EPIDEMICS-D-24-00094
Flusion: Integrating multiple data sources for accurate influenza predictions
Response to Reviewer Comments

September 19, 2024

We thank the editor and reviewers for their comments on this manuscript. We provide a detailed response to these comments below. Reviewer comments are given in **bold font** with a bar at the left margin, and our responses are in regular font. Where appropriate, we show changes to the manuscript with quoted text from the manuscript in a box, new text shown in blue, and deleted text shown in red. Note that in quoted text, reference numbers have changed from the original submission to the revised submission due to the inclusion of additional references.

1 Reviewer 1 Comments

> **Reviewer Comment 1: The authors describe a novel model for forecasting influenza hospitalizations, and show that their method produces forecasts that are more accurate than an ensemble of several competing models. This is important work, with the potential to inform improved forecasting not only of influenza, but of other infectious diseases as well. However, the manuscript would benefit from clarifying several of the concepts it discusses. There are also many details that feel superfluous, in particular about the details of the FluSight forecasting challenge. In general, the writing could be much more concise, particularly in the introduction and methods portions of the manuscript - as it stands, ideas are often repeated several times.**

We thank the reviewer for their comments, and in particular for their constructive suggestions about how to make the writing clearer. We address these individual suggestions below.

> **Reviewer Comment 2: If possible, could the authors include line numbers in future submissions? That would make it easier to comment on specific parts of the manuscript.**

We have added line numbers to the revised submission.

1.1 Introduction:

> **Reviewer Comment 3: In general, I would have liked to see more discussion of influenza forecasts and their utility, rather than a detailed introduction into the CDC's flu prediction challenge. As it stands, the first two paragraphs make it sound as though the goal of the authors' model is simply to win a challenge, but in reality, I assume the ultimate goal is to accurately predict flu activity in a way that is useful to decision makers.**

We have revised the first paragraph of the manuscript to place a greater focus on how influenza forecasts may be used to support public health decision making, removing information about the FluSight challenge that is not of direct relevance to this manuscript or relocating it to other paragraphs later in the introduction:

Since the early 2010s, short-term forecasting for infectious diseases has become an increasingly common activity, often through collaborations between governmental and industry or academic partners. Starting during the 2013/2014 influenza season, the FluSight collaborative forecasting exercises organized by the US Centers for Disease Control and Prevention (CDC) have brought together teams of academic, industry and governmental researchers to produce probabilistic forecasts of influenza activity in the United States [5]. ~~After a pause during the first two years of the COVID-19 pandemic, this forecasting exercise restarted in the spring of 2022 [47]. FluSight typically involves over 20 different teams submitting forecasts each week, using different methodologies and sometimes different data sources. Ensemble forecast techniques have been used by CDC and other groups to combine individual team submissions into a single consensus forecast, which typically has shown some of the most accurate performance overall [30,41].~~ A primary motivation for the FluSight challenges is to carefully evaluate forecasts against real data and to use them to improve situational awareness and yield more effective public health actions ~~[27]~~in realms such as the distribution of limited medical resources or the implementation of non-pharmaceutical interventions [58, 32, 34].

> **Reviewer Comment 4: In paragraph 4, you write that your work "closely follows the approach of [23]." Could you clarify the ways in which your model improves upon the one in [23]? In general, I found it difficult to understand from the introduction what is novel about your work compared to past work.**

We have added a sentence to this paragraph clarifying the ways in which we adapted the methods of Lander et al. to influenza forecasting. Note that the second of these adaptations, the use of multiple data sources for model training, is our primary contribution in this work and is the topic of the following two paragraphs.

In particular, our work closely follows the approach of [29], which was the leading approach in several time series forecasting competitions including strong performance for forecasting COVID-19 [33]. Our methods adapt that work to forecasting seasonal influenza by including features intended to help the model identify seasonal trends of disease activity and by incorporating multiple data sources measuring influenza activity.

> **Reviewer Comment 5: "We are aware of two studies that investigated forecasting a pathogen of interest using data from other pathogens" - It's unclear to me how this connects to the work you've done here, especially since the next sentence instead discusses a study where data on the same pathogen from different locations was used. Maybe this can be removed?**

We feel that these studies are relevant to include in the literature review since they are examples of the use of transfer learning in the domain of infectious disease forecasting. We have added some more detail about these studies, and also added language to the next sentence to clarify that it discusses a separate line of work:

We are aware of two studies that investigated forecasting a pathogen of interest using data from other pathogens to inform model estimation~~[45, 6].~~ : [53] considered using historical data on dengue to inform Zika predictions as well as the use of past influenza data to inform COVID-19 forecasts, while [10] used models trained on data measuring dengue incidence to generate predictions for Chikungunya and Zika. In another line of work, Zou et al. used multitask learning to forecast influenza-like illness by training models jointly on data for multiple geographic locations, and demonstrated that this was helpful in settings where some locations had only limited data but others had a longer history of observations [65].

Reviewer Comment 6: Again, the end of the second to last paragraph makes it sound as though there are several alternative methods that are just as good as the one used here, but from the rest of the manuscript this doesn't seem to be the case.

We have added an additional sentence to the introduction giving more specific details about the performance of these other approaches and where they placed relative to Flusion and other FluSight contributors (we also added a citation to another similar method and split this content into a separate paragraph):

> We note that contemporaneously with our development of Flusion, ~~another group~~ two other groups also explored methods for forecasting influenza hospitalizations as reported in NHSN using other data streams to extend the amount of available training data ~~[32]. Their approach~~ [40, 4] . Their approaches differed from ours in that they used these other data streams to impute additional historical observations of hospitalizations, rather than training a model directly on data from multiple surveillance signals. Although those methods did not outperform Flusion or the FluSight ensemble, they were ranked 2nd and 13th among the 29 individual models included in our evaluation of FluSight forecasts for the 2023/24 season (see Section 6.2). The success of the Flusion model and ~~similar methods~~ these other methods that incorporated multiple data sources in the 2023/2024 FluSight season suggests that this approach may be a promising direction, especially in the context of public health data modernization initiatives that may replace long-standing surveillance data systems with new data streams.

