Ensemble^2: scenarios ensembling for communication and performance analysis

Clara Bay[1], Guillaume St-Onge[1], Jessica Davis[1], Matteo Chinazzi[1,2], and Alessandro Vespignani[1]

[1]Laboratory for the Modeling of Biological and Socio-technical Systems, Northeastern University, Network Science Institute, Boston, MA, USA
[2]The Roux Institute, Northeastern University, Portland, ME, USA

Introduction

During the COVID-19 pandemic, scenario modeling informed public health policy and decision-making. Different from forecasts that aim at predicting the most likely outcome based on current data and trends, scenario models are built on specific assumptions about human behavior, changing environmental conditions, or the emergence of new pathogens [8]. This fundamental difference makes it difficult to directly evaluate the accuracy of multiple scenarios in the same way as forecast models.

In this work, we propose a novel ensembling procedure to aggregate projections and encompass the epistemic uncertainty associated with multiple scenarios. Our aim is to provide a methodology to assess the performance of scenario projections that remains independent from the a-posteriori identification of the most plausible scenarios, which may be clouded by specific and non-transparent additional modeling and parameter assumptions [4].

Methods

We present a quantitative analysis of our procedure using the Scenario Modeling Hub (SMH) projections for COVID-19 in the US. For each projection round, the SMH framework defines a matrix of four distinct scenarios, which identify specific epidemic indicators or drivers of interest tailored to each round [5], and we use them to define the **scenario ensemble** of each model by computing the median of each quantile across all scenarios, shown in Fig. 1.

The SMH integrates individual models' projections into a unified ensemble projection through three distinct methodologies: a modified version of the Vincent averaging technique using the median (Ensemble_vincent) [7], and the linear opinion pool method [6] with (Ensemble_LOP) and without (Ensemble_LOP_untrimmed) excluding the highest and lowest quantiles [3, 5].

We term the scenario ensemble of the SMH-reported ensemble models as **Ensemble^2**. This scenario ensemble procedure includes in the performance assessment: a) the ability of the defined scenarios assumptions to encompass the future trajectory of the epidemic, assessing if both upper and lower bounds are enveloping the realized epidemic trajectory; and b) assess whether the models are well calibrated simultaneously.

This approach also acknowledges that the future epidemic evolution should be viewed as a continuum of potential scenarios, with interpolations occurring between the specific ones identified in each round's quadrant. We assess performance using metrics such as prediction interval coverage, and weighted interval score (WIS).

The prediction interval coverage is defined as the percentage of times the actual outcome falls within the prediction interval across multiple predictions [2]. A well-calibrated model should demonstrate a strong correspondence between the forecast probabilities and the observed frequencies. The WIS accounts for the size of the prediction intervals, the placement of the intervals relative to the true outcome, and the weights assigned to the intervals [1]. Lower WIS values indicate better forecast performance. We calculate the standardized rank of the WIS as a straightforward way to compare model performance. It is computed by ranking the models, then the rank is reported on a 0 to 1 scale, with 1 being attributed to the best model and 0 to the worst.

Constructing the scenario ensemble using a median over all scenarios.

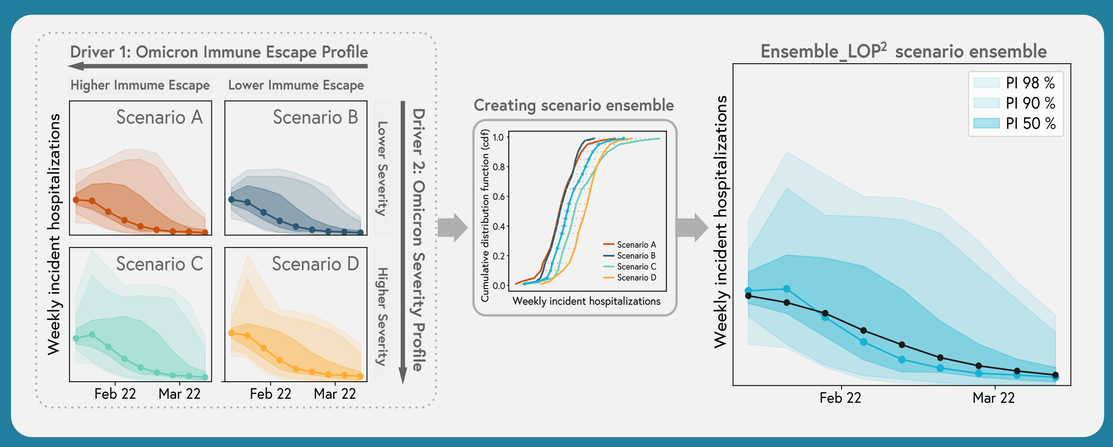
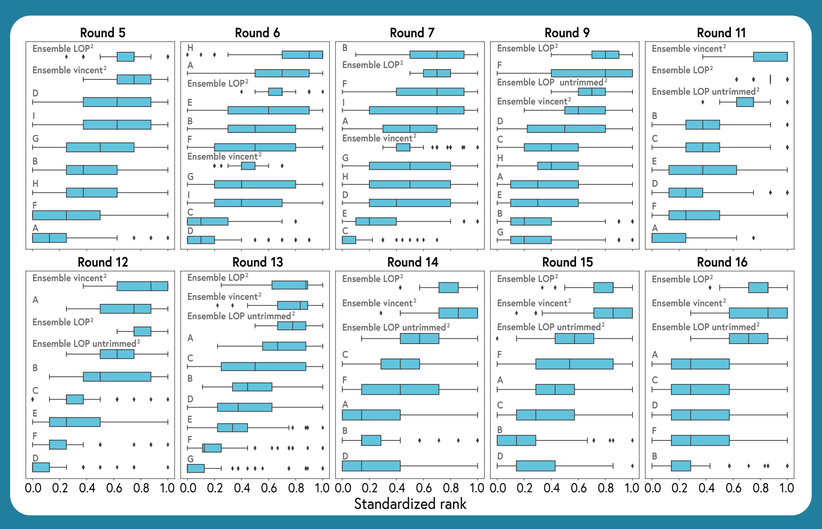


