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

DOI: xxx/xxxx

RESEARCH ARTICLE

Accounting for Reporting Revisions in Influenza Data in the United States

Graham C. Gibson* | Evan L. Ray | Tom McAndrew | Nicholas G. Reich

[1]University of Massachusetts, Amherst

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 weighted influenza-like-illness (wILI) at the national and regional 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 at a given time are revised later on. Initial reports of wILI can be revised either upwards or downwards as additional data from reporting facilities are processed. In order to accurately forecast wILI percentages, accounting for these revisions is critical. Most currently used methods focus on a separate signal (e.g. internet search data) in order to estimate the true wILI at a given time. Unfortunately, relevant search query data are not always available and not always well correlated with the final reported wILI. We present a method that relies solely on historical revisions which shows significant improvements in the seasonal targets defined by the CDC as measured by average log score. Short-term incidence forecasts are less affected by revisions to reporting, and accounting for these revisions using our methods offers no substantial gains in performance.

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 vulnurable 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. The FluSight challenge is part of a larger epidemic prediction initiative put forth by the CDC to increase infectious disease forecasting infrastructure.[4] They see accurate forecasts of infectious disease as critical to preventing illness, allocating hospital resources, and assess economic burden.[5] 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.[6] Five teams and seven models were submitted in the 2017/2018 season.[7] Some teams have also incorporated external data to improve forecasts.[8 9 10]

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. This happens for a variety of reasons, and initial treports 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 assess model performance is by evaluating forecasts made from unrevised data against the revised final data at the end of the season.[11] Revisions may occur up to 10 weeks after the final week of the season, after which the CDC fixes the currently observed data as the "truth". 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 has 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 some partially observed signal.[12 13 14] Early attempts at correcting for reporting revisions focused solely on the under-reporting aspect.[15] 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.[16] Recently, Nunes et al. employed a hidden markov model to estimate the reporting revisions to wILI data from Portugal with success.[17] Stone et al. also investigated the application of state space models to the reporting delay problem with count data in a hierarchical setting.[18]

With respect to wILI 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 marginal gains in forecasting accuracy, and requires access to reliable real time data, which is not always available.[19] It is also conceivable that adjusting the current observed data by historical revisions may capture the true wILI better than a noisy signal.[13]

We find that a method that samples historical revisions to create a distribution of possible wILI values to forecast from performs the best over simply using an average historical revision to adjust the currently reported data. This method is able to capture the uncertainty inherent in the reported wILI, which is especially helpful with seasonal targets where revisions can alter the true seasonal peak wILI or season onset drastically. With a large history of model building already established, a key strength of our method is the ability to "plug-in" existing process models into the delay framework without altering the disease transmission dynamics or tuning parameters.

To this end, we have developed a novel set of methods to harness historical reporting revisions by building a probabilistic model over the currently observed data. We begin by developing a general framework for the problem of reporting revisions. We then propose specific methods for reporting revisions in the US wILI data. In section 2 we introduce the wILI data available to us and the nature of reporting revisions. We propose a set of methods to adjust currently reported data using estimated revisions. In section 3 we evaluate our proposed models on a variety of seasons and across all regions using log score. In section 4 we discuss the results, generalizability, and limitations of the methods proposed.

2 | METHODS

2.1 | Forecasting Models

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

$$f(z_t|y_1, ..., y_t, \theta) \tag{1}$$

for some observed data $y_1, ..., y_t$ and a vector of parameters θ. For example, y_t may be a measure of disease incidence at time t. We use Z_t to indicate an arbitrary forecast target relative to time t. For example, $Z_t = Y_{t+1}$ would be a 1-step ahead prediction and $Z_t = argmax_{t \in S}(Y_1, ..., Y_t)$ 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 $(Y_1, ..., Y_t)$ as random, not fixed. We denote the revised wILI for time t at time $t + l$ as $Y_{t,l}$. Borrowing from the notation of HÃűhle [16] 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. We then consider a joint distribution over the finally reported data and the forecast target, conditional on the currently reported data.

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

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 restrict our attention to joint distributions that can be factored into a forecast distribution conditional on some observed data, and an "observed data distribution" that captures our uncertainty over the currently reported data.

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

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

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

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

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

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.

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

where

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

.

We explore various models for G to account for specific revision processes that occur in the influenza data. To see how this effects our original forecast distribution, we examine some properties of our modified forecast density for an arbitrary observed data distribution. We can use the law of total expectation to arrive at the expected value of $\tilde{f}$

$$E_{\tilde{f}}(Z_t) = E_g(E_f(Z_t|Y_1, ..., Y_t)) \tag{10}$$

and similarly use the law of total variance to obtain the marginal variance of our forecast distribution.

$$Var_{\tilde{f}}(Z_t) = E_g[Var_f(Z_t|Y_1, ..., Y_t)] + Var_g[E_f(Z_t|Y_1, ..., Y_t)] \tag{11}$$

Therefore, we can see that g is able to alter both the expected value and the variance of our original forecast distribution. Particular choices of g are able to increase or decrease the variance of the forecast. In what follows we develop a probability model for g that will allow us to incorporate the uncertainty in the observed data. In practice we do not know the true observed data distribution g so we estimate it with $\hat{g}$, which could introduce either bias or variance depending on how close $\hat{g}$ is to g.

2.2 | U.S. Influenza Surveillance Data

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 extends 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).

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.[20]

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).

2.3 | Reporting revision ratios

We now introduce more specific notation and methods that are tailored to our application to influenza in the United States.

2.3.1 | Notation

Tto estimate the finally revised data $Y_{w,\infty}$ we need a historical estimate of how previous wILI values have been revised. We choose the ratio of the partially revised wILI value to the finally revised wILI value as the key scale-free unit. More specifically, at a given region r, season s, week w, lag l as a central concept in our model.

$$a_{r,s,w,l} = \frac{Y_{r,s,w,l}}{Y_{r,s,w,\infty}} \tag{12}$$

We choose this unit as it is of roughly consistent magnitude across seasons and season weeks with different absolute magnitude.

2.3.2 | Mean scale up

The simplest model for the expected revision to partially revised wILI value is to simply take the mean over the observed reporting revisions. Specifically, if the current time is $w + l$ and we are revising week w in region r in season s we use the estimator

where we average over region, seasons, and weeks.

$$\hat{a}_{.,.,.,l} = \frac{1}{N} \sum_{r,s,w} a_{r,s,w,l}$$

We are therefore ignoring any region or season deviations from the average reporting revision, and only asserting that estimated revision should match the lag value we are trying to revise.

We can frame this in the general forecasting model notation as a degenerate g distribution using a delta function that denotes a point mass around the argument.

$$g(y_{r,s,w,\infty} | \vec{y}_{r,s,w,l}, \vec{a}_{r,s,w,l}) = \delta \left(\frac{y_{r,s,w,l}}{\hat{\vec{a}}_{r,s,w,l}} \right) \tag{13}$$

We can see from this definition of g, using equation 7 we have

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

$$E_{\tilde{f}}(z_{r,s,w,\infty}) = E_f(z_{r,s,w,\infty} | \frac{\vec{y}_{r,s,w,l}}{\hat{\vec{a}}_{r,s,w,l}}) \tag{14}$$

Therefore, we can see that the expected value of our forecast target under the mean revision model is simply the observed data scaled by the expected reporting revision ratio.

