

Challenges in ensemble infectious disease modelling — when can you trust your forecast?

Thao P. Le, Sheryl Chang, Tom Harris, Camelia Walker, and Chris Baker

Abstract During the COVID-19 pandemic, forecasts of disease incidence and burden were used by policy-makers around the world to assist in decision-making. These forecasts were made by various models and were often combined into an ensemble to produce an average forecast across several models. These ensemble forecasts are often perceived to be more trustworthy than the forecast from a single model. However, not all models produce good forecasts during all times in the pandemic, and the unweighted averages used to form ensemble forecasts will, therefore, not be optimal. How can we better weigh individual models in an ensemble? When can we trust individual models, and when can we trust the output of the overall ensemble? This perspective examines COVID-19 ensemble forecasts and the challenges that need to be solved before the next pandemic.

1 Introduction

Forecasts of natural phenomena have become embedded in many people's daily lives, and one of the most prominent regular forecasts is of the weather. Using complex models with large data inputs, weather forecasting models can predict the temperature, rainfall, wind, and humidity of the upcoming days, across different spatial regions. *Ensemble modelling* is used in weather forecasting, where results from different models are combined. These models have different assumptions, initial conditions (incorporating the uncertainty in the starting state), and different errors; combining them often results in a more accurate forecast [4, 13]. Many people routinely check the weather forecast—and trust its predictions, even if only as a guide—and use it to inform decision-making [12].

Thao P. Le, Tom Harris, Camelia Walker, Chris Baker
The University of Melbourne, Melbourne, Australia

Sheryl Chang
The University of Sydney, Sydney, Australia

Forecasting for infectious diseases has been an area of research for at least several decades (e.g., [1, 15, 17]), but the COVID-19 forecasting greatly increased the interest in and visibility of infectious disease forecasts. COVID-19 forecasts were used for a range of applications, including to predict disease incidence and health system burden over timeframes of days and weeks. Like with weather forecasting, estimates from different COVID-19 models were often pooled to form ensemble forecasts [2, 6, 7, 16, 19].

Despite the broad uptake of ensemble forecasting, the reliability of forecasts throughout the COVID-19 pandemic was variable. For instance, the CDC reported that ensemble forecasts, although usually more reliable than singular forecasts, have been unable to predict rapid or sudden changes in the epidemic [5] (see Figure 1). At times of high volatility, such as when there are many new emerging variants, large mobility changes in the population, policy interventions, or major events, many models perform poorly as they do not or cannot account for the effects of these factors. In weather forecasting, this would correspond to periods of high and low *forecast skill* [10, 18].

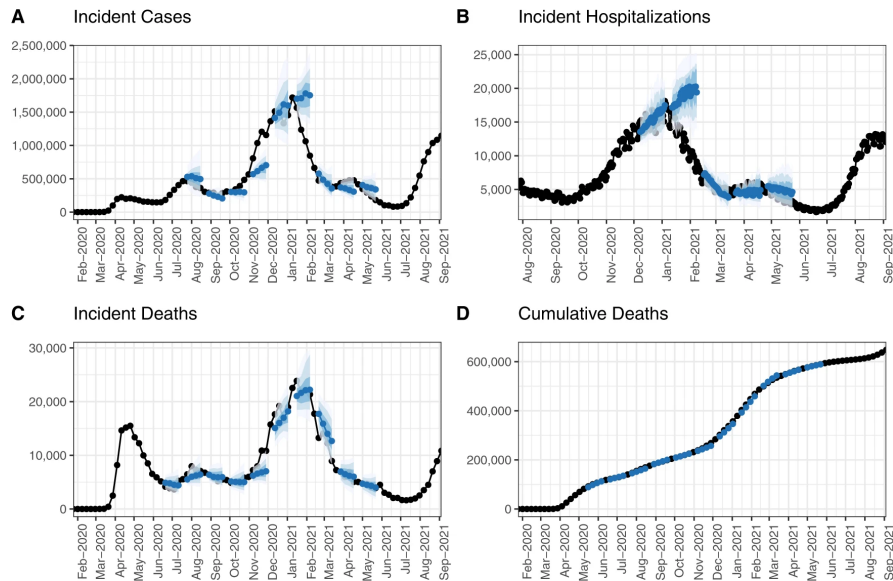


Fig. 1 Figure 1 from Cramer et al. [6]: times series of weekly (A) incident cases, (B) hospitalizations, (C) deaths and (D) cumulative deaths, in the US. Some forecasts from the US COVID-19 Forecast Hub ensemble model are shown in blue (with 50%, 80% and 95% prediction intervals) and the ground-truth data are shown in solid black.