Reviewer Comment 7: Overall, while I think most of the key information is here, I felt that the introduction bounced back and forth between topics quite a bit, and sometimes presented redundant information, making it difficult to actually grasp all of the key information without reading through it several times. The manuscript would benefit from a more concise, well-organized introduction.

We are grateful to the reviewer for their suggestions for how the writing in the introduction could be improved, and we hope that the revised text is clearer than the original submission.

1.2 Data Sources:

Reviewer Comment 8: For the NHSN data, a more comprehensive description is needed. For example, are all cases tested for influenza, or is it sometimes diagnosed based on symptoms alone? Does this data source cover all cases in a state or a subset of hospitals? Which states and geographic levels are included in the dataset? I see that some of this information is in the supplement, but it should be moved to the main text.

We have added more detail about the NHSN data to the text where this data source is first discussed in Section 2:

> Our model used three measures of influenza activity (Figure 1). The first of these was weekly influenza hospital admissions reported to the National Healthcare Safety Network (NHSN, [17]), which was the target signal for FluSight (see section 3 for more detail). NHSN reports counts of hospital admissions where the patient tested positive for influenza at the time of admission or within the previous two weeks. Requirements for facilities reporting to NHSN have changed over time, but for the time period included in our analyses reporting was generally mandatory for all hospitals registered with the Centers for Medicare and Medicaid Services. We worked with reported data that had been aggregated to the state level for each of the 50 US states, the District of Columbia, Puerto Rico and the national level.

Reviewer Comment 9: Could you clarify what the difference is between the NHSN data and the FluSurv-NET data? Both are described as data on patients hospitalized with influenza.

We have added information about the specific sites with FluSurv-NET data that we used to the text, quoted below (this passage also includes more detail about the ILI signal). It is our hope that in combination with the additional information we added about the NHSN data in response to the previous comment, the differences between the two data sources will be clearer.

> To address this limitation of the target surveillance signal, our model used two other measures of weekly influenza activity that have a longer reporting history. The first of these was a measure of hospital admissions where the patient has a positive influenza test as reported by FluSurv-NET, expressed as a rate per 100,000 population in the catchment areas of selected hospital facilities [15]. Our models used data from 14 sites in 13 states (including two sites in New York, which we aggregated to the state level), as well the "Entire Network", which we use as an estimate of influenza activity at the national level. The other signal was ILI+, which is an approximate measure of the proportion of outpatient doctor visits where the patient has influenza that is derived by combining data from ILINet~~and WHO/NREVSS[16, 12].~~, the World Health Organization (WHO), and the National Respiratory Enteric Virus Surveillance System (NREVSS) [21, 17]. Our models use ILI+ data at the state level and national level, as well as at the regional level for the ten Health and Human Services regions, which are groups of between 4 and 10 states and/or territories of the US.

Reviewer Comment 10: For the FluSurv-NET and ILI+ data, could you also include a bit more description? Again, what is the geographic and temporal coverage of these data? Are these also weekly data, like the NHSN data?

We have added statements on the geographic and temporal coverage for both of these data sources to the text; please see the passage that was quoted in response to the previous comment.

Reviewer Comment 11: Could you spell out what NREVSS stands for the first time the acronym is used?

We have moved the definitions of the acronyms for WHO and NREVSS up to their first appearance in the paper.

Reviewer Comment 12: How is ILI defined? Is the definition used consistent across states?

We have given a more specific definition of ILI in the revised manuscript. This definition is used system-wide.

> ILINet reports a measure of influenza-like illness (ILI), as the percent of outpatient doctor visits where the patient has ~~symptoms consistent with influenza without another known cause~~a fever with a temperature of at least 100°F (37.8°C) and a cough and/or a sore throat [17].

Reviewer Comment 13: Since the GBQR and GBQR-no-level models haven't been introduced yet, the last paragraph of this section will probably cause some confusion. Could the information on what data the models are trained on either be moved to a later section, or could this paragraph be written without naming specific models?

We have removed the specific mention of those models from this section.

> Additionally, ~~the GBQR and GBQR-no-level models (see section 5.2)~~ some of our models were not trained on the influenza off-season (see section 5.2 for more detail on our models).

Reviewer Comment 14: The definition of the influenza season is confusing - why not simply state what weeks of the year are included in the season? Also, why was week 31 chosen as the season start?

We have re-worked this text to explain why we chose to use week 31 as the season start and to frame the discussion in terms of the timing of the on-season:

> Designating US Epidemic week 31 as season week 1 (generally falling in early August, a time of consistently low influenza activity), we define the ~~off-season~~ on-season as season weeks ~~less than~~ 10 ~~or greater than~~ through 40 and the off-season to be the remaining weeks, which fall during the summer months [13].

Reviewer Comment 15: Could you define purging and embargoing?

We have added a brief definition of purging to the manuscript where it is mentioned:

> Omission of the off-season from the training set forms a natural implementation of purging/embargoing, a method that is popular in econometric time series forecasting in which the time series is split into blocks for cross-validation with a buffer between the time blocks to prevent leakage of information across the cross-validation folds. [12, 29].

