25 January 2024

Dear Katrina,

Many thanks for considering our manuscript and for your recommendation. We would like to thank the two reviewers for their comments. We very much appreciate the time they have taken to provide such thoughtful and detailed feedback.

We provide a point by point response to these reviews below, where we have combined reviewer comments for clarity. To summarise, we have made two substantial changes to the work, as follows:

1. We have added a quantitative evaluation of an ensemble using weighted trajectories. This was not requested but we believe this is necessary to support the conclusions of the work recommending weighting trajectories by past performance. This now forms a larger portion of the overall work as the key novel aspect of this piece.
2. We have added a Linear Opinion Pool ensemble to reflect this important method of model combination. At the same time we have re-balanced the article to focus away from aggregation/combination and towards method of model output collection. We believe this addresses the main concern that the paper conflated aggregation with methods of collection.

Best wishes,

Kath Sherratt

**Combined reviewer responses**

**Main reviewer rubric**

- *Additional details about peak calculations would be helpful. For example, in a 5-week window of strictly decreasing incidence, was the first week designated a "peak" (I believe this would technically be the "local maxima")? What does "capture distinct weight" mean?*

We have added information on peak calculations. We used a sliding 5-week window, where the local maxima would not be observed in a strictly decreasing period. This method for peak calculations was decided ad-hoc. We fully appreciate more data-driven and robust approaches exist. However, this work reflects the real-time policy response work that we conducted in real time, and this has been added as a comment in the discussion. We believe that further work on peak calculations is beyond the scope of the present piece.

- *Additional clarification on Figure 3 would be helpful. Does a single ribbon represent one forecast made for 4-29 weeks into the future? Or a series of 4-week ahead forecasts made? Were the trajectories used for "16 weeks ago", "8 weeks ago" and "4 weeks ago" all weighted using performance against the same number of observations, and horizon is the only thing that changes?*

We have clarified the figure caption and text; each ribbon represents one forecast for e.g. 4+ weeks into the future. Each ribbon shows a continuously updating sequence of the past 4, 8, or 16 weeks' accuracy data used to weight trajectories, so each ribbon uses a different number of observations for weighting.

- *I believe the current analyses confound two potential factors contributing to information loss: aggregation method and type of information collected (quantiles vs. trajectories). Please see my full review for details.*

Thank you for your very thoughtful comments. We agree with your view and have substantially updated the paper to reflect this - see full response to reviewer comments below.

- *I believe the manuscript would benefit from additional discussion about the contexts in which the new proposed performance-weighted sample aggregate might be used. Please see additional discussion in my full review.*

We have added further discussion of the use of performance-weighted samples, with a full response to this point below.

**Additional reviewer comments**

- *Distinguishing between representations and statistical ensemble methods*
  *It is important to be clear about whether observed differences in the qualitative behavior of ensemble predictions or quantitative measures of their performance are driven by the quantity that is used to represent a predictive distribution, or the ensemble method that is used. Currently, these are conflated. Currently, a linear pool ensemble is used for the trajectory-based ensemble, while a median-based Vincent ensemble is used for the quantile-based ensemble. These are distinct ensemble methods with known and well-studied differences in their behavior, but the writing attributes those differences to the choice of quantity used to represent the prediction rather than the statistical method used for ensembling. It would be possible to calculate a linear pool ensemble from predictive quantiles, and I would expect such an ensemble to have performance consistent with the linear pool that is calculated from the samples. Indeed, this direction has already been explored by Howerton et al. (2023). Summing up: I think the discussion in this manuscript incorrectly attributes differences in ensemble characteristics to differences in whether quantiles or samples are used rather than differences in ensemble methodology, and if we turn to the question of what ensemble method to use, I don't really see any novel insights here that were not already provided in previous work. To address these comments, I think the discussion around these points would need to be substantively reworked, and some additional analyses added to add something novel. Alternatively, it may make sense to just remove this part of the discussion to focus the paper more tightly on things that really are different between quantile and trajectory-based representations of predictions (i.e., points 1 and 3 in the list I gave in the summary, where I do think the paper adds new ideas).*

Thank you for raising this and your discussion of this point. We agree that the ideas of how model results are collected versus aggregated are distinct, and were not well separated in the original manuscript. We have reworked the text to focus on how results are collected, and substantially scaled back claims comparing ensemble methods. We created a linear opinion pool ensemble, and as suspected we find that this form of aggregation does not lose the information around uncertainty. We include this in the methods and results, and particularly highlight the potential role of LOP ensembles in the discussion. We have particularly focussed results on exploring ensembles of weighted trajectories, where we believe this work provides the most value.