We can also investigate the variance using equation 8

$$Var_{\tilde{f}}(z_{r,s,w,\infty}) = E_g[Var_f(z_{r,s,w,\infty}|\vec{y}_{w,\infty})] + Var_g[E_f(z_{r,s,w,\infty}|\vec{y}_{w,\infty})]$$

Here $Var_g(X) = 0$ regardless of X because g is a point mass.

$$E_g[Var_f(z_{r,s,w,\infty}|\vec{y}_{r,s,w,\infty})] = Var_f(z_t|\frac{\vec{y}_{r,s,w,l}}{\hat{\vec{a}}_{r,s,w,l}}) \tag{15}$$

following the same steps as above. Note that there is additional uncertainty in practice, where we need to estimate g by $\hat{g}$ since we do not know the true observed data distribution.

We consider alternative models for a degenerate g distribution in the appendix through different models for $a_{w,l}$. However, we found no appreciable differences in performance among the methods using estimates of the mean revision ratio. To streamline presentation, in the main manuscript we discuss results for the simple mean method to estimate $\hat{a}_{w,l}$. Results for the alternative models for mean revisions are presented in the supplementary materials.

2.3.3 | Sampling

We also consider a non-parametric form of g based on on historical reporting revision ratios.

$$g(\vec{y}_{r,s,w,\infty}; \vec{y}_{r,s,w,l}, \vec{a}_{w,l}) = \frac{1}{n}\sum_{i=1}^{n} \delta\left(\frac{\vec{y}_{r,s,w,l}}{\vec{a}^{(i)}_{r,s,w,l}}\right) \tag{16}$$

Sampling from g amounts to simply drawing $\vec{a}^{(i)}_{r,s,w,l}$ from historical reporting revisions and then dividing the observed data $\vec{Y}_{r,s,w,l}$ by the sampled revision trajectory. 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 using eqn 7.

$$\begin{aligned}
E_{\tilde{f}}(Z_{r,s,w,\infty}) &= \int_{\vec{Y}_{r,s,w,\infty}} E_f(Z_{r,s,w,\infty}|\vec{y}_{r,s,w,\infty}))dG_{\vec{Y}_{r,s,w,\infty}} \\
E_{\tilde{f}}(Z_{r,s,w,\infty}) &= \int_{\vec{Y}_{r,s,w,\infty}} \frac{1}{n}\sum_{i=1}^{n} E_f(Z_{r,s,w,\infty}|\frac{\vec{y}_{r,s,w,l}}{\vec{a}^{(i)}_{r,s,w,l}}))dG_{\vec{Y}_{r,s,w,\infty}} \\
E_{\tilde{f}}(Z_{r,s,w,\infty}) &= \frac{1}{n}\sum_{i=1}^{n} E_f(Z_{r,s,w,\infty}|\frac{\vec{y}_{w,l}}{\vec{a}^{(i)}_{w,l}})) \int_{\vec{Y}_{w,\infty}} dG_{\vec{Y}_{r,s,w,\infty}} \\
E_{\tilde{f}}(Z_{r,s,w,\infty}) &= \frac{1}{n}\sum_{i=1}^{n} E_f(Z_{r,s,w,\infty}|\frac{\vec{y}_{r,s,w,l}}{\vec{a}^{(i)}_{r,s,w,l}}))
\end{aligned} \tag{17}$$

where the last integral is 1 since it is a valid probability density. Therefore, the expected value of the sampling method is simply the average over the forecasts from the sampled revision trajectories.

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

$$E_g[Var_f(Z_{r,s,w,\infty}|\vec{y}_{r,s,w,\infty})] = \int_{\vec{Y}_{r,s,w,\infty}} Var_f(Z_{r,s,w,\infty}|\vec{y}_{r,s,w,\infty})dG_{\vec{Y}_{r,s,w,\infty}} \tag{18}$$

$$= \frac{1}{n}\sum_{i=1}^{n} Var_f(Z_{r,s,w,\infty}|\frac{\vec{Y}_{r,s,w,l}}{\vec{a}^{(i)}_{r,s,w,l}}) \tag{19}$$

The second term in eq. 8 is a variance with respect to the observed data distribution, and is therefore always positive. We now see that if the following inequality holds, the sampling method variance is strictly greater than the variance of forecasting from unrevsied data.

$$\frac{1}{n}\sum_{i=1}^{n} Var_f(Z_{r,s,w,\infty}|\frac{\vec{Y}_{r,s,w,l}}{\vec{a}^{(i)}_{r,s,w,l}}) \geq Var_f(Z_{r,s,w,\infty}|\vec{Y}_{r,s,w,l}) \tag{20}$$

That is, if the average forecast variance obtained by sampling revisions and applying them to the observed data is greater than or equal to the variance of forecasting from the unrevised data, then the sampling method will increase the variance of the predictive distribution.

2.4 | Forecasting Model

In order to forecast wILI across all HHS regions we use the SARIMATD method from the sarimaTD package.[21] We use the default options that include seasonal differencing and Box-Cox transformation of the underlying wILI data to normality. The SARIMA model fit is of the form

$$Y_{r,s,w,\infty} = \alpha_0 + \alpha_1 \cdot Y_{r,s,w-1,\infty} + ... + \alpha_c \cdot Y_{r,s,w-c,\infty} + \beta_1 \cdot Y_{r,s-1,w,\infty} + ... + \theta_1 e_{t-1} + ... + \theta_{t-q} e_{t-q-1} + \epsilon_{r,s,w,\infty}$$

See SARIMATD for further details.

We consider this as a canonical forecasting algorithm to perform our experiments. We do this because we are not particularly interested in the exact properties of the forecast distribution $f(Y_{r,s,w+1,\infty}|Y_{r,s,1,w}, ..., Y_{r,s,w,0})$ but rather the relative change in performance between the unrevised forecast distribution and our altered distribution $\tilde{f}(Y_{r,s,w+1,\infty}|Y_{r,s,1,w}, ..., Y_{r,s,w,0})$. Note that the forecast variance is constant with respect to the data used to forecast in a SARIMA model, so the variance of the forecast is strictly greater than forecasting from fixed observed data.

$$Var(Z_t) = \frac{1}{n}\sum_{i=1}^{n} Var_f\left(Z_t|\frac{\vec{Y}_{r,s,w,l}}{\vec{a}^{(i)}_{r,s,w,l}}\right) \tag{21}$$

$$Var(Z_t) = \frac{1}{n}\sum_{i=1}^{n} \sigma^2_{Z_t} = \sigma^2_{Z_t} \tag{22}$$

This choice of process model forces the variance to be larger than forecasting from fixed observed data and the theoretical properties may change depending on the process model behavior, a fact we use in the discussion.

