

# 1 Optimizing the number of models included in outbreak 2 forecasting ensembles

3 Spencer J. Fox<sup>1,2,3,\*</sup>

4 Minsu Kim<sup>4</sup>

5 Lauren Ancel Meyers<sup>5,6,7</sup>

6 Nicholas G. Reich<sup>4</sup>

7 Evan L. Ray<sup>4</sup>

## 8 Affiliations:

9 <sup>1</sup> - Department of Epidemiology & Biostatistics, University of Georgia

10 <sup>2</sup> - Institute of Bioinformatics, University of Georgia

11 <sup>3</sup> - Center for Ecology of Infectious Diseases, University of Georgia

12 <sup>4</sup> - Department of Biostatistics and Epidemiology, University of Massachusetts Amherst

13 <sup>5</sup> - Department of Integrative Biology, University of Texas at Austin

14 <sup>6</sup> - Department of Statistics and Data Science, University of Texas at Austin

15 <sup>7</sup> - Department of Population Health, Dell Medical School

16 \* - Corresponding author: 1-706-542-9394, [sjfox@uga.edu](mailto:sjfox@uga.edu), Spencer Fox 120 B.S. Miller Hall, Health  
17 Sciences Campus, 101 Buck Road, Athens, GA 30602

## 18 Running Title

19 Optimizing ensemble size for outbreak forecasts

## 20 Keywords

21 Infectious disease forecasting, ensemble forecasting, COVID-19, influenza, hospitalizations

## 22 Abstract

23 Based on historical influenza and COVID-19 forecasts, we quantify the relationship between the number  
24 of models in an ensemble and its accuracy and introduce an ensemble approach that can outperform the  
25 current standard. Our results can assist collaborative forecasting efforts by identifying target participation  
26 rates and improving ensemble forecast performance.

## 27 Text

28 Pioneered by the Centers for Disease Control and Prevention's (CDC's) 2013-2014 Influenza  
29 Season Challenge, real-time, collaborative forecast efforts have become the gold standard for  
30 generating and evaluating forecasts for infectious disease outbreaks (1,2). Individual component  
31 forecasts are aggregated into ensemble predictions that are the primary external communication  
32 provided by the organizing hubs and have consistently outperformed individual models (3–5).  
33 The current COVID-19 and influenza ensemble forecasts use the median across all eligible  
34 forecasts for each requested target, though other strategies that weight individual forecasts based  
35 on historical performance may further improve performance (6). To assist public health decision-  
36 makers considering target participation rates and the optimal design of ensemble forecast  
37 models, we retrospectively analyzed data from recent US-based collaborative outbreak forecast  
38 efforts to identify how the number of models included in an ensemble impacts performance.

39 We analyzed forecasts from five recent public collaborative forecast efforts including  
40 forecasts for influenza-like illness (ILI) from 2010-2017 (5), for COVID-19 reported cases,  
41 hospital admissions, and mortality from 2020-2023 (7), and for influenza hospital admissions  
42 from 2021-2023 (8). For each, we identified time periods with maximal model participation,

43 created training and testing time periods, and obtained forecasts for individual, non-ensemble,  
44 models that produced at least 90% of all possible forecasts throughout those periods (Table S1).  
45 We created ensemble forecasts of size  $n \in (1, \dots, N)$ , where  $n$  is the number of individual  
46 models included in a given ensemble and  $N$  is the total number of available individual models,  
47 using three strategies: (1) randomly sampling combinations of  $n$  models (Random), (2) choosing  
48 the top individually performing  $n$  models from a training period (Individual rank), or (3)  
49 choosing the top performing ensemble of size  $n$  from a training period (Ensemble rank). We  
50 compared performance of all ensembles against a baseline model (Baseline) that produces  
51 forecasts based on historical seasonality for ILI (5) or flat forecasts for all other metrics (3), and  
52 an unweighted ensemble composed of all submitted models that is currently used in real-time as  
53 the gold standard forecast (Published ensemble). We summarized probabilistic ensemble forecast  
54 skill (Figure 1A) using the log score for ILI forecasts and the weighted interval score (WIS) for  
55 all others (9,10), and transformed scores as needed so that lower numbers indicate better  
56 performance (Figure 1B). Further methodological details are provided in the supplement.