Ensemble forecasts are typically unweighted (i.e., projections of all ensemble models were considered equally likely), but it is worth considering whether there are ways to improve accuracy through careful weighting. Weighted ensembles have been implemented (e.g., weighting forecasts by their weighted interval score based

on each individual model's recent past performance [6]). Unfortunately, weighting does not necessarily impart greater performance; Refs. [3, 19, 7] found that some unweighted averages did no worse than weighted methods (noting that there are a variety of equal-weighting methods available, in particular, medians outperformed means [19]).

This leads to several questions. Why does equal-weighting often perform better than using weighted interval scores? What is the cause of this discrepancy? How can we better weight individual models in an ensemble? Can we formalise a model selection process that could be done in real time? When can we trust individual models, and when can we trust the output of the overall ensemble?

Since improving the weighting mechanism could enhance forecasting accuracy, there has been a lot of retrospective work [11, 14, 16, 20] examining different methods to combine forecasts and calculations of forecast skill for COVID-19 models. There are many ways to combine and create an ensemble. Aside from an unweighted average, individual forecasts can also be combined via a geometric mean and median, with symmetric or asymmetric trimming, with weights based on continuous ranked probability scores and other various scores based on historical performance [20]. There are also formal Bayesian methods for model selection, model evaluation, and model averaging. Representative clustering has also been proposed [11]. Each of these methods has its strengths and weaknesses and does not always do better than a simple average or median. For example, weighting based on historical performance becomes complicated if information is not available for every model over the same time period [20], and Bayesian methods suffer from challenges related to computational complexity and sensitivity to priors [9]. Note also that while weighting by models' forecast skill is an attractive prospect, the skill of the overall ensemble is different from the skills of the individual models, and weighting based only on the skills of the individual models may not produce the most skilled ensemble [8].

We propose that beyond ensemble weighting, there needs to be an additional measure of the 'trustability' of the ensemble, which is informed not only by the models but also by external factors. Sometimes, the ensemble truly cannot forecast reliably and it is not possible to derive model weights that can solve this. In cases like this, we need to be able to detect potential model failure. Therefore, we propose that in addition to considering ensemble skill, we also need to consider *ensemble trustworthiness*, where the latter requires evaluation using information that may not be included in the historical data or incorporated ensemble models. A trustworthy ensemble should account for model performance, which can vary through time, *and* should be accompanied by a measure of "trustability" for each point in time.

2 Proposed directions for a trustability measure

Ideally a trustability measure will indicate, at various points in time, the degree to which reliable forecasting is likely to be accurate. This trustability measure would

be derived from known heuristics about infectious disease model forecasts. It has been observed that models perform better at different times or different points on the epidemic curve: for example, models perform well during stable periods, but are unable to predict peaks, troughs, and sudden growth [14]. Thus, in unstable periods, the trustability of the ensemble should be low.

There are a number of exogenous dynamics that may impact disease transmission and control to various extents and cause instability: changes in mobility, human behaviours, phylodynamics, and public health policy. Abrupt changes in these dynamics disrupt epidemic patterns and negatively affect modelling outcomes. A trust measure should be able to evaluate the impact of these dynamics on the ensemble forecast.

Conceptually, we propose that the trust measure should have three attributes: output extremity, output subjectivity, and external inconsistency (Table 1).

Table 1 Attribute description of trustability.

Attribute	Description
Output extremity	Measuring the extent of the extremity of the modelling scenario (e.g., does the ensemble forecast indicate an extreme shift in epidemic dynamics?)
Output subjectivity	Measuring the extent of subjectivity (or objectivity) of the modelling scenario (e.g., is the ensemble forecast responding to a specific modelling scenario?)
External inconsistency	Measuring the discrepancy between the modelling scenario and the exogenous dynamics (e.g., is there evidence from exogenous dynamics that suggests alternate modelling scenarios and parameter setup?)

These attributes evaluate the relevance of the ensemble forecast in the face of exogenous dynamics and shed light on the design of the modelling scenario. For example, if genomic evidence suggests a new variant gaining dominance in the population, the parameters used to model existing variants may no longer be representative of the emerging variant. In this case, the ensemble forecast generated using parameters for existing variants will receive a low trust score due to the higher external inconsistency and output subjectivity, and calls for re-design of epidemic scenario and parameter setup.