1.3 The FluSight collaborative forecasting exercise:

Reviewer Comment 16: Ultimately I leave it up to the authors, but this section feels superfluous to me. Unless I have misunderstood, the point of the manuscript is to share your forecasting method, not to describe the CDC's flu forecasting challenge, which has been described before in other manuscripts. In particular, information on forecasting targets that were not attempted by the authors (as in the final sentence of paragraph 1), as well as information on the specifics of how forecasts were submitted to the CDC, is unnecessary. Even if this section is kept, I would recommend shortening it significantly to convey only the information that is necessary for understanding the comparisons between the authors' model and the FluSight ensemble that are made later in the manuscript.

The information in the final paragraph seems relevant, but could be moved to the next section.

We have removed the final sentences of the first paragraph, on the targets for which we did not produce forecasts, as well as a sentence in the next paragraph on hindcasts at a horizon of -1 weeks, which again we did not submit:

> For the 2023/24 season, the primary target for forecasts collected in the FluSight forecasting exercise organized by CDC was weekly hospital admissions with confirmed influenza as reported in the NHSN data set for each of the 50 US states, the District of Columbia, Puerto Rico, and in total at the national level (Figure 1, top panel). ~~Predictions for a second target representing a categorical measure of the direction and magnitude of change in admissions were also collected by FluSight, but we did not make predictions for that target with the Flusion model.~~
>
> Predictions were submitted to FluSight on Wednesday each week. The week of submission is anchored relative to the Saturday after the submission date, which corresponds to the final day of that US epidemic week and is denoted as the *reference date* for the predictions. Predictions were made for hospital admissions from Sunday to Saturday in the current week and each of the three following weeks, corresponding to forecast horizons of 0, 1, 2, and 3 weeks ahead relative to the week of submission. ~~Initially, FluSight also collected predictions at a horizon of -1, representing a "hindcast" of admissions in the week before submission, but these hindcasts were discontinued a few weeks into the season and we do not analyze them here.~~

Our feeling is that the remainder of the discussion in this section is relevant to the reader as it gives specific details about how the prediction targets are defined and what data were available when the predictions for a given reference date were made (which is relevant when describing our models later in the manuscript).

1.4 Notation and evaluation metrics:

> **Reviewer Comment 17: A brief (one sentence) description of how the pairwise tournament approach works would be helpful in better understanding the rMAE and rMWIS metrics. (Or is this what is described in the paragraph after it is first mentioned?)**

We have rearranged some of the text describing the rMAE and rMWIS metrics to describe how they are computed immediately after they are introduced, followed by a description of their interpretational advantages:

We also compute relative versions of MAE and MWIS, denoted rMAE and rMWIS, using the "pairwise tournament" approach outlined in [11]. ~~The primary purpose of this procedure is to correct for the varying level of difficulty of the predictions submitted by different forecasters in settings where some forecasters did not provide predictions for all locations or time points. This is relevant to the evaluation of submissions to FluSight in section 6, but in the experimental results in section 7 all forecasts were provided. A secondary goal is to standarize scores relative to a baseline model with known behavior, in our case the flat baseline described in section 6. Smaller values of rMAE or rMWIS indicate better performance relative to the other models in the comparison pool, and in particular values less than 1 indicate performance that is better than the baseline.~~ In a comparison of forecast accuracy among M models, computation of rMAE (rMWIS) has two steps. First, we obtain a summary of the average performance for model m relative to each other model m', denoted θ^m. This summary is computed as the geometric mean of the ratio of the MAE (MWIS) for model m to the MAE (MWIS) for each other model m', where for each model pair the MAE averages across the set $\mathcal{I}_{m,m'}$ of locations, dates, and forecast horizons for which both models submitted predictions. The rMAE (or rMWIS) then normalizes this geometric mean relative to the value of θ^m for a baseline model:

$$rMAE^m = \frac{\theta^m}{\theta^{baseline}}, \text{ where } \theta^m = \left(\prod_{m' \neq m} \frac{MAE^m_{\mathcal{I}_{m,m'}}}{MAE^{m'}_{\mathcal{I}_{m,m'}}} \right)^{1/(M-1)}$$

Here, $MAE^m_{\mathcal{I}}$ denotes the MAE for model m across all predictions for tasks i in the index set $\mathcal{I}$.

The primary purpose of this procedure is to correct for the varying level of difficulty of the predictions submitted by different forecasters in settings where some forecasters did not provide predictions for all locations or time points. In this situation, comparing MAE or MWIS could be misleading because a forecaster could achieve good MAE or MWIS by omitting forecasts for difficult locations or time points, such as near a peak in disease incidence. The relative metrics account for this by evaluating each pair of forecasters on the prediction tasks they shared in common, enabling a direct comparison of their scores. However, ultimately the rMAE and rMWIS metrics may also be susceptible to gaming by a forecaster who selectively omits forecasts that may have poor performance. Of note, considerations about missingness are relevant to the evaluation of submissions to FluSight in section 6, but in the experimental results in section 7 all forecasts were provided. A secondary goal of the relative scores is to standarize scores relative to a baseline model with known behavior, in our case the flat baseline described in section 6. Smaller values of rMAE or rMWIS indicate better performance relative to the other models in the comparison pool, and in particular values less than 1 indicate performance that is better than the baseline.

> **Reviewer Comment 18: It would also be useful to have a brief explanation of why it is important to have a metric capable of comparing your method to all other forecasts submitted to FluSight, either here or in the introduction. It is clear after reading the results, but on my first pass through this section, I wondered why comparison to all other models was necessary if the point of the paper was to evaluate your own model.**

We have added some signposting to the introduction clarifying that a comparison to other real-time FluSight submissions will be a part of the evaluation, to help motivate the need for suitable evaluation metrics:

> We establish notation in section 4 and describe our modeling approaches in section 5, and then we discuss the performance of our real-time forecasts relative to other contributors to FluSight in the 2023/24 season in section 6.

See also the response to reviewer comment number 35 below.

