Epidemics 47 (2024) 100767

Contents lists available at ScienceDirect

Epidemics

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

When do we need multiple infectious disease models? Agreement between projection rank and magnitude in a multi-model setting

La Keisha Wade-Malone [a,1], Emily Howerton [a,*,1], William J.M. Probert [b], Michael C. Runge [c], Cécile Viboud [d], Katriona Shea [a]

[a] Department of Biology and Center for Infectious Disease Dynamics, The Pennsylvania State University, University Park, PA, USA
[b] Big Data Institute, University of Oxford, UK
[c] US Geological Survey, Eastern Ecological Science Center at the Patuxent Research Refuge, Laurel, MD, USA
[d] Fogarty International Center, National Institutes of Health, Bethesda, MD, USA

ARTICLE INFO

Keywords:
Multi-model predictions
Outbreak control
Decision support

ABSTRACT

Mathematical models are useful for public health planning and response to infectious disease threats. However, different models can provide differing results, which can hamper decision making if not synthesized appropriately. To address this challenge, multi-model hubs convene independent modeling groups to generate ensembles, known to provide more accurate predictions of future outcomes. Yet, these hubs are resource intensive, and how many models are sufficient in a hub is not known. Here, we compare the benefit of predictions from multiple models in different contexts: (1) decision settings that depend on predictions of quantitative outcomes (e.g., hospital capacity planning), where assessments of the benefits of multi-model ensembles have largely focused; and (2) decisions settings that require the ranking of alternative epidemic scenarios (e.g., comparing outcomes under multiple possible interventions and biological uncertainties). We develop a mathematical framework to mimic a multi-model prediction setting, and use this framework to quantify how frequently predictions from different models agree. We further explore multi-model agreement using real-world, empirical data from 14 rounds of U.S. COVID-19 Scenario Modeling Hub projections. Our results suggest that the value of multiple models could be different in different decision contexts, and if only a few models are available, focusing on the rank of alternative epidemic scenarios could be more robust than focusing on quantitative outcomes. Although additional exploration of the sufficient number of models for different contexts is still needed, our results indicate that it may be possible to identify decision contexts where it is robust to rely on fewer models, a finding that can inform the use of modeling resources during future public health crises.

1. Introduction

Policy makers are increasingly leveraging mathematical models to support public health planning and outbreak response (Biggerstaff et al., 2022; Egger et al., 2018). Models can be used to estimate key biological parameters, predict what will happen in the future, or anticipate the effectiveness of potential intervention strategies (Biggerstaff et al., 2022; Metcalf et al., 2020). However, during an infectious disease outbreak, there are many unknowns (e.g., about pathogen biology or human behavior), and these unknowns can hamper decision making and public health response. For example, uncertainties can lead to differing models that offer conflicting results and contradictory policy recommendations (den Boon et al., 2019; Reich et al., 2022).

In the face of uncertainty, approaches that leverage multiple models can be used to better support decision making (den Boon et al., 2019; Reich et al., 2022; Shea et al., 2020). In particular, aggregating results from multiple models into an ensemble has been shown to provide more robust and reliable estimates of future outcomes across a range of settings (Clemen, 1989; Timmermann, 2006), including for infectious diseases such as Dengue fever (Johansson et al., 2019), Ebola (Viboud et al., 2018), influenza (Reich et al., 2019), and SARS-CoV-2 (Cramer et al., 2022; Howerton et al., 2023a). Ensembles are most useful when independent model predictions provide different insights about what could happen in the future, and therefore collectively represent

* Corresponding author.
E-mail address: ehowerton@psu.edu (E. Howerton).
[1] co-first authors

https://doi.org/10.1016/j.epidem.2024.100767
Received 21 June 2023; Received in revised form 27 March 2024; Accepted 8 April 2024
Available online 17 April 2024

uncertainty and estimate future outcomes better than a single model alone (Pennell and Reichler, 2011).

Multi-model hubs that leverage the power of ensembles are becoming increasingly common to support the management of infectious diseases (Reich et al., 2022), yet these efforts can be costly to initiate and maintain. A large part of this cost arises because hubs require effort from many modeling teams, and existing inequities in disease modeling mean that in some settings, multiple models may not even be available. There is likely a balance between convening enough models to gain the benefits of an ensemble while not wasting effort of teams that could instead address other important questions. But how do we find this balance?

Here, we provide perspective on this question by addressing multi-model agreement (or, conversely, disagreement) in different contexts. If there are contexts where we expect models are more likely to agree, we may then be able to proceed with fewer models and reserve multi-model ensembles for other contexts. Similarly, these may be the contexts we choose to support if only a few models are available. We consider situations that focus on predictions of quantitative epidemiological outcomes such as incident cases or peak timing; these are the settings in which multi-model ensembles have been repeatedly shown to be effective. We contrast these cases with settings where the primary decision depends on the ranking of alternative possible future scenarios. For example, a decision maker may compare outcomes under potential intervention scenarios and choose to implement the strategy that best achieves some objective. For these types of decision contexts, a few multi-model infectious disease studies have found different models to be largely consistent in scenario ranking, despite disagreement on the expected magnitude of the outbreak (Li et al., 2017; Prasad et al., 2023; Probert et al., 2018; Shea et al., 2023). Similar findings have also emerged in other ecological settings (Bauer et al., 2019). The generality of this pattern could have important implications for the allocation of modeling resources, especially in settings where modeling resources are limited. Could it be that fewer models are needed to accurately rank the severity of outcomes across multiple epidemic scenarios, including possible interventions to control infectious disease outbreaks?

We start to approach this question from both theoretical and real-world, empirical perspectives. First, we propose a simulation framework that mimics a multi-model setting and allows us to consider the potential factors that could drive various patterns of model agreement. Using this framework, we span a large universe of possible models and compare their predictions of epidemic size and corresponding ranks across alternative intervention scenarios, to assess the probability of model agreement in situations with varied degrees of biological uncertainty (e.g., an emergent vs. recurring pathogen). This simple framework can also be extended to address policies (i.e., suites of actions) rather than single intervention decisions. Second, we explore agreement between models across 14 rounds of real-time scenario projections from the U.S. COVID-19 Scenario Modeling Hub (SMH), a multi-model effort to generate 3- to 12- month ahead projections of cases, hospitalizations, and deaths throughout the evolving pandemic. SMH projections, which included four epidemic scenarios per round, were used to inform decisions about pandemic planning (Borchering et al., 2021; Truelove et al., 2022) and control (Borchering et al., 2023; Rosenblum, 2022). Across these rounds, an ensemble of SMH models demonstrated marked improvement in projecting the magnitude of future public health outcomes compared to the 4–9 component models (Howerton et al., 2023a). For the same 14 rounds spanning a variety of pandemic phases, we evaluate the agreement of individual models on scenario ranking. Taken together, this work further motivates the importance of understanding the benefits of multiple models in different decision contexts.

2. Methods

First, we propose a controlled epidemiological decision setting in which to theoretically investigate multi-model agreement. We define specific ways that models can differ in their assumptions and control all the other uncertainties so they do not complicate our results and interpretation. The goal here is to enumerate a large set of plausible models and assess the characteristics of agreement between models in this set. We deliberately focus on a decision setting where models are asked to rank three simple epidemic scenarios, to allow us to develop the approach, but such an analysis could be replicated for many different types of questions (e.g., see the case study in Howerton et al. (2023b)). In the following sections, we describe this simple decision setting and the proposed mathematical framework, discuss implementation of the analysis, and explain two potential methods for estimating model agreement. Second, we describe our application of these methods to real-world, empirical COVID-19 Scenario Modeling Hub (SMH) projections.

2.1. Hypothetical decision setting and mathematical framework

We consider the case where a decision maker is choosing between two possible interventions to control a hypothetical outbreak: (1) non-pharmaceutical interventions (NPIs), such as masking, or (2) vaccination, which immunizes individuals and prevents future infections. The decision maker can implement one, but not both, interventions (owing, say, to budget constraints or public tolerance for intervention). Multiple modeling groups generate scenario projections of future disease outcomes under each intervention and rank the scenarios accordingly to provide recommendations about which intervention will be most effective. Here, we set up a mathematical framework to mimic this multi-model setting.

Most multi-model efforts rely on independent modeling groups to make predictions (e.g., about future outcomes, intervention effectiveness), and each of these models will represent disease transmission processes differently because of uncertainties about pathogen biology, human behavior, etc. These uncertainties could lead to different model parameters (e.g., transmission rate), different model structure (e.g., including asymptomatic transmission), or different modeling approaches (e.g., agent-based vs. compartmental). In real-world multi-model settings, all these factors are present to some degree. However, this complicated set of differences can be hard to enumerate and untangle in model results, so here we intentionally simplify the sources of uncertainty driving differences between models. We assume the primary source of uncertainty is about biological model parameters, and let the structure and approach be consistent across models. Importantly, we also let assumptions about the effectiveness of interventions be shared across all models; in practice these assumptions would be based on literature estimates or expert opinion and would be uncertain to some degree. A similar approach, where well-defined alternative interventions make up different modeling scenarios and biological uncertainties are left to be handled by different models, has been used by multi-model scenario projection efforts to evaluate the impact of new interventions for a range of diseases (Borchering et al., 2023, 2021; Flasche et al., 2016; Prasad et al., 2023). We also assume that models generate projection point estimates rather than probabilistic distributions, although the same logic would apply to studying between-model agreement in different quantiles.

We represent different “modeling teams” with different combinations of parameters in a Susceptible-Infected-Recovered (SIR) model (Keeling and Rohani, 2008). The SIR model assumes that all individuals in the population can be classified as susceptible (S, can be infected), infected (I, currently infected), or recovered (R, immune to future infections). In the SIR model, susceptible individuals are infected through contact with infected individuals at some transmission rate, β. Infected individuals recover at rate, γ. We let all modeling teams assume the NPI intervention reduces transmission by some amount, d, and vaccines are distributed to susceptible individuals at some rate, ν. The total population size is assumed to be N. Then, the SIR model can be represented as a system of ordinary differential equations:

$$\frac{dS}{dt} = -(1-d)\beta SI/N - \nu S$$

$$\frac{dI}{dt} = (1-d)\beta SI/N - \gamma I$$

$$\frac{dR}{dt} = \gamma I + \nu S$$

2.2. Implementation of multi-model mathematical framework

We assume NPIs will reduce the transmission rate by 30% (i.e., $d = 0.3$) and the attainable vaccination rate (of fully immunizing susceptible individuals) is 0.01 (i.e., $\nu = 0.01$), approximating 1% of the susceptible population per day (Bjørnstad, 2018). In other words, the decision maker asks teams to model and rank scenarios where (1) NPIs reduce transmission by 30% or where (2) approximately 1% of the susceptible population is vaccinated per day. We assume we have a perfectly effective vaccine with sterilizing immunity. If either intervention is not implemented, then the respective parameter (d or ν) is set to zero. Interventions will be implemented for 50 days, which is sufficient time under these parameters to deplete the susceptible population and for the epidemic to fade out (Fig. S1). Then, we consider outcomes under each potential intervention from modeling teams with a range of assumptions about transmission and recovery rates in the absence of intervention. Specifically, we test models with all combinations of β between 0.75 and 1.25 and γ between 0.1 and 0.5, both in increments of 0.01. These ranges were chosen to span a wide range of plausible disease characteristics and basic reproduction numbers (here, $R_0 = \frac{\beta}{\gamma}$ values are between 1.5 and 12.5). These combinations yielded a total of 2091 possible models. Note, $\gamma = 0.1$ can be interpreted as an average recovery time of 10 days.

We assume the decision maker is interested in the intervention that minimizes cumulative infections (defined as the number of new infections that occur over the 50-day period). So, each model estimates the number of cumulative infections for three scenarios in total: under each of the two interventions, and a case without any intervention (i.e., $d = 0$, $\nu = 0$). Then, ranks are determined based on these projections, where the best ranked scenario is the one with the lowest projected cumulative infections. All models assume the outbreak occurs in a closed population of 1000 individuals, where 995 individuals are susceptible and 5 are infected at the start of the simulation. We implement each model numerically using `lsoda` integrator from the `deSolve` package in `R` version 4.2.0 (R Core Team, 2018; Soetaert et al., 2010).

We also assess the sensitivity of our results to the choice of objective (Probert et al., 2016) and the assumed effectiveness of each intervention. First, we perform a sensitivity analysis assuming the objective is to minimize the peak number of infected individuals (defined as the maximum number of individuals infected in a single day over the 50-day period). Second, we assess how results are affected by different levels of NPIs and vaccination (all combinations of $d \in \{0.1, 0.2, 0.3, 0.4\}$ and $\nu \in \{0.005, 0.01, 0.015, 0.02\}$). For these additional analyses on intervention effectiveness, we use fewer models for computational efficiency (i.e., a step size for β and γ of 0.1).

