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. 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.

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, 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. We use this approach because SMH scenarios are often designed to define upper and lower limits of epistemic uncertainty, with the reality expected to fall somewhere in between. Consequently, we make use of an ensembling approach that effectively interpolates between the scenario projections.

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), and the linear opinion pool method with (Ensemble_LOP) and without (Ensemble_LOP_untrimmed) excluding the highest and lowest quantiles.

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 for the plausible range of outcomes 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, mean absolute percentage error (MAPE), and weighted interval score (WIS).

Constructing the scenario ensemble

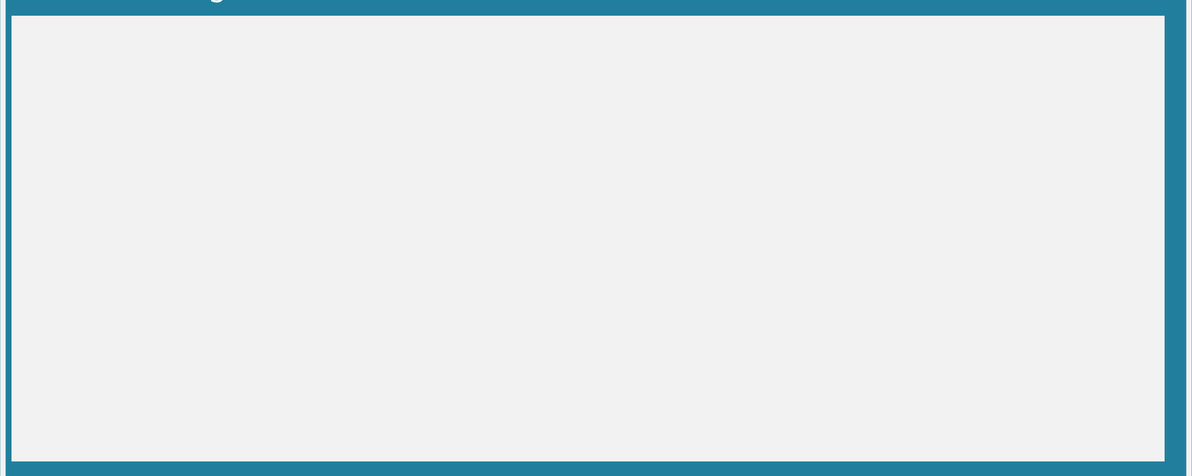

Fig. 1. Constructing the Ensemble_LOP^2 scenario ensemble for weekly incident hospitalization projections at the national level in the United States for round 12 of the Scenario Modeling Hub, which addresses the Omicron wave. All 23 quantiles of each of the scenario projection A-D of the Ensemble LOP 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.

Acknowledgements & References

acknowledgements?

[1] Bracher, J., Ray, E.L., Gneiting, T., Reich, N.G., 2021. Evaluating epidemic forecasts in an interval format. PLOS Comput. Biol. 17, 1–15. doi:10.1371/journal.pcbi.1008618. **[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. doi:10.1073/pnas. 2113561119. **[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. doi:10.1098/rsif.2022.0659. **[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 URL: https://www.medrxiv.org/content/early/ 2023/07/03/2023.06.28.23291998, doi:10.1101/2023.06.28.23291998. **[5]** Scenario Modeling Hub, 2023. COVID-19 Scenario Modeling Hub. https:// covid19scenariomodelinghub.org/. Accessed: 2023-11-13. **[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.

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

netsi | Northeastern University *Network Science Institute*

Distribution of MAPE and WIS for all scenario ensembles

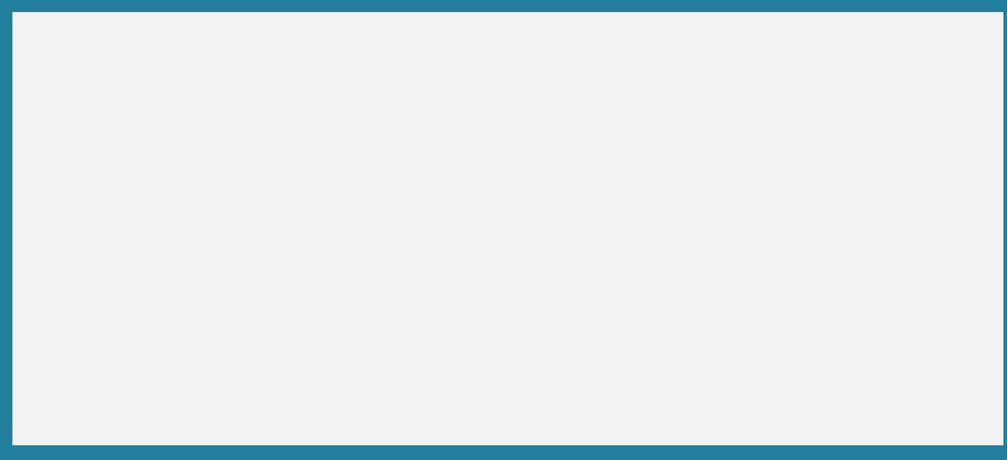

Fig. 2. Distribution of the A) the MAPE scores and B) the WIS scores rescaled by the standard deviation over all models and averaged over all weeks for each scenario ensemble model for round 12. Each point in the scatter plot refers to a specific target (cases, deaths, and hospitalizations) and location (US states), ordered by median MAPE/WIS value.

The distribution of the MAPE and rescaled WIS across all states and targets, where the median values suggest that the Ensemble^2 models outperform the scenario ensemble models in round 12. This confirms the efficacy of the Ensemble^2 approach, which offers enhanced accuracy, and improved estimation of uncertainty than what is typically achieved with single models.

Coverage for all 3 Ensemble^2 models

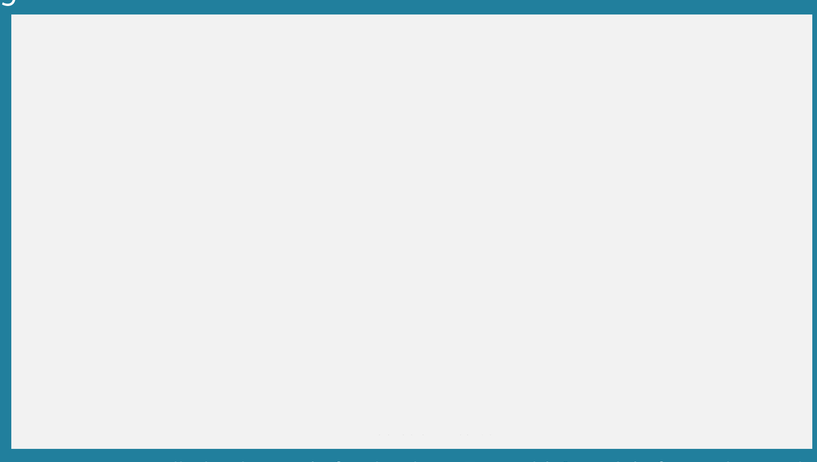

Fig. 3. Coverage versus prediction intervals for the three Ensemble^2 models for each round of Scenario Modeling Hub projections. The coverage is taken over all targets (cases, deaths, and hospitalizations) and locations (US states). The Ensemble_LOP model started being reported in round 5, and the Ensemble_LOP_untrimmed model in round 9.

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.

Standardized rank distribution for WIS

Fig. 4. Distribution fo standardized rank values for WIS value averaged over all weeks of the corresponding projection round of each of the scenario ensemble models in each round. WIS ranked so the model with the larger standardized rank has a better prediction.

We see in Fig. 4 that the Ensemble_vincent^2 and the Ensemble_LOP^2 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.

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.