Epidemics 46 (2024) 100748

Contents lists available at ScienceDirect

Epidemics

journal homepage: www.elsevier.com/locate/epidemics

# Ensemble[2]: Scenarios ensembling for communication and performance analysis

Clara Bay [a,1], Guillaume St-Onge [a,1], Jessica T. Davis [a], Matteo Chinazzi [a,b], Emily Howerton [c], Justin Lessler [d,e,f], Michael C. Runge [g], Katriona Shea [c], Shaun Truelove [f,h], Cecile Viboud [i], Alessandro Vespignani [a,b,*]

[a] Laboratory for the Modeling of Biological and Socio-technical Systems, Northeastern University, Network Science Institute, Boston, MA, USA
[b] The Roux Institute, Northeastern University, Portland, ME, USA
[c] Department of Biology and Center for Infectious Disease Dynamics, The Pennsylvania State University, University Park, PA, USA
[d] Department of Epidemiology, University of North Carolina Gillings School of Public Health, Chapel Hill, NC, USA
[e] Carolina Population Center, University of North Carolina, Chapel Hill, NC, USA
[f] Department of Epidemiology, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, USA
[g] U.S. Geological Survey, Eastern Ecological Science Center, Laurel, MD, USA
[h] Department of International Health, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, USA
[i] Division of International Epidemiology and Population Studies, Fogarty International Center, National Institutes of Health, Bethesda, MD, USA

ARTICLE INFO

*Keywords:*
Ensemble method
Scenario projections
COVID-19 models

ABSTRACT

Throughout the COVID-19 pandemic, scenario modeling played a crucial role in shaping the decision-making process of public health policies. Unlike forecasts, scenario projections rely on specific assumptions about the future that consider different plausible *states-of-the-world* that may or may not be realized and that depend on policy interventions, unpredictable changes in the epidemic outlook, etc. As a consequence, long-term scenario projections require different evaluation criteria than the ones used for traditional short-term epidemic forecasts. Here, we propose a novel ensemble procedure for assessing pandemic scenario projections using the results of the Scenario Modeling Hub (SMH) for COVID-19 in the United States (US). By defining a "scenario ensemble" for each model and the ensemble of models, termed "Ensemble[2]", we provide a synthesis of potential epidemic outcomes, which we use to assess projections' performance, bypassing the identification of the most plausible scenario. We find that overall the Ensemble[2] models are well-calibrated and provide better performance than the scenario ensemble of individual models. The ensemble procedure accounts for the full range of plausible outcomes and highlights the importance of scenario design and effective communication. The scenario ensembling approach can be extended to any scenario design strategy, with potential refinements including weighting scenarios and allowing the ensembling process to evolve over time.

## 1. Introduction

During the COVID-19 pandemic, scenario modeling played a critical role in shaping public health policy decision-making by exploring possible future trajectories of the pandemic and to better understand the potential consequences of interventions (Jewell et al., 2020; Borchering et al., 2021; Biggerstaff et al., 2022; Rosenblum et al., 2022; Reich et al., 2022; Truelove et al., 2022; Borchering et al., 2023). Different from forecasts that aim to predict as accurately as possible future outcomes based on current data and trends, the projections generated with scenario models depend on specific assumptions about human behavior, changing environmental conditions, or the emergence of new pathogens or variants (Vollmar et al., 2015; Runge et al., 2023) that are generally designed around policy-making questions, and may never be exactly realized in the future—e.g., expectation for vaccine coverage or whether or not nonpharmaceutical interventions will be relaxed. This fundamental difference makes it difficult to directly evaluate the performance of typically long-term scenario projections in the same way as short-term forecasts. In forecasts, the degree to which the predicted values match the actual outcomes is crucial; in scenario projections, evaluating the performance requires a different set of criteria that

* Corresponding author at: Laboratory for the Modeling of Biological and Socio-technical Systems, Northeastern University, Network Science Institute, Boston, MA, USA.
*E-mail address:* a.vespignani@northeastern.edu (A. Vespignani).
[1] Authors contributed equally.

https://doi.org/10.1016/j.epidem.2024.100748
Received 17 August 2023; Received in revised form 19 December 2023; Accepted 6 February 2024
Available online 8 February 2024

consider not only the accuracy of model outputs in matching reality but also how well the range of scenario assumptions is able to bound the realized epidemic drivers and capture the range of possible outcomes. This consideration assumes that the scenarios are designed to achieve this goal, i.e., that they aim to provide a "bracketing" of the epistemic uncertainties (Runge et al., 2023).

In this paper, we present a quantitative analysis of a novel ensemble procedure for assessing epidemic scenario projections. Ensemble methods are commonly utilized to consolidate predictions from numerous models. This process has proven to generate more accurate results and significantly improve the representation of uncertainty (Bates and Granger, 1969; Krishnamurti et al., 1999; Biggerstaff et al., 2018; McGowan et al., 2019; Reich et al., 2019a; Cramer et al., 2022a; Lutz et al., 2019; Reich et al., 2022). Here, we propose an alternative use of ensembles where we aggregate model outputs across multiple scenarios using the results generated by the Scenario Modeling Hub (SMH) in 10 rounds of projections for the trajectory of the COVID-19 pandemic in the United States (US). The SMH framework consistently defines a matrix of four distinct scenarios and for each model, we define a *scenario ensemble* that generates an overall projection. We refer to the *scenario ensemble* of the ensemble of models as "Ensemble$^2$". Subsequently, the Ensemble$^2$ performance is assessed using standard metrics, such as coverage, mean absolute percentage error (MAPE), and the weighted interval score (WIS) to provide a comprehensive evaluation across all projection rounds.

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) the ability to 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. This perspective is crucial when scenarios are designed not to predict specific trajectories but to explore and bound the uncertainties inherent in pandemic progression, such as transmission rates of emerging variants or vaccine uptake. By interpolating between these scenarios, Ensemble$^2$ offers a nuanced view of potential futures, enhancing our understanding of the pandemic's trajectory within the bounds of defined uncertainties. Finally, this methodology remains independent from the *a-posteriori* identification of the most plausible scenarios (Howerton et al., 2023b), which may be clouded by specific and non-transparent additional modeling and parameter assumptions.

The performance assessment of the SMH projections indicates that the Ensemble$^2$ models are generally outperforming the scenario ensemble of individual models. The approach is able to identify specific rounds where the Ensemble$^2$ models are miscalibrated, potentially indicating that the scenario specifications are not enveloping the future trajectories and/or that most of the models are not providing an adequate representation of the epidemic dynamics, thus highlighting the importance of the scenario design process. The proposed scenario ensemble procedure provides a more coherent representation of possible epidemic outcomes and holds significant importance when it comes to communication with the public and policymakers.

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.

## 2. Materials and methods

In the following, we consider COVID-19 scenario projections in the US at both national and state levels, as coordinated by the SMH. The SMH has coordinated 16 rounds of projections as of November 2022. For our analysis, we look at projection rounds 5–16, with rounds 8 and 10 excluded due to their use as internal training rounds. We exclude early rounds 1 to 4 that used slightly different approaches in the data aggregation and reporting. A typical round of scenario projections is composed of four distinct scenarios. These scenarios are organized into a $2 \times 2$ matrix, each representing potential trajectories of the epidemic, based on different assumptions (see Fig. 1, left panel). Each round of scenario projections focused on specific epidemic indicators or drivers of interest varying along each axis of the matrix. Examples of these drivers included the availability and uptake of vaccines, the application/relaxation of non-pharmaceutical interventions, and the uncertainty surrounding the factors contributing to the growth advantage of emerging variants. The specific drivers considered and assumptions made about their variability are tailored to each round of projections. For each round, the modeling teams supply target projections comprised of 23 quantiles (0.01, 0.025, 0.05, every 5% up to 0.95, 0.975, and 0.99) for each week of the projection period. These quantiles represent anticipated incident cases, incident hospitalizations, and incident deaths. To visualize and evaluate probabilistic estimates, quantile projections are transformed into central *prediction intervals* (PIs), which encapsulate a model's confidence that future observations will land within a specified range of values; for instance, the 50% PI is derived from the interquartile range. See the SMH website (Scenario Modeling Hub, 2023) for further visualizations of scenario projections.