2.3. Two potential methods to estimate model agreement

Quantifying and comparing agreement in rank and magnitude is not a straightforward task, and we are not aware of any existing methods in the modeling literature that do so. This difficulty arises in part because definitions of "agreement" will vary based on the decision context. For example, what degree of difference among projection magnitudes is tolerable for the decision at hand? Here, we propose two potential methods as a starting point for thinking about this problem. The first method compares ranks and magnitudes for projections under a single scenario, and the second attempts to quantify agreement of rank and magnitude both across and within scenarios. We outline each method below.

2.3.1. Single scenario agreement: the "tolerance" method

The "tolerance" method focuses on model projections for a single scenario and compares the number of models that agree about (1) the ranking of that scenario projection (relative to the other scenarios) versus (2) the magnitude of that scenario projection. Here, we suppose models agree on scenario rank when their assigned rank matches exactly. In other words, for a single scenario, we count the number of models that have returned each rank, and we report the largest of these values. For example, a set of 10 models may have 4 models that rank intervention A as best, and 6 models that rank intervention A as second best; we always take the largest number of agreeing models (so 6 in this example).

For agreement on projection magnitude, we presume a decision maker has some tolerance, τ, within which projections of varying magnitudes are largely equivalent. So, for a given scenario projection p, we define a window of width τ, and again count the number of model projections that fall within this window. Practically, we calculate this by defining two windows for each projection, p: $[p, p+\tau]$ and $[p-\tau, p]$.We calculate the number of projections across the set that fall within these windows, and again choose the window that contains the most projections. We test various window sizes, τ. In the supplement, we also report results with a window size relative to the projection magnitude, i.e., window size $\tau = rp$, for some relative change r, but other definitions of "agreement" for rank and magnitude could also be used.

2.3.2. Agreement across scenarios: Inter-rater reliability method

The first method compares agreement of outbreak rank and magnitude in only a single epidemic scenario. However, our goal may also be to assess the consistency of projections across multiple epidemic scenarios. This problem is similar to evaluating the reliability of different measurements, and therefore we propose to use multiple statistics from the inter-rater reliability literature (Field, 2005; Liljequist et al., 2019) as a first attempt at quantifying rank and magnitude agreement across scenarios.

First, Kendall's Coefficient of Concordance, or Kendall's W could be used to measure rank agreement. This statistic compares the sum of squared error (SSE) of the observed rank totals to the SSE of the expected rank totals (Field, 2005). Kendall's W is bounded between 0, denoting no concordance in ranks, and 1, denoting perfect concordance in ranks. In other words, larger Kendall's W implies more agreement among models on the rank of different epidemic scenarios.

To illustrate this calculation, consider the following example presented in Table 1, where we have 4 models that rank 4 scenarios (so the mean of rank totals, $\bar{x}$, is 10). First, we calculate SSE of the observed rank totals, where

$$SSE_{observed} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$
$$= (8-10)^2 + (7-10)^2 + (15-10)^2 + (10-10)^2 = 38$$

Then, for m models and n scenarios, the Kendall's W statistic is calculated using the following formula (Field, 2005):

Table 1
Example of 4 scenarios ranked across 4 models. The "Rank Total" column is the sum for a given scenario across model ranks, or x_i (for scenario i). The mean of rank totals, $\bar{x}$, is 10.

	Model 1	Model 2	Model 3	Model 4	**Rank Total**
Scenario A	2	1	2	3	**8**
Scenario B	1	2	3	1	**7**
Scenario C	4	3	4	4	**15**
Scenario D	3	4	1	2	**10**

$$W = \frac{12 * SSE_{observed}}{m^2(n^3 - n)} = \frac{12 * 38}{4^2(4^3 - 4)} = 0.475$$

We also correct for the relatively rare case of ties between ranks, by calculating a correction factor $T_j = \sum_{i=1}^{k_i}(t_i^3 - t_i)$ for each model j that includes k_i distinct tied ranks across scenarios, each containing t_i scenarios tied for that rank. Then, we calculate Kendall's W statistic as $W = \frac{12*SSE_{observed}}{m^2(n^3-n)-m\sum_{j=1}^{m}T_j}$. Finally, we can estimate the probability that the observed agreement occurred by random chance using a p-value. To estimate the p-value, we use a chi-square test with $\chi^2 = m(n-1)W$ and $n-1$ degrees of freedom. In this example, with four models and four scenarios, the p-value is 0.127. For other settings with four scenarios, a Kendall's W value of 0.65 is required for a p-value of less than or equal to 0.05 if there are four models, 0.43 if there are six models, and 0.33 if there are eight models.

Next, the intraclass correlation coefficient (ICC) could be used to measure agreement across models in projection magnitude. This metric compares the variance of projections between scenarios to the total variance across projections from all models and scenarios (Liljequist et al., 2019). There are multiple versions of the ICC calculation, and here we use the "two-sided" model for measuring agreement. Following Liljequist et al. (2019), this model assumes a projection, p_{ij} from model j in scenario i can be decomposed into $p_{ij} = \mu + r_i + c_j + v_{ij}$, where μ is the mean value across models and scenarios, r_i is random variance across scenarios, c_j is random variance across models, and v_{ij} is additional error. Then, the ICC calculation measures the proportion of variance explained by differences between scenarios (as measured by ${\sigma_r}^2$), compared to the total variance (${\sigma_r}^2 + {\sigma_c}^2 + {\sigma_v}^2$), or $ICC = \frac{{\sigma_r}^2}{{\sigma_r}^2+{\sigma_c}^2+{\sigma_v}^2}$.

To estimate this value from a set of projections, we use various mean square relationships. Again let m be the number of models and n the number of scenarios; $x_{i,j}$ is a projection for scenario i from model j, $\overline{x}$ is the mean of projections across all models and scenarios, $\overline{s_i}$ is the mean of projections in scenario i, and $\overline{m_j}$ is the mean of projections from model j. Then, the total sum of squares is $SST = \sum_{j=1}^{m}\sum_{i=1}^{n}(x_{ij} - \overline{x})^2$, the mean squared errors between scenarios is $MSBS = \frac{m}{n-1}\sum_{i=1}^{n}(\overline{s_i} - \overline{x})^2$, the mean squared errors between models is $MSBM = \frac{n}{m-1}\sum_{j=1}^{m}(\overline{m_j} - \overline{x})^2$, and the mean squared error is $MSE = \frac{SST-(n-1)MSBS-(m-1)MSBM}{(n-1)(m-1)}$. Using these quantities, we can estimate ICC as

$$ICC = \frac{\text{MSBS} - \text{MSE}}{\text{MSBS} + (\text{k}-1)\text{MSE} + (\frac{\text{k}}{\text{n}})(\text{MSBM} - \text{MSE})}.$$

See Liljequist et al. (2019) for derivations.

Overall ICC values are generally limited to the range of 0–1, where higher values again imply higher agreement. When models agree about the magnitude of projections within each scenario, we expect the variance between models to be low and thus the variance between scenarios to compose the majority of the observed variance (i.e., ICC is close to one, and thus high ICC implies that projections between models within a single scenario are largely consistent). However, there can be cases where both variance between models and variance between scenarios are low (say, for example, because the scenarios themselves are not inherently different); in these cases, ICC can be estimated to be low, or even negative, even though variance between models is small. In the rare cases where ICC was estimated to be negative, we set it to zero following Bartko (1976).

Directly comparing these metrics assumes that they can be interpreted in similar ways (i.e., 0.5 has a similar meaning for both metrics), but this may not be the case. So, we also propose the comparison of Kendall's W and ICC metrics to a "null" model (discussed in 2.5) to understand the joint behavior of the two metrics under controlled assumptions and to quantify the significance of results relative to this joint behavior. We implement Kendall's W and ICC calculations using functions from the `irr` package (Gamer et al., 2019) in the `R` statistical software version 4.2.0 (R Core Team, 2018).

2.4. Application to projections from the theoretical multi-model setting

The theoretical simulation results span a large set of possible models, and from this set, we can estimate the probability that any subset of those models will agree using the methods defined in 2.3. However, with 2091 models from which to choose, it is not computationally feasible to explicitly enumerate all possible combinations and calculate the probability of agreement exactly. So instead, we approximate this probability by randomly drawing sets of models and calculating agreement for each set. We then summarize how many within the drawn set agree on the ranking of scenarios or the size of the uncontrolled outbreak (i.e., cumulative infections without intervention). Our algorithm to randomly select a set of models is based on our idea of how a multi-model set is generated in practice.

Typically, teams build their models and calibrate parameters based on available information, however limited that information may be. This available information, then, will inform where models are more likely to fall within the uncertainty space. Within our mathematical framework, we introduce this idea using the concept of a "neighborhood", or a subset of the uncertainty space from which a set of models is drawn. The center of the neighborhood is determined by available information and the size of the neighborhood is driven by how strong that information is. In our case, the uncertainty space is defined only by model parameters for transmission and recovery rates.

Here, we do not explicitly define what information is available, but instead randomly select neighborhoods of varying sizes and test how likely models are to agree for neighborhoods of that size. To implement this, we draw one model at random and find all models within neighborhood size $\pm\eta$ of that model's parameters. Because transmission and recovery rates are on relatively similar scales, we use the same η for both model parameters (i.e., $\beta \pm \eta$ and $\gamma \pm \eta$). This defines our neighborhood. Then, from the neighborhood, we draw $m-1$ additional models (the first model drawn is included in the set of m models), and we calculate agreement following our two proposed methods. We generate results for 10,000 sets of models from a neighborhood of size $\eta = 0.1$ for both parameters, and we test alternative neighborhood sizes, $\eta \in \{0.05, 0.2\}$ for both parameters.

2.5. Application to U.S. COVID-19 Scenario Modeling Hub projections

To complement our simulation study, we use the two proposed methods in an attempt to quantify patterns of agreement across a large-scale multi-model projection collaboration used to guide policy during the COVID-19 pandemic. Data from over two years of U.S. COVID-19 SMH projections (Howerton et al., 2023a, https://covid19scenariomodelinghub.org/) provide instances of multi-model projections across multiple future scenarios (covering both possible policy actions as well as key uncertainties about drivers of disease dynamics), generated across a range of pandemic contexts. Although each round is not entirely independent (e.g., many rounds include projections from some of the same models), SMH projections across multiple rounds provide a unique opportunity to test multi-model agreement in a real-world setting.

Between February 2021 and November 2022, SMH produced 16 rounds of multi-model projections of incident and cumulative cases, hospitalizations, and deaths. We focus our analysis on 14 rounds that were released publicly (i.e., we exclude Rounds 8 and 10). In each round, models made projections for four distinct scenarios, which typically included two axes of uncertainty. Each axis focused on drivers of disease dynamics that were uncertain at the time of scenario design and projection, including implementation of potential control strategies and uncertainties about pathogen biology or human behavior. Models made projections on horizons of 12 weeks to 52 weeks, depending on the goals of the round. SMH projections were generated for 52 locations in total, including U.S. national projections and all U.S. states. For more

details on the SMH process, see Loo et al. (2023).

Thirteen teams participated over the first 16 rounds, with each round including projections from 4 to 9 models. SMH modelers have come from a diversity of backgrounds, ranging from well-established epidemiological modeling groups to groups from other fields. Some models only made projections in some rounds, and others only made projections for specific states. The approach of each model was different, including mechanistic models and agent-based models across a variety of spatial structures and fitting schemes (summary of each model provided in (Howerton et al., 2023a)). Some submitted projections did not comply with basic SMH standards, and we exclude those here (following the inclusion criteria of (Howerton et al., 2023a)).

We analyze agreement for each set of SMH projections again using the two methods we have proposed. A single set of projections are made for one location (e.g., U.S. national projections or projections for a single state), target (e.g., cumulative hospitalizations), horizon (e.g., 26 weeks into the future), and SMH round. For the tolerance method, we redefined projections relative to the population size of a given location (i.e., cumulative hospitalizations per 100,000 population) so that an absolute window size would be comparable across locations. We provide results for a relative window size in the supplement.

Then, to better understand the possible joint behavior of Kendall's W and ICC values for different sets of SMH projections, we generate 1000 "null" projection sets for each round-target-location assuming no agreement between models and accounting for inherent variation between scenarios observed in the SMH projections. To do so, for a given round-target-location, we calculate the range of median projections across models for each scenario (i.e., a lower bound of all projections for scenario i across models j, $l_i = \min_{j=1,..,m} x_{ij}$, and a corresponding upper bound $u_i = \max_{j=1,..,m} x_{ij}$). Then, null projection sets are drawn uniformly from these ranges, $n_{ij} \sim U(l_i, u_i)$, where the same number of projections are drawn from each scenario as models that made SMH projections. These null projections retain the differences between scenarios observed in SMH projections (which can influence measures of agreement as discussed in 2.3.2). Importantly, the null projections also sample each scenario independently, explicitly excluding potential consistencies within a model (for instance, a model that would systematically project low outcomes in all scenarios, compared to all other models). Note that these consistencies may or may not exist in SMH projections. We then use these null projection sets to calculate the relative change in ranking and magnitude agreement. We estimate Kendall's W and ICC relative to the mean value of Kendall's W and ICC and the percent of null projection sets that had higher Kendall's W and ICC.