57       When using random sampling for choosing component ensemble models, we found that  
58 for all forecasting exercises, including more models yielded better average forecast performance  
59 and all ensembles outperformed the Baseline model after the inclusion of at least four models  
60 (Figure 1B). Increasing the ensemble size above four models only slightly improved the average  
61 forecast performance, but substantially decreased the variability of performance across randomly  
62 assembled models. For example, for influenza hospital admission forecasts, increasing the  
63 number of models in the ensemble from four to seven improved the average ensemble  
64 performance by 2%, but reduced the interquartile range across possible ensembles by 56.5%. In

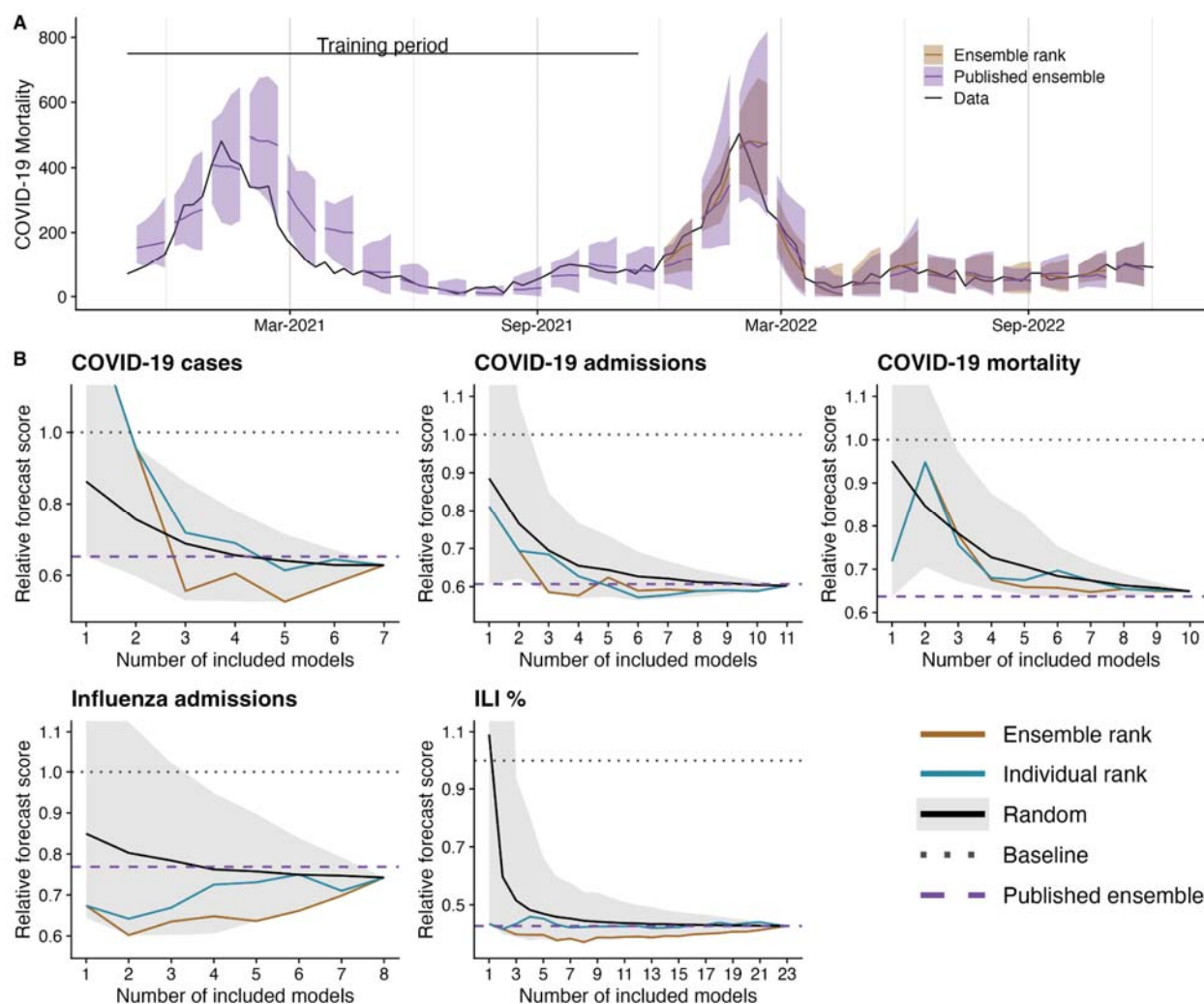
65 essence, increasing the ensemble size increases the likelihood that a randomly chosen ensemble  
66 performs well.

