILI at lag l and the final wILI value by

$$d_{r,s,w,l} = Y_{r,s,w,l} - Y_{r,s,w,\infty} \tag{10}$$

$$g(\vec{y}_{r,s,w,\infty}; \vec{y}_{r,s,w,l}, \vec{d}_{w,l}) = \frac{1}{n}\sum_{i=1}^{n} \delta\left(\vec{y}_{r,s,w,l} - \vec{d}^{(i)}_{r,s,w,l}\right) \tag{11}$$

Sampling from g amounts to drawing $\vec{d}^{(i)}_{r,s,w,l}$ from historical reporting revisions and then subtracting the sampled revision from the observed data $\vec{Y}_{r,s,w,l}$. This allows us to create a non-parametric distribution around the observed data by borrowing information from historical reporting revisions.

We can see how this modifies the expected value of a forecast by using the law of total expectation.

$$E_{\tilde{f}}(Z_{r,s,w,\infty}) = \int_{\vec{Y}_{r,s,w,\infty}} E_f(Z_{r,s,w,\infty}|\vec{y}_{r,s,w,\infty}))dG_{\vec{Y}_{r,s,w,\infty}}$$

$$E_{\tilde{f}}(Z_{r,s,w,\infty}) = \int_{\vec{Y}_{r,s,w,\infty}} \frac{1}{n}\sum_{i=1}^{n} E_f(Z_{r,s,w,\infty}|\vec{y}_{r,s,w,l} - \vec{d}^{(i)}_{r,s,w,l}))dG_{\vec{Y}_{r,s,w,\infty}}$$

$$E_{\tilde{f}}(Z_{r,s,w,\infty}) = \frac{1}{n}\sum_{i=1}^{n} E_f(Z_{r,s,w,\infty}|\vec{y}_{w,l} - \vec{d}^{(i)}_{w,l})) \int_{\vec{Y}_{w,\infty}} dG_{\vec{Y}_{r,s,w,\infty}}$$

$$E_{\tilde{f}}(Z_{r,s,w,\infty}) = \frac{1}{n}\sum_{i=1}^{n} E_f(Z_{r,s,w,\infty}|\vec{y}_{r,s,w,l} - \vec{d}^{(i)}_{r,s,w,l}))$$

The expected value of the sampling method is the average over the forecasts from the sampled revisions.

We can also examine the variance of the resulting altered forecast distribution using equation 8. The first term is a weighted sum of the forecast distribution variance weighted by the probability of the observed data.

$$E_g[Var_f(Z_{r,s,w,\infty}|\vec{y}_{r,s,w,\infty})] = \int_{\vec{Y}_{r,s,w,\infty}} Var_f(Z_{r,s,w,\infty}|\vec{y}_{r,s,w,\infty})dG_{\vec{Y}_{r,s,w,\infty}} \tag{12}$$

$$= \frac{1}{n}\sum_{i=1}^{n} Var_f(Z_{r,s,w,\infty}|\vec{Y}_{r,s,w,l} - \vec{d}^{(i)}_{r,s,w,l}) \tag{13}$$

The second term in equation. 8 is a variance with respect to the observed data distribution.

$$Var_g(E_f(Z_{r,s,w,\infty}|\vec{Y}_{r,s,w,l} - \vec{d}^{(i)}_{r,s,w,l})) \tag{14}$$

which will generally be positive unless g is a point-mass around the observed data. In our particular choice of g, this term is clearly positive, demonstrating where the additional variation around the observed data comes from.

3.2 | Effects of reporting revisions on 1-4 step ahead forecasts

While adding uncertainty around the observed data may protect against seasonal target revisions, it creates a problem for short term forecasts. As demonstrated, the above model for g increases the variance of the forecasts. This is a desirable property for seasonal targets, but not for short term targets. This is because process models have been developed to have well calibrated predictive distributions for short term targets, and inflating the variance will decrease the amount of probability placed on the true target value. Therefore, in order to correct for the effect of reporting revisions on short term targets, we focus on modelling the bias induced by reporting revisions, while maintaining the well calibrated predictive distributions under the process model of choice. To directly model the bias we again model the difference between the available wILI (or ILI) and the final wILI (or ILI). We allow the reporting revision bias to vary with time in season. This is because we assume a greater difference in the available and final wILI (or ILI) during the peak of the season, when activity is highest. We also make use of the how many ILINet health care providers have submitted information. This can be thought of as a mechanistic indicator of reporting revisions. If only a small number of the total health care providers have submitted their information, there is a high likelihood of revision. Unfortunately, we do not currently have access to the total number of health care providers and instead use the un-normalized number of providers. This allows us to model the bias in the reporting delay as follows:

$$\begin{aligned}\mathbf{E}[Y_{w,0,r} - Y_{w,\infty,r}] = \beta_0 + \beta_1 \text{spline(season week)} + \beta_2 \text{num providers} + \\ b_{1,r}\text{spline(season week)} + b_{0,r}\text{num providers} \\ b_0, b_1 \sim MVN(0, \Sigma)\end{aligned}$$

where we have introduced the subscript r to indicate region. We use a spline in the season week to allow for a greater impact of reporting delay at higher levels of incidence (near the peak week) than at the beginning and end of the season. This hierarchical model allows us to both borrow information about the bias of the reporting delay across regions (states) but also allow the effect

of the season week and number of providers to vary by region (states). This is desirable because the total number of providers varies by state along with the effect of season week. This could be because different states have different reporting protocols and disparate effects of high wILI (ILI) on reporting infrastructure burden.