The SMH integrates individual models' projections into a unified ensemble projection through three distinct methodologies. The first of these is a modified version of the Vincent averaging technique (Vincent, 1912; Howerton et al., 2023a). In this approach, each reported quantile $Q_i$ corresponds to the median of the quantiles $Q_i^m$ over all individual models $m$. This composite model is simply known as the Ensemble model, but we refer to it as the *Ensemble_vincent* model here to distinguish it from other approaches. The underlying assumption behind Vincent averaging is that all predictions are flawed approximations of a single target distribution. However, it posits that the random noise across these predictions can be averaged out, thereby producing an appropriate, aggregated distribution (Howerton et al., 2023a). This also implies that the actual outcome of the epidemic trajectory is expected to fall in between the different projections. The second method employed by the SMH is grounded in probability averaging, also known as the Linear Opinion Pool (Stone, 1961; Howerton et al., 2023a). Rather than aggregating the quantiles, this technique averages the cumulative probabilities of the individual models. These probabilities are reconstructed from the quantile predictions via linear interpolation. The resulting model is dubbed the *Ensemble_LOP_untrimmed* model. This technique considers that all predictions are distinct plausible alternative futures and that the uncertainty should be preserved, resulting in a higher (or equal) variance compared to the Vincent averaging technique (Howerton et al., 2023a). Finally, the last method (*Ensemble_LOP*) is identical to the second one, but the highest and lowest quantiles at a given value are excluded beforehand, which reduces the variance of the resulting ensemble.

### 2.1. Construction of the scenario ensemble

In order to succinctly communicate the performance of scenario projections, we propose the construction of the *scenario ensemble*, aggregating the four different scenarios, for each individual model and the ensemble models, termed "Ensemble$^2$". More specifically, we refer to the latter as "Ensemble_vincent$^2$", "Ensemble_LOP_untrimmed$^2$", and "Ensemble_LOP$^2$" according to each methodology used for the construction of the multi-model ensemble. Our approach to scenario ensembling mirrors the process used by the Scenario Modeling Hub in creating the Ensemble_vincent model. In this method, we compute the median of each quantile across all scenarios (with linear interpolation for ties). This procedure is exemplified in Fig. 1, where we illustrate the projections of the Ensemble_LOP model for each of the four scenarios in

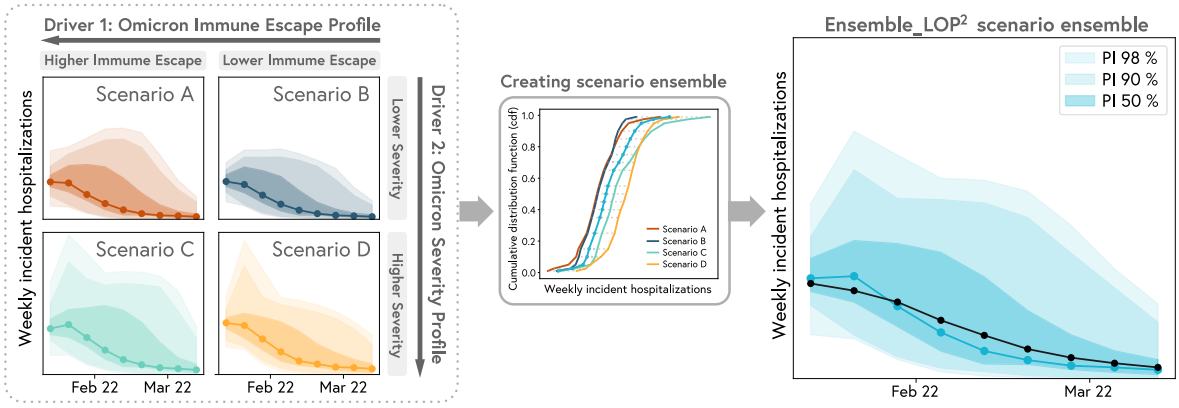


**Fig. 1.** Construction of the Ensemble_LOP$^2$ scenario ensemble for weekly incident hospitalization projections at the national level in the United States (US) for round 12 of the Scenario Modeling Hub (SMH). Note this is in 2022 and addresses the Omicron wave. All 23 quantiles of each of the scenario projections A-D of the Ensemble_LOP model, an SMH-reported ensemble over models, (left) are 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. Each colored line in the middle panel represents a scenario from the left panel, with the blue line corresponding to the median, which is used to create the ensemble. In the right panel, black circles represent the observed hospitalizations in the US at this time, whereas the colored circles correspond to the median of the model prediction. Prediction intervals (PIs) are represented by the shaded regions. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

round 12, as well as the projection of the resulting scenario ensemble model.

Traditional approaches often limit their focus to individual scenarios, each exploring specific outcomes under certain assumptions. However, when scenarios are structured to define the upper and lower limits of scenario uncertainties, such as the transmission advantage of a new variant or the vaccine intake in the population, reality is often a complex interplay between these extremes. The Ensemble$^2$ goes beyond simple probability averaging or selecting plausible scenarios. It dynamically integrates various projections, providing a synthesized view that is more informative for understanding the range of potential pandemic trajectories. Consequently, we make use of an ensembling approach that effectively interpolates between the scenario projections. This approach is particularly valuable in acknowledging the inherent uncertainties in pandemic progression and scenario planning.

In cases where scenarios do not aim to bracket uncertainty, Ensemble$^2$ approaches may not be suited to provide the needed information or should be adapted using probability averaging or excluding certain scenarios to prevent bias in the ensemble. See Figs. S11 and S12, and Table S3 in the Supplementary Material for an analysis on the ordering of ensemble steps in the construction of an Ensemble$^2$ model.

### 2.2. Performance metrics

In order to evaluate the efficacy of the scenario ensemble projections, we employ a range of scoring techniques. These measures compare the point and probabilistic estimates of a model with the observed ground truth values. The scores are computed on a weekly basis throughout the projection period for each specified target, which include cases, hospitalizations, and deaths. In our analysis, it is important to acknowledge that the reported cases, hospitalizations and deaths used as 'ground truth' are imperfect proxies for the true extent of the pandemic. These data are inherently subject to biases and noise, such as reporting delays, changes in testing rates, and other factors influencing case detection and recording. Despite these limitations, they represent the best available data for retrospective evaluation and are essential for our long-range scenario modeling.

#### 2.2.1. Prediction interval coverage

The first metric we use is the prediction interval coverage (or the coverage for short), defined as the percentage of times that the actual outcome falls within the PI across multiple predictions (Cramer et al., 2022a). More precisely, for $n$ time points, and a given $(1-\alpha)\times 100\%$ PI, the coverage is calculated as follows:

$$\text{Coverage}_{\alpha} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(l_{\alpha,i} \le y_i \le u_{\alpha,i}) \,,$$

where $l_{\alpha,i}$ represents the lower bound of the PI, $u_{\alpha,i}$ denotes the upper bound of the PI, and $y_i$ signifies the observed value at each time point $i$. Using this expression, we are able to calculate the coverage for each of the 11 PIs (98%, 95%, 90%, 80%, ... 10%) based on the 23 quantiles submitted by the modeling teams (Bracher et al., 2021; Cramer et al., 2022a). The coverage is a key measure in assessing the calibration of a model. Calibration in the context of probabilistic forecasting refers to the alignment between forecast probabilities and the frequencies observed in reality. For instance, if a model yields 20%, 50%, and 80% prediction intervals, we would anticipate that the actual values reside within these respective intervals 20%, 50%, and 80% of the time.

A well-calibrated model should demonstrate a strong correspondence between the forecast probabilities and the observed frequencies. If a model lacks proper calibration, we may detect low coverage (frequencies less than the corresponding prediction intervals) or high coverage (frequencies greater than the corresponding prediction intervals). Low coverage is indicative of an overconfident model, wherein the actual data points often fall outside the PI. Conversely, high coverage suggests an underconfident model, which tends to generate overly broad PIs.

Coverage holds particular significance in the context of scenario analysis, where we anticipate that the projections should encapsulate all potential epidemic trajectories should scenario assumptions materialize as stipulated. In essence, the prediction interval is expected to encompass the ground truth data, and overconfident projections could significantly mislead the policy-making process by prematurely excluding specific outcomes.