2.5 | Evaluation

2.5.1 | Experimental setup

In order to compare forecasts made from the unrevised data against 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, 2017/2018 as model test seasons. We fit the model to each region separately. Unfortunately, because of limited data availability (specifically with respect to HHS region data), our model validation is limited to three seasons. Along with model fitting, we also use the data from 2010/2011 through 2015/2016 to estimate the reporting revision ratios. We assume that underlying reporting revision process that occurred in the training seasons also occurs in the testing seasons. Although we do not explicitly evaluate the estimation of the reporting revision ratios, the reporting revision estimation performance is tied into the log score of the wILI target forecast.

2.5.2 | 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 against the truth as simply the log of the probability assigned to the truth. In order to avoid $-\infty$ when the probability assigned to the target is 0 we truncate at -10.

$$\text{log score}_{r,s,w,Z_t} = max(-10, log(p_{r,s,w,Z_t,i})) \tag{23}$$

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.

$$\text{mutlibin log score}_{r,s,w,Z_t} = log(\sum_{i \in I} max(-10, p_{r,s,w,Z_t,i})) \tag{24}$$

We log the sum of the probability assigned to each point in the set of true values. 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 is used when making a forecast for a particular target during the testing phase.

3 | RESULTS

3.1 | Data revisions usually don't matter, but when they do they have a big impact

The difference in log scores for 1 week-ahead forecasts made from revised and unrevised data in a particular week were between [-.348,.398] 95% of the time. This shows that the revisions play a small role in 1-4 week ahead forecasts. On the probabiity scale, this corresponds to a multiplicative difference of [0.706,1.48] on the probability scale. Contrast this with the difference in log scores for forecasts made from revised and unrevised data for the season onset target, which were between [-.44,10] 95% of the time. This corresponds to a multiplicative change of [.6,22026.7], a much larger change. This is further illustrated in Figures 4 ,5 ,6 However, for certain targets in certain regions the difference in log score is quite large. This suggests that reporting revisions usually don't affect forecasts, but when they do they cause a large difference in log scores. This is highlighted in Figure 7 B, where the difference in log score between the revised and unrevised data is usually very small, regardless of the amount of revision. We quantify the amount of revision through the variance of initially reported revision ratio, since reporting revision ratios are centered around the currently observed data. However, we notice a few extreme outliers, specifically region 2 in the 2015 season. This makes correcting for reporting revisions a particularly difficult task. We need to both model a rare event and ensure we do not hurt forecasts when the event does not occur.

The fact that reporting revisions usually don't affect the log score of forecasting seems to stem from two properties of the revisions. First, as we can see in Figure 2 , the revisions are centered around the observed data. This means that within a single season and region, data are revised both upwards and downwards, making prediction of the direction of revisions particularly difficult. The second complicating factor is the multibin scoring adopted the CDC for the flusight challenge. In order for a revision to have an appreciable impact on log score, it would have to shift the forecast distribution such that the total probability mass assigned to I changes. There are multiple predictive distributions that map to the same amount of probability mass assigned to the set I. For example, consider the predictive distribution of 1 step ahead forecast that assigned .5 to the truth and .1 to the bins to the two bins to the left and two bins to the right. Therefore, $P(I) = .1 + .5 + .1 = .7$. Correcting for reporting revisions may simply shift the forecast one bin to the right, yielding a total probability of $P(I) = .1 + .1 + .5 = .7$. We can see that the importance of reporting revisions is mitigated under multibin scoring. This is evident in Figure 8 B.

In order to evaluate the difference in log scores formally, we employ the Diebold Mariano test.[22] This is shown in Table 2, where we extract the pairwise p-value of the DM test between methods for all weeks in all regions by target. Most importantly, we see that for almost all week ahead targets (1,2,4) there is no significant difference ($\alpha = .05$) between forecasting from the revised versus the unrevised data. However, for all seasonal targets we see a significant difference between forecasting from the revised and unrevised data.

3.2 | The impact of reporting revisions is target specific

Forecasts made from the revised data on 1-4 week ahead targets show minimal difference from forecasts made based on unrevised data, highlighting that revisions are not the main driver of 1-4 week ahead log score. This is confirmed in Figure 7 A where the majority of low (< 3) log scores for 1 step ahead forecasts fall within a reporting revision ratio of [.9,1.1], indicating that the lowest log scores come from the inherent difficulty of forecasting wILI, rather than the reporting revisions that occur. However, Seasonal targets tell a different story. As indicated in Figure 4 , while most regions display little to no difference in seasonal target log score, some regions display a big gain in log score after revising the data. This is most apparent in the season onset and season peak percentage targets. As noted above, 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 8 , the season

onset can be revised below the baseline, moving the truth more than 1 week (multibin scoring limit) away from the currently reported season onset. This results in a log score of negative infinity (truncated to negative ten) when predicting season onset after it has been initially observed. 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. If we are at the end of the season, 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 (negative ten) log score. Both of these extreme situations highlight when reporting revisions matter, but the probability of being in either of these situations is low.

3.3 | Sampling method can improve forecast accuracy for seasonal targets, but hurt others

Sampling algorithm improves the log score of forecasts for seasonal targets. Using the sampling method we are able to avoid the extreme cases (-10 log score) since we do not treat the current observed data as finally revised. 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 revsied 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 8 , 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 and peak week as shown in Figure 6.

Almost all of the statements and results about the sampling method are model agnostic. The benefits outlined in Figure 8 B are process model agnostic since they concern situations in which the seasonal target has already been observed under the currently reported data. Adding uncertainty to the observed data via our choice of F_g does not rely on any process model forecasts. Therefore, the results on season targets do not depend on the choice of process model and log scores should improve regardless of the particular form of the predictive density.

However, the sampling method negatively affects 1-4 week ahead forecasts. This can be seen from the theoretical properties derived above (eq 22) with regards to the SARIMATD process model. The expected value of the sampling method will be close the forecast made from unrevised data, since the $\hat{a}_{w,l}$ are centered around 1 and the variance is strictly larger. This means we are spreading out our 1-4 week ahead predictive distributions while keeping them centered, on average, around the currently observed data. This explains the slight decrease in log score consistently observed in Figure 4 .

4 | 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 ratios 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. However, the effects of revisions on short term forecasts are mitigated by both the inherent difficulty of forecasting wILI and the FluSight specific multbin scoring rule.

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.

Further investigation into an ensemble approach combining external data sources with historical reporting revision ratios may yeild even more benefit during forecasting. Research seems to suggest that an external signal may help improve nowcasting, so combining this with hisotrical 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.

Even after accounting for reporting revisions, accurate forecasting of wILI reamins a difficult task. The complex transmission dynamics and limited data availability mean the main source of forecast error is simply the underyling model, not the reporting revisions. In addition, wILI is not a perfect signal of the true level of influenza in the population.[11] 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.

5 | APPENDIX

Below we describe additional mean models that take into account possible reporting revision ratio differences with regards to week in season and across region.