> **Reviewer Comment 19: Is it correct that the rMAE and rMWIS metrics will only include forecasts targets that were submitted for both models in a given pair of models? If so, is this a limitation of this metric, in that it might leave out targets for which one of the models did either particularly well or particularly poorly?**

It is correct that for each model pair, the ratio of their MAE or MWIS scores is computed based on the subset of forecast targets that were submitted by both models. However, the overall rMAE and rMWIS metrics aggregate these pairwise scores across all model pairs, and several models submitted predictions for all targets – so the overall score includes all forecast targets.

The central challenge here is how to address a situation where one or more models are missing some forecasts. The MAE and MWIS metrics simply average the scores for the available forecasts from each model, but this could give an advantage to models that do not submit forecasts for challenging locations or time points. The idea behind the relative metrics is to address the problem of missing forecasts by the comparing forecast skill for each pair models on the subset of predictions where no forecasts are missing, so that the resulting scores can be directly compared. Our understanding is that neither approach is wholly satisfactory without an assumption along the lines that forecasts are missing completely at random, but seeing consistent model rankings using both methods offers some reassurance.

We have added some additional discussion of these challenges to the paragraph describing the rMAE and rMWIS metrics; please see the text that was quoted above in response to reviewer comment number 17.

> **Reviewer Comment 20: Could you expand on why the metrics you've used were chosen? If it is just because they were the metrics used by FluSight, it would still be useful to briefly describe their benefits.**

We have added a note on the reasons for choosing these metrics to the paragraph where they are introduced:

> All of these metrics have been used by FluSight for forecast evaluation. The absolute error and weighted interval score are proper scoring rules for the predictive median and the full collection of predictive quantiles collected by FluSight, respectively [20, 6]. This means that in expectation under a forecaster's predictive distribution, the scores are optimized by submitting the predictive median or the collection of elicited predictive quantiles of that distribution.

1.5 Model:

> **Reviewer Comment 21: Why a fourth root specifically?**

We added a sentence describing our procedure for selecting the fourth root transform:

> We took a fourth root transformation to stabilize the variance of the signal across times of low and high influenza activity. The fourth root was selected through visual inspection of data plots with various power transforms, using the criteria that the amount of variability around a local trend should be fairly consistent throughout all phases of the influenza season.

> **Reviewer Comment 22: In Figure 2, because there is so much state-level data, it is difficult to pull much meaning from the left panels, since there are so many overlapping lines. It might be better to present just the national and HHS-level data, with lines colored by HHS region. Alternatively, a few (¡=10) representative states could be chosen.**

We agree that it is challenging to distinguish the individual lines in the left side of the plot, but we elected to leave this figure as-is in the resubmission after experimenting with some alternatives. Ultimately, we believe that although the figure is imperfect, including a display of data at the state level is preferable to showing data from only the regional and national levels. Additionally, we feel that a primary advantage of this figure is that it conveys the different volumes of data that are available through different reporting systems (with much more data for ILI+ than for NHSN admissions, for example). The figure is still challenging to read with only a few representative states (e.g., consider the 10 regions on the right hand side), and is less effective at showing the full extent of the available ILI+ data.

> **Reviewer Comment 23: For readers who aren't aware, a short description of what gradient boosting and bagging are should be included.**

We have added descriptions of boosting and bagging to the text:

~~The~~ In gradient boosting, the estimated function $f_{\alpha_k}(x_i)$ takes the form of a sum of regression trees which are estimated sequentially. In each step of estimation, a new regression tree is added by fitting to pseudo-observations which are the gradient of the objective function (here, average quantile loss) with respect to the predictions for each data point after the previous iteration. At each quantile level, the final prediction was obtained using bagging, ~~by taking~~ a procedure in which predictions are formed by ensembling models fit to different random samples drawn from the full data set. In our implementation, the predictions were computed as the median of the predictive quantiles from 100 separate fits that were each based on a different randomly selected 70% of the seasons in the training set (including partial data for the current season).

> **Reviewer Comment 24: Some additional justification is also needed - why was a gradient boosting approach chosen? What are the advantages of this method? And why were 70% of the seasons in the training set specifically chosen for obtaining the final predictions?**

We have added a note explaining our choice of gradient boosting based on its strong forecasting performance:

The first two models used gradient boosting for quantile regression, which we abbreviate as GBQR; this methodological choice was motivated by the success of gradient boosting in several recent forecasting exercises as described in the introduction.

We also added text to clarify that each bagged model was fit to a different randomly selected 70% of the training set seasons, rather than selecting 70% of the training set seasons to use for all model training (see the text that was quoted in response to reviewer comment number 23).

> **Reviewer Comment 25: I'm a little confused by the sentence ”the features x_i contained information only about influenza activity for the particular data source and location.” Could you explain a bit more what this means, and how it is consistent with the fact that information from multiple locations and data sources was included in the training data?**

We have added some additional text to this paragraph attempting to clarify that while the feature vector x_i for a particular prediction task only had information about the location $l(i)$ and data source $s(i)$ relevant to that task, the full model training set included examples from all available combinations of location and data source:

> The models were trained jointly on data for all data sources, locations, dates, and forecast horizons. However, the features x_i for a particular prediction task indexed by i contained information only about influenza activity for the particular data source $s(i)$ and location $l(i)$. Inclusion of multiple locations and data sources in the training data set (corresponding to different prediction tasks with different indices i) allowed the model to use past examples from multiple locations and data sources to learn a mapping from x to y.

Reviewer Comment 26: If the model is trained simultaneously on data from all locations at the state, HHS-region, and national level, does that mean that the model is trained on the same data multiple times, since states are included within the HHS regions, as well as within the national-level data?