Fig. 1. *Constructing the Ensemble_LOP^2 scenario ensemble for weekly incident hospitalization projections at the national level in the United States for round 12 of the Scenario Modeling Hub, which addresses the Omicron wave. All 23 quantiles of each of the scenario projection A-D of the Ensemble_LOP model (left) is used to construct the scenario ensemble Ensemble_LOP^2 model (right). The middle panel shows the method of constructing the scenario ensemble for one date, where we take the median over scenarios A-D for each quantile.*

Ensemble^2 models outperform individual model's scenario ensemble.

We see in Fig. 2 that the Ensemble_vincent^2 and the Ensemble_LOP^2 models are outperforming all other models in six over ten rounds of projections and one of them ranks across the top three models in all rounds. This corroborates the results found in several studies: ensemble models are overall better calibrated and performing than individual models [2], and this extends into the ensemble across scenarios.

Fig. 2. *Distribution of standardized rank values for WIS averaged over all weeks of the corresponding projection round of each of the scenario ensemble models. The WIS is ranked so the model with the larger standardized rank has a better prediction.*



Ensemble^2 models are well-calibrated.

Fig. 3 shows that the scenario ensembles of all individual models display overconfidence in their predictions- their prediction intervals are excessively narrow- with the exception of the Ensemble_LOP^2 and Ensemble_LOP_untrimmed^2. Moreover, from the coverage of Ensemble_LOP^2 for a few selected states, we see that while it occasionally exhibits slight overconfidence and at other times slight underconfidence, the model is generally well-calibrated.

In Fig. 4 we observe that the coverage is fairly good for all Ensemble^2. Overall, the Ensemble_vincent^2 is slightly overconfident, the Ensemble_LOP_untrimmed^2 underconfident, and the Ensemble_LOP^2 is generally well calibrated, with the exceptions of rounds 6 and 7, where all models underestimated the Delta variant.

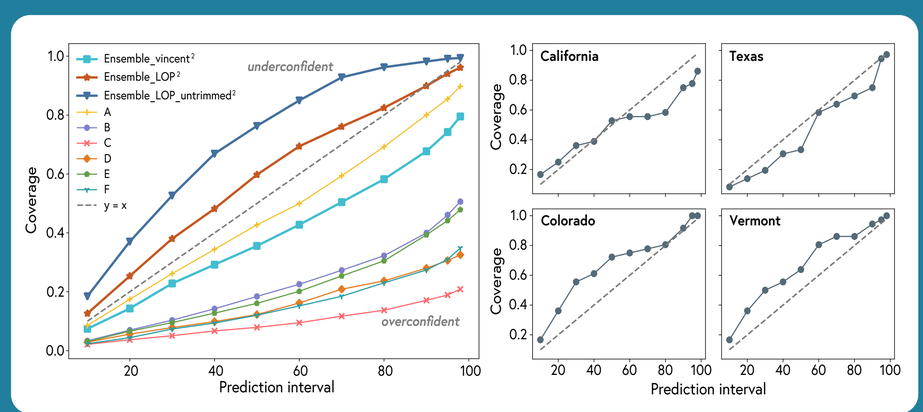


Fig. 3. *(left) Coverage over different prediction intervals for the scenario ensemble predictions for all targets (cases, deaths, and hospitalizations) and all US states during round 12 for each model. A to F correspond to the scenario ensemble of individual models. (right) Coverage over different prediction intervals for the Ensemble_LOP^2 for round 12 and for all targets in 4 different US states, ordered by descending population size. The dotted line represents the case where the coverage perfectly matches the expectation from the prediction interval. Coverages below (above) this line represent overconfidence (underconfidence).*