Mean reporting revision: region specific

We extend this basic model by allowing revisions to vary over weeks of the current region *r* and season *s*. This allows for different reporting revisions in peak parts of the season versus low-level regions of the season. Although we have gained flexibility in the reporting revisions, we have decreased the number of historical revisions used to estimate each parameter.

$$\hat{a}_{r,s,w,l} = a_{\cdot,\cdot,w,l}$$

Mean reporting revision: week specific

We instead assume that reporting ratios vary over region more than they vary over the week of the season. While this gains us flexibility in the reporting ratios we also limit the amount of data used to estimate each reporting ratio.

$$\hat{a}_{r,s,w,l} = a_{r,\cdot,\cdot,l}$$

Hierarchical random effects

We depart from the non-parametric estimates and instead adopt a hierarchical linear regression model where we assume that the reporting ratio depends on the week of the season and region, where the effect of region is allowed to vary according to our random effect.

$$\hat{a}_{r,s,w,l} \sim N(\alpha + \beta_w + b_r, \sigma^2),\ \tilde{b} \sim N(0, \Sigma)$$

This allows for flexibility in the reporting by varying both over week and region, but in a structured way. Note that the support of the reporting ratio is $[0, \infty]$ but we use a normal approximation based on the observed ratios in Figure2B which are approximately normally distributed around 1 ($N(1, \sigma^2)$).

Non-linear

We then relax the linear effect of week and region assumption by employing a basic feed-forward neural network to estimate the reporting revision. We again assume that the reporting revision is a function of the week in season and the region.

$$\hat{a}_{r,s,w,l} = g(\alpha, \beta_w, \gamma_r)$$

where

$$g(\alpha, \beta_w, \gamma_r) = \sigma_2(W_2 \cdot \sigma_1(W_1 \cdot [\alpha, \beta_w, \gamma_r]^T))$$

where W_1 is a 20x4 weight matrix of parameters and W_2 is a 20x1 weight matrix of parameters updated using back-propagation. Here we choose $\sigma_1 =$ sigmoid and $\sigma_2 =$ identity function.

For all of the models above we employ the estimated revision ratio to estimate the fully revised data as

$$\hat{Y}_{r,s,w,\infty} = \frac{Y_{r,s,w,l}}{\hat{a}_{r,s,w,l}}$$

That is, we take the current data and scale it by an expected future revision.

6 | FIGURES

TABLE 1 p-values from Diebold Mariano Test of Difference of Log Score Between Methods. Notice that for almost all 1-4 week ahead targets there is no statistically significant (at .05 level) difference between forecasts made from the revised data and the unrevised data. We also see that on seasonal targets there is a statistically significant difference between the log score of forecasts made from the unrevised data and from the revised data. However, only the sampling method (not the mean scale up method) is able to similarly reject the null that the difference in log score between the methods is 0. We also note that the sampling method is statistically signficiantly different from the unrevised data when forecasting 1-4 week ahead. This means there is a statistically signficant reduction in log score.

	Sampling v Unrevised	Mean v Unrevised	Revised v Unrevised
1 Wk Ahead	0.01051	0.3192	0.3652
2 Wk Ahead	0.1261	0.3089	0.01916
3 Wk Ahead	0.0663	0.9228	0.726
4 Wk Ahead	0.2878	0.8729	0.2526
Peak Week	0.05918	0.1533	0.002415
Peak Week Percentage	1.136e-11	0.8015	3.244e-17
Season onset	4.745e-10	0.5355	3.848e-12

Week	Data Used	K-step Ahead Target	Season onset	Season peak week	Season peak percentage
1	$Y_{r,s,1,0}$	$Y_{r,s,1+k,\infty}$	$w\ s.t. Y_{r,s,w:w+1:w+2,\infty} \geq$ onset	$argmax_w Y_{r,s,w,\infty}$	$max_w Y_{r,s,w,\infty}$
2	$Y_{r,s,1:2,1:0}$	$Y_{r,s,2+k,\infty}$	$w\ s.t. Y_{r,s,w:w+1:w+2,\infty} \geq$ onset	$argmax_w Y_{r,s,w,\infty}$	$max_w Y_{r,s,w,\infty}$
...	...	...	...	...	...
52	$Y_{r,s,1:52,51:0}$	$Y_{r,s,52+k,\infty}$	$w\ s.t. Y_{r,s,w:w+1:w+2,\infty} \geq$ onset	$argmax_w Y_{r,s,w,\infty}$	$max_w Y_{r,s,w,\infty}$

TABLE 2 Forecast template for a given region r and season s, outlining what data is used to forecast each target.

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. Chretien Jean-Paul, Swedlow David, Eckstrand Irene, et al. Advancing epidemic prediction and forecasting: a new US government initiative. *Online journal of public health informatics.* 2015;7(1).

5. ;.

6. 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.

7. 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.

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

9. 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.

10. 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.

11. 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.

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

13. 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 -.

14. 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.

15. 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.

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

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

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

19. 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.

20. *Influenza (Flu).* 2018.

21. https://github.com/reichlab/sarimaTD. ;.

22. Diebold Francis X, Mariano Robert S. Comparing predictive accuracy. *Journal of Business & economic statistics.* 2002;20(1):134–144.