In the main text, we focus our analysis on agreement of median cumulative projections of incident hospitalizations over the maximum projection horizon for each round, in order to maximize the differences between scenarios. This avoids instances where agreement is spuriously low because scenario projections are not sufficiently different. For example, Round 11 and 12 scenarios focused one uncertainty axis on the severity of the emerging Omicron variant. Because severity scenarios focused on risk of hospitalization and death, we do not expect agreement about ranking of scenarios for cumulative cases to be meaningful. We provide additional results for alternative quantiles (Q25, Q75) and horizons (4, 8, 12, 16, 20, 26 weeks) in the supplement. For more information on SMH scenario projections and participating models, see other papers in this special issue or visit https://covid19scenariomodelinghub.org/.

3. Results

3.1. Simulation study

Our simulation study illustrates how models with different parameters predict different epidemic trajectories and can offer different rankings of alternative epidemic scenarios, which here corresponds to divergent intervention recommendations (Fig. 1). For example, one model (that assumes a transmission rate of 1 and a recovery rate of 0.2 in the absence of interventions, model 1 in Fig. 1A) estimates 961 cumulative infections with NPIs and 919 cumulative infections with vaccination, therefore recommending vaccination to minimize infections. A second model assumes a transmission rate of 1.15 and a recovery rate of 0.4 (model 2 in Fig. 1A) in the absence of interventions; this model estimates 797 and 847 cumulative infections under NPIs or vaccination, therefore recommending NPIs.

Whether a model ranks NPIs or vaccination scenarios as best to minimize infections depends on both transmission and recovery rates (Fig. 1B), with assumptions about recovery rate being the more meaningful driver. Models that assume recovery is fast (i.e., high recovery rates) rank NPI scenarios as best, and models that assume recovery is slow (i.e., low recovery rates) rank vaccination scenarios as best. Across all combinations of transmission and recovery rates, predictions of cumulative infections without intervention vary dramatically (Fig. 1C). As expected, models predict the outbreak will be small when R_0 is low (583 cumulative infections in the smallest predicted outbreak), whereas models with large R_0 predict nearly all individuals in the population will be infected.

Despite the substantial heterogeneity in predicted outbreak magnitude, ranking of intervention scenarios are largely consistent (i.e., models with small R_0 recommend NPIs, and models with large R_0 recommend vaccination). However, there is a subset of models that make similar biological assumptions but rank intervention scenarios differently (45% of pairs of models with R_0 between 3 and 4 will have differing intervention scenarios ranked as best). In these instances, the models predict similar cumulative infection outcomes under the two possible interventions. The largest difference in cumulative infections is 44 (or 4.4% of the population) for models with R_0 between 3 and 4, compared to 250 (or 25% of the population) which is the largest difference overall.

Within our simulation framework, a randomly drawn set of models is more likely to agree on the ranking of intervention scenarios than generate highly similar estimates of cumulative infections (Fig. 2). The probability that 4 of 6 randomly selected models agree on ranking of intervention scenarios is 68%, whereas the probability that they agree on estimates of cumulative infections within small bounds is less likely (18% probability of agreement within 20 infections, which is 2% of the population and 5% of the range of possible cumulative infection outcomes; 50% agreement within 50 infections, which is 5% of the population and 12% of the range of possible cumulative infection outcomes). However, for all numbers of models in the set, the probability at least 66% of those models agree on infection estimates within 100 infections is similar to or greater than the probability of agreeing on intervention scenario ranking. A window of one hundred infections covers almost 40% of the range of possible cumulative infection outcomes. In all cases, probability of agreement is higher when models make similar assumptions.

These patterns are also demonstrated by ICC and Kendall's W statistics (Fig. 3). Model agreement on intervention rank is high and significant for all sets of models. Kendall's W is always above 0.7 and is 0.8 on average across model set size (IQR: 0.75–0.81) for randomly selected models and is 0.93 on average (IQR: 0.85–1.0) for similar models. In contrast, agreement in magnitude is more variable (average ICC of 0.24 (IQR: 0.13–0.28) across randomly selected model sets of all sizes, and 0.68 (IQR: 0.54–0.82) across similar models). Presumably the high agreement on ranking of intervention scenarios is due in part to the inclusion of the "no intervention" scenario, which is the worst of the three intervention scenarios in every model. When we assessed ranking of the two intervention scenarios (without the counterfactual, "no intervention" scenario), Kendall's W dropped to an average of 0.19 (IQR: 0.12–0.25) for randomly selected models and 0.75 (IQR: 0.40–1.00) for similar models (Fig. S7).

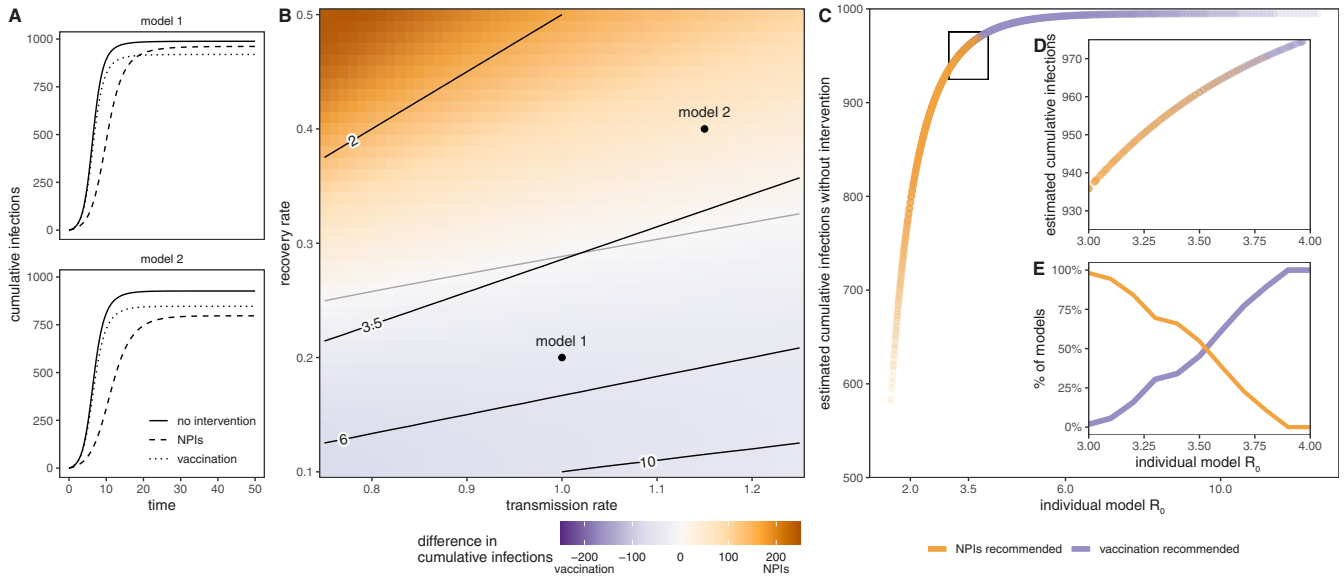

Fig. 1. Model projections of a generic infectious disease outbreak under two potential interventions. (A) Two models that differ in recommended intervention. Model 1 (top) assumes transmission rate is 1 and recovery rate is 0.2 (i.e., average time to recovery is 5 days) in the absence of interventions. Model 2 (bottom) assumes transmission rate is 1.15 and recovery rate is 0.4 (i.e., average time to recovery is 2.5 days) in the absence of interventions. Both models predict the cumulative number of infections over a period of 50 days under three intervention scenarios: no intervention (solid), non-pharmaceutical interventions (NPIs) that reduce transmission by 30% (dashed), and vaccination of approximately 1% of the susceptible population per day (dotted). In Model 1, vaccination is recommended as it minimizes infections, whereas NPIs minimize infections and are recommended by Model 2. (B) Spanning the space of possible models across biological uncertainties (transmission rate, β, and recovery rate, γ, in the absence of interventions), each position on the graph represents an individual model, with the color of the tile representing the recommended intervention (orange: non-pharmaceutical interventions, NPIs; purple: vaccination), and the intensity of the color recommending the magnitude of the projected difference in cumulative infections. The gray line shows where the difference in cumulative infections is 0, or where neither intervention is recommended over the other. Black contours show sample values of individual model $R_0 = 2$, 3.5, 6, and 10. (C) Projected magnitude of outbreak without intervention, by model R_0. Each point represents projections from a different model, and the color of each point represents the intervention that model recommends (orange: non-pharmaceutical interventions, NPIs; purple: vaccination). (D) Zoom in on results for individual models with R_0 between 3 and 4. (E) Percent (%) of models with R_0 between 3 and 4 that recommend NPIs (orange) or vaccination (purple).

The agreement between models depends on the intervention scenarios that are being compared, especially with respect to differences in effectiveness and coverage of interventions (Fig. S2). In some instances, the same intervention scenario is ranked as best by all models. For example, if NPIs will only reduce transmission by 10%, vaccination of at least 0.5% of the population per day is universally best to minimize cumulative infections. The agreement between models also depends on the objective of interest (i.e., the outcome we are minimizing). Of the NPI efficacies and vaccination rates we considered, there were no instances where NPIs were universally best in all models. However, for minimizing peak infections, NPIs were ranked best by almost all models (except when NPI efficacy was low and the vaccination rate was high) (Fig. S3).

3.2. *Empirical study based on SMH scenario projections for COVID-19*

We analyzed 14 rounds of SMH projections (the 2 non-public rounds were excluded), which included 4 scenarios in each round from 4 to 9 independent modeling teams. The scenarios modeled by SMH and the number of teams participating varied across rounds. For example, the earliest SMH rounds (Round 1 – Round 4) focused on the early rollout of vaccination and NPIs, whereas the later rounds (Round 13 – Round 16), generated more than two years later, addressed booster vaccination and continuing SARS-CoV-2 evolution. We assessed the agreement across models on scenario rank and projection magnitude for 52 locations per round, totaling 728 sets of projections overall.

Visually inspecting the SMH projections reveals a spectrum of possible outcomes across this large set. For example, there were some SMH projections where models appear to largely agree about the rank of interventions but are less consistent in the projected magnitude for at least some scenarios (such as the projection shown in Fig. 4A). Other instances with largely consistent ranking across models have projection magnitude more clearly aligned within scenarios (example in Fig. 4B). There are also instances where rank across models is less consistent, both when projected magnitude across models is relatively similar (example in Fig. 4C) or variable (example in Fig. 4D). Here, we provide a few illustrative examples from a single SMH round, but these patterns also vary across all SMH rounds and locations (Figs. S8-S21).

For most SMH projections, more than 50% of models agree about the ranking of scenarios (707/728 total sets of projections, or 97%, for Scenario A; 659/728, 91% for Scenario B; 676/728, 93% for Scenario C; and 708/728, 97% for Scenario D). Yet, it is less common for 75% of models to agree about scenario rank, especially in Scenarios B and C (513/723, 70% for Scenario A; 272/728, 37% for Scenario B, 243/728, 33% for Scenario C, and 508/728, 70% for Scenario D). The higher rank agreement in Scenarios A and D likely reflects the "optimistic" and "pessimistic" definitions that were typical of these scenarios in most SMH rounds; Scenarios B and C were typically defined as intermediate to Scenarios A and D. Similar or greater levels of agreement for projection magnitude can be obtained with sufficiently large windows to define "agreement" (Fig. 5). For example, in Scenario B, the number of models agreeing on projection magnitude (i.e., falling within a window of a particular size) is greater than or equal to the number of models agreeing on projection rank in 379/728 (52%) sets of projections when the window size is 200 cumulative hospitalizations, 581/728 (80%) when the window size is 500 cumulative hospitalizations, and 710/728 (98%) when the window size is 1000 cumulative hospitalizations. Similar tradeoffs can be found for the other three SMH scenarios (Fig. S22) and can be examined relative to the projection rather than using an absolute window size (Fig. S23).