In order to predict and correct for the bias we construct a series of regression models, fit independently, for each value of the lag. Thus, if we are in epiweek 50, we use the estimates from 10 independent regression models to correct the bias for all weeks from the start of the season to the current week. If we notate the collection of predicted corrections as $d_{t,l}$ We then correct the bias as follows

$$\hat{Y}_{w,\infty} = Y_{w,l} - d_{w,l} \tag{15}$$

We can treat this model in our framework as a point density g that takes the available data and maps it to the bias corrected data. That is,

$$g(Y_{w,\infty} = y_{w,\infty}) = \begin{cases} 1 \text{ if } y_{w,\infty} = y_{w,l} - d_{w,l} \\ 0 \text{ otherwise} \end{cases}$$

Using the law of total variance we can see that we not change the forecast variance, since the variance under g is 0, and the expected value is simply

$$E_g(Var(Z_{w,\infty}|Y_{w,l})) = Var(Z_{w,\infty}|Y_{w,l} - d_{w,l})$$

4 | EVALUATION

4.1 | Model Scoring

In order to score the probabilistic forecasts made by SARIMATD under each of the revision models we employ the multibin log score used by the CDC in the FluSight challenge. In order to score forecasts produced by models we discretize the continuous predictive distributions for each target by binning values. If we index each of the predictive distributions for target Z_t at week w for bin i by region r and season s we obtain a discrete distribution of the form,

$$p_{r,s,w,Z_t,i} = P_{r,s,w}(Z_t = i)$$

For example, if $Z_t = 1$ Wk Ahead then $i = \{0, .1, .2, ..., 13, 13+\}$ and if $Z_t =$ Season Onset then $i = \{1, ..., 52\}$. We therefore have that $\sum_i p_{r,s,w,Z_t,i} = 1$. We compute the log score of a forecast as the log of the probability assigned to the observed outcome In order to avoid $-\infty$ when the probability assigned to the target is 0 we truncate at -10, following convention set by the CDC.

$$\text{log score}_{r,s,w,Z_t} = max(-10, log(p_{r,s,w,Z_t,i})) \tag{16}$$

We can extend this to multibin scoring by expanding the set of values that are considered correct (the true wILI), from a point i to a set I.

$$\text{mutlibin log score}_{r,s,w,Z_t} = log(\sum_{i \in I} max(-10, p_{r,s,w,Z_t,i})) \tag{17}$$

For example, under the multibin scoring, the season onset truth set is $\{i-1, i, i+1\}$ and for 1-4 week ahead the truth set is $\{i-.5, i+.5\}$.

Table 1 explains exactly what data are used when making a forecast for a particular target during the testing phase. ## Experimental setup

4.1.1 | National & Region Level

In order to compare forecasts made from the available data to those made by our revision models and the revised data, we train a SARIMATD model on final reported wILI data ($Y_{r,s,w,\infty}$) from 2010/2011 to 2014/2015. We reserve 2015/2016, 2016/2017, and 2017/2018 as test seasons. We fit the model to each region separately. Along with model fitting, we also use the data from 2010/2011 through 2015/2016 to estimate the reporting revision differences. Although we do not explicitly evaluate the estimation of the reporting revision differences, the reporting revision estimation performance is tied into the log score of the wILI target forecast. This is because our end goal is improvement in forecasts.

4.1.2 | State Level

We use the same process model and evaluation scheme as the national level. However, we are limited to only three seasons of data for the state level ILI, and only two seasons where the reporting delay was monitored. Therefore, we use 2016/2017 and 2017/2018 to fit both the process model and delay models, and use only 2018/2019 as a test season.

5 | RESULTS

The results for the both national/regional and state levels are shown in Tables 1 & 2.

5.1 | Sampling method can improve forecast accuracy for seasonal targets

Using the sampling method we are able to avoid the extreme cases (-10 log score) since we do not treat the available data as fixed. As noted above, the sampling method is able to correct the misidentified season onset by placing some probability on the event that the current observed season onset will be revised. Similarly, the sampling method assigns some probability to the event that the currently reported peak percentage may be revised downwards. In fact, this benefit of the sampling method is model agnostic, since all models that don't explicitly account for revisions would treat a currently observed season onset as the truth, without accounting for some uncertainty in the reported data. Therefore, the season onset specific results should extend to all other process models. A specific example of the benefits of the sampling method is illustrated in Figure 7 , where the -10 values all appear later in the season, when the season onset has been observed. The sampling method removes these -10s, regardless of prospective forecast score by simply placing probability on wILI values below the currently observed season onset. We see a similar effect on season peak percentage. As noted above, we see little benefit to the peak week target, since revisions usually respect the relative ordering of wILI (or ILI)

5.2 | The bias correction method is able to improve the MSE of our estimate of the true data

As we can see from Table 1 & 2 we are able to get closer to the final reported data (in terms of MSE) using our bias correction method. This means the data that we pass into SARIMA to forecast from is closer to the finally reported data that will be used to score the model at the end of the season. This clearly is a desirable attribute of any bias correction model, but still leaves the link between reduction in MSE of available data and log-score of forecasts to be examined. Note that the results of the DM test are relatively uninformative because of the small sample size (only 1 test season).

5.3 | The bias correction method is able to improve 1-4 step ahead log scores on average

As expected, we see the biggest gain in log score for 1 step ahead forecasts. This is because the 1 step ahead forecast is most sensitive to corrections in the recent data under the SARIMA process model. By 4 steps ahead our forecasts have become sufficiently spread out that correcting for the bias has little effect. These results may vary under the chosen process model. For instance, for a seasonal historical average model that ignores the current season data, the revision process will have no effect on forecasts, and therefore correcting it will have no effect on forecasts. However, most models in the FluSight competition do make use of recent data. As seen in Tables 1 & 2, the log-score increases are quite small, as reflected in the Diebold-Mariano p-values. They are, however, consistently better, which in the realm of influenza forecasting is enough to warrant discussion. We also computed the p-values for the test of significance between forecasts made from the available data and forecasts made from the revised data. For 1-4 step ahead targets, these were all $>.05$. Therefore, it is not clear that either modelling 1-4 step ahead forecasts will lead to statistically significant improvement in log score of forecasts or that the DM test is suitable to detect such small differences. Further investigation into potential uses of internet search data or bio-sensor data may improve the bias correction models.

5.4 | The results vary heavily by region

The bias correction method requires that the direction of the revisions (either over or under reported) is consistent across seasons. If we examine Figure 8 , we can see that stats such as Arizona (az) are both under and over reported, violating the consistency

assumption. It may be beneficial to consider an average level of bias significantly different from 0 before applying revision correction. At the national level, reporting revisions are relatively minor and inconsistent in their direction, and therefore correcting for the bias actually increases the MSE of the adjusted data as an estimate of the final data. This loss in MSE is offset by hhs regions with a larger level of reporting revisions, but further investigation into indicators of the amount of revision per region may help identify those regions which do not need reporting revision modelling.