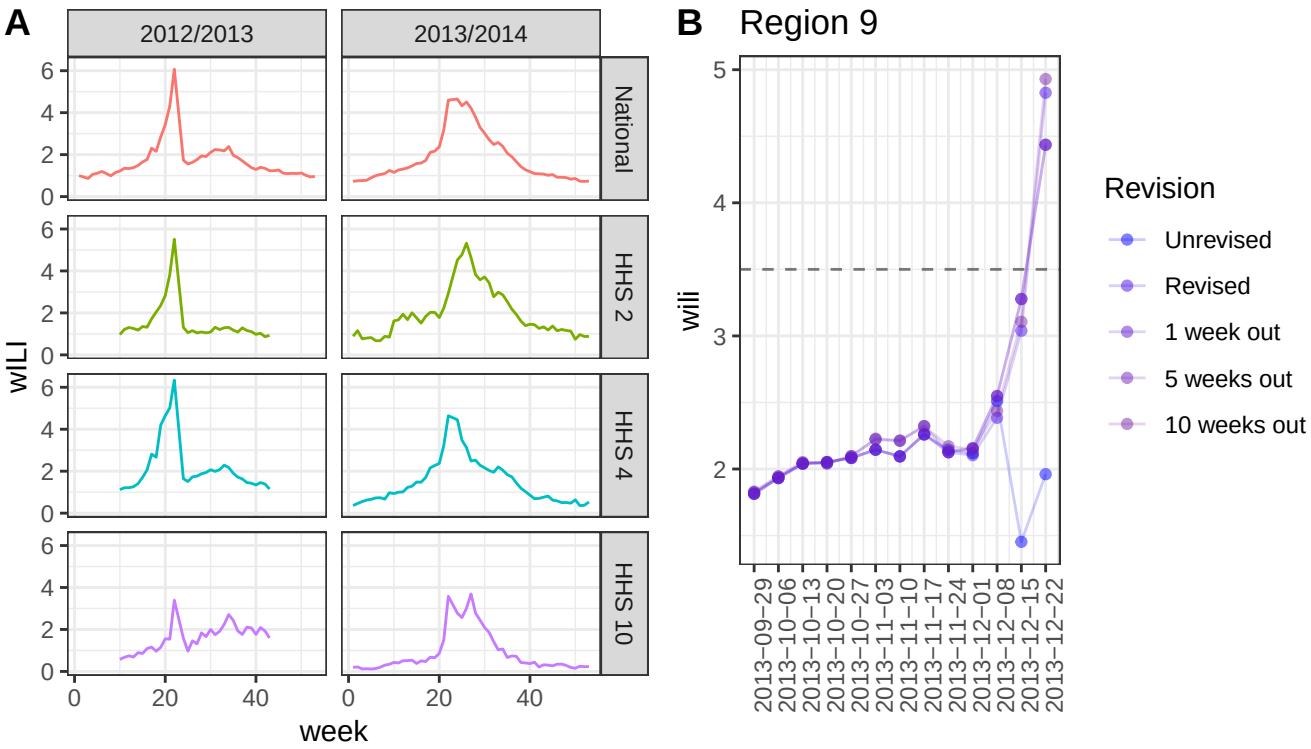

FIGURE 1 A: wILI data from the 2012/2013 and the 2013/2014 season across 4 example regions. Notice the regional variability and the seasonal structure with a peak usually (but not always) occuring somewhere between week 20 and week 30. B: Data from 2013-09-29 (week 1 of the 2013/2014 season) to 2013-12-22 (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.

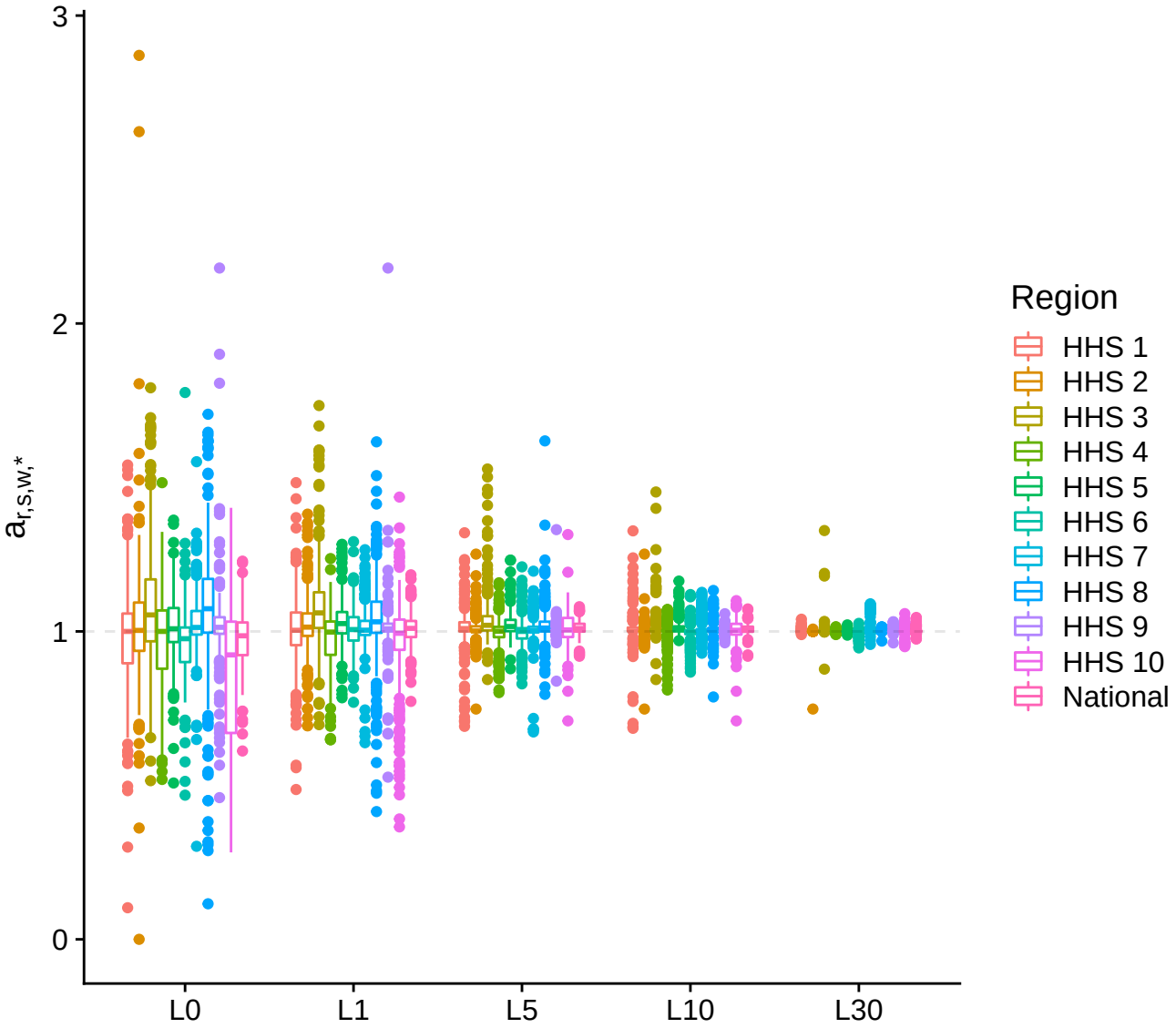

FIGURE 2 Reporting revision ratios broken down by region. The x-axis represents the lag value at discrete intervals. We see that for all regions the revision ratio converges to 1 as the lag increases and that for the most part revision ratios are centered around the currently reported data.