Yes. We have added a note about this to the manuscript:

> Of note, the inclusion of data for the same past season from multiple data sources and the inclusion of data that have been aggregated to the regional and national levels means that the model sees data related to any given time period and location multiple times during training.

Reviewer Comment 27: Why were the features listed in Table 1 included? How did you make decisions about which features to include? Were there multiple iterations of the model using different combinations of features? In particular, the inclusion of the difference between the week of the most reported data and Christmas is confusing without justification.

This model was developed in a short time span before the start of the influenza season. Given those time constraints, we did not do any formal experimentation with different combinations of features; indeed, part of the purpose of the analyses described in the present manuscript is to investigate the value of these features. We have added text to the paragraph where the features are described specifically mentioning and motivating the inclusion of a feature measuring the difference between the week of the most recent reported data and Christmas:

> These features included information about the data source and location being forecasted, the time of season when the forecast was generated, the difference between the time of season when the forecast was generated and Christmas week (intended to help the model adjust for a consistent peak in reported influenza activity near the holidays), the forecast horizon, and measures of the local level, trend, and curvature of the surveillance signal in the weeks leading up to the time $d(i)$.

Reviewer Comment 28: In the legend for Table 1, it says that the GBQR-no-level model did not include features from groups 8-12. Does this mean that this model did not know the most recently reported data value?

It is correct that that model did not see the most recently reported value as a feature, but it did have some access to this information through the formulation of the prediction target. We have added text clarifying this:

> Starting in the eighth week of the season, on the reference date of December 2, 2023, we included a second variation on the GBQR model that was not allowed to see these "local level" features. This was motivated by two considerations: (1) a model fit without features that had high importance in the primary GBQR model might introduce more model diversity to the Flusion ensemble; and (2) in seasons with particularly high or low incidence, measures of local level might not be a reliable indicator of the magnitude and direction of changes in future values of influenza activity. Although this model variation did not have access to any local level features (including the most recent data report), the formulation of the prediction target y_i as the difference between the most recent data value and the data value on the target date meant that predictions from this model were still placed in the correct location relative to recent observations.

Reviewer Comment 29: What was the rationale behind including the autoregressive model (ARX) in your ensemble? It seems much less sophisticated than the other methods, and less accurate, based on Table 3.

We have added text to the ARX section explaining our motivations for including this model in the Flusion ensemble:

> Our third component model was a Bayesian auto-regressive time series model with covariates (ARX). This model is simpler than the models based on GBQR, but we included it in the Flusion ensemble for two reasons. First, simple AR models have a history of strong performance for infectious disease forecasting [e.g., 48]. Second, past theoretical and applied investigations have shown that even relatively weak individual models can be valuable ensemble members if their prediction errors are not highly correlated with the other ensemble members, and in particular including methodologically diverse forecasters can be beneficial [e.g., 3].

Again, we note that investigation of the value of the ARX model as a member of the Flusion ensemble is one of the purposes of the present manuscript and was not conducted prior to the start of the season due to time constraints.

Reviewer Comment 30: "...this behavior was likely not ideal" - You mention changes to remedy this behavior, but could you briefly discuss why these changes weren't made before?

We have added text explaining that these changes were not made due to time constraints:

> In our modeling setting, where $x_{l,t}$ was a deterministic function of the season week, this behavior was likely not ideal. Remedying this to allow for the provision of known future values of covariates is on a short list of model improvements to make, though we did not make this update prior to the 2023/24 season due to time constraints.

Reviewer Comment 31: The second to last subsection (5.5) again seems a bit superfluous. If the goal is to assess the quality of your model, I don't think it is necessary to describe minor tweaks made throughout the season. Rather, it makes more sense to assess the model as it was employed for the bulk of the influenza season. At the very least, I think this subsection could be reduced substantially in length/detail.

We believe that it is important to evaluate prospectively generated and registered real-time predictions, which is the approach to evaluation that we took in section 6 of the manuscript. At the same time, we agree with the reviewer that in many ways, an assessment of the model as it was specified for the bulk of the season has more value in terms of generating insights about strong modeling practice, and this is the approach that we took in section 7 of the manuscript. Our concern is that without some context, the juxtaposition of these two different approaches to analysis may raise questions in the mind of some readers, such as (1) "Why didn't you submit all forecasts in the real-time exercise?", and (2) "Why are the aggregated scores for the Flusion model different in sections 6 and 7?" We feel that the primary value of section 5.5 is to answer these questions, and so we have opted to keep the bulk of the section describing ways in which the methods we used for real-time submissions differed from our finalized methods. That said, in the revision we have removed text where we felt that we could do so without compromising clarity of communication:

> As we described above, the GBQR and ARX component models were used throughout the full season, but the GBQR-no-level model was introduced starting in the eighth week. In the first week of the season only, we used an additional model that obtained a predictive median using the same method as GBQR, but obtained predictions at other quantile levels by bootstrapping out-of-sample residuals. ~~We discontinued use of this model from the second week on. Although we did not investigate formally, our anecdotal sense was that predictions from this model were~~

~~too conservative (with wide prediction intervals), likely due to a strategy of sharing bootstrapped residuals across locations and surveillance signals with different signal-to-noise ratios.~~

In the first week of the season, as a hedge against possible reporting revisions we formed our submitted predictions by combining forecasts based on all available data and forecasts based on data up to the second-to-last observation (i.e., omitting the final reported value). ~~This was because it was indicated that the latest available data were tentative and were subject to reporting corrections. In that instance, we used an equally-weighted linear pool (or distributional mixture) to combine the predictions based on the full data set and the partial data. From the second submission on, we submitted only the predictions based on all available data.~~

Reviewer Comment 32: In subsection 5.6, the final link is broken.

Thanks for reporting this! We had inadvertently left the GitHub repository on a private setting. It is now public, and we have confirmed that the link works.