6 | CONCLUSION

We have presented a general framework to account for reporting revisions in a statistically principled way. By treating the observed data as random we are able to introduce uncertainty to capture the effect of reporting revisions. By sampling historical reporting revision differences and applying them to the currently observed data we have protected against the scenarios where revisions cause large negative effects on log score for seasonal targets. By building well informed reporting delay bias correction models, we are able to improve the 1-4 step ahead log score of forecasts and the MSE of the available data as an estimate of the final data.

Although we chose a canonical process model to forecast wILI (SARIMA), the main benefit of treating the observed data as random occurs when an initial season onset has already been observed, or an unrevised season peak percentage above the revised peak percentage has already been observed. In this way, the benefits offered by our approach are irrespective of the choice of process model. Their effect does not rely upon any process model forecast values, but is simply based on imparting uncertainty into the currently observed data.

Lack of data for proper cross-validation of our methods is a significant limiting factor of the above analysis. While the benefits of the sampling method are grounded in specific reporting revision scenarios, the probability of those scenarios remains low. It could be that the reporting process of the ILINet network is improving over time, meaning the chance of an extreme reporting revision situation is continually decreasing. A larger set of training and testing seasons to analyze would help address this question. The lack of data also impacts the ability to detect statistical significance in the small improvements made by the bias correction model.

Further investigation into an ensemble approach combining external data sources with historical reporting revision differences may yield even more benefit during forecasting. Research seems to suggest that an external signal may help improve nowcasting, so combining this with historical revisions may outperform either model on their own. Nevertheless, this requires access to external data that shows strong correlation with the wILI signal and is itself not prone to reporting revisions. In many infectious disease settings, this data does not exist. There is also room for improvement of the bias modelling by using covariates that mechanistically drive reporting delay, such as the number of providers that have reported out of total number of providers.

Even after accounting for reporting revisions, accurate forecasting of wILI remains a difficult task. The complex transmission dynamics and limited data availability mean the main source of forecast error is simply the underlying model, not the reporting revisions. In addition, wILI is not a perfect signal of the true level of influenza in the population.[9] However, wILI forecasts still have an actionable value for public health officials. Effective risk assessment is crucial in vaccine allocation, and well calibrated forecasts are helpful to that end.

7 | FIGURES

TABLE 1 National/Regional results for 2015/2016,2016/2017,2017/2018 season averaged over all regions. Seasonal target log-scores are evaluated under the non-parametric sampling method described in section 3.1. Short term forecast target log-scores are evaluted under the bias correction method described in section 3.2. MSE of both the available data and the adjusted data under the bias correction method as an estimate of the final data is reported. Diebold-Mariano tests of forecasts made from the available data versus the adjusted data under each model are reported.

Method	SO	PWP	PW	MSE	1	2	3	4
Available	-0.9943556	-2.571994	-1.314099	0.1435912	-0.8863	-1.181	-1.43	-1.601
Adjusted	-0.7947232	-2.399926	-1.330966	0.1338544	-0.8675	-1.179	-1.429	-1.61
Perfect	-0.5784056	-1.967124	-1.226389	0	-0.7959	-1.139	-1.394	-1.586

Method	SO	PWP	PW	MSE	1	2	3	4
DM p-value	0.02761	3.848e-12	2.2e-16	0.002415	0.1412	0.4033	0.3592	0.6128

TABLE 2 State results for the 2017/2018 season averaged over all regions. Seasonal target log-scores are evaluated under the non-parametric sampling method described in section 3.1. Short term forecast target log-scores are evaluted under the bias correction method described in section 3.2. MSE of both the available data and the adjusted data under the bias correction method as an estimate of the final data is reported. Diebold-Mariano tests of forecasts made from the available data versus the adjusted data under each model are reported.

Method	PWP	PW	MSE	1	2	3	4
Available	-3.282336	-1.869723	0.6009343	-2.624	-3.079	-3.558	-3.879
Adjusted	-3.272363	-1.867507	0.5800103	-2.575	-3.048	-3.544	-3.85
Perfect	-1.967124	-2.872685	0	-1.802	-3.031	-3.554	-3.861
DM p-value	0.1877	0.9266	0.6688	0.06581	0.05104	0.1185	0.4033

References

1. Lafond Kathryn E, Nair Harish, Rasooly Mohammad Hafiz, et al. Global role and burden of influenza in pediatric respiratory hospitalizations, 1982–2012: a systematic analysis. *PLoS medicine.* 2016;13(3):e1001977.

2. Skowronski Danuta M, Chambers Catharine, De Serres Gaston, et al. Early season co-circulation of influenza A (H3N2) and B (Yamagata): interim estimates of 2017/18 vaccine effectiveness, Canada, January 2018. *Eurosurveillance.* 2018;23(5).

3. Mylius Sido D, Hagenaars Thomas J, Lugnér Anna K, Wallinga Jacco. Optimal allocation of pandemic influenza vaccine depends on age, risk and timing. *Vaccine.* 2008;26(29-30):3742–3749.

4. Kandula Sasikiran, Yamana Teresa, Pei Sen, Yang Wan, Morita Haruka, Shaman Jeffrey. Evaluation of mechanistic and statistical methods in forecasting influenza-like illness. *Journal of The Royal Society Interface.* 2018;15(144):20180174.

5. Biggerstaff Matthew, Johansson Michael, Alper David, et al. Results from the second year of a collaborative effort to forecast influenza seasons in the United States. *Epidemics.* 2018;24:26–33.

6. Dugas Andrea Freyer, Jalalpour Mehdi, Gel Yulia, et al. Influenza forecasting with Google flu trends. *PloS one.* 2013;8(2):e56176.

7. Araz Ozgur M, Bentley Dan, Muelleman Robert L. 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.

8. Volkova Svitlana, Ayton Ellyn, Porterfield Katherine, Corley Courtney D. Forecasting influenza-like illness dynamics for military populations using neural networks and social media. *PloS one.* 2017;12(12):e0188941.

9. Reich Nicholas G, McGowan Craig J, Yamana Teresa K, et al. A Collaborative Multi-Model Ensemble for Real-Time Influenza Season Forecasting in the US. *bioRxiv.* 2019;:566604.

10. Lawless JF. Adjustments for reporting delays and the prediction of occurred but not reported events. *Canadian Journal of Statistics.* 1994;22(1):15–31.