#### 2.2.2. Mean absolute percentage error

A second metric we use is the classic mean absolute percentage error (MAPE), which is a relative measure that assesses the accuracy of a point prediction and does not account for the probabilistic uncertainty (Makridakis et al., 1982). The MAPE is calculated with observed values $y_i$ and predicted point estimates $P_i$ over $n$ time points with

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - P_i}{y_i}\right| .$$

This metric is undefined if the observation is equal to zero, so we discard these cases from our analysis. The modeling teams provide a

point prediction for each scenario, but for the ensemble models, we use the median value of the resulting prediction as the point estimate.

#### 2.2.3. Weighted interval score

The third measure we have considered is the weighted interval score (WIS) which 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. Lower WIS values indicate better forecast performance. The WIS is calculated for each time point with prediction $P$ and observed values $y$ as

$$\mathrm{WIS}_{\alpha_{0:K}}(P,y) = \frac{1}{K+0.5}\left(w_0|y-m| + \sum_{k=1}^{K} w_k \mathrm{IS}_{\alpha_k}(P,y)\right),$$

where $\mathrm{IS}_\alpha$ corresponds to the interval score of the $(1-\alpha)\times 100\%$ PI,

$$\mathrm{IS}_\alpha(P,y) = (u_\alpha - l_\alpha) + \frac{2}{\alpha}(l_\alpha - y)\mathbb{1}(y < l_\alpha) + \frac{2}{\alpha}(y - u_\alpha)\mathbb{1}(y > u_\alpha),$$

$K$ is the number of PIs used (in our case, 11), $m$ is the median of $P$, and $l_\alpha$ ($u_\alpha$) is the lower bound (upper bound) of the PI. The standard weights are chosen such that $w_k = \frac{\alpha_k}{2}$ and $w_0 = \frac{1}{2}$. The indicator function $\mathbb{1}$ in the interval score is used to penalize observations that lie outside the PI and the term $(u_\alpha - l_\alpha)$ penalizes wider intervals (Bracher et al., 2021).

When aggregating WIS scores from all weeks in an SMH round, we take the mean over the values for each week to get one average WIS value for a given SMH round, target and location. We define for each model $m$, location $\ell$, target $t$, and round $r$ the average WIS

$$\mathrm{WIS}^m_{\ell,t,r} = \frac{\sum_w \mathrm{WIS}^m_{\ell,t,r,w}}{N_w}$$

where $\mathrm{WIS}^m_{\ell,t,r,w}$ is the WIS for a given model, location, target, round, and week and $N_w$ is the number of weeks in the projection round. We discard cases where the calculated WIS score is less than zero, which occurred in a few cases where the prediction intervals reported by modeling teams included a lower bound that was greater than the upper bound.

In our analysis, we compare and aggregate the WIS across different prediction locations and scenario rounds. The WIS is an absolute measure and therefore it depends on the magnitude of the observation and prediction values that changes across locations, rounds, and weeks within a round. For instance, the various targets (death, hospitalization, cases) are affected by the size of the population of each state, and so is the WIS. Similarly, during different rounds, the magnitude of the targets may change because of the different phases of the epidemic. Because of this, we either compare pairs of models through a ratio of their average WIS, or we generate a rescaled weighted interval score $\mathrm{WIS}_{\mathrm{rescaled}}$. To compute the latter, we calculate the standard deviation of the WIS across all models reported for the corresponding $\ell, t, r, w$, i.e.,

$$\sigma_{\ell,t,r,w} = \sqrt{\frac{\sum_m \left(\mathrm{WIS}^m_{\ell,t,r,w} - \overline{\mathrm{WIS}}_{\ell,t,r,w}\right)^2}{N_m}},$$

where $N_m$ is the number of reported models, $\mathrm{WIS}^m_{\ell,t,r,w}$ is the WIS for a given model, location, target, round, and week, and $\overline{\mathrm{WIS}}_{\ell,t,r,w}$ is the average of $\mathrm{WIS}^m_{\ell,t,r,w}$ over all models. To rescale this metric, we divide the WIS value of a given model by this calculated standard deviation (Howerton et al., 2023b),

$$\mathrm{WIS}_{\mathrm{rescaled}} = \frac{\mathrm{WIS}_{\alpha_{0:K}}(P,y)}{\sigma_{\ell,t,r,w}}.$$

We aggregate rescaled WIS scores using the same method as discussed above, whereas we take the average over all weeks in a projection round to get one rescaled WIS value for a given model, location, target, and round. This rescaling, adopted in climate prediction (Pennell and Reichler, 2011), accommodates the specific scale of each week, target, and round, thereby enabling a fair comparison of the WIS. In the Supplementary Material (section 3), we detail the outcomes obtained from an alternative rescaling method. This alternate approach characterizes a relative WIS, adjusted each week according to the target magnitude. This revised definition assigns equivalent significance to identical relative deviations, irrespective of the target's magnitude. However, it is worth noting that this adjustment does not significantly alter the overall performance evaluation of the models.

## 3. Results

To illustrate the outcomes achieved through the scenario ensemble procedure, we begin by detailing the results of a single scenario round. This step facilitates a comparative analysis between the Ensemble[2] and the scenario ensemble drawn from individual models. Following this, we provide a comprehensive evaluation of the results obtained across ten different projection rounds. This thorough analysis is aimed at assessing the performance variations under distinct scenario designs and during different phases of the epidemic. This analysis not only provides insight into the overall efficacy of the scenario ensembling procedure, but it also aids in determining its robustness and adaptability under changing epidemic phases and scenarios. In the following we report the analysis aggregated over all the targets (deaths, hospitalizations, and cases), however, we report the results specific to each target in Figs. S13–S18 of the Supplementary Material.

### 3.1. Single round analysis

In this section, we delve into the specifics of round 12, which saw contributions from six modeling teams providing three-month projections spanning January 15, 2022, to April 2, 2022. The main objective of this round was to assess the implications of the Omicron wave, by analyzing four scenarios that varied along two dimensions:

- The extent of immune evasion by the Omicron variant, which was assumed to increase the risk of infection among those with prior immunity to SARS-CoV-2 by 50% to 80%.
- The severity of the variant in terms of its impact on hospitalization and death rates, with the risk speculated to reduce by 30% to 70% compared to the Delta variant.

More detailed information on the round 12 projections can be found in the publicly available COVID-19 SMH's Github repository.

In the left panel of Fig. 2, we show the coverage of the scenario ensembles of all individual and ensemble models, consolidated across all targets and all states in the US. It is apparent that all models display overconfidence in their predictions – their PIs are excessively narrow – with the exception of the Ensemble_LOP[2] and the Ensemble_LOP_untrimmed[2]. In the right panel of Fig. 2, we analyze the coverage of the Ensemble_LOP[2] for a few selected states and see that while it occasionally exhibits slight overconfidence and at other times slight underconfidence, the model is generally well-calibrated (see Fig. S1 in the Supplementary Material for the coverage in all states).

In Fig. 3A, we show the distribution of the MAPE across states, including all targets, of all scenario ensemble models for round 12. The median of the MAPE suggests that the three Ensemble[2] models perform better than the scenario ensemble of the individual models.

In Fig. 3B, we show the distribution of the rescaled WIS across states, including all targets, for all scenario ensemble models in round 12. The median of the rescaled WIS suggests that the Ensemble[2] models outperform the scenario ensembles of the individual models. Similar results are obtained in the Supplementary Material (Fig. S6–S7) with another rescaling procedure where we divide the WIS by the observation. In Table 1, we summarize aggregate results for all measures performed on round 12. The three distinct projections derived from the Ensemble[2] demonstrate superior performance in comparison to the scenario ensemble of individual models. This confirms the efficacy of the Ensemble[2] approach, which offers not only enhanced accuracy, but also improved estimation of uncertainty than what is typically achieved with single models.