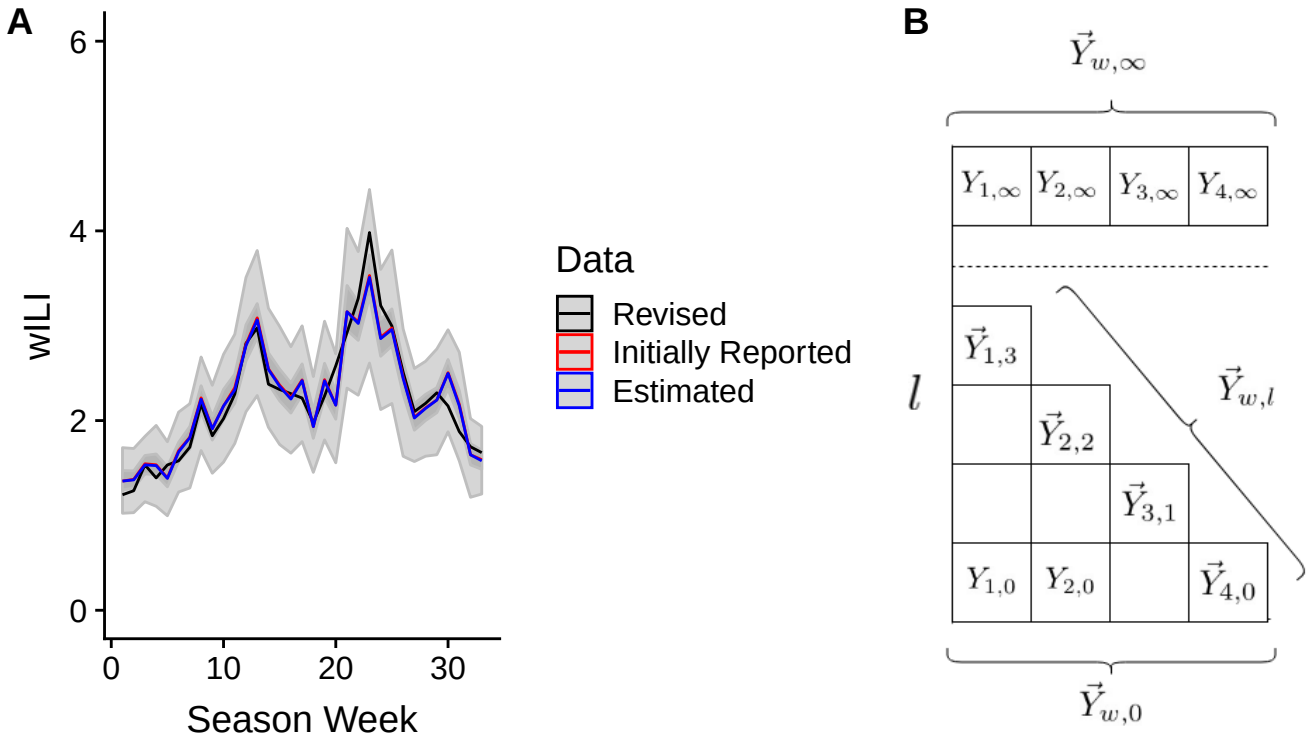

FIGURE 3 A. Example of observed data distribution g under the empirical distribution induced by sampling historical reporting revision ratios and applying them to the currently observed data. Notice the uncertainty around the currently observed data as represented by both an 80 and 50 CI around the true 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 $\vec{Y}_{4,0}$.

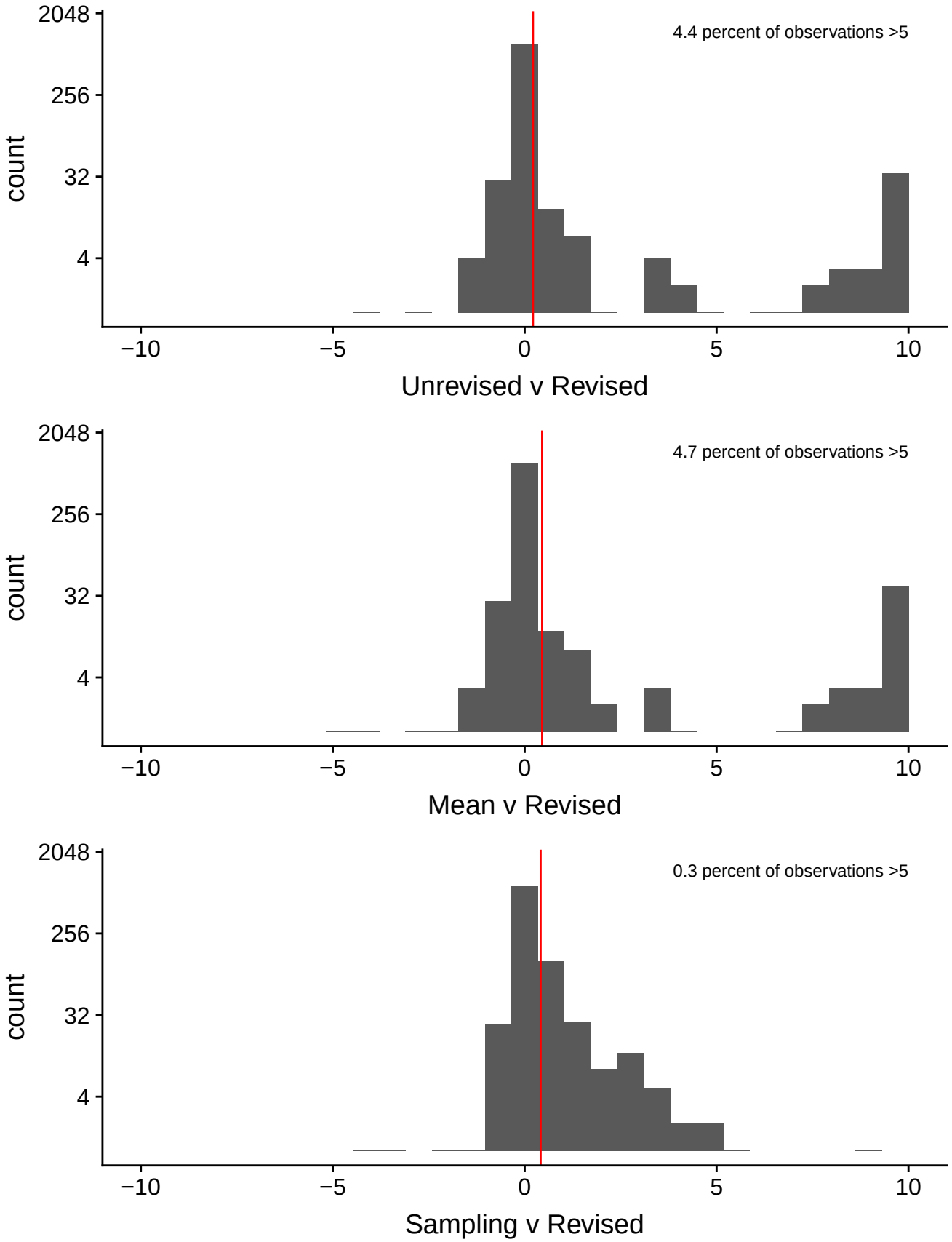

FIGURE 4 Log score histograms of the difference in scores between forecasts made from the method noted and forecasts made from the revised data for the season onset target. Mean difference is displayed by the red line. The sampling method is able to remove the tail of the histogram for the season onset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