11. Lampos Vasileios, Miller Andrew C., Crossan Steve, Stefansen Christian. Advances in nowcasting influenza-like illness rates using search query logs. *Scientific Reports.* 2015;5:12760 EP -.

12. Johansson Michael A, Powers Ann M, Pesik Nicki, Cohen Nicole J, Staples J Erin. Nowcasting the spread of chikungunya virus in the Americas. *PloS one.* 2014;9(8):e104915.

13. Kalbfleisch JD, Lawless Jerald F. Inference based on retrospective ascertainment: an analysis of the data on transfusion-related AIDS. *Journal of the American Statistical Association.* 1989;84(406):360–372.

14. Höhle Michael, Heiden Matthias. Bayesian nowcasting during the STEC O104: H4 outbreak in Germany, 2011. *Biometrics.* 2014;70(4):993–1002.

15. Nunes Baltazar, Natário Isabel, Lucília Carvalho M. Nowcasting influenza epidemics using non-homogeneous hidden Markov models. *Statistics in medicine.* 2013;32(15):2643–2660.

16. Stoner Oliver, Economou Theo. Multivariate Hierarchical Frameworks for Modelling Delayed Reporting in Count Data. *arXiv preprint arXiv:1904.03397.* 2019;.

17. Osthus Dave, Daughton Ashlynn R, Priedhorsky Reid. Even a good influenza forecasting model can benefit from internet-based nowcasts, but those benefits are limited. *PLoS computational biology.* 2019;15(2):e1006599.

18. Brooks Logan C, Farrow David C, Hyun Sangwon, Tibshirani Ryan J, Rosenfeld Roni. Nonmechanistic forecasts of seasonal influenza with iterative one-week-ahead distributions. *PLoS computational biology.* 2018;14(6):e1006134.

19. *Influenza (Flu).* 2018.

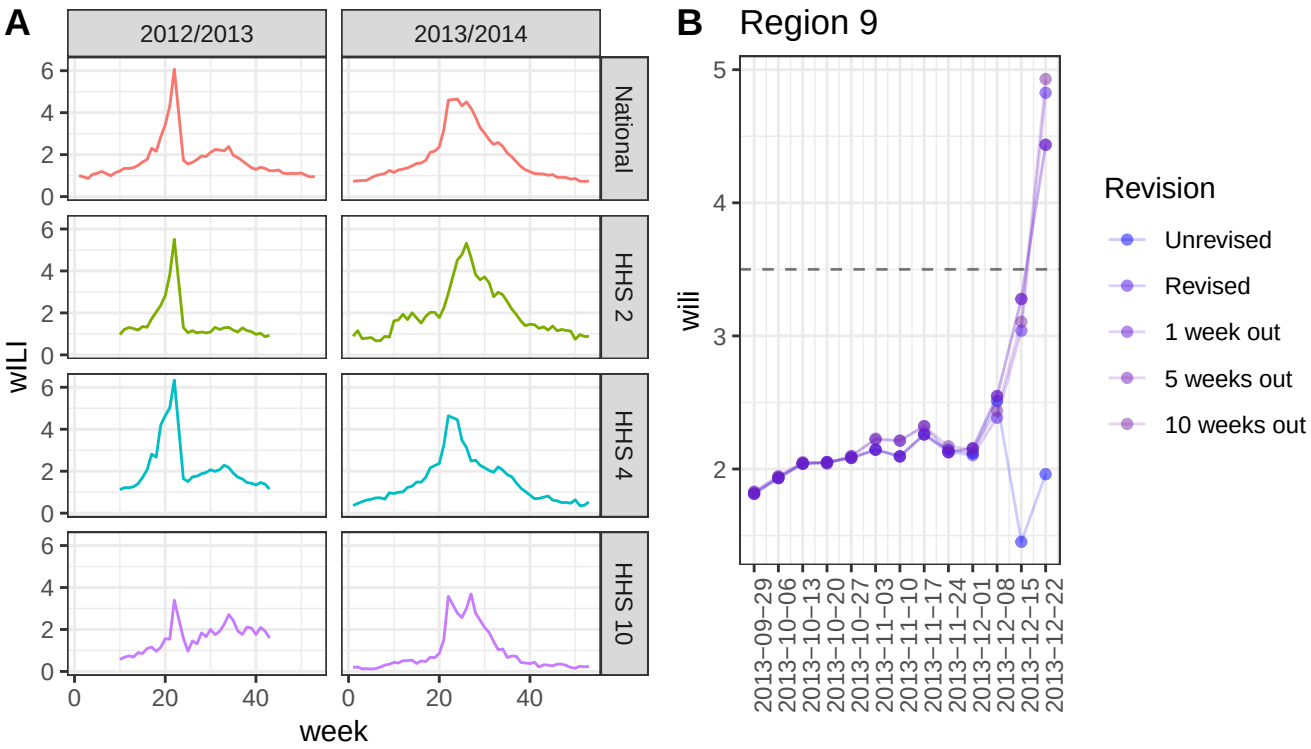

FIGURE 1 A: wILI data from the 2012/2013 and the 2013/2014 season across 4 example regions. Notice the regional variability and the seasonal structure with a peak usually (but not always) occuring somewhere between week 20 and week 30. B: Data from 2013-09-29 (week 1 of the 2013/2014 season) to 2013-12-22 (week 12 of the 2013/2014 season) from HHS region 9. The 'revised' data is a snapshot of wILI values for the listed weeks if the current time is the end of the 2013/2014 season. The 'unrevised' data is a snapshot of wILI values for the listed weeks if the current time were 2013-12-22. Similarly, the lag 5 data is a snapshot of wILI values for the listed weeks if the current time were 5 weeks after 2013-12-22. Notice that unrevised data is both over and under reported relative to the revised data at different epiweeks. Dashed line represents the season onset baseline.