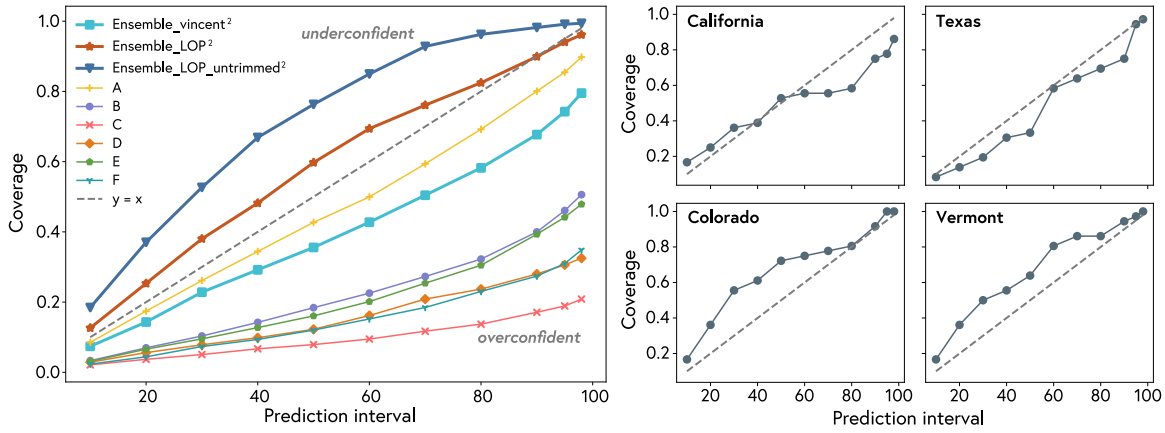


**Fig. 2.** (left) Coverage over different prediction intervals for the scenario ensemble predictions for all targets (cases, deaths, and hospitalizations) and all US states during round 12 for each model. Each data point shows the coverage value over all prediction time points, locations, and targets. A to F correspond to the scenario ensemble of individual models and we only present the scenario ensemble of models providing projections in all locations. (right) Coverage over different prediction intervals for the Ensemble_LOP² for round 12 and for all targets (cases, deaths, and hospitalizations) in 4 different US states, ordered by descending population size. The dotted line represents the ideal case where the coverage is perfectly matching the expectation from the prediction interval. Coverages below (above) this line represent overconfidence (underconfidence).

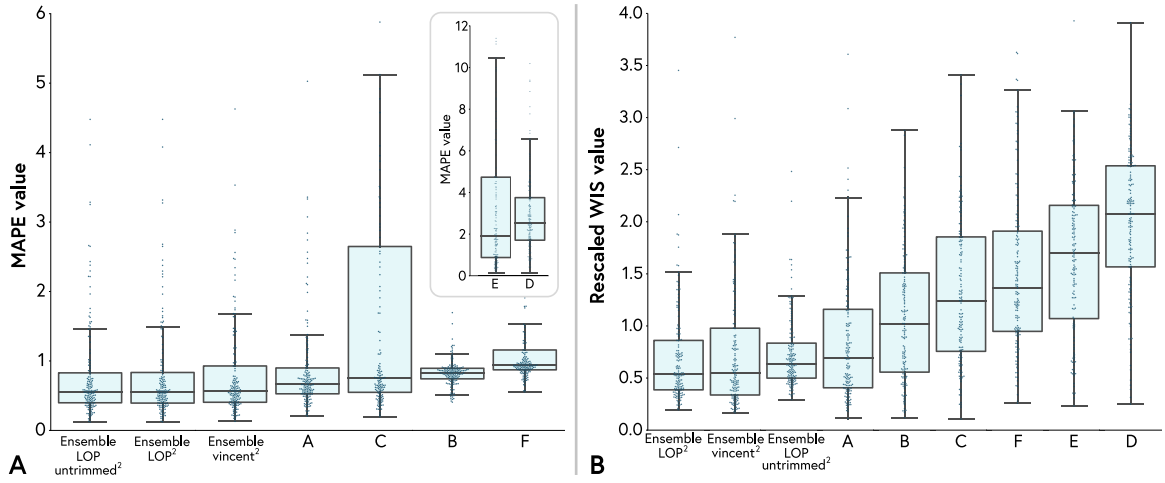


**Fig. 3.** Distribution of (A) the MAPE scores and (B) the WIS scores rescaled by the standard deviation and averaged over all weeks for each scenario ensemble model for round 12. Each point in the scatter plot refers to a specific target (cases, deaths, and hospitalizations) and location (US states), ordered by median MAPE/WIS value. We only present the scenario ensemble of models providing projections in all locations. For the sake of visualization, outliers above the (A) 90% and (B) 99% quantile are not shown.

**Table 1**
Table showing the median (25%, 75% percentile) averaged and rescaled WIS score and MAPE, and the 50%, and 95% coverage values over all targets for the scenario ensemble of each model averaged over all locations (US states) for round 12. The rescaled WIS divides the raw WIS score by the standard deviation across all models at each time point, and is averaged over all weeks in round 12. The boldface values represent the smallest WIS and MAPE values, which corresponds to the best prediction of the observed values, and the bold coverage represents the values closest to the associated prediction interval.

| Model | WIS | MAPE | 50% coverage | 95% coverage |
|---|---|---|---|---|
| Ensemble_vincent² | 0.55 (0.34, 0.98) | 0.57 (0.41, 0.93) | 0.36 | 0.74 |
| Ensemble_LOP² | **0.53 (0.39, 0.86)** | 0.55 (0.39, 0.83) | 0.60 | **0.94** |
| Ensemble_LOP_untrimmed² | 0.64 (0.50, 0.84) | **0.55 (0.40, 0.83)** | 0.76 | 0.99 |
| A | 0.69 (0.41, 1.16) | 0.66 (0.53, 0.90) | **0.43** | 0.86 |
| B | 1.02 (0.56, 1.51) | 0.83 (0.74, 0.89) | 0.18 | 0.46 |
| C | 1.24 (0.76, 1.85) | 0.76 (0.55, 2.65) | 0.08 | 0.19 |
| D | 2.08 (1.57, 2.54) | 2.53 (1.71, 3.75) | 0.12 | 0.31 |
| E | 1.70 (1.07, 2.16) | 1.91 (0.88, 4.74) | 0.16 | 0.44 |
| F | 1.36 (0.95, 1.91) | 0.94 (0.87, 1.16) | 0.12 | 0.31 |

### *3.2. Overall performance assessment of the Ensemble²*

In order to provide a thorough performance assessment of the scenario ensemble approach we considered the projection results from rounds 5 to 16, excluding rounds 8 and 10 because they were internal training rounds, for a total of 10 rounds. We excluded rounds 1 to 4 because the Ensemble_LOP was not reported. To make a fair assessment of these projections, we excluded dates from the projection period when a new variant emerged and diverged considerably from the scenarios and their assumptions (see Howerton et al., 2023b and Table S1 in the Supplementary Material). See Table S2 in the Supplementary Material for a list of models included in our analysis for each round.

We focus our analysis on the three Ensemble² models and investigate the performance across rounds and geographic locations. In Fig. 4

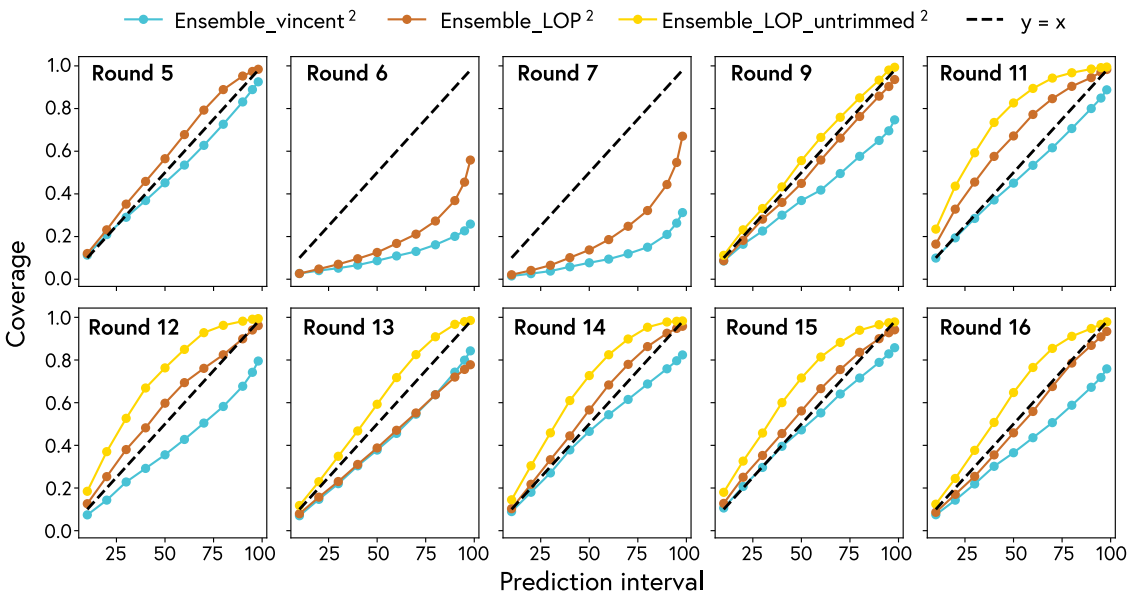


**Fig. 4.** Coverage versus prediction intervals for the three Ensemble[2] models for each round of Scenario Modeling Hub projections. The coverage is taken over all targets (cases, deaths, and hospitalizations) and locations (US states). The Ensemble_LOP model started being reported in round 5, and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