We can also attempt to summarize agreement across all scenarios simultaneously. A substantial part of this analysis involves

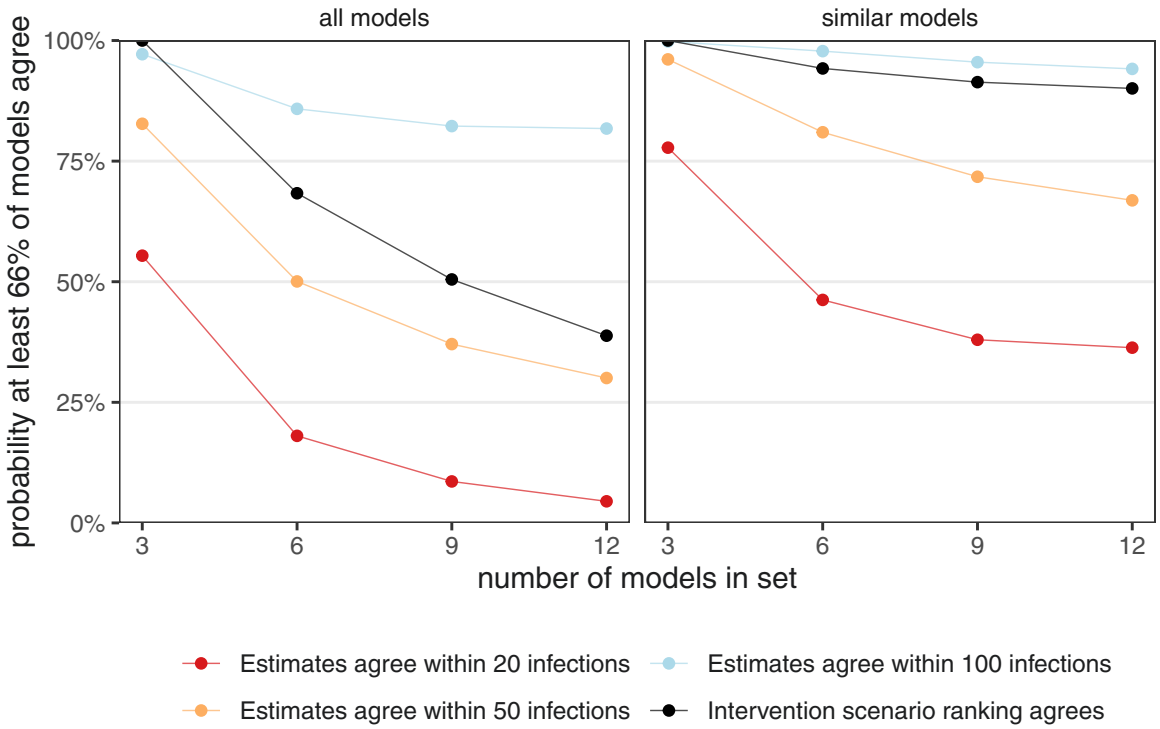

Fig. 2. Probability of at least 66% of models agreeing about intervention recommendations or estimates of cumulative infections without intervention. Agreement probability is shown for estimates of cumulative cases within 20 infections (2% of population, red), within 50 infections (5% of population, yellow), within 100 infections (10% of population, blue), or recommended intervention (black). Probabilities are calculated for agreement of 2 models out of a set of 3, 4 out of a set of 6, 6 out of set of 8, and 8 out of a set of 12. Agreement on intervention recommendations is 100% for a set of 3 models because for a binary task (i.e., recommend vaccination or non-pharmaceutical interventions), agreement of 2 models is guaranteed. All probabilities were calculated both when choosing randomly among all models or among models with "similar" assumptions. "Similar" models are defined as those with transmission and recovery rate assumptions within ±0.1 of a given model. See Fig. S4 for the probability that at least 2,4, 6, or 8 models agree across alternative definitions of "similar" and see Fig. S5 for agreement probabilities for all combinations of model set size and percentage of models agreeing. See Fig. S6 for results when magnitude agreement is defined relative to the projected magnitude. Note that the space of model parameters is the same as in Fig. 1B.

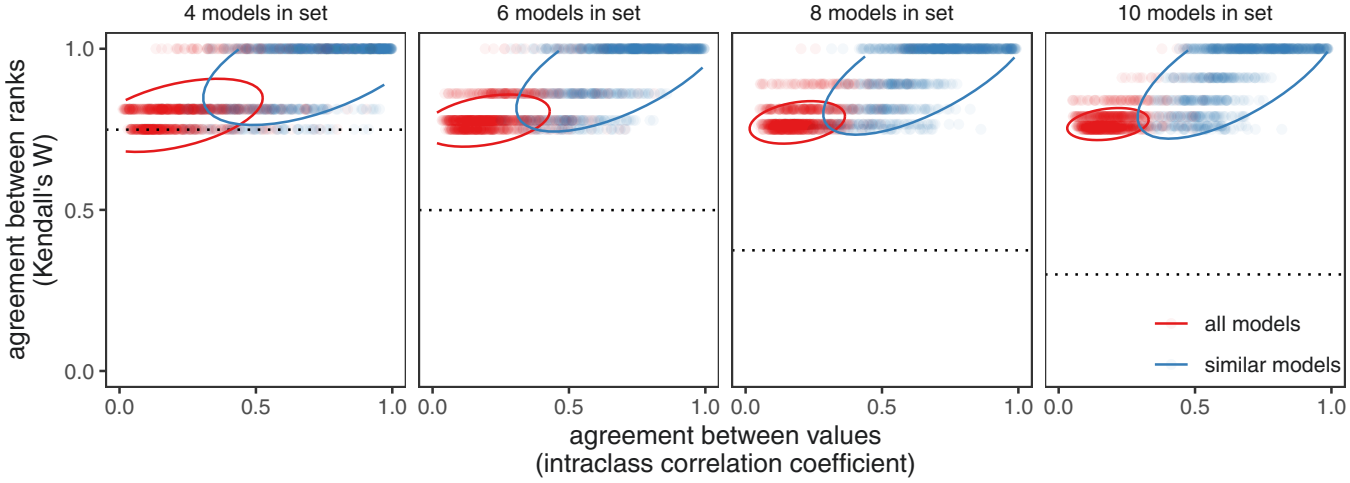

Fig. 3. Agreement between models in theoretical framework. For a varying number of models in a set, agreement is measured for projection magnitude across intervention scenarios (using intraclass correlation coefficient, ICC) and ranking of intervention scenarios (using Kendall's W). Results are shown when models in the set come from all possible models considered (red), and when models are similar (blue). Each point represents one set of randomly selected models from either all possible models considered (red) or a neighborhood of "similar" models (blue, those with transmission and recovery rate assumptions within ±0.1 of a given model). Ellipses show the area within which 95% of points fall. Dotted horizontal lines show agreement that is significant with a p-value of 0.05. See Fig. S7 for results with only vaccination and NPI scenarios included.

understanding how projection variance is partitioned across models and scenarios. Results show that differences between models drove the variance of SMH projections. The mean squared error between models composed 72% of the total sum of squares on average (35%-95% range across rounds), whereas mean squared error between scenarios composed 23% on average (4%-53% range across rounds) (Fig. 6C). This could be a result of comparatively low model agreement about the magnitude of projections (if agreement about projection magnitude was

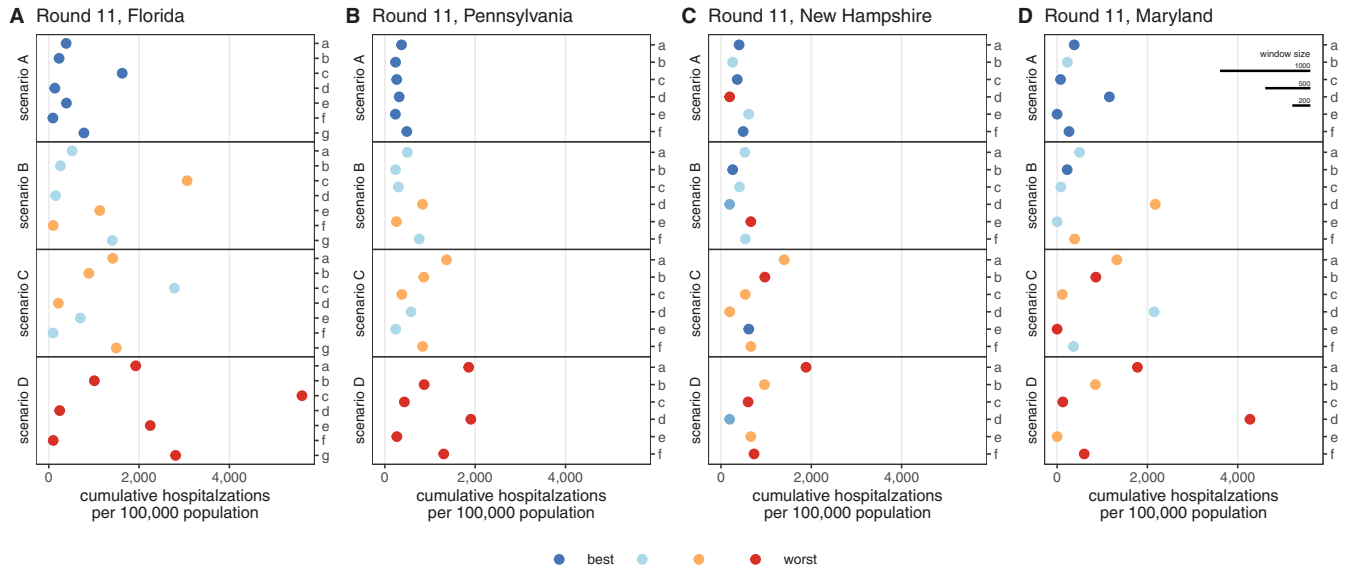

Fig. 4. Four examples of U.S. COVID-19 Scenario Modeling Hub multi-model projections. Panels (A)-(D) show examples of SMH projections from four different locations in SMH Round 11. Each SMH round consists of four scenarios (scenarios A-D), shown in a different segment of each panel, and for each scenario, median projections from 6 models (models a-f) are shown as points. The color of each point represents how each model ranked that scenario (blue indicates best ranked scenario, and red indicates worst ranked scenario). SMH Round 11 was released on December 25, 2022 at the outset of the Omicron variant wave; scenarios in this round varied low and high levels of immune escape and transmissibility of Omicron, and projections were made 12 weeks into the future (i.e., the x-axis shows cumulative hospitalization per 100,000 population on March 12, 2023). For reference, in the top right corner of panel D, we show three example "window" sizes for defining magnitude agreement that are used for results in Fig. 5 (window size = 200, 500, and 1000 hospitalizations per 100,000 population).

high, variance would be driven primarily by differences between scenarios) or minimal inherent differences between the scenarios. We can use the null projections to help interpret these findings and distinguish between the two hypotheses, since the null projections control for (low) variation between scenarios. We find that SMH projections had low ICC values compared to the null projections, indicating low model agreement about magnitude. ICC values for SMH projections were on average 71% lower than the average of null simulations in the same location and for the same round (range across rounds: 35%-92%) (Fig. 7A). In 234 locations (out of 728, 32%), ICC was lower than 90% of simulations (Fig. 7B).

In the null projections, ICC values were highly correlated with Kendall's W values ($r^2 > 0.8$ in 431/728 locations, 59%; example in Fig. 6B, all results in Figs. S21-S34). In other words, null projections that have low ICC (i.e., agreement of scenario magnitude, which is largely observed for SMH projections) are also expected have low Kendall's W (i.e., agreement on scenario ranking). However, observed SMH projections have high Kendall's W values compared to the null (Fig. 7A). Averaged across rounds, Kendall's W for SMH projections was 1.25 times higher than the average null projections (range across rounds: 1.22–4.3) (Fig. 7A). In 562 locations (77%), Kendall's W for SMH projections was higher than 90% of null projections (Fig. 7B). This suggests that that SMH models agree more frequently about ranking than would be expected based on the variability we see between the projections from these models.

Agreement of SMH projections also varied by round. With the exception of Round 1, agreement relative to the null projections was high in early SMH rounds. In Rounds 2–6, 86% of locations (269/312) had Kendall's W greater than 90% of null projections and 23% of locations (73/312) had ICC lower than 90% of null projections (Fig. 5B). These early scenarios focused on uncertainties around early vaccine supply and uptake, NPI adherence, and the emergence of the Alpha and Delta variants. The scenarios modeled likely had an *a priori* expectation of ranking (e.g., optimistic vaccination scenarios are expected to be better than pessimistic vaccination scenarios with an efficacious vaccine in a largely susceptible population). Later SMH rounds (Rounds 13–16) addressed the emergence of immune escape variants and booster vaccination, scenarios in which the *a priori* ordering is less clear (e.g., due to complex interactions with the rate of immune waning). These rounds demonstrated comparatively low levels of agreement relative to the null; 76% of locations (159/208) had Kendall's W greater than 90% of null projections and 27% of locations (58/208) had ICC lower than 90% of null projections (Fig. 5B).

4. Discussion

Leveraging predictions from multiple models via an ensemble is more robust than relying on a single model, especially when uncertainty is high (Clemen, 1989; Howerton et al., 2023b; Timmermann, 2006). Yet, multi-model efforts require substantial resources, and we lack clear theoretical or empirical guidance on the sufficient number of models needed to address a particular intervention or planning decision. Moreover, a multitude of models is not always available. Here, we have proposed the idea that the incremental value of adding a model to a multi-model set depends on the decision context. In particular, we considered decisions that depend on ranking a discrete set of well-defined epidemic scenarios (e.g., choosing between alternative interventions based on some objective). For our theoretical simulations and empirical results across 14 rounds of COVID-19 Scenario Modeling Hub projections, agreement between models on scenario ranking was relatively common. Similar levels of agreement could be obtained for projection magnitudes with sufficiently large tolerance windows; whether such windows are appropriate will depend on the decision context. Better understanding these tradeoffs, and how general such tradeoffs are, could have important implications for understanding the decision contexts in which using fewer models may be tolerable.

