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

September 12, 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.

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 have followed many of the suggestions made in the comments 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 [**?**]. ~~After a pause during the first two years of the COVID-19 pandemic, this forecasting exercise restarted in the spring of 2022 [?]. 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 [?, ?].~~ 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 [**?**].

> **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.

> **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.

> **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:

> **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.**

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. 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.

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:

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

We have move 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.

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.

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:

Reviewer Comment 15: Could you define purging and embargoing?

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

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. 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 are 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:

> **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:

> **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 is indicative of consitent trends.

We have added some additional discussion of these challenges to the paragraph describing the rMAE and rMWIS metrics:

> **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:

1.5 Model:

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

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

> **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 does less to show 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 a description of boosting to the

> **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 as was reviewed 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.

> **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.

> **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 of this to the manuscript:

> **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:

> **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:

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:

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:

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:

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:

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:

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:

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:

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. Both of these ideas were actually mentioned in the discussion previously, but we didn't make the specific connection with the limitations of our methods. We have now updated the text in the discussion to make this connection clearer:

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:

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:

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 ** (two up).

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:

The purpose of the sensitivity analysis is to confirm that the overall evaluations in Table 2 of the manuscript are reflective of differences in model forecast skill in the presence of good quality data. 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. We also feel that this would be secondary to the main purpose of this sensitivity analysis.

> **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:

> **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 table caption 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:

> **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 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; the full Flusion ensemble was not reconstructed.

> **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.

> **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?**

> **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 nationwide 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 ** 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:

> **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:

> **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/flu/fluvax by-season.htm. 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:

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 this model in real time after the season was already underway. We have added to the text explaining the reasoning behind our claim that the introduction of this model would have had only a limited impact on the overall performance of Flusion:

> **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.