we observe that the coverage is fairly good for all Ensemble[2] models. 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. Round 6 was the first to explicitly account for the spreading of the Delta variant. Round 7 was an update of round 6 taking into consideration updated data on the Delta variant and the vaccine hesitancy. The Delta variant was assumed 20% to 60% more transmissible than the Alpha variant in round 6, and 40% to 60% more transmissible in round 7. However, for both rounds, even the more pessimistic scenarios, which consider that the Delta variant has a high increase in transmissibility, led to an underestimation of the incident cases, hospitalizations, and deaths for all models (Scenario Modeling Hub, 2023). This explains the overconfidence of the Ensemble[2] models for these rounds in Fig. 4.

To assess the overall performance of Ensemble[2] models, we show in Fig. 5 the distribution of the MAPE and WIS across different states for each of the considered projection rounds (here only for the Ensemble_LOP[2]). The WIS and MAPE performance show a similar behavior across rounds with rounds 6 and 7 having the highest median values, and thus a lower quality for the projections than in the other rounds. This feature can again be attributed to the underestimation of the targets following the emergence of the Delta variant. See Figs. S3 and S5 in the Supplementary Material for a similar analysis with the Ensemble_vincent[2] model.

Figs. 4 and 5 are good examples of the Ensemble[2] models being able to flag anomalies in the round 6 and 7 projections. While the transmissibility and the vaccine assumptions (round 6 only) seem to correctly bound reality (Howerton et al., 2023b), the models have underestimated the targets. There are multiple factors that could explain this: a more rapid waning of vaccine protection, other epidemiological differences of the Delta variant not taken into account, and changes in human behavior (Howerton et al., 2023b). While this can be in part attributed to a miscalibration from the modeling teams, some of these elements could also fall within the scope of scenario design, even though they were not the drivers of interest at the time. Altogether, the discussion among modeling teams and the coordination team in the light of the Ensemble[2] results presents a valuable chance to enhance both the model implementation and scenario design processes.

In order to compare the performance of the Ensemble[2] models with the scenario ensemble of single models, we analyzed the distribution of the standardized rank across all considered projection rounds according to the WIS. The standardized rank 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. Remarkably, we see in Fig. 6 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 finding corroborates the results found in several studies: ensemble models are overall better calibrated and performing than individual models (Bates and Granger, 1969; Krishnamurti et al., 1999; Viboud et al., 2018; McGowan et al., 2019; Reich et al., 2019b,a; Johansson et al., 2019; Cramer et al., 2022a). See Fig. S4 in the Supplementary Material for a similar analysis using the MAPE.

The performance analyses conducted thus far are relative, comparing across different models and rounds. To have a more absolute assessment of the projection quality, it is necessary to establish a reference point for comparison, which can serve as a minimum performance standard or highlight improvements over a recognized state-of-the-art approach. In the context of scenario projections, identifying an appropriate reference point is a challenging task. Therefore, we opted to use as reference two well-regarded models in forecasting generated by the COVID-19 Forecast Hub, a platform that aggregates and visualizes COVID-19 forecasts from various predictive models (Cramer et al., 2022b; Forecast Hub, 2023). The first is a naive baseline forecast, where the median of the prediction mirrors the last observed value (here we use the four-week-ahead forecast) and the uncertainty is based on previous changes in the weekly incidence. The second is the four-week-ahead ensemble forecast that aggregates the predictions of the modeling teams.

In Fig. 7 we compare the overall performance of the Ensemble[2] models with the four-week-ahead baseline and ensemble forecast for the predicted incident deaths. For each round and location, we calculate the ratio of the raw WIS score for a given Ensemble[2] model divided by the raw WIS score obtained by the reference model. A ratio smaller than 1 indicates that the Ensemble[2] model provided better predictions. We see that generally, the Ensemble[2] models perform better than the naive baseline with a median ratio between 0.87 and 0.94. However, the four-week-ahead ensemble forecast is generally better than the Ensemble[2] models, with a median ratio between 1.36 and 1.42. See Fig. S2 in the Supplementary Material for a comparison of the coverage with these reference models and Figs. S8–S10 for a comparison of the WIS by

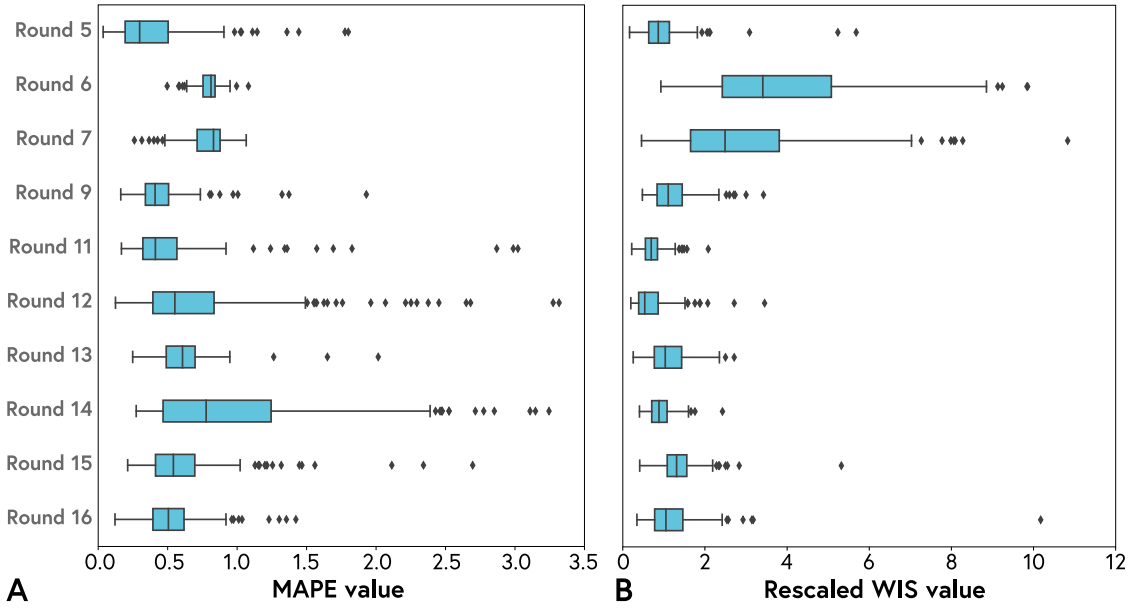


**Fig. 5.** (A) Distribution of MAPE values for the Ensemble_LOP$^2$ model for each round, over all targets and locations. (B) Distribution of averaged, rescaled WIS values for the Ensemble_LOP$^2$ model for each round, for all targets and locations. For the sake of visualization, we do not show outliers above the 98% quantile.

round. It is crucial to recognize that neither the naive baseline forecast nor the four-week-ahead ensemble should be regarded as the minimum performance benchmark for scenario projections. Scenario projections are typically formulated for a timeframe of three to six months, whereas both the baseline and the Forecast Hub ensemble are updated on a weekly basis, incorporating ongoing information about the trajectory of the epidemic.