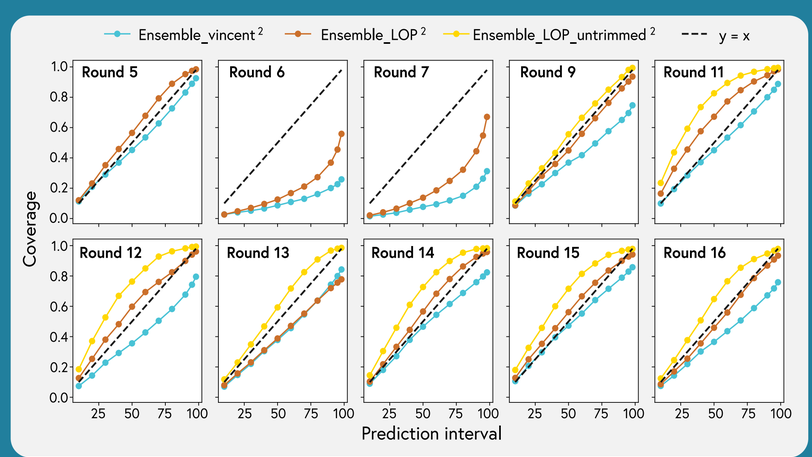


Fig. 4. *Coverage versus prediction intervals for the three Ensemble^2 models for each round of Scenario Modeling Hub projections. The coverage is taken over all targets (cases, deaths, and hospitalizations) and locations (US states). The Ensemble_LOP model started being reported in round 5, and the Ensemble_LOP_untrimmed model in round 9.*

Discussion

Through the examination of 10 rounds of SMH scenario projections, we find that the Ensemble^2 models generally outperform the scenario ensemble of individual models, and yield well-calibrated projections capable of enveloping epidemic trajectories, even when individual models or scenarios fall short. The inability of a scenario ensemble to offer sufficient coverage can serve as an effective indicator of issues with scenario specifications and/or model definitions. In turn, this approach contributes to a more efficient yet transparent communication of scenario projections to the public, along with more informed and effective decision-making in the face of epidemics.

The performance assessment proposed here is not limited to the SMH scenario modeling framework and can be potentially extended to consider any scenario design strategy. Finally, it is possible to envision refinement of this approach in which the scenarios are weighted according to specific priors, and the Ensemble^2 can evolve over time.

Acknowledgements & References

CB, MC, JTD, GS, and AV acknowledge support from CDC-HHS-6U01IP001137-01 and Cooperative Agreement no. NU38OT000297 from the Council of State and Territorial Epidemiologists (CSTE). GS additionally acknowledges financial support from the Fonds de recherche du Que bec – Nature et technologies (project 313475). MC additionally acknowledges support from CDC-JHU-2005702123.

[1] Bracher, J., et al., 2021. Evaluating epidemic forecasts in an interval format. PLOS Comput. Biol. 17, 1–15. **[2]** Cramer, E.Y., et al., 2022a. Evaluation of individual and ensemble probabilistic forecasts of covid-19 mortality in the united states. Proc. Natl. Acad. Sci. U.S.A. 119, e2113561119. **[3]** Howerton, E., et al., 2023a. Context-dependent representation of within- and between-model uncertainty: aggregating probabilistic predictions in infectious disease epidemiology. J. R. Soc. Interface 20, 20220659. **[4]** Howerton, E., et al., 2023b. Informing pandemic response in the face of uncertainty. an evaluation of the u.s. covid-19 scenario modeling hub. medRxiv. **[5]** Scenario Modeling Hub, 2023. COVID-19 Scenario Modeling Hub. https:// covid19scenariomodelinghub.org/. **[6]** Stone, M., 1961. The opinion pool. Ann. Math. Stat. 32, 1339–1342. **[7]** Vincent, S.B., 1912. The Functions of the Vibrissae in the Behavior of the White Rat. Ph.D. thesis. University of Chicago. **[8]** Vollmar, H.C., et al., 2015. Using the scenario method in the context of health and health care – a scoping review. BMC Medical Res. Methodol. 15, 89.

MOBS LAB | Laboratory for the Modeling of Biological + Socio-technical Systems

netsi | Northeastern University Network Science Institute