1.6 Real-time performance: the 2023/24 FluSight season:

Reviewer Comment 33: What do you mean by ”distortion”? How would national-level results cause this distortion?

We have updated the description here to remove the word ”distortion” and clarify our reasoning for removing the national level predictions for our summaries of forecast skill:

~~To avoid distortion of WIS and AE results, we did not evaluate~~ We only evaluated forecasts that were made at the ~~national level. Although FluSight originally allowed for collection of predictions at a horizon of -1 week, these were discontinued; our~~ state level; MWIS and MAE can be dominated by locations with large populations, and we have found that averages that include scores for national level predictions often only reflect predictive skill at the national level. Our analysis includes predictions made at horizons of 0 weeks (nowcasts), and predictions at horizons of 1, 2, and 3 weeks ahead relative to the reference date.

Reviewer Comment 34: Instead of AE, do you mean MAE?

We have corrected this acronym (and also changed WIS to MWIS).

Reviewer Comment 35: Again, I think it would be useful to justify earlier on why you've included the results from so many other models, especially when the bulk of your results simply compare your model to the overall FluSight ensemble. A brief summary of some of the most common model types submitted would also be useful, to better understand the types of models that make up the FluSight ensemble.

We believe that including the other models strengthens the results by giving a more complete picture of variation in individual model performance, and we have added text explaining this:

We included all models that contributed forecasts for at least two thirds of the combinations of state-level locations, reference dates, and non-negative horizons for which FluSight collected forecasts over the course of the season. Because a comprehensive evaluation of all FluSight contributors is not the aim of this manuscript, we anonymized the names of other individually-contributed models in these results to focus attention on the comparisons that are of interest for our purposes. ~~FluSight produced two ensemble forecasts during the season: one using a quantile averaging approach and one using a linear pool. These two ensembles had very similar performance, though the linear pool had slightly better marginal calibration. However, the quantile averaging ensemble was~~ However, we included these other models in the analysis to give a more complete sense of how Flusion compared to all other contributing forecasters. We also included the ensemble forecasts

> that were produced by FluSight and were used by CDC as the source of official communications throughout the season~~, and so we include results from only that ensemble here~~. These ensemble forecasts were calculated by taking the median of the predictions at each quantile level across all contributing models.

> **Reviewer Comment 36: I also don't think you need to discuss the linear pool ensemble here, since you do not include its results in the manuscript.**

We have edited this paragraph to remove the description of the linear pool ensemble. Please see the updated text quoted in response to the previous reviewer comment.

> **Reviewer Comment 37: Why is it important that the second baseline model was called UMass_trends_ensemble? I'm confused about what this indicates - was this a model that your group created? If not, who did create it?**

The original name of this model was not an important detail, and we have removed it from the description of the method. This is a baseline model that our group created, and we have added this information to the description of that method:

> The second baseline method, Baseline-trend, was produced by our team. It followed a similar process to the Baseline-flat model, with a few modifications that were designed so that the resulting forecasts tended to follow the trend of recent observations. It was a quantile averaging ensemble of 16 variations on the baseline method. Most importantly, it incorporated variations that did not symmetrize the past differences, and rather than using all available history, it collected differences in a rolling window of the past few weeks. The 16 variations were obtained by using different options for the rolling window size, the temporal resolution of data used as an input (daily or weekly), a data transformation that was applied (no transformation or square root), and whether or not symmetrization was used. We emphasize that although this baseline was more methodologically involved than Baseline-flat, it produced an epidemiologically naive forecast that pushed forward a local estimate of the trend observed over the few most recent weeks. ~~This model was named UMass-trends_ensemble in real-time submissions to FluSight.~~

> **Reviewer Comment 38: The first paragraph in subsection 6.2 is interesting. Were there any other notable situations where your model performed particularly well, or, conversely, where it tended to struggle? If you have any particular intuition into why the model performed well or struggled in certain situations, this could be an interesting conversation for the discussion section.**

We have not noticed any other specific moments where our model's performance was particularly better or worse than that of the ensemble. Our sense is that the most likely candidate for improving forecast skill in settings where there are two peaks (one near the holidays and another second peak) are (a) better handling of holiday effects, and (b) the use of data broken out by influenza strain to help predict a second peak that may be partially driven by a different strain that is dominant later in the season. Some of these ideas were actually mentioned in the discussion previously, but we didn't make the specific connection with the limitations of our methods around turning points. We have now updated the text in the discussion to expand on these ideas and make this connection clearer:

> ~~Our model could also be extended~~ Forecasting the turning point of an epidemic wave is a challenging problem [9], and as we saw in Section 6.2 both Flusion and the FluSight ensemble often failed to accurately predict the turaround after a local peak in disease activity or the increase leading into a second peak within the same season. We conjecture that it may be possible to improve forecasting performance in these instances through better handling of holiday effects or by extending our model to take into account epidemiological understanding of disease transmission. Improved handling of holiday effects, e.g. through features designed to measure trends accounting

> for the typical magnitude of holiday effects, could help models to predict a decline in activity immediately after the holidays and might allow the model to better see a second wave coming.
>
> Features derived from mechanistic insights, such as measures of vaccine uptake and efficacy or the circulation of multiple strains of the influenza virus at different times over the course of the season may also be helpful may also be helpful [27]. For instance, a second wave of influenza later in the season is often at least partially driven by a different influenza type or subtype than was circulating early in the season. A model that saw data on disease incidence broken down by flu strain might be better able to predict a second wave in these situations. Information about vaccine uptake and efficacy might help a model predict season severity more accurately. A challenge with these approaches is that timely data may not be available for use in a real-time forecasting exercise.