Within the simulation framework, one key mechanism driving agreement was the similarity of model assumptions. In practice, we expect the similarity of model assumptions to be governed in part by the quality of our existing information and the independence of the models. For example, R_0 for an endemic pathogen will be much more certain than for an emerging pathogen. However, the set of models will not necessarily capture all uncertainties (e.g., unknown unknowns will not be accounted for), and model independence may be difficult to obtain in

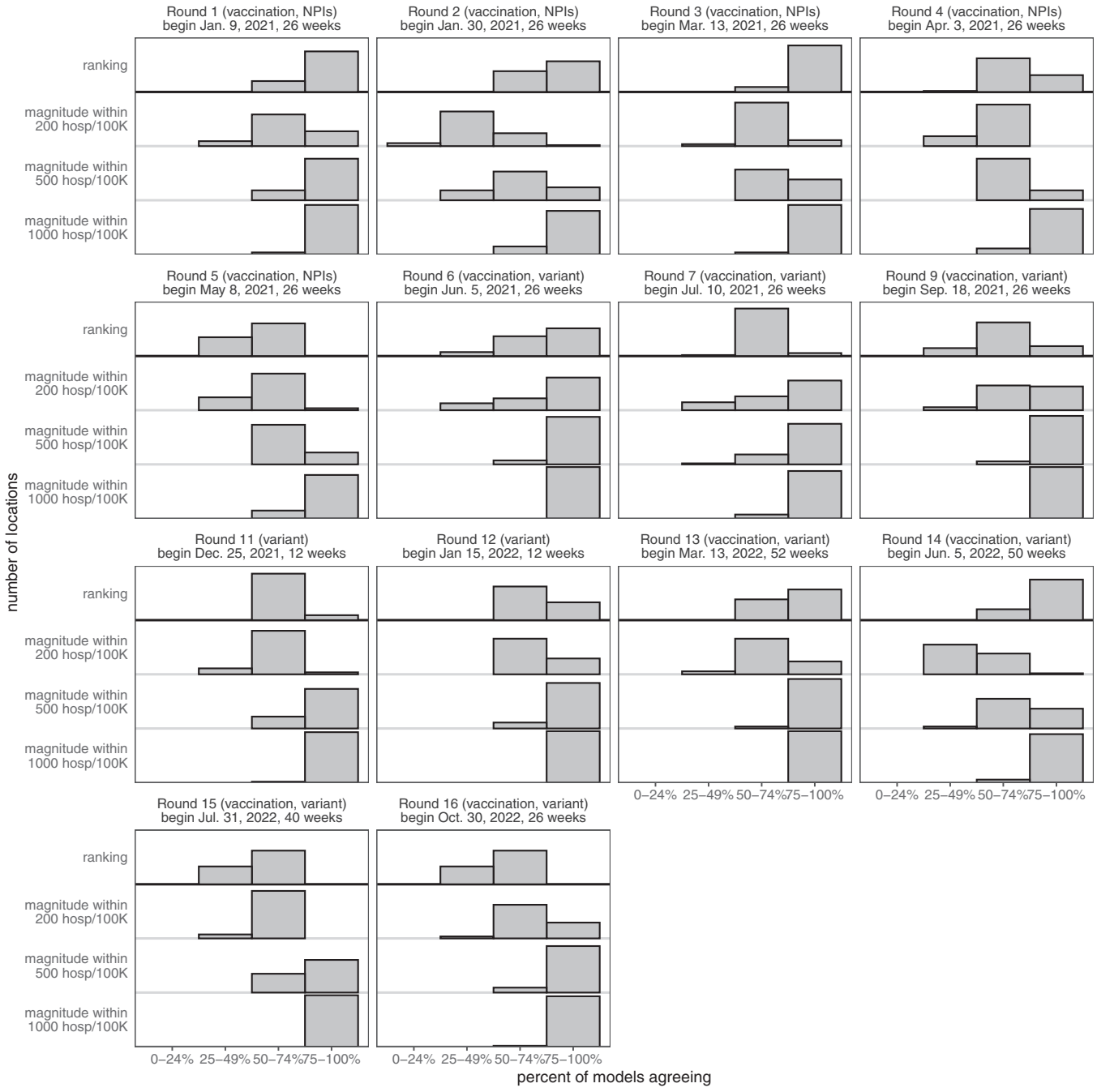

Fig. 5. Agreement of U.S. COVID-19 Scenario Modeling Hub projections assessed using the "tolerance" method. For each round of COVID-19 projections, we calculated the percent of all models that agreed on scenario ranking and on projection magnitude, where projection magnitude agreement was defined as the number of projections falling within a window of *n* hospitalizations per 100,000 population. For a given round (panel), the height of each bar shows the number of locations that fall within a given bin of percent of models agreeing. Here, we show results for Scenario B, as Scenario B has lower overall levels of rank agreement; see Fig. S22 for results across all scenarios and Fig. S23 for results when the tolerance window is defined relative to the projected magnitude (tolerance windows under a relative definition are shown in Fig. S24). Note, the number of contributing models varies across rounds (R1: 4, R2: 5, R3: 4, R4: 6, R5: 7, R6: 8, R7: 8, R9: 8, R11: 6, R12: 6, R13: 8, R14: 8, R15: 6, R16: 6 models). This list reports the number of models that is most frequent across all locations in a given round, although occasionally the number of contributing models for a particular location would vary slightly (e.g., some models only submitted for a subset of locations).

practice (Knutti et al., 2013; Pennell and Reichler, 2011). Further, agreement between models was higher when the effectiveness of the interventions considered was greater. In other words, the expected agreement will depend not only on the similarity of the underlying model assumptions and approach but also on the scenarios modeled. This provides one hypothesis for the relatively high rank agreement observed for SMH projections: there was some inherent expectation about the ranking of SMH scenarios which was shared across models (e.g., a counterfactual scenario should be worst). There are many possible goals in scenario design (Runge et al., 2023), and future work could consider whether agreement is more or less likely for different kinds of designs. Model agreement on scenario ranking also depended strongly on the objective of interest (i.e., recommendations were largely consistent for minimizing peak cases, but different from recommendations targeted at minimizing final epidemic size), emphasizing the importance of having a clearly defined objective (Probert et al., 2016).

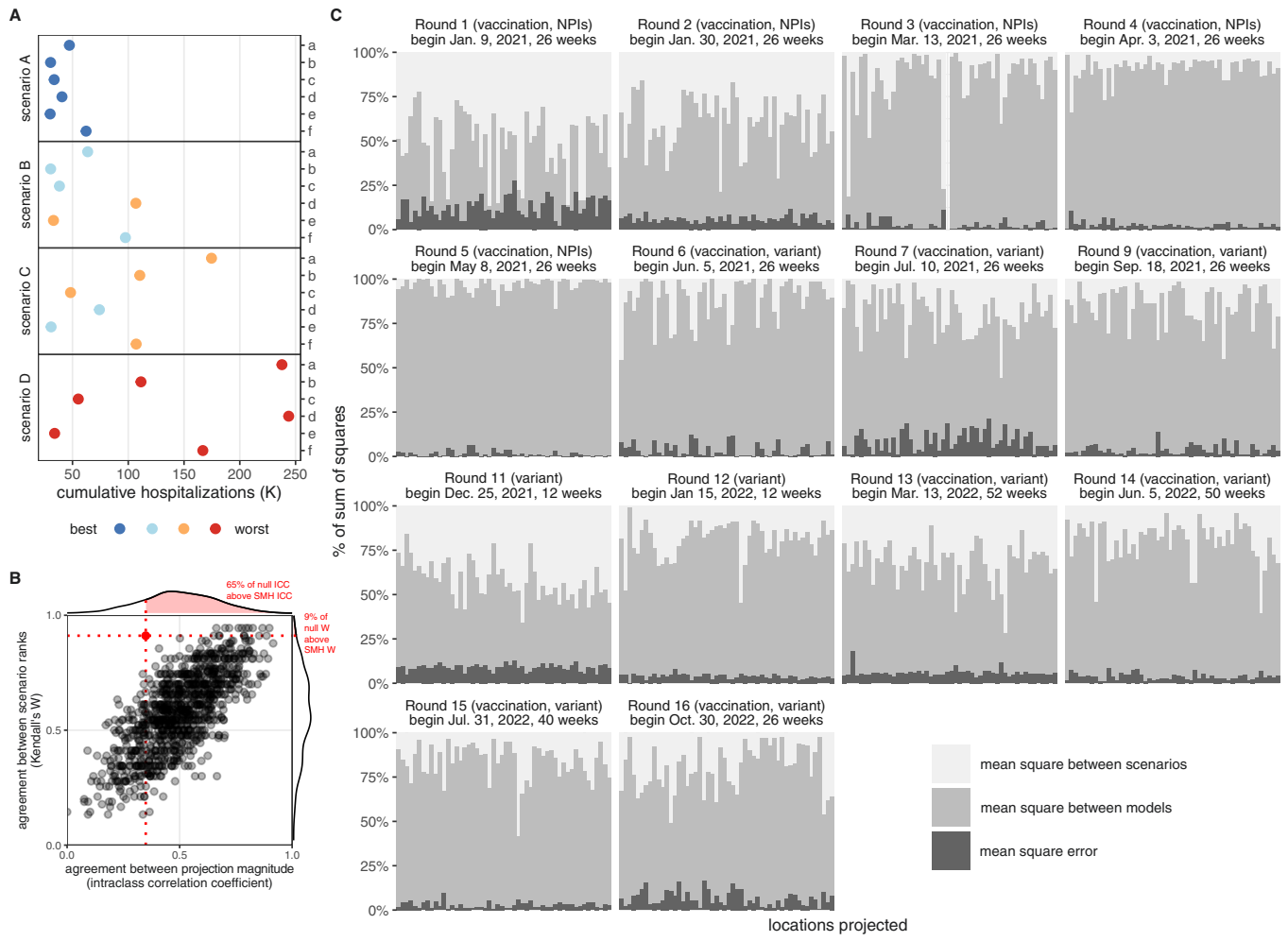

Fig. 6. U.S. COVID-19 Scenario Modeling Hub (SMH) projections. (A) Example of SMH multi-model projections for four scenarios (scenarios A-D) from Round 11 released on December 25, 2022 at the outset of the Omicron variant wave. Median projections from 6 models (models a-f) of cumulative hospitalizations in Pennsylvania (after 12 weeks (on March 12, 2023). Color of each point represents how each model ranked that scenario (e.g., model a ranked scenario A best, and scenario D worst). Scenarios in this round varied low and high levels of immune escape and transmissibility of Omicron. This example is also shown in Fig. 4B. (B) Results from null model for the round, location, and target presented in (A). The agreement of projection magnitude (as measured by intraclass correlation coefficient, ICC) and agreement of scenario rank (as measured by Kendall's W) for each of the 1000 null simulations shown as black points. The observed ICC and Kendall's W for the SMH projections (shown in (A)) is shown with a red point. The densities above either axis show the distribution of ICC and Kendall's W for the null model, and the red filled regions represent those null simulations with higher values than the SMH projections. (C) Components of variance for SMH projections of incident hospitalizations across all locations and rounds (mean squares between scenarios in light gray, mean squares between models in gray), and mean squared error in dark gray; see Methods for formulas). Each panel shows results for a single SMH round, and the header for that panel also include the axes of uncertainty addressed by the scenarios (e.g., Round 1 scenarios varied levels of vaccine supply in early rollout phases and levels of NPIs in the community), as well as the projection start date (Round 1 projections began on Jan 9, 2021) and the projection horizon (Round 1 projections were made 26 weeks into the future). Each bar within a panel represents one location. For more discussion of scenario specifications in each SMH round, see Runge et al. (2023) and Howerton et al. (2023a). For all SMH projections, see Figs. S8-S21 and for all null model results, see Figs. S25-S38.

In our simulation framework, there was a subset of similar models that were more likely to disagree about the ranking of intervention scenarios (models with R_0 between 3 and 4 in this example, but this range is likely context-specific). In these instances, the estimated differences in outcomes between scenarios (e.g., the benefits of a particular intervention) were small, suggesting one scenario was not meaningfully better than the other. Alternative considerations not explicitly modeled, for example about economic, social, or political costs of an intervention may also affect decisions, especially when projected epidemic outcomes are similar. Relative effect size (Prasad et al., 2023), the uncertainty associated with such estimates, and the relative cost of implementation (e.g., Castonguay et al., 2023) are important alternative considerations and could be explored further within our framework.