67 Ensemble creation based on the individual rank order performance from the training  
68 period gave mixed forecast results, while creation based on historical ensemble performance  
69 consistently selected high-performing ensembles that prospectively beat or matched the  
70 Published ensemble (Figure 1B, Table 1). For the Ensemble rank method, performance generally  
71 plateaued or declined when more than four models were included for both the testing (Figure 1)  
72 and training period (Figure S1). The Ensemble rank of size four had relative forecast  
73 performance against the Published ensemble of 0.94 and 0.84 for ILI and influenza hospital  
74 admissions, respectively, and 0.93, 0.95, and 1.06 for COVID-19 cases, hospital admissions, and  
75 mortality, respectively, where values less than 1 indicate performance improvements. While the  
76 Ensemble rank model did not always match the prediction interval coverage of the Published  
77 ensemble (Figure S2), its average rank for individual prediction tasks was always better than that  
78 of the Published ensemble (Figure S3-S7). We found that relative forecast performance is  
79 consistent when viewed across the different locations, dates, and targets (Figure S8-S16).

80

81

82



83

84 **Figure 1: Forecast performance on recent influenza and COVID-19 collaborative forecast efforts comparing**  
 85 **the number of models included in the ensemble and different ensemble methodologies. (A)** Weekly COVID-19  
 86 mortality data for Massachusetts (black line) with forecasts from the published forecast that summarizes across all  
 87 individual contributed forecasts (Published ensemble) and the best performing ensemble of size four from the  
 88 training period (Ensemble rank). The line indicates the four-week point predictions and the shaded region indicates  
 89 the 95% prediction interval for each ensemble. Ensemble rank forecasts require training and therefore are only  
 90 shown during the testing period. **(B)** Summarized ensemble forecast scores from the collaborative forecast efforts  
 91 for the weekly influenza-like illness (ILI) data provided by the CDC (ILI %), COVID-19 weekly case and mortality  
 92 counts provided by JHU (COVID-19 cases and COVID-19 mortality), and COVID-19 and Influenza daily hospital  
 93 admissions provided by HHS (COVID-19 admissions and Influenza admissions). Scores correspond to the average  
 94 forecast performance during the respective testing periods across all dates, locations, and forecast horizons (Table  
 95 S1). We plot the minimum (Grey region, lower), maximum (Grey region, upper), and mean (Solid black line) scores  
 96 of random ensemble combinations of a given size (Random), and the trained ensembles composed of the top  $n$   
 97 individual performing models from the training period (Individual rank) or the best performing ensemble of size  $n$   
 98 from the training period (Ensemble rank). All scores are standardized by the baseline forecast model for that metric  
 99 (horizontal dotted line), and the horizontal dashed line corresponds to the Published ensemble that is the unweighted  
 100 ensemble across all models that submitted for a specific date and forecast target and is used as the gold-standard  
 101 forecast prediction. Relative scores less than 1 indicate better accuracy than the Baseline. On average across the  
 102 testing phase, the Published ensemble included 15 models for COVID-19 cases, 17 models for COVID-19  
 103 admissions, 19 models for COVID-19 deaths, 21 models for influenza admissions, and 23 models for ILI.

104 **Table 1:** Forecast performance of ensembles of size four relative to the Published ensemble  
105 forecast model for each of the respective collaborative forecast efforts, where values less than 1  
106 indicate improved forecast performance. On average across the testing phase, the Published  
107 ensemble included 15 models for COVID-19 cases, 17 models for COVID-19 admissions, 19  
108 models for COVID-19 deaths, 21 models for influenza admissions, and 23 models for ILI.