In Fig. 8, we compare the performance of the Ensemble_LOP$^2$ model against the Ensemble_LOP of individual scenarios for all targets and locations for each projection round. A value less than 1 indicates that the Ensemble$^2$ model gives better predictions, and a value greater than 1 suggests that the projection for the individual scenario provides better predictions than the scenario ensemble. We observe that, overall, the Ensemble$^2$ model consistently performs better than the poorest performing scenario in each round and is competitive against the best, as indicated by the median WIS ratio falling between 0.76 and 1.25 for individual scenarios in each SMH round. When compared to only the individual scenarios with the highest WIS ratio (best-performing scenarios) in each projection round, the WIS ratio falls between 0.98 to 1.25. It is also worth noting that in three rounds, the Ensemble$^2$ model performs better than all other scenarios (median WIS ratio smaller than one), illustrating the ability of the Ensemble$^2$ models to interpolate between scenarios, potentially providing a better fit to reality.

This analysis highlights the effectiveness of the Ensemble$^2$ approach in generating reliable projections prior to utilizing ground truth data to evaluate the plausibility of scenarios. Essentially, the Ensemble$^2$ approach captures both the inherent uncertainty in model design and the uncertainty inherent in scenario design, prior to any posterior considerations.

## 4. Discussion

The task of outlining an appropriate framework for the performance assessment of epidemic scenario projections carries substantial implications for policy-making, long-term planning, and the evaluation of scenario designs. Our analysis underscores the merits of a scenario ensemble procedure in evaluating the performance of scenario modeling. The coverage of the scenario ensemble quantifies to which extent the defined scenarios as a whole encompass the future trajectory of the epidemic, conditional on the appropriate models' calibration. This aspect is particularly pertinent in assessing the quality of scenario projections in relation to policy-making, which necessitates accurate estimation of both upper and lower bounds of plausible outcomes. Through the examination of 10 rounds of SMH scenario projections, we found that the scenario ensemble typically yields well-calibrated projections capable of enveloping epidemic trajectories, even when individual models or scenarios fall short. Furthermore, 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. This deficiency implies that the range of scenarios under consideration does not comprehensively encompass the possible epidemic trajectories, suggesting that other epidemiological aspects warrant revision by the coordination and modeling teams. In other words, the performance of a scenario ensemble can provide valuable feedback to both the scenario design team and the modeling teams. It can guide adjustments to the scenario specifications, leading to a more comprehensive representation of potential epidemic paths. Simultaneously, it can prompt a critical reevaluation of the models' underlying assumptions, helping to refine and improve their predictive power.

The scenario ensemble approach also fully acknowledges that future epidemic developments should be viewed as a continuum of potential scenarios, with interpolations occurring between the specific ones. Every individual scenario is intended to explore plausible future outcomes under different assumptions that may never fully materialize. The information provided by each scenario is valuable to the policy-making process, but their comparison with ground truth data as if they would be forecasts does not adequately capture their true value. While this issue may be assuaged with the a-posteriori identification of the most plausible scenario to compare with the ground truth data, the scenario ensemble organically considers the ability of multiple scenarios to span a wide range of plausible outcomes. This approach has also the advantage of not relying on the a-posteriori identification of the most plausible scenarios that are generally dependent on the geographical scale, the time window, and may be clouded by additional modeling and parameter assumptions. Furthermore, the quantity of scenarios significantly influences the degree of alignment between the most plausible scenario and reality. It is crucial to identify the optimal number of scenarios required for optimum results. Also, considering the interdependencies among them, the effective number of scenarios employed warrants careful evaluation. The scenario ensembling approach

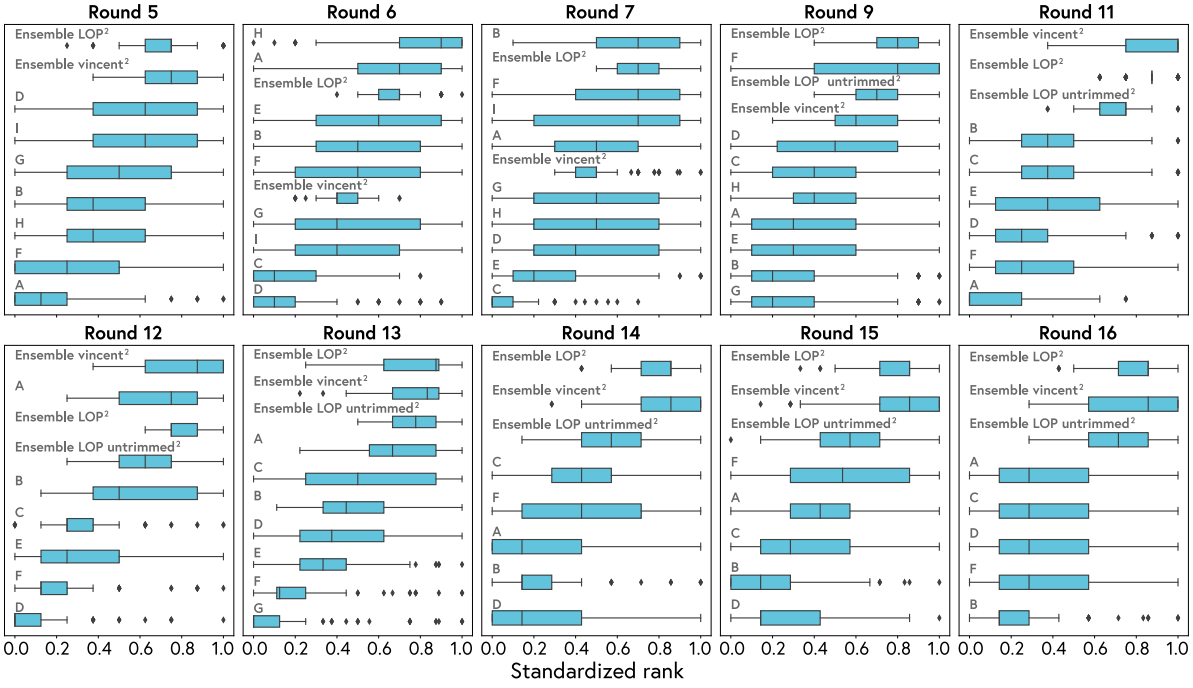


**Fig. 6.** Distribution of standardized rank values for the raw WIS value averaged over all weeks of the corresponding projection round of each of the scenario ensemble models included in each round. WIS values are ranked such that the model with the smallest WIS value for a given target and location is ranked as 1, and so on. These values are then normalized such that the rankings are between 0 and 1, with a larger standardized rank corresponding to a model with a better prediction. The standardized ranks for each round are ordered by median value. These distributions are shown for ranking models over all targets and locations. We only present the scenario ensemble of models providing projections in all locations.

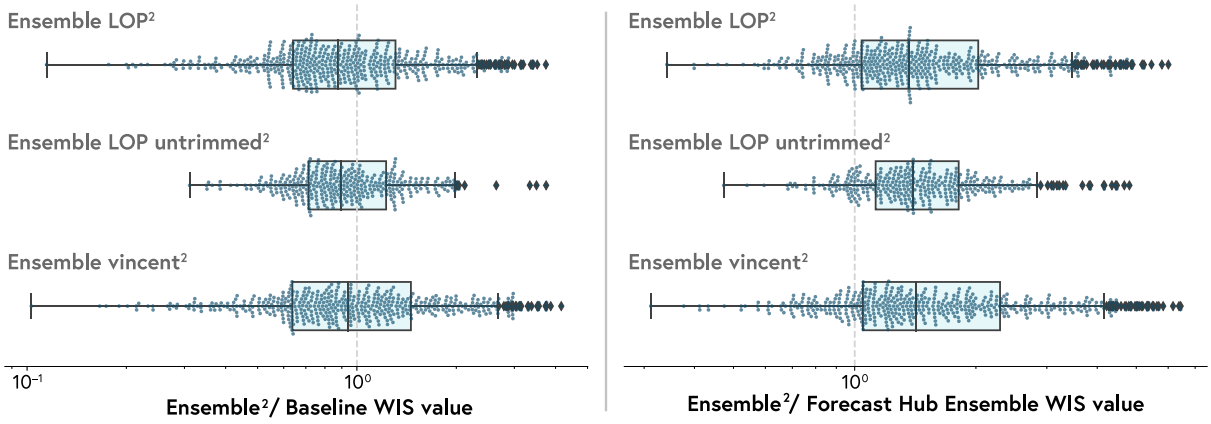


**Fig. 7.** Average WIS value of Ensemble² models divided by the average WIS value of a reference model for the prediction of incident deaths for all projection rounds and locations. A ratio greater (smaller) than 1 indicates that the reference model (Ensemble² model) performs better. Each data point represents the WIS ratio for a round of the Scenario Modeling Hub at a particular US state. Rounds with WIS < 0 are discarded. (A) The reference is the COVID-19 Forecast Hub naive baseline model (four-week-ahead prediction). (B) The reference is the COVID-19 Forecast Hub ensemble model (four-week-ahead prediction).

represents progress towards the creation of a framework that can also help to refine our scenario generation and selection processes, ensuring we capture a truly representative spectrum of potential outcomes.

