Characterising information gains and losses when collecting multiple epidemic model outputs

Response to reviewers

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. 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 method of aggregation/combination, and re-balanced the article to focus away from aggregation 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

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. As we use a sliding 5-week window, the local maxima would not be observed in a strictly decreasing period.

- *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.*

- *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.

Manuscript Number: EPIDEMICS-D-23-00094

Characterising information loss due to aggregating epidemic model outputs

Dear Ms Sherratt,

Thank you for submitting your manuscript to Epidemics.

I have completed my evaluation of your manuscript. The reviewers recommend reconsideration of your manuscript following major revision. I invite you to resubmit your manuscript after addressing the comments below. Please resubmit your revised manuscript by Oct 24, 2023.

When revising your manuscript, please consider all issues mentioned in the reviewers' comments carefully: please outline every change made in response to their comments and provide suitable rebuttals for any comments not addressed. Please note that your revised submission may need to be re-reviewed.

To submit your revised manuscript, please log in as an author at https://www.editorialmanager.com/epidemics/, and navigate to the "Submissions Needing Revision" folder.

Research Elements (optional)
This journal encourages you to share research objects - including your raw data, methods, protocols, software, hardware and more – which support your original research article in a Research Elements journal. Research Elements are open access, multidisciplinary, peer-reviewed journals which make the objects associated with your research more discoverable, trustworthy and promote replicability and reproducibility. As open access journals, there may be an Article Publishing Charge if your paper is accepted for publication. Find out more about the Research Elements journals at https://www.elsevier.com/authors/tools-and-resources/research-elements-journals?dgcid=ec_em_research_elements_email.

Epidemics values your contribution and I look forward to receiving your revised manuscript.

Kind regards,
Katrina Lythgoe
Editor-in-Chief

Epidemics

Editor and Reviewer comments:

Reviewer's Responses to Questions

Note: In order to effectively convey your recommendations for improvement to the author(s), and help editors make well-informed and efficient decisions, we ask you to answer the following specific questions about the manuscript and provide additional suggestions where appropriate.

1. Are the objectives and the rationale of the study clearly stated?

Please provide suggestions to the author(s) on how to improve the clarity of the objectives and rationale of the study. Please number each suggestion so that author(s) can more easily respond.

Reviewer #1: Yes, the introduction provides a thorough description of the problem and relevant context.

Reviewer #2: Yes

---

2. If applicable, is the application/theory/method/study reported in sufficient detail to allow for its replicability and/or reproducibility?

Please provide suggestions to the author(s) on how to improve the replicability/reproducibility of their study. Please number each suggestion so that the author(s) can more easily respond.

Reviewer #1: Mark as appropriate with an X:
Yes [X] No [] N/A []
Provide further comments here:
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?

Reviewer #2: Mark as appropriate with an X:
Yes [X] No [] N/A []
Provide further comments here:

---

3. If applicable, are statistical analyses, controls, sampling mechanism, and statistical reporting (e.g., P-values, CIs, effect sizes) appropriate and well described?

Please clearly indicate if the manuscript requires additional peer review by a statistician. Kindly provide suggestions to the author(s) on how to improve the statistical analyses, controls, sampling mechanism, or statistical reporting. Please number each suggestion so that the author(s) can more easily respond.

Reviewer #1: Mark as appropriate with an X:
Yes [X] No [] N/A []
Provide further comments here:

Reviewer #2: Mark as appropriate with an X:
Yes [] No [] N/A [X]
Provide further comments here:

---

4. Could the manuscript benefit from additional tables or figures, or from improving or removing (some of the) existing ones?

Please provide specific suggestions for improvements, removals, or additions of figures or tables. Please number each suggestion so that author(s) can more easily respond.

Reviewer #1: 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?

Reviewer #2: No

---

5. If applicable, are the interpretation of results and study conclusions supported by the data?

Please provide suggestions (if needed) to the author(s) on how to improve, tone down, or expand the study interpretations/conclusions. Please number each suggestion so that the author(s) can more easily respond.

Reviewer #1: Mark as appropriate with an X:
Yes [] No [X] N/A []
Provide further comments here:
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.

Reviewer #2: Mark as appropriate with an X:
Yes [X] No [] N/A []
Provide further comments here:

---

6. Have the authors clearly emphasized the strengths of their study/theory/methods/argument?

Please provide suggestions to the author(s) on how to better emphasize the strengths of their study. Please number each suggestion so that the author(s) can more easily respond.

Reviewer #1: Yes, the authors provide a comprehensive discussion of the strengths and weaknesses of each potential information source (quantiles vs. trajectories).

Reviewer #2: yes

---

7. Have the authors clearly stated the limitations of their study/theory/methods/argument?

Please list the limitations that the author(s) need to add or emphasize. Please number each limitation so that author(s) can more easily respond.

Reviewer #1: Yes, the authors provide adequate limitations and useful future directions. 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.

Reviewer #2: yes

---

8. Does the manuscript structure, flow or writing need improving (e.g., the addition of subheadings, shortening of text, reorganization of sections, or moving details from one section to another)?

Please provide suggestions to the author(s) on how to improve the manuscript structure and flow. Please number each suggestion so that author(s) can more easily respond.

Reviewer #1: The manuscript is very well written; it is well organized and easy to follow.

Reviewer #2: no

---

9. Could the manuscript benefit from language editing?

Reviewer #1: No

Reviewer #2: No

Co-Guest Editor comments: While both reviewers agree that the paper presents an important and timely study worthy of publication, they have highlighted specific issues that need to be addressed before final decision. They raise concerns about conflating the choice of representation (trajectory vs quantiles) with the ensembling schemes. They also have questions about the proposed trajectory weighting schemes and how they will account for obsoletion of certain scenarios.

Reviewer #1: Minor comments:
1. 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.
2. 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.
3. I believe reference [6] and [15] are duplicated.

[Full review attached]

Reviewer #2: # EPIDEMICS-D-23-00094

Summary

This article explores advantages of using a trajectory (or sample-based) representation of predictive distributions in a scenario projection task rather than quantiles summarizing the predictive distributions. The paper explores three angles on this:
1. trajectories enable principled calculation of a variety of quantities that would not be possible to directly calculate from quantiles, such as cumulative counts over a time span or the timing of local peaks.
2. trajectories facilitate the calculation of linear pool ensembles, which capture between-model uncertainty better than Vincent ensembles (though I note that the authors did not frame the discusion quite this way).
3. trajectories enable novel weighting schemes

Overall, I think the paper raises some important points. In particular, I am sympathetic to the ideas in points 1 and 3.

However, I do have questions about the discussion related to point 2, which I describe in more detail below. Briefly, I think the second of these claims, about ensemble methods, conflates the format used to represent predictions with the statistical methods used to calculate ensembles. I do not find this part of the discussion convincing, and I think clearer discussion of related issues has been given elsewhere in the literature already (e.g., in "Context-dependent representation of within- and between-model uncertainty: aggregating probabilistic predictions in infectious disease epidemiology" by Howerton et al. (2023), which is already cited by this manuscript). I would prefer to see the discussion here substantively reworked or just removed.

That said, I think the ideas in points 1 and 3 are interesting and important, and deserve a place in the public record.

Major Comments

1. 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).

Minor Comments

2. 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?

3. 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.

4. Misc. minor points
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.
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?
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?).