Underlying our results, at least in part, is the premise that ranking is an easier task than quantifying continuous outcomes. For most practical decisions, there are usually fewer possibilities (in other words, the size of the intervention space is smaller than the size of the uncertainty space), and therefore ranking agreement will be more likely probabilistically. As the number of intervention scenarios increases, the probability of agreement by chance will decrease. Further, infectious disease control decisions are often multifaceted and complex (interventions depend on both effectiveness and uptake; none of which can be fully known when projections are made). The principles presented in this paper can be extended beyond comparisons of a few epidemic scenarios, to comparisons of a multitude of control policies representing suites of actions.

The methods and results presented here are subject to a number of limitations and represent only an initial step toward understanding how the value of predictions from multiple models could vary across different decision contexts or prediction targets. Our simulation framework

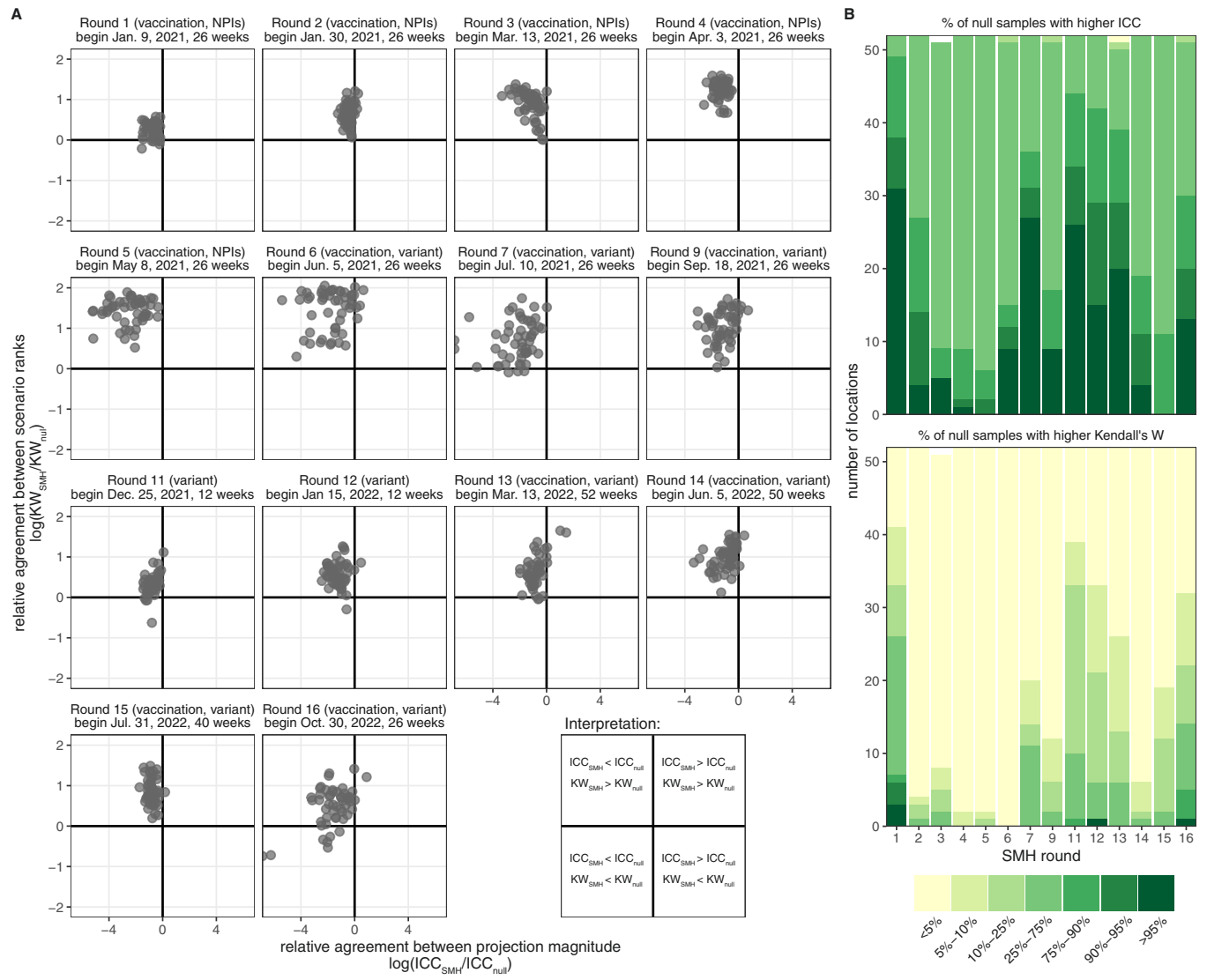

Fig. 7. Agreement between models on ranking of scenarios and projection magnitude from U.S. COVID-19 Scenario Modeling Hub (SMH). (A) Relative agreement between SMH projections and mean of null model projections for agreement about projection magnitude (as measured by intraclass correlation coefficient, ICC) and agreement about scenario ranking (as measured by Kendall's W). Each panel shows results for a single SMH round, and the header for that panel also includes the axes of uncertainty addressed by the scenarios, as well as the projection start date and the projection horizon. Each point represents a single location. Relative agreement between projection magnitude is calculated as $\log(ICC_{SMH}/ICC_{null})$ and relative agreement between scenario ranks is calculated as $\log(KW_{SMH}/KW_{null})$ where ICC_{null} and KW_{null} are mean values across all 1000 null simulations for that location. Interpretation is shown in the bottom right panel; negative relative agreement implies the value observed for SMH projections were smaller than the mean value for the null model. (B) Percent of null simulations that have ICC and Kendall's W greater than the observed SMH value. Each bar plot shows the number of locations falling into each discrete class of percentages (<5%, 5%-10%, 10%-25%, 25%-75%, 75%-90%, 90%-95%, and >95%, where the highest value in the range is not included, i.e., 5% exactly would be included in the 5%-10% range). For SMH agreement for other quantiles, horizons, and targets see Figs. S39-S42.

deliberately controlled for many aspects that may be important in real-world multi-model settings. We used highly simplistic disease models and assumed the models only differed in their assumed biological parameters. In practice, many other modeling decisions (e.g., about model structure, initial conditions) will vary and data sources used as part of model calibration or model fitting may vary across approaches, as seen in our empirical (SMH-based) analyses. We also did not include operational uncertainty (e.g., about intervention effectiveness) in the models, which may affect intervention rank more strongly than biological uncertainty. Future work could build operational uncertainties into this framework (Li et al., 2019), and investigate how model agreement is affected by interactions between operational uncertainty and biological uncertainty. Given the potential importance of these interactions, our conclusions should be validated under such complexities before being generalized. Second, our simulation framework considered a simple decision at one point in time. However, outbreak response and infectious disease management is a complex, ongoing task, with changing biological context (e.g., changes in human behavior, pathogen evolution) and acquisition of new information. Questions about how model agreement changes across time, and as more information is acquired, could inform better strategies for passive and active learning within and across outbreaks (Atkins et al., 2020; Shea et al., 2014). Building upon our simulation framework and exploring the implications of these real-world complexities could extend the generality of our conclusions.

Additionally, our ICC analysis comes with significant assumptions and limitations, which should be reconsidered with future methodological developments. Kendall's W and ICC are conceptually distinct statistical metrics, with differing definitions of "agreement". This makes the interpretation of their comparison difficult. We have attempted to overcome this limitation by quantifying their joint behavior via null

projections; however, this approach still requires that we assume changes in each metric relative to the null can be interpreted in similar ways. In addition, ICC may not be the best metric for quantifying magnitude agreement; ICC depends in part on projection rank, it is not well suited for sets of projections with small differences between scenarios, and it may not clearly correlate with meaningful levels of agreement in epidemiological decision contexts. More generally, the definitions we have created here for "ranking agreement" and "magnitude agreement" may be somewhat artificial, and in fact the decision context should ultimately determine what is most important to predict and what constitutes "agreement" for these predictions. Other analytical approaches will be required to soundly compare agreement of rank and magnitude across multiple scenarios. By building off SMH, other multi-model efforts (e.g., Li et al., 2017; Prasad et al., 2023; Shea et al., 2023), and the multitude of existing studies that consider different intervention scenarios across uncertainties, we can further test our hypotheses, understand model agreement, and make informed decisions about when we need multiple models for predicting and controlling infectious disease outbreaks. These conclusions may translate into other fields that use multiple models to inform decisions, such as climate science or ecology.

This work provides a first step in helping us better balance the tradeoffs between the resources required to obtain predictions from multiple models, the risks of under-expressing uncertainty, and the potential consequences of being wrong. When predicting quantitative future outcomes, discordant results from multiple models may be problematic for decision makers if interpreted arbitrarily. However, when combined into a multi-model ensemble, this diversity of opinions becomes a key asset and allows the ensemble to provide more accurate and reliable information about the future. The same is true for ranking alternative epidemic scenarios; whenever possible, opinions from multiple models should be solicited. Agreement from multiple independent sources builds confidence in the conclusions. However, as seen in our results, multi-model agreement on scenario ranking is by no means guaranteed. In cases where models disagree about scenario ranking, a decision maker could use vote processing methods to combine rankings from each model into a consensus, much like an ensemble combines quantitative predictions (Probert et al., 2022). There are also many settings where projections from multiple models are not available, due for example to inequities in disease modeling resources (Heesterbeek et al., 2015). In these instances, decision makers may choose to focus on qualitative model results or the ranking of model projections, rather than quantitative outcomes.

5. Conclusions

Understanding what mathematical models can effectively predict is essential to using modeling resources wisely, including during public health crises like an infectious disease outbreak or in low-resource settings where modeling teams are scarce. Multi-model ensembles are known to be an important tool to generate accurate and robust predictions of future outcomes, overcoming inconsistency and disagreement in predictions from individual models. Much of the work on multi-model ensembles has focused on predictions of quantitative outcomes (e.g., incident deaths), but quantitative outcomes are not the only type of information a decision maker may glean from model predictions. Here, we considered decision contexts that depend on the ranking of alternative epidemic scenarios. In both a simple simulation context and an empirical setting that includes 14 rounds of real-world COVID-19 projections, our results suggest that multi-model agreement may depend on the decision context, and it may thus be possible to identify decision contexts where predictions from only a few models, or possibly even a single model, may suffice. When few models are available, these are the kinds of decisions we can robustly support. Further exploring the conditions under which models disagree will be important to understand when to initiate resource-intensive, multi-model predictions, and in what circumstances we can use models to support decision making, if only a single or a few models are available.

Funding Acknowledgments

L. Wade-Malone was supported by two NSF REU grant supplements to EEID grant DEB-1911962. E. Howerton and K. Shea were supported by NSF RAPID awards DEB-2028301, DEB-2037885, DEB-2126278 and DEB-2220903 and the Huck Institutes of the Life Sciences at The Pennsylvania State University. E. Howerton was supported by the Eberly College of Science Barbara McClintock Science Achievement Graduate Scholarship in Biology at the Pennsylvania State University.

Disclaimers

Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government. The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the U.S. National Institutes of Health or Department of Health and Human Services.

CRediT authorship contribution statement

La Keisha Wade-Malone: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Conceptualization. **Emily Howerton:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Methodology, Formal analysis, Conceptualization. **William J. M. Probert:** Writing – review & editing, Validation. **Michael C. Runge:** Writing – review & editing, Validation. **Cecile Viboud:** Writing – review & editing, Validation. **Katriona Shea:** Writing – review & editing, Validation, Supervision, Funding acquisition, Conceptualization.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

All code required to reproduce the analyses in this manuscript can be found at https://github.com/eahowerton/scenario-magnitude-vs-ranking, https://zenodo.org/records/10888271. COVID-19 Scenario Modeling Hub data are provided at https://github.com/midas-network/covid19-scenario-hub_evaluation and are available on the COVID-19 Scenario Modeling Hub website, https://covid19scenariomodelinghub.org.

Acknowledgements

The authors thank Matt Biggerstaff for useful discussion on these topics, Matt Ferrari for REU project support and general feedback, and all of the U.S. COVID-19 Scenario Modeling Hub contributors.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.epidem.2024.100767.

References

Atkins, B.D., Jewell, C.P., Runge, M.C., Ferrari, M.J., Shea, K., Probert, W.J.M., Tildesley, M.J., 2020. Anticipating future learning affects current control decisions: a comparison between passive and active adaptive management in an epidemiological setting. J. Theor. Biol. 506, 110380 https://doi.org/10.1016/j.jtbi.2020.110380.