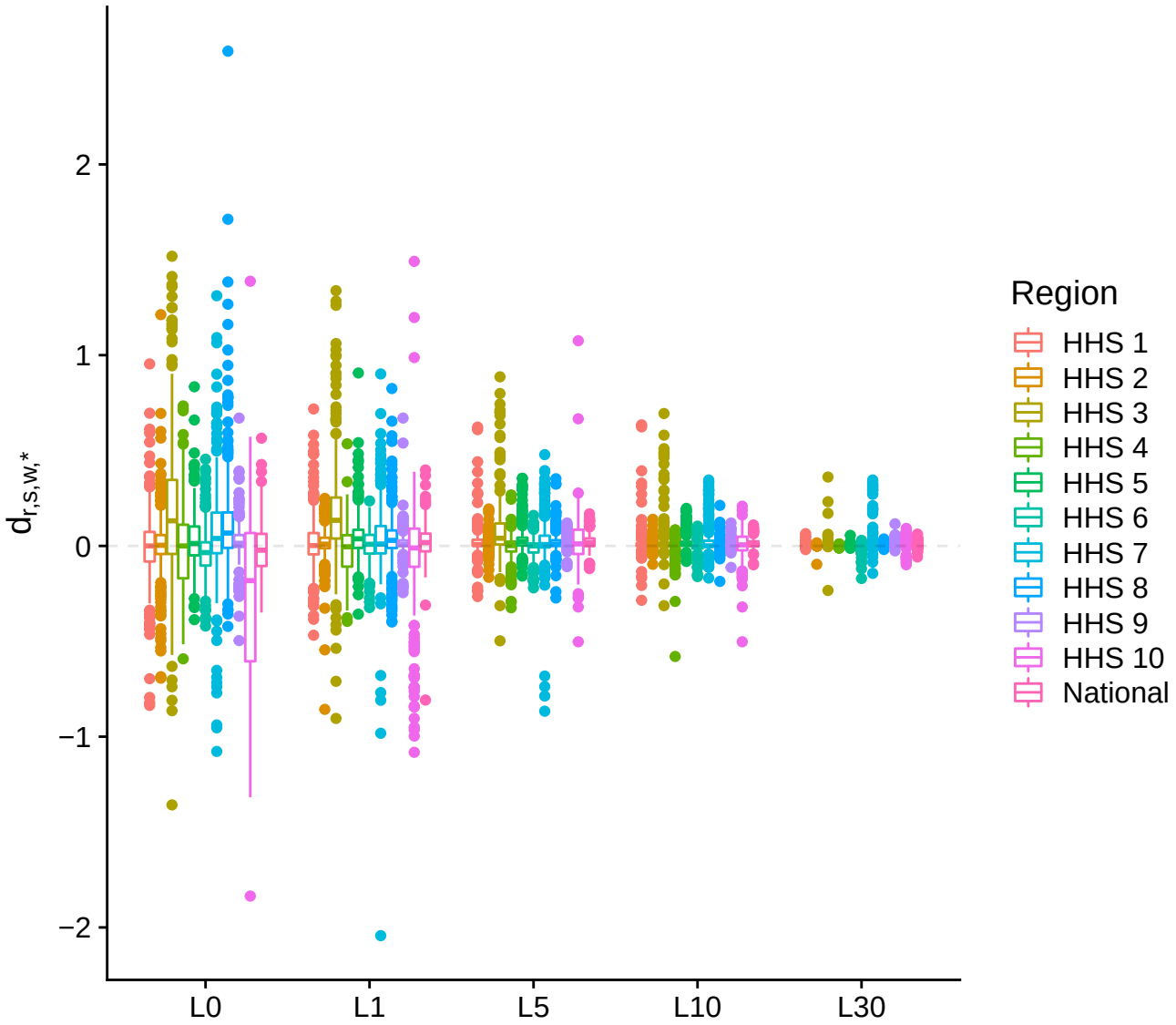

FIGURE 2 Reporting revision differences broken down by region. The x-axis represents the lag value at discrete intervals. We see that for all regions the revision differences converges to 0 as the lag increases and that for the most part revision differences are centered around the currently reported data.

