16 January 2022

Dear Amy Wesolowski, Neil Ferguson,

Many thanks for reviewing our article on European multi-model forecasting, and many thanks to Jeffrey L Shaman and Sen Pei for their very thoughtful and helpful comments.

We include below a detailed summary of our responses to the reviewers' comments, and the action we have taken to address them. We have made several revisions to the manuscript and we believe this has improved the clarity of the work, particularly in articulating its added value.

We have also fixed an error in the code for this paper associated with processing anomalous data. This has increased the number of models included by 1, and does not have any substantial effect on the main results. Separately, we fixed an error in the methods for evaluating the retrospective weighted ensembles. This created some changes in the exact numbers in Table 1 but does not affect the results, which only concern relative difference.

We note that the manuscript places the Methods directly following the Background, rather than as a final section. We believe this improve the flow of the paper as many of the models discussed in the Results and Discussion sections are explained in the Methods section, helping understanding of the context and the details of the models and results.

In response to the journal's formatting recommendations, we have removed the Supplementary Information. We have instead incorporated Supplementary Figure 1 into the main text (now Figure 1), and removed Supp. Fig. 2. The supplementary tables (team metadata, and EPIFORGE guidelines) are uploaded as individual supplementary files.

Please find attached a copy of the revised manuscript along with a track-changes comparison to the first submitted version.

Best wishes,

Katharine Sherratt, on behalf of all co-authors

Summary of author responses

Reviewer comments are grouped by theme with reviewer number in square brackets, followed by our response and relevant actions.

Novelty

[eLife] There are other papers reporting very similar findings/work in this setting (and others) but the added value of this work, in particular, was not clear.

[#1] I guess my main question is: do we need another report on multi-model 'ensembling'?

[#2] These findings provide practical guides for generating real-time forecasts for infectious diseases and novel insights into coordinating international forecasting efforts during a public health emergency.

[#2] A parallel effort of real-time COVID-19 forecasting in the US (i.e., the US COVID-19 Forecast Hub) reported similar findings on the use of ensemble models. This study from Europe provides independent validation that shows the robustness of these findings. While both studies followed similar guidelines and used the same evaluation metrics (coverage and WIS), I believe there should be unique challenges associated with forecasting for multiple countries (as opposed to forecasting in a single country). As a result, it might be worthwhile to discuss those challenges and potential solutions to inform similar efforts in the future.

- Response: We agree with reviewers that our findings add depth rather than breadth to the evidence base for multi-model ensembles in real time infectious disease forecasting. As mentioned by Reviewer #2, this work was unique and unprecedented for European policy makers in spanning multiple countries while aiming to inform continent-wide public health and we believe holds particularly strong value in highlighting the relevance of forecasting at multiple policy-making scales (national, regional, international).
- Action: We have added commentary on the specific value of this effort to European policy makers as well as forecast producers in both the background and discussion sections.

Methodological limitations

[#1] ‘Teams could also submit a single-point forecast.' Were there any issues arising from this?

- Response: Several teams did submit point forecasts, or forecasts with less than the full set of quantiles (5 of 29 models evaluated here). We have historically reported absolute error for all models in real time but in this paper they are excluded from the evaluations using the interval score, which rely on the full set of quantiles.
- Action: The exclusion from the ensemble of forecasts without the full set of quantiles was not clear from the current paper text. We have now updated the text to make this exclusion

explicit (Methods section, under “Forecast evaluation”, and re-stated for clarity in the Results).

[#1] Note that in US flu forecasting, there is an expectation of observation revisions. Forecasts are validated against final revised observations, despite what was available in real time.

- Response: As discussed in the text (Discussion, page 14), we excluded forecasts with revised observations. As we noted: “*More generally it is unclear if the expectation of observation revisions should be a feature built into forecasts. Further research is needed to understand the perspective of end-users of forecasts in order to assess this.”* In the context of this paper we felt the fairest approach to evaluation was to exclude forecasts made for revised observations, while recognising that evaluating forecasts against updated data is a valid approach.
- Action: We have added a note on this alternative approach in the discussion.

[#2] Data in the tables and figures were used to compare forecasts. It would be great to have a formal statistical test for comparing model performance, if possible.

- Response: We have note used a formal statistical test for modelling model scores. Tests for comparing forecast performance are available such as the Diebold-Mariano test. However it is not clear how we might apply this to a situation with multiple models as well as multiple forecasts, in a way that is relative and includes a baseline reference forecast.
- Action: We have noted this in the discussion of methodological limitations. To add some slight further support to the comparisons we have added interquartile ranges around the median ensemble and model scores compared in the Results.