Our findings indicate that even in its simplest implementation, the scenario ensemble approach demonstrates comparable or superior performance to four-week-ahead forecast models from the COVID-19 Forecast Hub (ensemble and baseline models) in terms of fundamental metrics such as MAPE and WIS. Moreover, the scenario ensemble methodology paves the way for more sophisticated strategies that dynamically adjust scenario weights over time, enabling the ensemble to evolve (Raftery et al., 2005; Ray and Reich, 2018). This evolution could be guided either by empirical evidence derived from fits to historical data (Johnson et al., 2015) when the scenarios are designed or by initiating a process of expert elicitation for differently formed scenarios. Additionally, mechanisms can be devised to detect when the scenario ensemble ceases to bracket reality (Runge et al., 2016), regularly checking the goodness-of-fit of the weighted ensemble to determine when the scenario design process – potentially beyond the drivers initially identified – needs revisiting.

Nevertheless, it is important to acknowledge the inherent limitations of any quantitative performance assessment of scenario projections. The utility of scenario projections for decision-making should be a primary criterion for their evaluation. The question should be: do they assist policy-makers in comprehending the potential range of outcomes and the impacts of varying interventions? Furthermore, the long-term nature of the scenario projections implies that their performance assessment is not solely about appraising the accuracy of the models employed. It also involves assessing the performance of

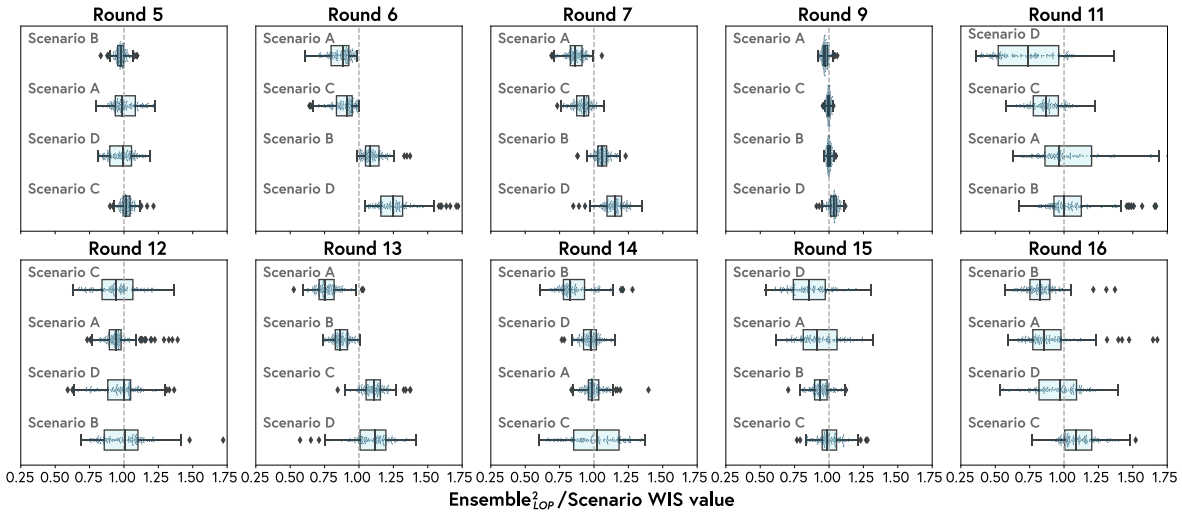


**Fig. 8.** Distribution of the ratio of the average WIS of the Ensemble_LOP$^2$ and the Ensemble_LOP of each specific scenario for all targets and locations for each projection round of SMH. Vertical lines at a ratio of one show where the performance between the Ensemble_LOP$^2$ and a scenario is the same. For values less than one, the Ensemble_LOP$^2$ performs better, whereas for values greater than one the individual scenario has a better WIS score. For the sake of visualization, we exclude outliers above the 95% quantile.

the scenario design itself—i.e., the assumptions and conditions under which the models function. While comparisons with ground truth data form part of performance assessment, they are not the sole or even the primary metric for evaluating the performance of scenario projections. For instance, contrasting scenario projections (e.g., percentage of cases or deaths averted) under different assumptions is a metric that remains useful for policy decisions even for projections that do not follow accurately the epidemic trajectory. A more comprehensive, nuanced approach is essential to thoroughly understand and appreciate the value of these tools in epidemic modeling and public health decision-making.

Let us also stress that while in this work we focused on scenario projections aimed at bounding the realized epidemic drivers, such collaborative efforts might have other goals (Runge et al., 2023). For instance, some scenario designs might include counterfactuals that are not expected to happen but serve to compare and contrast the results of different policy-making decisions. These counterfactual scenarios should be excluded from the ensembling procedure presented here, or appropriately weighted prior to their inclusion.

To summarize, we suggest that the scenario ensemble procedure offers a synthesis of potential epidemic outcomes, compensating for the uncertainties and limitations inherent in individual scenarios. This approach could be especially useful for planning purposes (e.g., determining how many beds or treatment courses are needed over the coming months), so policymakers do not need to analyze and interpret multiple distinct scenario projections. 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.

## CRediT authorship contribution statement

**Clara Bay:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Guillaume St-Onge:** Conceptualization, Formal analysis, Investigation, Methodology, Software, Supervision, Validation, Writing – original draft, Writing – review & editing. **Jessica T. Davis:** Investigation, Methodology, Software, Writing – review & editing. **Matteo Chinazzi:** Investigation, Writing – review & editing. **Emily Howerton:** Investigation, Methodology, Writing – review & editing. **Justin Lessler:** Investigation, Writing – review & editing. **Michael C. Runge:** Investigation, Writing – review & editing. **Katriona Shea:** Investigation, Writing – review & editing. **Shaun Truelove:** Investigation, Writing – review & editing. **Cecile Viboud:** Investigation, Writing – review & editing. **Alessandro Vespignani:** Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Supervision, Writing – original draft, Writing – review & editing.

## Declaration of competing interest

None.

## Data and code availability

As ground truth data, we use the weekly incident deaths and reported cases from the JHU CSSE group (Dong et al., 2020, 2022) and the weekly incident hospitalizations from HealthData.gov (2023). We use more specifically the formatted version provided by the COVID-19 Forecast Hub (2023). All models' projections we use are made available by the Scenario Modeling Hub (2023). Our code for the project is publicly available on Zenodo (Bay and St-Onge, 2023).

## Acknowledgments

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 Québec – Nature et technologies (project 313475). MC additionally acknowledges support from CDC-JHU-2005702123. EH and KS acknowledge support from NSF RAPID awards DEB-2028301, DEB-2037885, DEB-2126278 and DEB-2220903, and EH additionally acknowledges support from the Eberly College of Science Barbara McClintock Science Achievement Graduate Scholarship in Biology at the Pennsylvania State University. The findings and conclusions in this study are those of the authors and do not necessarily represent the official position of the funding agencies, the National Institutes of Health, or the U.S. Department of Health and Human Services. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

## Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.epidem.2024.100748.

## References


Bates, J.M., Granger, C.W.J., 1969. The combination of forecasts. J. Oper. Res. Soc. 20 (4), 451–468. http://dx.doi.org/10.1057/jors.1969.103.

Bay, C., St-Onge, G., 2023. gstonge/ensemble-square. http://dx.doi.org/10.5281/zenodo.10309030.