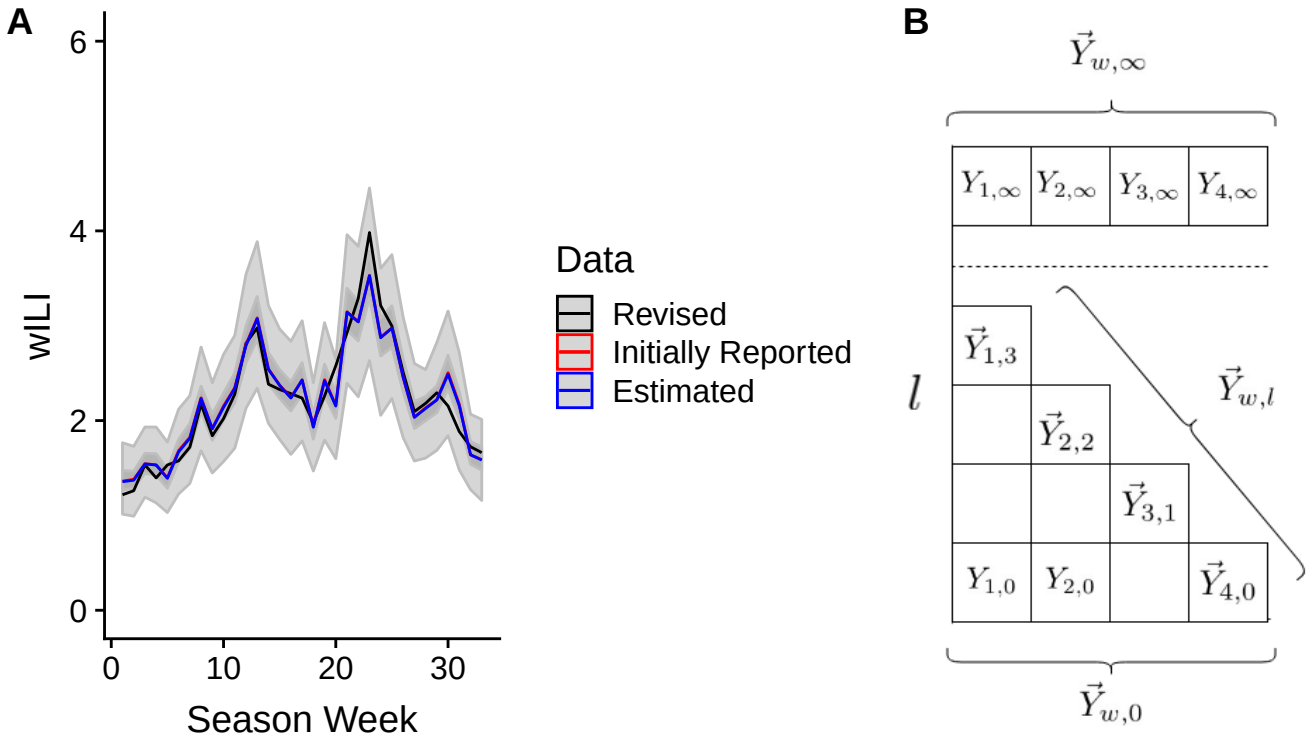

FIGURE 3 A. Example of observed data distribution g under the empirical distribution induced by sampling historical reporting revision differences and applying them to the currently observed data. Notice the uncertainty around the currently observed data as represented by both an 80 and 50 CI around the true observed data. The sampling method is able to put some positive probability on the finally revised data, but remains centered around the currently observed data. B Notation schematic highlighting the cross sections of data used in the experiments. Of primary interest are the three vectors: the set of revised data $\vec{Y}_{4,\infty}$, the most recent set of data $\vec{Y}_{4,l}$ and the initially reported set of data $\vec{Y}_{4,0}$.

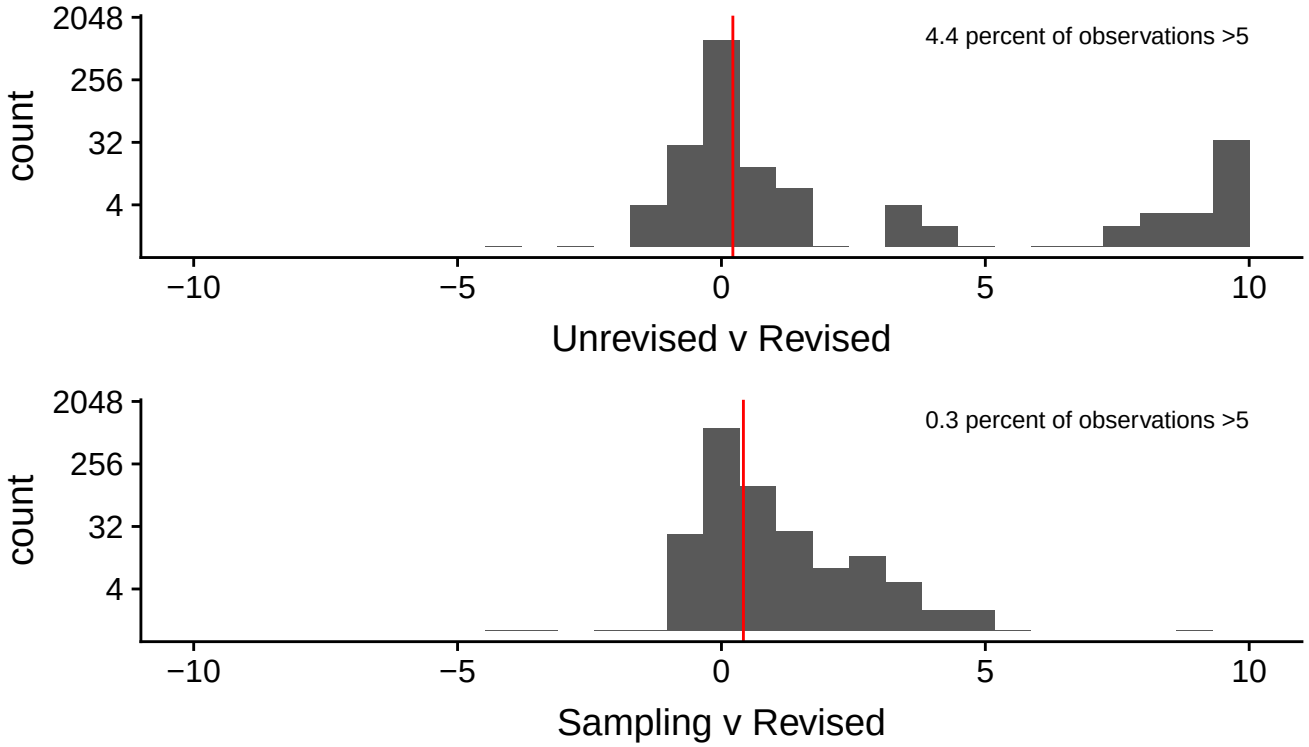

FIGURE 4 Log score histograms of the difference in scores between forecasts made from the method noted and forecasts made from the revised data for the season onset target. Mean difference is displayed by the red line. The sampling method is able to remove the tail of the histogram for the season onset target.Histograms are presented with log scaled y-axis.