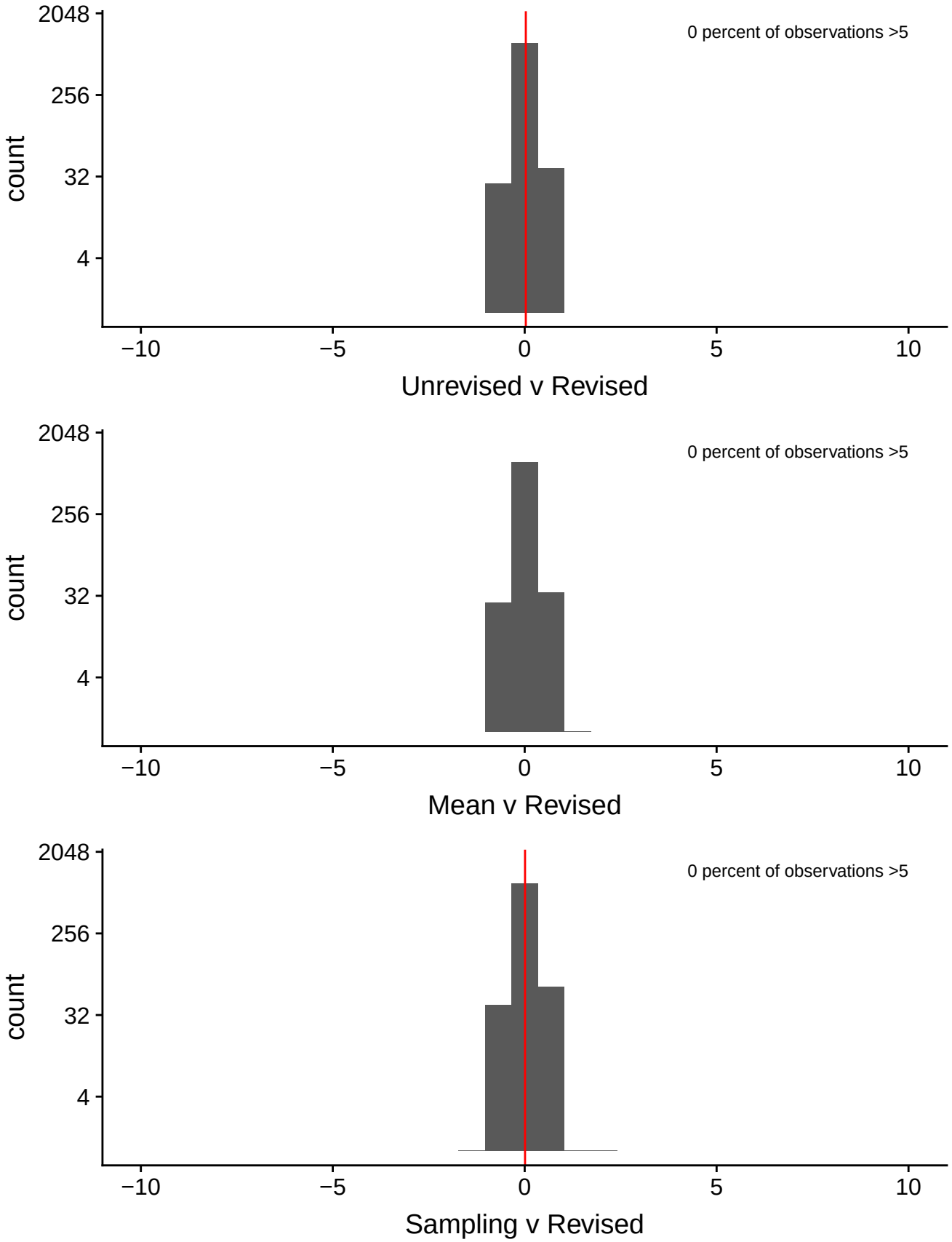

FIGURE 5 Log score histograms of the difference in scores between forecasts made from the method noted and forecasts made from the revised data for the 1 week ahead target. Mean difference is displayed by the red line. There is very little difference between the histograms. Histograms are presented with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

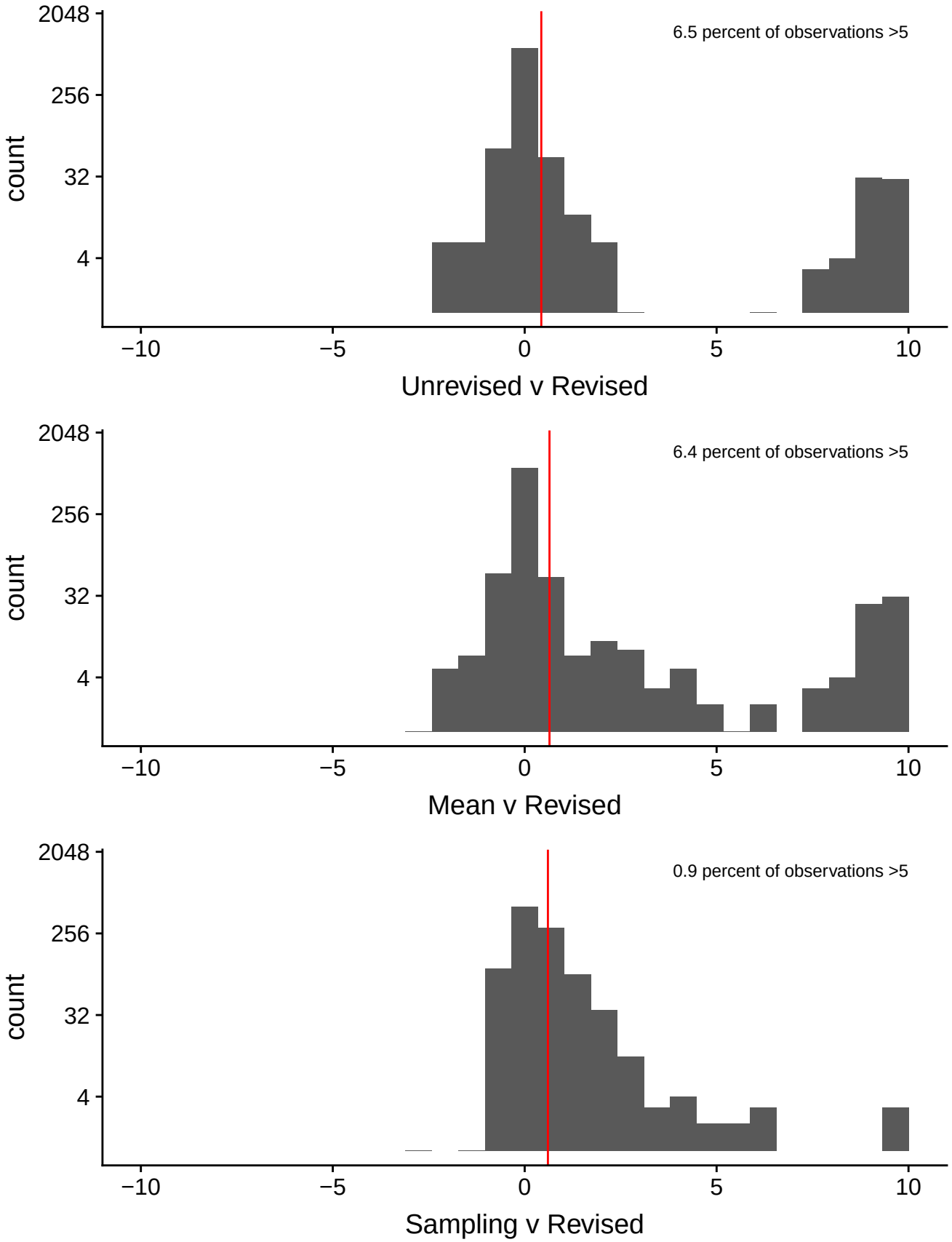

FIGURE 6 Log score histograms of the difference in scores between forecasts made from the method noted and forecasts made from the revised data for the peak week percentage. Mean difference is displayed by the red line. There is very little difference between the histograms. Histograms are presented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