- *Interpretability of weighted ensembles*
  *Methodologically, the proposed trajectory weighting method is interesting. However, it raises for me a question about what it means to condition on varying sets of data and*

> *the setting specified in a scenario. Suppose that a data-generating system produces outcomes $Y_t$ (here, disease incidence at time $t$) as well as variables $Z$ that are specified in a scenario setting (e.g. target population for a booster campaign, vaccine effectiveness). The data generating process has some joint distribution for these quantities, $F_{Y, Z}(y, z)$. My understanding is that the set up for scenario projections is that projections generated at time $T$ are generated by conditioning on a value $z$ specified in the scenario and the observed data up to that time, and estimating the conditional distribution of future values of the outcome $Y$: $F_{Y_{T+1:T+H} \mid Y_{:T}=y_{:T}, Z=z}(y_{T+1:T+H})$. Now suppose that time passes and at some point $S>T$ we wish to update the projections. What if the vaccine has already been administered targeting a specified age group, with a certain level of effectiveness, so that (some of) the intervening data points $y_{T+1:S}$ were actually generated under some value $z'$ which may not be equal to $z$? It would seem dubious to update projections generated under scenario $z$ when some of the data came from scenario $z'$... In other words, it seems like we can't do this kind of updating if there is any potential that the specified scenario is "obsolete" by the time we want to make the update?*
>
> *These questions are related to the following sentence in the discussion: "This could be used for ongoing evaluation of scenario projections, increasing the useful life of data from a single cross-sectional collection of multiple model output." Is this really possible? Under what conditions?*

Thanks to the reviewer for raising this, and we agree it is useful to expand on this area of the work in the discussion. For the context of this work, model results had originally been created based on a set of four scenarios relevant to European vaccine policy decisions in 2022. As the reviewer notes, no future scenario is likely to accurately predict eventual reality, as only some or none of the original scenario assumptions are realised and others are made obsolete. Given we started with four scenarios that deliberately contrasted in their assumptions, most of these assumptions would have been disproven by observation over time.

However, we do not believe that this means the individual model trajectories cannot be used. We note that we have not collected any new results from updating the original models; in this work we only took a single set of simulations and re-used them. The aim of this was to treat each trajectory as a single independent simulation, ignoring the scenario assumptions, modelling technique, parameter values, etc., that were originally involved in its creation. We wanted to explore whether it were possible to use trajectories without this context. We therefore treat all trajectories equally by only using the observed accuracy of each simulated trajectory. We can then create a weighted ensemble that can be updated over time (as observations become available, changing the accuracy weight of each trajectory). We suggest

that this abstracted use of trajectories is a key advantage of collecting trajectories, rather than quantile representations, of model results. We have added these points to the discussion.

- *Fourth paragraph of introduction, "standardise epistemic uncertainty across different models"*
  *I wonder if "standardise epistemic uncertainty" is the best description. Perhaps, the goal of these hubs is to reduce unwanted linguistic uncertainty (relating to differences in interpreting scenario assumptions, data sources, etc.) and retain important epistemic uncertainty (because the uncertainty captured across differing models is the primary benefit of ensembling predictions from multiple models). Nevertheless, the point about standardization and "like-for-like comparison" seems very important. My confusion may arise from differing definitions of "epistemic", "aleatory" and "stochastic" uncertainty, so providing definitions of each may be sufficient.*

We agree that there are varying aims and uses of uncertainty between hubs. The paragraph in question (and paper in general) comments on all of the various hub efforts, and on reflection these are not all explicitly standardised in their aims/goals/definitions around different types of uncertainty. We have reworded the paragraph to de-emphasise the speculation about different types of uncertainty, and focus on the more general like-for-like comparisons among direct model results.

- *Figure 1 provides a nice visualization of the alternative ensembles, as well as "information loss". It might be useful to reference the supplemental figures that show which trajectories were from which models.*

Thank you, we have included this reference.

- *I believe reference [6] and [15] are duplicated.*

We have checked that this is no longer the case.

- *Limitations of samples*
  *At some point in the manuscript, it would be nice to address what I see as the main limitation of collecting trajectory samples: it restricts the class of models that can be used to those that are able to estimate joint distributions of incidence across multiple weeks. This eliminates approaches such as quantile regression which have often had good performance in infectious disease modeling tasks. As a minor comment that is related, I note here a collision with text from the introduction stating "In several COVID-19 modelling hub efforts each modeller submits a common set of quantiles for each time point estimated from any number of trajectories." But as I've noted here,*

> *quantiles need not be obtained as a summary of trajectories from a probabilistic model, as direct application of quantile regression is possible.*

Thank you for noting this important point. We have included this limitation in the discussion and corrected the text in the introduction.

- *Abstract: "We found that collecting models' simulated trajectories, as opposed to collecting models' quantiles at each time point, enabled us to show ... performance against data". I think phrasing here could be clearer.*

Thanks, we have updated the abstract and hope it is both clearer and better reflects the reshaped work.

- *Introduction: "In contrast, scenarios are projections attuned to a particular context by being conditioned on specific qualitative factors...". It would seem that scenarios could condition on specific values of quantitative variables as well?*

Thanks, corrected.

- *Methods: "...with a focus on targets with three or more independent projections." Consider "...with a focus on targets with projections from three or more different models," to avoid use of the word "independent" which may be unclear (are they statistically/probabilistically independent?).*

Thanks, corrected.