Biggerstaff, M., Slayton, R.B., Johansson, M.A., Butler, J.C., 2022. Improving pandemic response: Employing mathematical modeling to confront coronavirus disease 2019. Clin. Infect. Dis. 74 (5), 913–917. http://dx.doi.org/10.1093/cid/ciab673.

Biggerstaff, M., et al., 2018. Results from the second year of a collaborative effort to forecast influenza seasons in the United States. Epidemics 24, 26–33. http://dx.doi.org/10.1016/j.epidem.2018.02.003.

Borchering, R.K., et al., 2021. Modeling of future COVID-19 cases, hospitalizations, and deaths, by vaccination rates and nonpharmaceutical intervention scenarios — United States, april–september 2021. Morb. Mortal. Wkly. Rep. 70 (19), 719–724. http://dx.doi.org/10.15585/mmwr.mm7019e3.

Borchering, R.K., et al., 2023. Impact of SARS-CoV-2 vaccination of children ages 5–11 years on COVID-19 disease burden and resilience to new variants in the United States, november 2021–march 2022: a multi-model study. Lancet Reg. Health - Am. 17, 100398. http://dx.doi.org/10.1016/j.lana.2022.100398.

Bracher, J., Ray, E.L., Gneiting, T., Reich, N.G., 2021. Evaluating epidemic forecasts in an interval format. PLoS Comput. Biol. 17 (2), 1–15. http://dx.doi.org/10.1371/journal.pcbi.1008618.

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. USA 119 (15), e2113561119. http://dx.doi.org/10.1073/pnas.2113561119.

Cramer, E.Y., et al., 2022b. The United States COVID-19 forecast hub dataset. Sci. Data 9 (1), 462. http://dx.doi.org/10.1038/s41597-022-01517-w.

Dong, E., Du, H., Gardner, L., 2020. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect. Dis. 20 (5), 533–534. http://dx.doi.org/10.1016/S1473-3099(20)30120-1.

Dong, E., et al., 2022. The Johns Hopkins university center for systems science and engineering COVID-19 dashboard: data collection process, challenges faced, and lessons learned. Lancet Inf. Dis. 22 (12), e370–e376. http://dx.doi.org/10.1016/S1473-3099(22)00434-0.

Forecast Hub, 2023. COVID-19 Forecast Hub. Accessed: 2023-05-17, https://covid19forecasthub.org/.

HealthData.gov, 2023. COVID-19 Reported Patient Impact and Hospital Capacity by State Timeseries. Accessed: 2023-05-17, https://healthdata.gov/Hospital/COVID-19-Reported-Patient-Impact-and-Hospital-Capa/g62h-syeh.

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 (198), 20220659. http://dx.doi.org/10.1098/rsif.2022.0659.

Howerton, E., et al., 2023b. Evaluation of the US COVID-19 scenario modeling hub for informing pandemic response under uncertainty. Nature Commun. 14 (1), 7260. http://dx.doi.org/10.1038/s41467-023-42680-x.

Jewell, N.P., Lewnard, J.A., Jewell, B.L., 2020. Predictive mathematical models of the COVID-19 pandemic: Underlying principles and value of projections. JAMA 323 (19), 1893–1894. http://dx.doi.org/10.1001/jama.2020.6585.

Johansson, M.A., et al., 2019. An open challenge to advance probabilistic forecasting for dengue epidemics. Proc. Natl. Acad. Sci. USA 116 (48), 24268–24274. http://dx.doi.org/10.1073/pnas.1909865116.

Johnson, F.A., Boomer, G.S., Williams, B.K., Nichols, J.D., Case, D.J., 2015. Multilevel learning in the adaptive management of waterfowl harvests: 20 years and counting. Wildl. Soc. Bull. 39 (1), 9–19. http://dx.doi.org/10.1002/wsb.518.

Krishnamurti, T.N., et al., 1999. Improved weather and seasonal climate forecasts from multimodel superensemble. Science 285 (5433), 1548–1550. http://dx.doi.org/10.1126/science.285.5433.1548.

Lutz, C.S., et al., 2019. Applying infectious disease forecasting to public health: a path forward using influenza forecasting examples. BMC Public Health 19 (1), 1–12. http://dx.doi.org/10.1186/s12889-019-7966-8.

Makridakis, S., et al., 1982. The accuracy of extrapolation (time series) methods: Results of a forecasting competition. J. Forecast. 1 (2), 111–153. http://dx.doi.org/10.1002/for.3980010202.

McGowan, C.J., et al., 2019. Collaborative efforts to forecast seasonal influenza in the United States, 2015–2016. Sci. Rep. 9 (1), 683. http://dx.doi.org/10.1038/s41598-018-36361-9.

Pennell, C., Reichler, T., 2011. On the effective number of climate models. J. Clim. 24 (9), 2358–2367. http://dx.doi.org/10.1175/2010JCLI3814.1, URL https://journals.ametsoc.org/view/journals/clim/24/9/2010jcli3814.1.xml.

Raftery, A.E., Gneiting, T., Balabdaoui, F., Polakowski, M., 2005. Using Bayesian model averaging to calibrate forecast ensembles. Mon. Weather Rev. 133 (5), 1155–1174. http://dx.doi.org/10.1175/MWR2906.1, URL https://journals.ametsoc.org/view/journals/mwre/133/5/mwr2906.1.xml.

Ray, E.L., Reich, N.G., 2018. Prediction of infectious disease epidemics via weighted density ensembles. PLoS Comput. Biol. 14 (2), 1–23. http://dx.doi.org/10.1371/journal.pcbi.1005910.

Reich, N.G., et al., 2019a. Accuracy of real-time multi-model ensemble forecasts for seasonal influenza in the U.S.. PLoS Comput. Biol. 15 (11), 1–19. http://dx.doi.org/10.1371/journal.pcbi.1007486.

Reich, N.G., et al., 2019b. A collaborative multiyear, multimodel assessment of seasonal influenza forecasting in the United States. Proc. Natl. Acad. Sci. USA 116 (8), 3146–3154. http://dx.doi.org/10.1073/pnas.1812594116, URL https://www.pnas.org/doi/abs/10.1073/pnas.1812594116.

Reich, N.G., et al., 2022. Collaborative hubs: Making the most of predictive epidemic modeling. Am. J. Public Health 112 (6), 839–842. http://dx.doi.org/10.2105/AJPH.2022.306831.

Rosenblum, H.G., et al., 2022. Interim recommendations from the advisory committee on immunization practices for the use of bivalent booster doses of COVID-19 vaccines — United States, october 2022. Morb. Mortal. Wkly. Rep. 71 (45), 1436–1441. http://dx.doi.org/10.15585/mmwr.mm7145a2.

Runge, M.C., Stroeve, J.C., Barrett, A.P., McDonald-Madden, E., 2016. Detecting failure of climate predictions. Nature Clim. Change 6 (9), 861–864. http://dx.doi.org/10.1038/nclimate3041.

Runge, M.C., et al., 2023. Scenario design for infection disease projections: Integrating concepts from decision analysis and experimental design. Epidemics In review.

Scenario Modeling Hub, 2023. COVID-19 Scenario Modeling Hub. Accessed: 2023-05-17, https://covid19scenariomodelinghub.org/.

Stone, M., 1961. The opinion pool. Ann. Math. Stat. 32, 1339–1342.

Truelove, S., et al., 2022. Projected resurgence of COVID-19 in the United States in july—december 2021 resulting from the increased transmissibility of the Delta variant and faltering vaccination. eLife 11, e73584. http://dx.doi.org/10.7554/eLife.73584.

Viboud, C., et al., 2018. The RAPIDD ebola forecasting challenge: Synthesis and lessons learnt. Epidemics 22, 13–21. http://dx.doi.org/10.1016/j.epidem.2017.08.002.

Vincent, S.B., 1912. The Functions of the Vibrissae in the Behavior of the White Rat. (Ph.D. thesis). University of Chicago.

Vollmar, H.C., Ostermann, T., Redaèlli, M., 2015. Using the scenario method in the context of health and health care – a scoping review. BMC Med. Res. Methodol. 15 (1), 89. http://dx.doi.org/10.1186/s12874-015-0083-1.