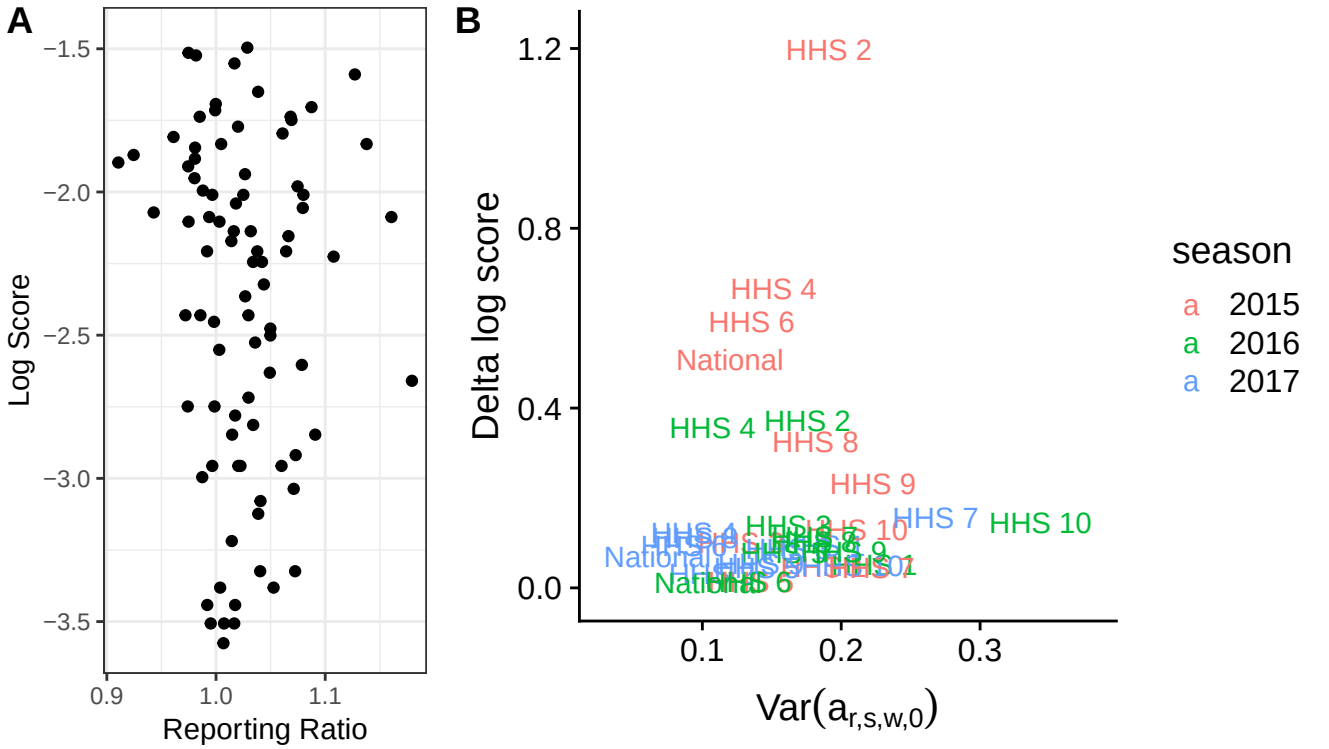

FIGURE 7 A: By examining the correlation between log score and reporting revisions we see that poor performing 1 week ahead forecasts are not highly correlated with extreme reporting ratios (further from 1). B Relationship between the sample based variance of the initially reported revision ratios over all test seasons and all test regions and the difference in the log score of forecasts based on the revised and the unrevised data. A single point represents a single region and season combination for all test seasons. Notice that as the variance of the reporting ratio decreases so does the difference in the log score, and therefore the room for improvement made by the revision algorithms diminishes. This is expected, as the more revisions occur during a season, the more the revision algorithms would help.

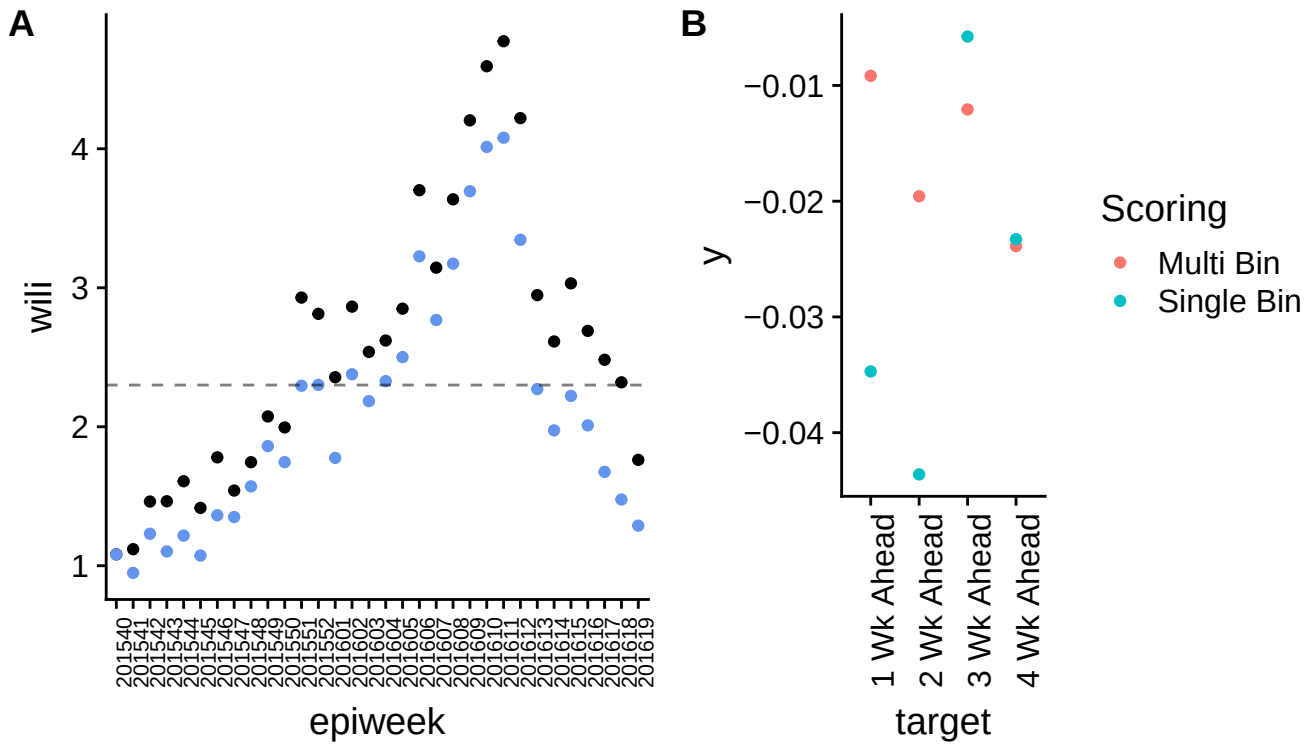

FIGURE 8 A. Example of HHS 2 seasonal target delay using currently reported data as of 2016 week 19 (black) against the fully revised data (blue). Notice that revisions are made to the sesaon onset at week 2016-01 that make initially reported season onset invalid. Similarly, season peak week is initially 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.