Figure 2 shows two alternative workflows of the proposed framework. Figure 2(a) shows a simple, linear process where the computed trust score for the ensemble forecast is fed directly into decision-making. Alternatively, the trust score can follow an iterative approach (Figure 2(b)) integrated into the ensemble weighting. This iterative process optimises the relevance of ensemble forecasts to ensure effective decision-making. For example, drastic changes in mobility patterns before public events reduce the trust score of existing ensemble forecasts. This suggests that models that capture changes in mobility should be given higher weight in the ensemble compared to those that only consider fixed mobility patterns, affecting the resultant ensemble forecast.

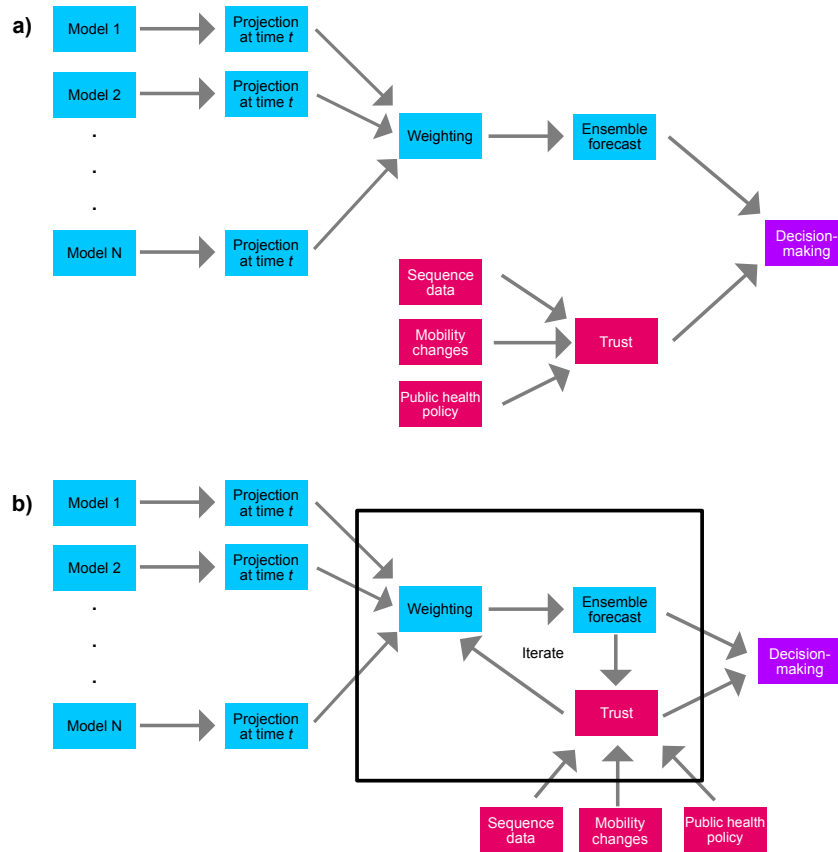


Fig. 2 Proposed workflows of ensemble forecasting incorporating a trust measure ('Trust'). Ensemble forecasts weight a collection of projections from some set of N predictive models. **(a)** Trust can capture the reliability of predictions in the broader forecast environment, calculated using a collection of external, real-time data sources (e.g., mobility). **(b)** An iterative approach where Trust is incorporated into ensemble weighting. Decision-making using both approaches can be informed by the ensemble forecast projections and the calculated Trust measure.

Functionally, this trust measure could be another model whose task is to predict instability (rather than to forecast future cases). A measure of trust could also be determined via expert elicitation by a range of experts with knowledge of upcoming events and policy changes. Regardless, it is important that it incorporates elements *beyond* what the ensemble itself contains, and provides a formal way to include all available information, not just the information available to the ensemble.

3 Conclusion

The COVID-19 pandemic has highlighted the usefulness of infectious disease forecasts, but it has also revealed their weaknesses. There is a lack of sophisticated and fast methods for evaluating model weights for infectious disease forecasting models. Critically, current methods do not integrate external information to flag whether ensembles are trustworthy. The development of these methods will improve the reliability of epidemic forecasts and prepare the modelling community for the next pandemic.

Acknowledgements We thank the mathematical research institute MATRIX in Australia where part of this research was performed.