Bartko, J.J., 1976. On various intraclass correlation reliability coefficients. Psychol. Bull. 83, 762–765. https://doi.org/10.1037/0033-2909.83.5.762.
Bauer, B., Horbowy, J., Rahikainen, M., Kulatska, N., Müller-Karulis, B., Tomczak, M.T., Bartolino, V., 2019. Model uncertainty and simulated multispecies fisheries management advice in the Baltic Sea. PLOS ONE 14, e0211320. https://doi.org/10.1371/journal.pone.0211320.
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, 913–917. https://doi.org/10.1093/cid/ciab673.
Bjørnstad, O.N., 2018. Epidemics: Models and Data using R, Use R! Springer International Publishing. https://doi.org/10.1007/978-3-319-97487-3.
Borchering, R.K., Mullany, L.C., Howerton, E., Chinazzi, M., Smith, C.P., Qin, M., Reich, N.G., Contamin, L., Levander, J., Kerr, J., Espino, J., Hochheiser, H., Lovett, K., Kinsey, M., Tallaksen, K., Wilson, S., Shin, L., Lemaitre, J.C., Hulse, J.D., Kaminsky, J., Lee, E.C., Davis, J.T., Mu, K., Xiong, X., Piontti, A.P. y, Vespignani, A., Srivastava, A., Porebski, P., Venkatramanan, S., Adiga, A., Lewis, B., Klahn, B., Outten, J., Hurt, B., Chen, J., Mortveit, H., Wilson, A., Marathe, M., Hoops, S., Bhattacharya, P., Machi, D., Chen, S., Paul, R., Janies, D., Thill, J.-C., Galanti, M., Yamana, T., Pei, S., Shaman, J., Espana, G., Cavany, S., Moore, S., Perkins, A., Healy, J.M., Slayton, R.B., Johansson, M.A., Biggerstaff, M., Shea, K., Truelove, S.A., Runge, M.C., Viboud, C., Lessler, J., 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 https://doi.org/10.1016/j.lana.2022.100398.
Borchering, R.K., Viboud, C., Howerton, E., Smith, C.P., Truelove, S., Runge, M.C., Reich, N.G., Contamin, L., Levander, J., Salerno, J., van Panhuis, W., Kinsey, M., Tallaksen, K., Obrecht, R.F., Asher, L., Costello, C., Kelbaugh, M., Wilson, S., Shin, L., Gallagher, M.E., Mullany, L.C., Rainwater-Lovett, K., Lemaitre, J.C., Dent, J., Grantz, K.H., Kaminsky, J., Lauer, S.A., Lee, E.C., Meredith, H.R., Perez-Saez, J., Keegan, L.T., Karlen, D., Chinazzi, M., Davis, J.T., Mu, K., Xiong, X., Porebski, P., Venkatramanan, S., Adiga, A., Lewis, B., Klahn, B., Outten, J., Schlitt, J., Corbett, P., Telionis, P.A., Wang, L., Peddireddy, A.S., Hurt, B., Chen, J., Vullikanti, A., Marathe, M., Healy, J.M., Slayton, R.B., Biggerstaff, M., Johansson, M.A., Shea, K., Lessler, J., 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, 719–724. https://doi.org/10.15585/mmwr.mm7019e3.
Castonguay, F.M., Blackwood, J.C., Howerton, E., Shea, K., Sims, C., Sanchirico, J.N., 2023. Optimal spatial evaluation of a pro rata vaccine distribution rule for COVID-19. Sci. Rep. 13, 2194. https://doi.org/10.1038/s41598-023-28697-8.
Clemen, R.T., 1989. Combining forecasts: a review and annotated bibliography. Int. J. Forecast. 5, 559–583. https://doi.org/10.1016/0169-2070(89)90012-5.
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, Yuxin, Jayawardena, D., Kanji, A.H., Khandelwal, A., Le, K., Mühlemann, A., Niemi, J., Shah, A., Stark, A., Wang, Yijin, Wattanachit, N., Zorn, M.W., Gu, Y., Jain, S., Bannur, N., Deva, A., Kulkarni, M., Merugu, S., Raval, A., Shingi, S., Tiwari, A., White, J., Abernethy, N.F., Woody, S., Dahan, M., Fox, S., Gaither, K., Lachmann, M., Meyers, L.A., Scott, J.G., Tec, M., Srivastava, A., George, G.E., Cegan, J.C., Dettwiller, I.D., England, W.P., Farthing, M. W., Hunter, R.H., Lafferty, B., Linkov, I., Mayo, M.L., Parno, M.D., Rowland, M.A., Trump, B.D., Zhang-James, Y., Chen, S., Faraone, S.V., Hess, J., Morley, C.P., Salekin, A., Wang, D., Corsetti, S.M., Baer, T.M., Eisenberg, M.C., Falb, K., Huang, Yitao, Martin, E.T., McCauley, E., Myers, R.L., Schwarz, T., Sheldon, D., Gibson, G.C., Yu, R., Gao, Liyao, Ma, Y., Wu, D., Yan, X., Jin, X., Wang, Y.-X., Chen, Y., Guo, L., Zhao, Y., Gu, Q., Chen, J., Wang, Lingxiao, Xu, P., Zhang, W., Zou, D., Biegel, H., Lega, J., McConnell, S., Nagraj, V.P., Guertin, S.L., Hulme-Lowe, C., Turner, S.D., Shi, Y., Ban, X., Walraven, R., Hong, Q.-J., Kong, S., van de Walle, A., Turtle, J.A., Ben-Nun, M., Riley, S., Riley, P., Koyluoglu, U., DesRoches, D., Forli, P., Hamory, B., Kyriakides, C., Leis, H., Milliken, J., Moloney, M., Morgan, J., Nirgudkar, N., Ozcan, G., Piwonka, N., Ravi, M., Schrader, C., Shakhnovich, E., Siegel, D., Spatz, R., Stiefeling, C., Wilkinson, B., Wong, A., Cavany, S., España, G., Moore, S., Oidtman, R., Perkins, A., Kraus, D., Kraus, A., Gao, Z., Bian, J., Cao, W., Lavista Ferres, J., Li, C., Liu, T.-Y., Xie, X., Zhang, S., Zheng, S., Vespignani, A., Chinazzi, M., Davis, J.T., Mu, K., Pastore y Piontti, A., Xiong, X., Zheng, A., Baek, J., Farias, V., Georgescu, A., Levi, R., Sinha, D., Wilde, J., Perakis, G., Bennouna, M.A., Nze-Ndong, D., Singhvi, D., Spantidakis, I., Thayaparan, L., Tsiourvas, A., Sarker, A., Jadbabaie, A., Shah, D., Della Penna, N., Celi, L.A., Sundar, S., Wolfinger, R., Osthus, D., Castro, L., Fairchild, G., Michaud, I., Karlen, D., Kinsey, M., Mullany, L.C., Rainwater-Lovett, K., Shin, L., Tallaksen, K., Wilson, S., Lee, E.C., Dent, J., Grantz, K.H., Hill, A.L., Kaminsky, J., Kaminsky, K., Keegan, L.T., Lauer, S.A., Lemaitre, J.C., Lessler, J., Meredith, H.R., Perez-Saez, J., Shah, S., Smith, C.P., Truelove, S.A., Wills, J., Marshall, M., Gardner, L., Nixon, K., Burant, J.C., Wang, Lily, Gao, Lei, Gu, Z., Kim, M., Li, X., Wang, G., Wang, Yueying, Yu, S., Reiner, R.C., Barber, R., Gakidou, E., Hay, S.I., Lim, S., Murray, C., Pigott, D., Gurung, H.L., Baccam, P., Stage, S.A., Suchoski, B.T., Prakash, B.A., Adhikari, B., Cui, J., Rodríguez, A., Tabassum, A., Xie, J., Keskinocak, P., Asplund, J., Baxter, A., Oruc, B.E., Serban, N., Arik, S.O., Dusenberry, M., Epshteyn, A., Kanal, E., Le, L.T., Li, C.-L., Pfister, T., Sava, D., Sinha, R., Tsai, T., Yoder, N., Yoon, J., Zhang, L., Abbott, S., Bosse, N.I., Funk, S., Hellewell, J., Meakin, S.R., Sherratt, K., Zhou, M., Kalantari, R., Yamana, T. K., Pei, S., Shaman, J., Li, M.L., Bertsimas, D., Skali Lami, O., Soni, S., Tazi Bouardi, H., Ayer, T., Adee, M., Chhatwal, J., Dalgic, O.O., Ladd, M.A., Linas, B.P., Mueller, P., Xiao, J., Wang, Yuanjia, Wang, Q., Xie, S., Zeng, D., Green, A., Bien, J., Brooks, L., Hu, A.J., Jahja, M., McDonald, D., Narasimhan, B., Politsch, C., Rajanala, S., Rumack, A., Simon, N., Tibshirani, R.J., Tibshirani, R., Ventura, V., Wasserman, L., O'Dea, E.B., Drake, J.M., Pagano, R., Tran, Q.T., Ho, L.S.T., Huynh, H., Walker, J.W., Slayton, R.B., Johansson, M.A., Biggerstaff, M., Reich, N. G., 2022. Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the United States. Proc. Natl. Acad. Sci. 119, e2113561119 https://doi.org/10.1073/pnas.2113561119.
den Boon, S., Jit, M., Brisson, M., Medley, G., Beutels, P., White, R., Flasche, S., Hollingsworth, T.D., Garske, T., Pitzer, V.E., Hoogendoorn, M., Geffen, O., Clark, A., Kim, J., Hutubessy, R., 2019. Guidelines for multi-model comparisons of the impact of infectious disease interventions. BMC Med 17, 163. https://doi.org/10.1186/s12916-019-1403-9.
Egger, M., Johnson, L., Althaus, C., Schöni, A., Salanti, G., Low, N., Norris, S.L., 2018. Developing WHO guidelines: time to formally include evidence from mathematical modelling studies. F1000Research 6, 1584. https://doi.org/10.12688/f1000research.12367.2.
Field, A.P., 2005. Kendall's Coefficient of Concordance. in: Encyclopedia of Statistics in Behavioral Science. John Wiley & Sons, Ltd, Chichester, pp. 1010–1011.
Flasche, S., Jit, M., Rodríguez-Barraquer, I., Coudeville, L., Recker, M., Koelle, K., Milne, G., Hladish, T.J., Perkins, T.A., Cummings, D.A.T., Dorigatti, I., Laydon, D.J., España, G., Kelso, J., Longini, I., Lourenco, J., Pearson, C.A.B., Reiner, R.C., Mier-Y-Terán-Romero, L., Vannice, K., Ferguson, N., 2016. The Long-term safety, public health impact, and cost-effectiveness of routine vaccination with a recombinant, live-attenuated dengue vaccine (Dengvaxia): a model comparison study. PLoS Med 13, e1002181. https://doi.org/10.1371/journal.pmed.1002181.
Gamer, M., Lemon, J., Fellows, I., Singh, P., 2019. irr: Var. Coeff. Inter. Reliab. Agreem.
Heesterbeek, H., Anderson, R., Andreasen, V., Bansal, S., De Angelis, D., Dye, C., Eames, K., Edmunds, J., Frost, S., Funk, S., Hollingsworth, D., House, T., Isham, V., Klepac, P., Lessler, J., Lloyd-Smith, J., Metcalf, J., Mollison, D., Pellis, L., Pulliam, J., Roberts, M., Viboud, C., 2015. Modeling infectious disease dynamics in the complex landscape of global health. Science 347, aaa4339. https://doi.org/10.1126/science.aaa4339.
Howerton, E., Contamin, L., Mullany, L.C., Qin, M., Reich, N.G., Bents, S., Borchering, R. K., Jung, S., Loo, S.L., Smith, C.P., Levander, J., Kerr, J., Espino, J., van Panhuis, W. G., Hochheiser, H., Galanti, M., Yamana, T., Pei, S., Shaman, J., Rainwater-Lovett, K., Kinsey, M., Tallaksen, K., Wilson, S., Shin, L., Lemaitre, J.C., Kaminsky, J., Hulse, J.D., Lee, E.C., McKee, C.D., Hill, A., Karlen, D., Chinazzi, M., Davis, J.T., Mu, K., Xiong, X., Pastore y Piontti, A., Vespignani, A., Rosenstrom, E.T., Ivy, J.S., Mayorga, M.E., Swann, J.L., España, G., Cavany, S., Moore, S., Perkins, A., Hladish, T., Pillai, A., Ben Toh, K., Longini, I., Chen, S., Paul, R., Janies, D., Thill, J.-C., Bouchnita, A., Bi, K., Lachmann, M., Fox, S.J., Meyers, L.A., Srivastava, A., Porebski, P., Venkatramanan, S., Adiga, A., Lewis, B., Klahn, B., Outten, J., Hurt, B., Chen, J., Mortveit, H., Wilson, A., Marathe, M., Hoops, S., Bhattacharya, P., Machi, D., Cadwell, B.L., Healy, J.M., Slayton, R.B., Johansson, M.A., Biggerstaff, M., Truelove, S., Runge, M.C., Shea, K., Viboud, C., Lessler, J., 2023a. Evaluation of the US COVID-19 Scenario Modeling Hub for informing pandemic response under uncertainty. Nat. Commun. 14, 7260. https://doi.org/10.1038/s41467-023-42680-x.
Howerton, E., Runge, M.C., Bogich, T.L., Borchering, R.K., Inamine, H., Lessler, J., Mullany, L.C., Probert, W.J.M., Smith, C.P., Truelove, S., Viboud, C., Shea, K., 2023b. Context-dependent representation of within- and between-model uncertainty: aggregating probabilistic predictions in infectious disease epidemiology. J. R. Soc. Interface 20, 20220659. https://doi.org/10.1098/rsif.2022.0659.
Johansson, M.A., Apfeldorf, K.M., Dobson, S., Devita, J., Buczak, A.L., Baugher, B., Moniz, L.J., Bagley, T., Babin, S.M., Guven, E., Yamana, T.K., Shaman, J., Moschou, T., Lothian, N., Lane, A., Osborne, G., Jiang, G., Brooks, L.C., Farrow, D.C., Hyun, S., Tibshirani, R.J., Rosenfeld, R., Lessler, J., Reich, N.G., Cummings, D.A.T., Lauer, S.A., Moore, S.M., Clapham, H.E., Lowe, R., Bailey, T.C., García-Díez, M., Carvalho, M.S., Rodó, X., Sardar, T., Paul, R., Ray, E.L., Sakrejda, K., Brown, A.C., Meng, X., Osoba, O., Vardavas, R., Manheim, D., Moore, M., Rao, D.M., Porco, T.C., Ackley, S., Liu, F., Worden, L., Convertino, M., Liu, Y., Reddy, A., Ortiz, E., Rivero, J., Brito, H., Juarrero, A., Johnson, L.R., Gramacy, R.B., Cohen, J.M., Mordecai, E.A., Murdock, C.C., Rohr, J.R., Ryan, S.J., Stewart-Ibarra, A.M., Weikel, D.P., Jutla, A., Khan, R., Poultney, M., Colwell, R.R., Rivera-García, B., Barker, C.M., Bell, J.E., Biggerstaff, M., Swerdlow, D., Mier-Y-Teran-Romero, L., Forshey, B.M., Trtanj, J., Asher, J., Clay, M., Margolis, H.S., Hebbeler, A.M., George, D., Chretien, J.-P., 2019. An open challenge to advance probabilistic forecasting for dengue epidemics. Proc. Natl. Acad. Sci. 116, 24268–24274. https://doi.org/10.1073/pnas.1909865116.
Keeling, M.J., Rohani, P., 2008. Modeling infectious diseases in humans and animals. Princeton University Press, Princeton.
Knutti, R., Masson, D., Gettelman, A., 2013. Climate model genealogy: Generation CMIP5 and how we got there. Geophys. Res. Lett. 40, 1194–1199. https://doi.org/10.1002/grl.50256.
Li, S.-L., Bjørnstad, O.N., Ferrari, M.J., Mummah, R., Runge, M.C., Fonnesbeck, C.J., Tildesley, M.J., Probert, W.J.M., Shea, K., 2017. Essential information: uncertainty and optimal control of Ebola outbreaks. Proc. Natl. Acad. Sci. 114, 5659–5664. https://doi.org/10.1073/pnas.1617482114.
Li, S.-L., Ferrari, M.J., Bjørnstad, O.N., Runge, M.C., Fonnesbeck, C.J., Tildesley, M.J., Pannell, D., Shea, K., 2019. Concurrent assessment of epidemiological and operational uncertainties for optimal outbreak control: ebola as a case study. Proc. R. Soc. B Biol. Sci. 286, 20190774. https://doi.org/10.1098/rspb.2019.0774.
Liljequist, D., Elfving, B., Roaldsen, K.S., 2019. Intraclass correlation – A discussion and demonstration of basic features. PLoS One 14, e0219854. https://doi.org/10.1371/journal.pone.0219854.
Loo, S.L., Howerton, E., Contamin, L., Smith, C.P., Borchering, R.K., Mullany, L.C., Bents, S., Carcelen, E., Jung, S., Bogich, T.L., van Panhuis, W.G., Kerr, J., Espino, J., Yan, K., Hochheiser, H., Runge, M.C., Shea, K., Lessler, J., Viboud, C., Truelove, S., 2023. The US COVID-19 and Influenza Scenario Modeling Hubs: delivering long-