Reviewer Comment 39: In Figure 3, you focus on states with the largest number of cases. However, it would be interesting to also see examples from some smaller states - I would assume that smaller outbreaks may often actually be harder to forecast. On that note, how did population size (and potentially other characteristics of a location) impact forecast accuracy?

We have added plots for all forecasted state-level locations to the supplement (supplemental figures 3 through 8), and included a reference to these figures in the main text. We did not observe any consistent differences in forecast accuracy between the states with large and small cumulative disease counts, and we have noted that in the main text as well:

> Forecasts from Flusion often appeared similar to forecasts from the FluSight-ensemble, though in several states (e.g. Florida, California, and New York) Flusion did a better job of capturing the increase in weekly hospital admissions in the early part of the season and a slightly better job of predicting the turnaround after the peak in late December (Figure 3, Supplemental Figures 3 through 8).
>
> ...
>
> These general trends in forecast performance also held in other states (Supplemental Figures 3 through 8).

Reviewer Comment 40: From the figure, it looks like your model tends to do a better job of not overshooting the peaks than the FluSight ensemble - was this true in general? It also looks like your model, like the FluSight ensemble, still tends to forecast relatively constant flu activity at the peak, rather than the decrease in activity that is seen in reality. (Although I understand that this is a difficult task for forecasting models in general.) Do you have any idea why either of these things are happening?

Flusion did a better job of not overshooting near the peak than the FluSight ensemble in most locations, but not all; this briefly noted in the first paragraph where we discuss evaluation results. We have also added text to this paragraph noting that in many locations the forecasts did not accurately predict the decrease in disease incidence following local peaks:

> Forecasts from Flusion often appeared similar to forecasts from the FluSight-ensemble, though in several states (e.g. Florida, California, and New York) Flusion did a better job of capturing the increase in weekly hospital admissions in the early part of the season and a slightly better job of predicting the turnaround after the peak in late December (Figure **??**, Supplemental Figures 3 through 8).
>
> ...
>
> There were also several instances, such as predictions made in late December in Florida, California, Pennsylvania, and Michigan, where predictions did not accurately capture the decline after the holiday peak. We return to these challenges in the discussion.

We have also extended and reframed some of the text in the discussion to suggest some possible approaches to correcting these problems, see the discussion text quoted in response to reviewer comment number 38.

> **Reviewer Comment 41: Could you expand on why you decided to also look at only those forecasts for targets were the latest available data were not revised by 10 or more admissions? I assume this was to check just those forecasts where the model had good-quality information, but it would be good to see this reasoning spelled out. As an alternative analysis, is it possible to regenerate all the forecasts using the data that were available at the end of the season, rather than the data available in real-time, in order to check the extent to which inaccuracies in the real-time data influenced forecast accuracy? Or is this too computationally intensive?**

We have added an explanation of the reasoning behind the sensitivity analysis to the description of the analysis in the supplement:

> We conducted a sensitivity analysis to confirm that the overall evaluations in Tables 2 and 3 of the manuscript are reflective of differences in model forecast skill in the presence of good-quality input data and our evaluation procedure does not differentially penalize models for predictions that were issued when the target data shifted dramatically.

Although we could re-run our own models based on the finalized data, we do not have the code for all models that contributed to FluSight and re-running all model forecasts would be infeasible.

> **Reviewer Comment 42: I would assume that, in general, nowcasts were the most accurate, while forecast accuracy degraded as the forecast horizon increased - was this true?**

We have added Supplemental Figure 9 which shows mean WIS for each model by forecast horizon, and added a note to the manuscript describing the phenomenon of reduced forecast accuracy at larger forecast horizons:

> Forecasts had decreasing accuracy at larger forecast horizons (Supplemental Figure 9).

> **Reviewer Comment 43: It looks like the models in Table 2 are ordered by rMWIS - is this correct? If so, it should be stated in the legend.**

This is correct, and we have updated the captions for Tables 2 and 3 to describe this ordering.

1.7 Post hoc model exploration:

> **Reviewer Comment 44: I wonder how common it was for other models submitted to FluSight to also train on multiple data sets and/or locations simultaneously - do you have this information?**

We added some text to the discussion addressing this question:

> Not all modelers that contributed to FluSight provided a description of their methods, but we reviewed the available model metadata [14] and identified at least 3 other models that in some way used data for multiple surveillance signals and trained jointly on data for multiple locations, along with an additional 13 that used multiple surveillance signals and another 8 that were trained jointly on data for multiple locations. As with previous forecast evaluation efforts that have not found consistent associations between individual model design decisions and forecast skill [e.g., 36], these methods had a range of performance rankings. This indicates the challenge of identifying

> important modeling decisions by comparing scores from models that differ in many aspects of their model formulations, and highlights a strength of our study: the use of ablation studies to investigate the value-added from specific aspects of our model setup while controlling for other parts of the model design. That said, the second-best individual FluSight model overall (denoted as "Other Model #1" in Table 2) also used gradient boosting and multiple surveillance signals [40], which lends additional support to our finding that these design choices were helpful.

Reviewer Comment 45: In subsection 7.2, were the GBQR-only-NHSN and GBQR-by-location models simply compared to the GBQR model, or were they used alongside the GBQR-no-level and ARX models to form an ensemble, then compared to Flusion?

We did not reconstruct the full Flusion ensemble with these variations on the GBQR model. We added text to the caption of Table 3 clarifying that the GBQR variations in sections 7.2 and 7.3 should be compared with the GBQR model:

> In Experiments B and C, we did not reconstruct the Flusion ensemble with the alternative GBQR formulations; we include Flusion as a reference, but appropriate comparisons are among variations on the GBQR model.

Reviewer Comment 46: It looks like the two models above, in addition to having reduced accuracy relative to the GBQR model, also have worse calibration, looking at the 50% and 95% cov. columns. Is this worth mentioning and explaining as well?