References

1. Bansal, S., Chowell, G., Simonsen, L., Vespignani, A., Viboud, C.: Big Data for Infectious Disease Surveillance and Modeling. *The Journal of Infectious Diseases* **214**(Suppl 4), S375–S379 (2016). DOI 10.1093/infdis/jiw400
2. Bicher, M., Zuba, M., Rainer, L., Bachner, F., Ripplinger, C., Ostermann, H., Popper, N., Thurner, S., Klimek, P.: Supporting COVID-19 policy-making with a predictive epidemiological multi-model warning system. *Communications Medicine* **2**(1), 157 (2022)
3. Brooks, L.C., Ray, E.L., Bien, J., Bracher, J., Rumack, A., Tibshirani, R.J., Reich, N.G.: Comparing ensemble approaches for short-term probabilistic COVID-19 forecasts in the US. *International Institute of Forecasters* (2020). URL <https://forecasters.org/blog/2020/10/28/comparing-ensemble-approaches-for-short-term-probabilistic-covid-19-forecasts-in-the-u-s/>
4. Casanova, S., Ahrens, B.: On the weighting of multimodel ensembles in seasonal and short-range weather forecasting. *Monthly Weather Review* **137**(11), 3811 – 3822 (2009). DOI 10.1175/2009MWR2893.1
5. Centers for Disease Control and Prevention: Covid-19 forecasts: Cases (2023). URL <https://www.cdc.gov/coronavirus/2019-ncov/science/forecasting/forecasts-cases.html>. Accessed July 2024
6. Cramer, E.Y., Huang, Y., Wang, Y., Ray, E.L., Cornell, M., Bracher, J., Brennen, A., Castro Rivadeneira, A.J., Gerding, A., House, K., Jayawardena, D., Kanji, A.H., Khandelwal, A., Le, K., Niemi, J., Stark, A., Shah, A., Wattanachit, N., Zorn, M.W., Reich, N.G., Consortium, U.C.F.H.: The United States COVID-19 Forecast Hub dataset. *Scientific Data* (2022). DOI 10.1038/s41597-022-01517-w
7. Cramer, E.Y., Ray, E.L., Lopez, V.K., Bracher, J., Brennen, A., Castro Rivadeneira, A.J., Gerding, A., Gneiting, T., House, K.H., Huang, Y., et al.: Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the United States. *Proceedings of the National Academy of Sciences* **119**(15), e2113561,119 (2022)
8. Herrmann, M., Marzocchi, W.: Maximizing the forecasting skill of an ensemble model. *Geophysical Journal International* **234**(1), 73–87 (2023). DOI 10.1093/gji/ggad020
9. Hollenbach, F.M., Montgomery, J.M.: Bayesian model selection, model comparison, and model averaging. *The SAGE Handbook of Research Methods in Political Science and International Relations* pp. 937–960 (2020)
10. Kalnay, E., Dalcher, A.: Forecasting forecast skill. *Monthly Weather Review* **115**(2), 349 – 356 (1987). DOI 10.1175/1520-0493(1987)115(0349:FFS)2.0.CO;2