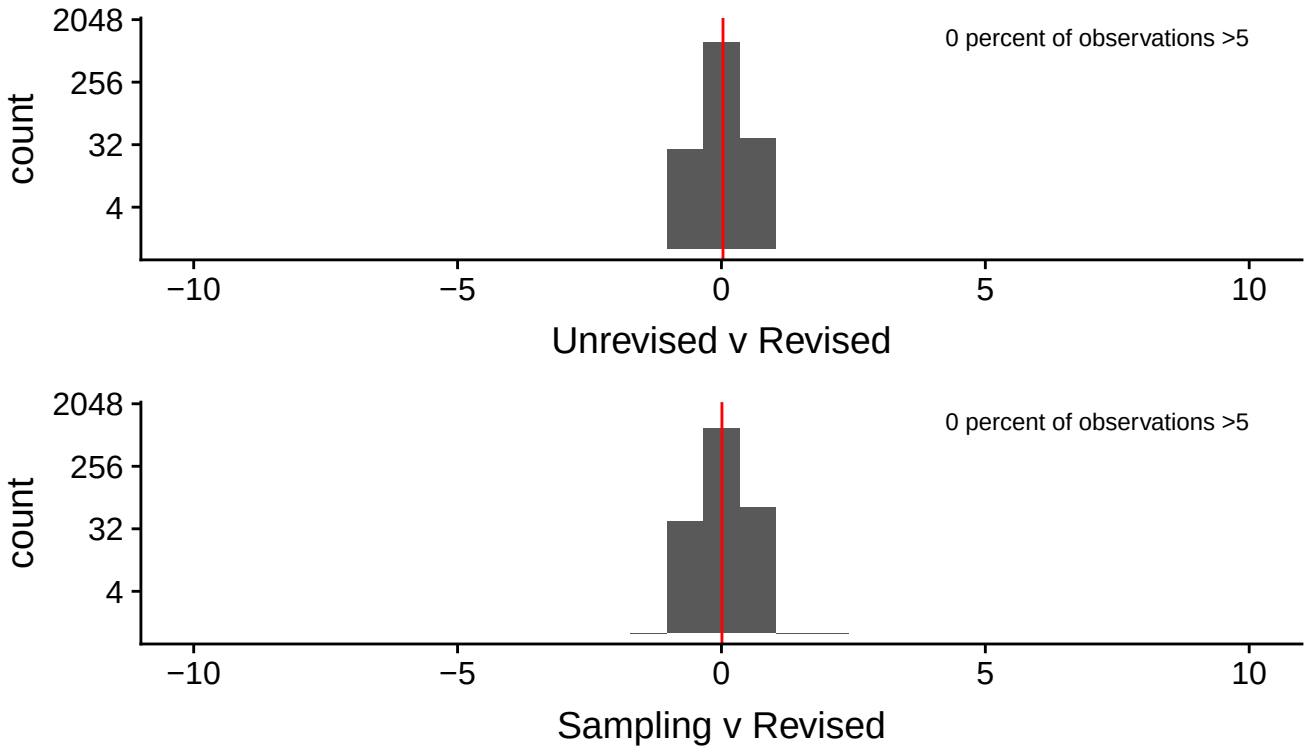

FIGURE 5 Log score histograms of the difference in scores between forecasts made from the method noted and forecasts made from the revised data for the 1 week ahead target. Mean difference is displayed by the red line. There is very little difference between the histograms. Histograms are presented with log scaled y-axis.

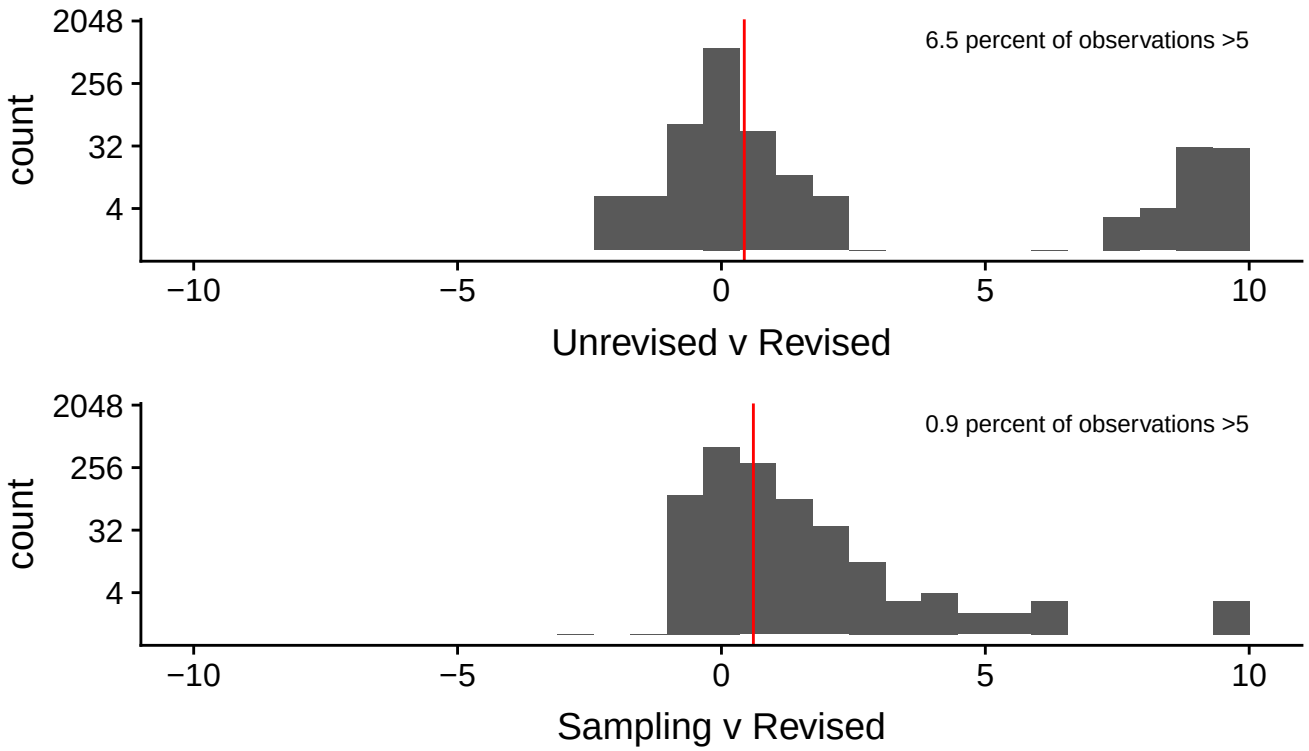

FIGURE 6 Log score histograms of the difference in scores between forecasts made from the method noted and forecasts made from the revised data for the peak week percentage. Mean difference is displayed by the red line. There is very little difference between the histograms. Histograms are presented with log scaled y-axis.