We have clarified that our statement that GBQR-only-NHSN and GBQR-by-location underperformed relative to GBQR held according to all evaluation metrics:

> Both of these alternatives underperformed relative to the GBQR model in terms of MWIS, MAE, and interval coverage rates (Table 3, Experiment B).

Reviewer Comment 47: It's surprising that the adjustments to the ILI and FluSurv-NET data actually seemed to reduce forecast quality. Do you have any idea why this may have happened? Would it be feasible to try models where, rather than leaving out the processing of both ILI and FluSurv-NET, you omit the adjustments to just one or the other, to see whether either of the datasets are more responsible for this change in accuracy?

We have two main guesses about why these data adjustments may have been harmful to forecast skill:

1. Using additional data streams may have introduced extra noise into the data. The hospitalization burden estimates and test positivity rates used in the data adjustments are not perfect measures, and it may be that the extra variability from multiplying noisy data signals led to less reliable measures of disease activity overall.

2. The test positivity data had many values that were zero or close to zero, meaning that ILI+ had many observations that were very close to zero. This differs from the behavior we see for NHSN admissions (compare Figure 1, top and bottom panels). Thus, it may be that ILI+ is a better measure of influenza activity among outpatient doctor visits than ILI, but that the behavior of influenza hospital admissions is actually more similar to ILI than ILI+.

We have not conducted the suggested additional analysis because it is our feeling that it would not shed much additional light on the ultimate driver of the phenomenon of reduced forecast accuracy when these data adjustments were used. That is, we might learn which data source or sources were associated with reduced forecast accuracy when these adjustments were used without gaining a deeper understanding of which of the hypotheses above (or other factors) are causing these changes in performance. Additionally, we view this line of inquiry as being at the margins of the scope of our work.

Reviewer Comment 48: The final paragraph in this section seems to no longer be about the data processing. Maybe this should be a new subsection (7.4)?

We view feature engineering as an aspect of data processing, and so have opted to leave this topic within section 7.4.

Reviewer Comment 49: Why was January 6, 2024 chosen as the reference date? Are results similar for other dates?

January 6, 2024 was an arbitrarily selected reference date to use for the purpose of investigating feature importances, and is located roughly midway through the 2023/24 season just after the nation-wide peak. Of note, the training data set is large, and is dominated by the many seasons of ILI+ data. Any given week of NHSN data from the 2023/24 season accounts for less than 0.1% of the training data; in that context, we would not anticipate that the addition or removal of a few weeks of data would have a meaningful impact on feature importance values. Anecdotally, the feature importance results reported in the manuscript are consistent with feature importance values that we saw in ad hoc analyses a few times throughout the season.

1.8 Discussion:

Reviewer Comment 50: Did any other models submitted to FluSight use gradient boosting? Did any others train to multiple surveillance signals or multiple locations?

We have added some text addressing these questions to the discussion; see text quoted in the response to reviewer comment 44 above.

Reviewer Comment 51: By ”contemporaneous observations of multiple signals,” do you mean data from the forecast target week? Would these be expected to be available in real time? If this isn't what you mean here, could you clarify?

This is a good question, and we have added a note on this challenge to the text:

> A challenge with any approach integrating multiple data streams in real time is that they may have different reporting lags and different revision behaviors.

Reviewer Comment 52: As other groups have previously explored forecasting using insurance claims and internet activity, it would probably be worth citing a couple examples using these approaches. In general, throughout the discussion you've listed several ideas to improve your model. Citing examples of work that uses similar approaches would help to better support your ideas, as it would show that the methods you've suggested are indeed potentially useful.

We have added citations illustrating applications of similar ideas throughout the discussion.

Reviewer Comment 53: What do you mean by "reconciling predictions made at multiple hierarchical levels"?

We have added text giving an example of how this might work:

> Alternatively, it may be possible to improve forecast accuracy by reconciling predictions made at multiple hierarchical levels, for example generating probabilistic forecasts at the state, regional, and national levels and subsequently adjusting the predictions so that the forecasts at those scales are consistent with each other [2].

Reviewer Comment 54: Are data on influenza vaccine uptake available? In the past I haven't been able to find any comprehensive data on this in the US.

We are aware of some estimates of vaccine uptake that are available through the CDC at https://www.cdc.gov/fluvaxview/coverage-by-season/. Unfortunately, it is not clear whether any data for this are available early in the season, which would be required to enable use of these data in forecasting models. We have added a note on this to the text; please see the last sentence of the text quoted in response to reviewer comment number 38.

2 Reviewer 2 Comments

> **Reviewer Comment 1: The manuscript nicely and clearly described the Flusion method for predicting a time series of flu hospitalizations with short historical data, empowered by other flu surveillance data. A series of post-hoc model exploration provided further insights on the key components contributed to good performance. The study would inform future development of these forecasting model.**
>
> **Introduction, the innovation of the proposed method was well described and highlighted.**

We are glad to hear of the reviewer's positive overall assessment of this work.

> **Reviewer Comment 2: Methods, the GBQR-no-level model was introduced starting in the eighth week. How was the eighth week decided? From post-hoc model exploration it is likely that the choice would have large impact on performance.**

We developed the GBQR-no-level model after the season was already underway. We have added to the text explaining the reasoning behind our claim that the introduction of this model had only a limited impact on the overall performance of Flusion:

> The experimental results below indicate that these changes had only a minor impact on forecast performance. For example, averaging over the full season, the MWIS of ensembles that included GBQR-no-level or omitted it differed by only 0.3 (Section 7, Table 3).

> **Reviewer Comment 3: The post-hoc model exploration section provides many insights which will be valuable to further development of forecast models.**

Again, we are glad that the reviewer sees value in these analyses.