Added interpretation

[#2] death forecasts are more stable than case forecasts [...] the time series of death is subject to less severe fluctuations compared to the case time series. Assuming a long-tailed delay distribution (e.g., gamma distribution) from case reporting to death (which is usually the case), the convoluted time series of death typically has a lower magnitude of fluctuation as the effect of sharp changes in infection is spread out over time.

- Response: This is a sensible addition to the interpretation
- Action: We have clarified the discussion comparing deaths and cases to specifically include this point.

[#2] WIS is a strictly proper score for evaluating forecast performance; however, it must rely on a reference forecast model. This may create difficulties in interpreting forecast accuracy for the general public who may not understand the concept of WIS. For instance, what is a WIS score good enough to trust? The authors may want to include a simple metric (e.g., mean absolute

error) as a supplement even though these metrics have some caveats. I presume the performance should be highly correlated using different evaluation metrics.

- Response: This is an important point and it's fair that WIS may be difficult to interpret for a lay audience. For example, in the Hub's real-time performance reports (on the website), we focus on visualising absolute error and report the absolute weighted interval score (under/over performance and dispersion, in absolute terms). In this paper, we include figure 1 as an example of absolute difference between forecasts and observations in one location at a particularly difficult forecasting period. Elsewhere, as the reviewer notes we have explicitly focused on relative performance to a flat-line reference model. Currently, the Discussion section includes some comments on the epidemiological dynamics influencing absolute performance, and our choice of reference. We have updated the results to point to the Hub website for a fuller overview of absolute performance.
- Action: We have also added a comparison of ensemble performance by relative absolute error in the Results section (though we note this is still relative, not absolute).

[#2] It might be helpful to elaborate more on the assumptions for near-term predictions in participating models (e.g., status quo, reactive change of transmission, etc.). Essentially all real-time predictions were generated based on assumptions, although sometimes those assumptions were not stated explicitly. For behavior-induced changing points (peaks or troughs), it might be challenging to predict using the status quo without considering a change in model states.

- Response: We asked modellers to include as much information as possible in their metadata, including linking to and citing any more detailed work surrounding their model. We include this in the supplement. We agree the variation of modeller methods and assumptions is a critical element of explaining model performance, which we have not yet investigated.
- Action: We have added a note to direct readers to metadata in the supplement and online, and we have added a paragraph noting the challenges of varying modeller assumptions in the discussion of methodological limitations.

Definitions & clarifications

[#1] 'Coverage' is an evocative term; in weather, they more typically use 'reliability', defined as the correspondence between the forecast probability of an event and the observed frequency of that event. Consider at least noting that coverage, as defined here, is reliability. Calibration is used to describe reliability, and I note this is used in the text.

- Response: Thank you for suggesting this as we had not observed this difference in usage. Here we used coverage in line with previous work discussing interval scores for quantile forecasts.

- Action: We have added a note and reference in the Methods section for our use of “coverage”.

[#1] I believe your definition of F^(-1) (α), which is confusing as I reflexively read this as a matrix inverse (perhaps use G(α) instead), is for the supremum (least upper bound), not the infimum (greatest lower bound), i.e. G(α)=sup{t:F_i (t){greater than or equal to}α}. If not, I think the {greater than or equal to} should be {less than or equal to}.

- Response: We apologise if the notation was perceived as ambiguous. We followed the notation of the original reference cited in the paper[1] (eq.1.1). As an example, the 0.05 quantile of a distribution with cumulative distribution function F(x) would be the greatest value of t that is a lower bound of the set of all t that fulfil F(t)>=0.05, which is the definition as written (it would be the supremum if the direction of the inequality was reversed). We could replace it with the minimum here (the lowest t with F(t) >=0.05) for all practical intents and purposes, but decided to stay with the original notation so that it can be tracked to the given reference.
- Action: We have clarified reference to notation in the text.

[#2] In constructing the weighted ensemble, it would be good to clarify the period for which each model's past performance was evaluated and how the value of S_i was computed.

- Response: Yes, we agree this needs clarification.
- Action: We have clarified that forecasting skill means relative WIS using past performance over all available model data.

[#2] In the last paragraph of page 13, a typo "although though this finding ..."

- Response: Thank you for noting this.
- Action: We have removed the typo.

[1] C. Genest, “Vincentization Revisited,” The Annals of Statistics, vol. 20, no. 2, pp. 1137–1142, 1992,Available: https://www.jstor.org/stable/2242003