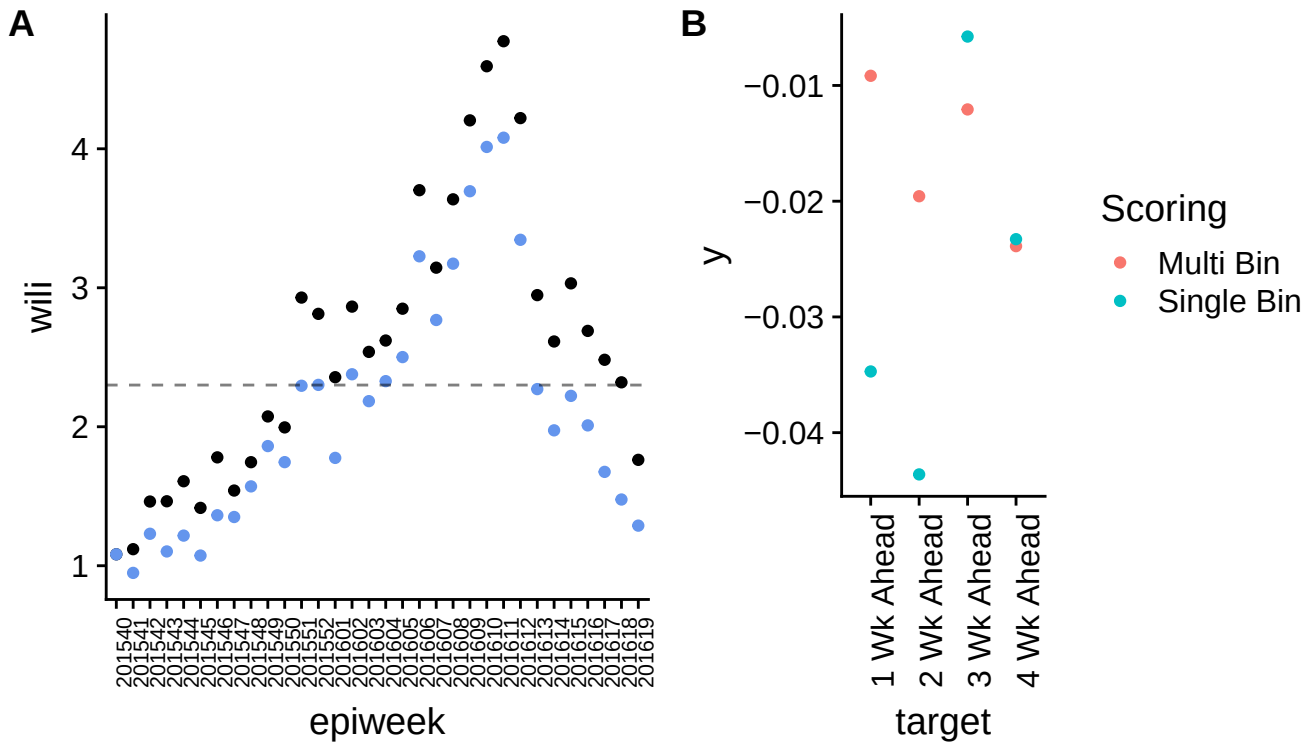

FIGURE 7 A. Example of HHS 2 seasonal target delay using currently reported data as of 2016 week 19 (black) against the fully revised data (blue). Notice that revisions are made to the sesaon onset at week 2016-01 that make initially reported season onset invalid. Similarly, season peak week is initially reported above the true value, so for all epiweeks after 2016-10 the model incorrectly places all denisty on or above the initally reported density. B Difference in log score between forecasts made from unrevised vs revised data for the week ahead targets under both multibin and single bin log scoring rules. We can see that, especially for 1-2 week ahead targets, there is a difference between the two scoring procedures. However, this difference is quite samll on the log score scale, suggesting the scoring rule is not masking the effect of revisions.

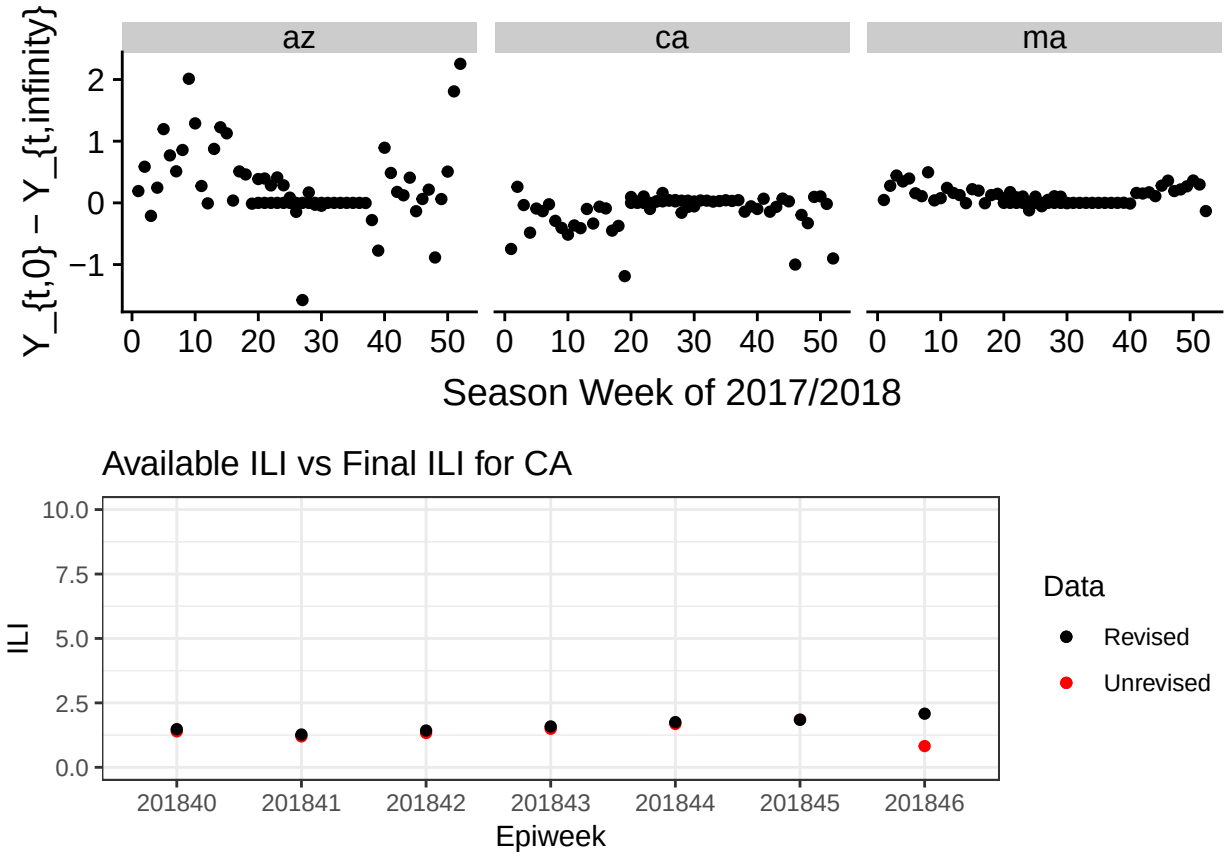

FIGURE 8 A Example of reporting delay for three states. Points are diffeences between initially observed values and finally reported values by season week. B Example of extreme reporting delay in state data, where initially reported value was well below finally revised data and denoted a large departure from the existing trend.