term projections to guide policy. Epidemics, 100738. https://doi.org/10.1016/j.epidem.2023.100738.
Metcalf, C.J.E., Morris, D.H., Park, S.W., 2020. Mathematical models to guide pandemic response. Science 369, 368–369. https://doi.org/10.1126/science.abd1668.
Pennell, C., Reichler, T., 2011. On the effective number of climate models. J. Clim. 24, 2358–2367. https://doi.org/10.1175/2010JCLI3814.1.
Prasad, P.V., Steele, M.K., Reed, C., Meyers, L.A., Du, Z., Pasco, R., Alfaro-Murillo, J.A., Lewis, B., Venkatramanan, S., Schlitt, J., Chen, J., Orr, M., Wilson, M.L., Eubank, S., Wang, L., Chinazzi, M., Pastore y Piontti, A., Davis, J.T., Halloran, M.E., Longini, I., Vespignani, A., Pei, S., Galanti, M., Kandula, S., Shaman, J., Haw, D.J., Arinaminpathy, N., Biggerstaff, M., 2023. Multimodeling approach to evaluating the efficacy of layering pharmaceutical and nonpharmaceutical interventions for influenza pandemics. Proc. Natl. Acad. Sci. 120, e2300590120 https://doi.org/10.1073/pnas.2300590120.
Probert, W.J.M., Jewell, C.P., Werkman, M., Fonnesbeck, C.J., Goto, Y., Runge, M.C., Sekiguchi, S., Shea, K., Keeling, M.J., Ferrari, M.J., Tildesley, M.J., 2018. Real-time decision-making during emergency disease outbreaks. PLoS Comput. Biol. 14, e1006202 https://doi.org/10.1371/journal.pcbi.1006202.
Probert, W.J.M., Nicol, S., Ferrari, M.J., Li, S.-L., Shea, K., Tildesley, M.J., Runge, M.C., 2022. Vote-processing rules for combining control recommendations from multiple models. Philos. Trans. R. Soc. Math. Phys. Eng. Sci. 380, 20210314 https://doi.org/10.1098/rsta.2021.0314.
Probert, W.J.M., Shea, K., Fonnesbeck, C.J., Runge, M.C., Carpenter, T.E., Dürr, S., Garner, M.G., Harvey, N., Stevenson, M.A., Webb, C.T., Werkman, M., Tildesley, M.J., Ferrari, M.J., 2016. Decision-making for foot-and-mouth disease control: objectives matter. Epidemics 15, 10–19. https://doi.org/10.1016/j.epidem.2015.11.002.
R Core Team, 2018. R: A language and environment for statistical computing (manual). R Foundation for Statistical Computing, Vienna, Austria.
Reich, N.G., Lessler, J., Funk, S., Viboud, C., Vespignani, A., Tibshirani, R.J., Shea, K., Schienle, M., Runge, M.C., Rosenfeld, R., Ray, E.L., Niehus, R., Johnson, H.C., Johansson, M.A., Hochheiser, H., Gardner, L., Bracher, J., Borchering, R.K., Biggerstaff, M., 2022. Collaborative hubs: making the most of predictive epidemic modeling. Am. J. Public Health 112, 839–842. https://doi.org/10.2105/AJPH.2022.306831.
Reich, N.G., McGowan, C.J., Yamana, T.K., Tushar, A., Ray, E.L., Osthus, D., Kandula, S., Brooks, L.C., Crawford-Crudell, W., Gibson, G.C., Moore, E., Silva, R., Biggerstaff, M., Johansson, M.A., Rosenfeld, R., Shaman, J., 2019. Accuracy of real-time multi-model ensemble forecasts for seasonal influenza in the U.S. PLOS Comput. Biol. 15, e1007486 https://doi.org/10.1371/journal.pcbi.1007486.
Rosenblum, H.G., 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, 1436–1441. https://doi.org/10.15585/mmwr.mm7145a2.
Runge, M.C., Shea, K., Howerton, E., Yan, K., Hochheiser, H., Rosenstrom, E., Probert, W.J.M., Borchering, R., Marathe, M.V., Lewis, B., Venkatramanan, S., Truelove, S.A., Lessler, J., Viboud, C., 2023. Scenario Design for Infectious Disease Projections: Integrating Concepts from Decision Analysis and Experimental Design. https://doi.org/10.1101/2023.10.11.23296887.
Shea, K., Borchering, R.K., Probert, W.J.M., Howerton, E., Bogich, T.L., Li, S.-L., van Panhuis, W.G., Viboud, C., Aguás, R., Belov, A.A., Bhargava, S.H., Cavany, S.M., Chang, J.C., Chen, C., Chen, J., Chen, S., Chen, Y., Childs, L.M., Chow, C.C., Crooker, I., Del Valle, S.Y., España, G., Fairchild, G., Gerkin, R.C., Germann, T.C., Gu, Q., Guan, X., Guo, L., Hart, G.R., Hladish, T.J., Hupert, N., Janies, D., Kerr, C.C., Klein, D.J., Klein, E.Y., Lin, G., Manore, C., Meyers, L.A., Mittler, J.E., Mu, K., Núñez, R.C., Oidtman, R.J., Pasco, R., Pastore y Piontti, A., Paul, R., Pearson, C.A.B., Perdomo, D.R., Perkins, T.A., Pierce, K., Pillai, A.N., Rael, R.C., Rosenfeld, K., Ross, C.W., Spencer, J.A., Stoltzfus, A.B., Toh, K.B., Vattikuti, S., Vespignani, A., Wang, L., White, L.J., Xu, P., Yang, Y., Yogurtcu, O.N., Zhang, W., Zhao, Y., Zou, D., Ferrari, M.J., Pannell, D., Tildesley, M.J., Seifarth, J., Johnson, E., Biggerstaff, M., Johansson, M.A., Slayton, R.B., Levander, J.D., Stazer, J., Kerr, J., Runge, M.C., 2023. Multiple models for outbreak decision support in the face of uncertainty. Proc. Natl. Acad. Sci. 120, e2207537120 https://doi.org/10.1073/pnas.2207537120.
Shea, K., Runge, M.C., Pannell, D., Probert, W.J.M., Li, S.-L., Tildesley, M., Ferrari, M., 2020. Harnessing multiple models for outbreak management. Science 368, 577–579. https://doi.org/10.1126/science.abb9934.
Shea, K., Tildesley, M.J., Runge, M.C., Fonnesbeck, C.J., Ferrari, M.J., 2014. Adaptive management and the value of information: learning via intervention in epidemiology. PLOS Biol. 12, e1001970 https://doi.org/10.1371/journal.pbio.1001970.
Soetaert, K., Petzoldt, T., Setzer, R.W., 2010. Solving differential equations in R: package deSolve. J. Stat. Softw. 33, 1–25. https://doi.org/10.18637/jss.v033.i09.
Timmermann, Allan, 2006. Chapter 4 Forecast Combinations. In: Elliott, G., Granger, C.W.J., Timmermann, A. (Eds.), Handbook of Economic Forecasting. Elsevier, pp. 135–196. https://doi.org/10.1016/S1574-0706(05)01004-9.
Truelove, S., Smith, C.P., Qin, M., Mullany, L.C., Borchering, R.K., Lessler, J., Shea, K., Howerton, E., Contamin, L., Levander, J., Salerno, J., Hochheiser, H., Kinsey, M., Tallaksen, K., Wilson, S., Shin, L., Rainwater-Lovett, K., Lemairtre, J.C., Dent Hulse, J., Kaminsky, J., Lee, E.C., Perez-Saez, J., Hill, A., Karlen, D., Chinazzi, M., Davis, J.T., Mu, K., Xiong, X., Pastore y Piontti, A., Vespignani, A., Srivastava, A., Porebski, P., Venkatramanan, S., Adiga, A., Lewis, B., Klahn, B., Outten, J., Orr, M., Harrison, G., Hurt, B., Chen, J., Vullikanti, A., Marathe, M., Hoops, S., Bhattacharya, P., Machi, D., Chen, S., Paul, R., Janies, D., Thill, J.-C., Galanti, M., Yamana, T.K., Pei, S., Shaman, J.L., Healy, J.M., Slayton, R.B., Biggerstaff, M., Johansson, M.A., Runge, M.C., Viboud, C., 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. https://doi.org/10.7554/eLife.73584.
Viboud, C., Sun, K., Gaffey, R., Ajelli, M., Fumanelli, L., Merler, S., Zhang, Q., Chowell, G., Simonsen, L., Vespignani, A., 2018. The RAPIDD ebola forecasting challenge: synthesis and lessons learnt. Epidemics, RAPIDD Ebola Forecast. Chall. 22, 13–21. https://doi.org/10.1016/j.epidem.2017.08.002.