11. Kim, J.S., Kavak, H., Züfle, A., Anderson, T.: COVID-19 ensemble models using representative clustering. *SIGSPATIAL Special* **12**(2), 33–41 (2020). DOI 10.1145/3431843.3431848
12. Lazo, J.K., Morss, R.E., Demuth, J.L.: 300 billion served: Sources, perceptions, uses, and values of weather forecasts. *Bulletin of the American Meteorological Society* **90**(6), 785–798 (2009)
13. Leutbecher, M., Palmer, T.: Ensemble forecasting. *Journal of Computational Physics* **227**(7), 3515–3539 (2008). DOI 10.1016/j.jcp.2007.02.014. Predicting weather, climate and extreme events
14. Lopez, V.K., Cramer, E.Y., Pagano, R., Drake, J.M., O’Dea, E.B., Adey, M., Ayer, T., Chhatwal, J., Dalgic, O.O., Ladd, M.A., et al.: Challenges of COVID-19 Case Forecasting in the US, 2020–2021. *PLoS computational biology* **20**(5), e1011200 (2024)
15. Lutz, C.S., Huynh, M.P., Schroeder, M., Anyatonwu, S., Dahlgren, F.S., Danyluk, G., Fernandez, D., Greene, S.K., Kipshidze, N., Liu, L., Mgbere, O., McHugh, L.A., Myers, J.F., Siniscalchi, A., Sullivan, A.D., West, N., Johansson, M.A., Biggerstaff, M.: Applying infectious disease forecasting to public health: a path forward using influenza forecasting examples. *BMC Public Health* **19**(1) (2019). DOI 10.1186/s12889-019-7966-8
16. Moss, R., Price, D.J., Golding, N., Dawson, P., McVernon, J., Hyndman, R.J., Shearer, F.M., McCaw, J.M.: Forecasting COVID-19 activity in Australia to support pandemic response: May to October 2020. *Scientific Reports* **13**(1), 8763 (2023)
17. Myers, M., Rogers, D., Cox, J., Flahault, A., Hay, S.: Forecasting disease risk for increased epidemic preparedness in public health. In: *Remote Sensing and Geographical Information Systems in Epidemiology, Advances in Parasitology*, vol. 47, pp. 309–330. Academic Press (2000). DOI 10.1016/S0065-308X(00)47013-2
18. Richardson, D., Black, A.S., Monselesan, D.P., II, T.S.M., Risbey, J.S., Schepen, A., Squire, D.T., Tozer, C.R.: Identifying periods of forecast model confidence for improved subseasonal prediction of precipitation. *Journal of Hydrometeorology* **22**(2), 371 – 385 (2021). DOI 10.1175/JHM-D-20-0054.1
19. Sherratt, K., Gruson, H., Grah, R., Johnson, H., Niehus, R., Prasse, B., Sandmann, F., Deuschel, J., Wolfram, D., Abbott, S., Ullrich, A., Gibson, G., Ray, E.L., Reich, N.G., Sheldon, D., Wang, Y., Wattanachit, N., Wang, L., Trnka, J., Obozinski, G., Sun, T., Thanou, D., Pottier, L., Krymova, E., Meinke, J.H., Barbarossa, M.V., Leithauser, N., Mohring, J., Schneider, J., Wlazlo, J., Fuhrmann, J., Lange, B., Rodiah, I., Baccam, P., Gurung, H., Stage, S., Suchoski, B., Budzinski, J., Walraven, R., Villanueva, I., Tucek, V., Smid, M., Zajicek, M., Perez Alvarez, C., Reina, B., Bosse, N.I., Meakin, S.R., Castro, L., Fairchild, G., Michaud, I., Osthus, D., Alaimo Di Loro, P., Mariotti, A., Eclerova, V., Kraus, A., Kraus, D., Pribylova, L., Dimitris, B., Li, M.L., Saksham, S., Dehning, J., Mohr, S., Priesemann, V., Redlarski, G., Bejar, B., Ardenghi, G., Parolini, N., Ziarelli, G., Bock, W., Heyder, S., Hotz, T., Singh, D.E., Guzman-Merino, M., Aznarte, J.L., Morina, D., Alonso, S., Alvarez, E., Lopez, D., Prats, C., Burgard, J.P., Rodloff, A., Zimmermann, T., Kuhlmann, A., Zibert, J., Pennoni, F., Divino, F., Catala, M., Lovison, G., Giudici, P., Tarantino, B., Bartolucci, F., Jona Lasinio, G., Mingione, M., Farcomeni, A., Srivastava, A., Montero-Manso, P., Adiga, A., Hurt, B., Lewis, B., Marathe, M., Porebski, P., Venkatramanan, S., Bartczuk, R.P., Dreger, F., Gambin, A., Gogolewski, K., Gruziel-Slomka, M., Krupa, B., Moszyński, A., Niedzielewski, K., Nowosielski, J., Radwan, M., Rakowski, F., Semeniuk, M., Szczurek, E., Zielinski, J., Kisielewski, J., Pabjan, B., Holger, K., Kheifetz, Y., Scholz, M., Przemyslaw, B., Bodych, M., Filinski, M., Idzikowski, R., Krueger, T., Ozanski, T., Bracher, J., Funk, S.: Predictive performance of multi-model ensemble forecasts of COVID-19 across European nations. *eLife* **12**, e81,916 (2023). DOI 10.7554/eLife.81916
20. Taylor, J.W., Taylor, K.S.: Combining probabilistic forecasts of COVID-19 mortality in the United States. *European Journal of Operational Research* **304**(1), 25–41 (2023). DOI 10.1016/j.ejor.2021.06.044. The role of Operational Research in future epidemics/ pandemics