	Random model (range)	Individual rank	Ensemble rank
COVID-19 Cases	1.01 (0.81-1.2)	1.06	0.93
COVID-19 Admits	1.08 (0.94-1.27)	1.03	0.95
COVID-19 Deaths	1.14 (1.02-1.37)	1.07	1.06
ILI %	1.13 (0.89-1.9)	1.08	0.94
Influenza Admits	0.99 (0.79-1.23)	0.94	0.84

109  
110  
111 While our results are constrained by a limited number of diseases, forecasting exercises, and  
112 models to draw upon, they have several implications for future collaborative forecast efforts: (1)  
113 increasing participation in collaborative forecast hubs increases model diversity, improves the  
114 average forecast performance, and decreases variability between possible ensemble  
115 combinations, (2) while optimizing ensemble forecasts based on historical performance does not  
116 guarantee optimal future performance, data-driven selection of models can improve forecast  
117 performance compared to unweighted ensembles, and (3) evaluating ensemble rather than  
118 individual performance selects for complementarity in forecasts and consistently improved  
119 forecast performance. As public health officials and researchers look to expand collaborative

120 forecast efforts and as funding agencies allocate budgets across methodological and applied  
121 forecast efforts, our results can be used to identify target participation rates, guide the  
122 interpretation and communication of ensemble forecasts, and improve forecast performance.  
123

## 124 References

- 125 1. Reich NG, Lessler J, Funk S, Viboud C, Vespignani A, Tibshirani RJ, et al. Collaborative Hubs:  
126 Making the Most of Predictive Epidemic Modeling. *Am J Public Health*. 2022 Jun;112(6):839–42.
- 127 2. Biggerstaff M, Alper D, Dredze M, Fox S, Fung ICH, Hickmann KS, et al. Results from the centers  
128 for disease control and prevention’s predict the 2013-2014 Influenza Season Challenge. *BMC Infect*  
129 *Dis*. 2016 Jul 22;16:357.
- 130 3. Cramer EY, Ray EL, Lopez VK, Bracher J, Brennen A, Castro Rivadeneira AJ, et al. Evaluation of  
131 individual and ensemble probabilistic forecasts of COVID-19 mortality in the United States. *Proc*  
132 *Natl Acad Sci U S A*. 2022 Apr 12;119(15):e2113561119.
- 133 4. Lutz CS, Huynh MP, Schroeder M, Anyatonwu S, Dahlgren FS, Danyluk G, et al. Applying  
134 infectious disease forecasting to public health: a path forward using influenza forecasting examples.  
135 *BMC Public Health*. 2019 Dec 10;19(1):1659.
- 136 5. Reich NG, Brooks LC, Fox SJ, Kandula S, McGowan CJ, Moore E, et al. A collaborative multiyear,  
137 multimodel assessment of seasonal influenza forecasting in the United States. *Proc Natl Acad Sci U*  
138 *S A*. 2019 Feb 19;116(8):3146–54.
- 139 6. Ray EL, Brooks LC, Bien J, Biggerstaff M, Bosse NI, Bracher J, et al. Comparing trained and  
140 untrained probabilistic ensemble forecasts of COVID-19 cases and deaths in the United States. *Int J*  
141 *Forecast*. 2023 Jul-Sep;39(3):1366–83.
- 142 7. Cramer EY, Huang Y, Wang Y, Ray EL, Cornell M, Bracher J, et al. The United States COVID-19  
143 Forecast Hub dataset. *Sci Data*. 2022 Aug 1;9(1):462.
- 144 8. Flusight-forecast-data [Internet]. Github; [cited 2023 Jul 12]. Available from:  
145 <https://github.com/cdcepi/Flusight-forecast-data>
- 146 9. Bracher J, Ray EL, Gneiting T, Reich NG. Evaluating epidemic forecasts in an interval format. *PLoS*  
147 *Comput Biol* [Internet]. 2021 Feb [cited 2023 Sep 6];17(2). Available from:  
148 <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7880475/>
- 149 10. Gneiting T, Raftery AE. Strictly Proper Scoring Rules, Prediction, and Estimation. *J Am Stat Assoc*.  
150 2007 Mar 1;102(477):359–78.

## 151 Acknowledgments

152 The authors acknowledge the helpful comments from the members of the CSTE, CDC, and MIDAS  
153 forecasting working groups as well as the Scenario Modeling Hub. The authors also acknowledge the  
154 Texas Advanced Computing Center (TACC) at The University of Texas at Austin for providing HPC  
155 resources that have contributed to the research results reported within this paper. URL:  
156 <http://www.tacc.utexas.edu>. SJF and LAM were supported by the Council for State and Territorial  
157 Epidemiologists (NU38OT000297) and the CDC (75D30122C14776). MK, ELR, and NGR were  
158 supported by the National Institutes of General Medical Sciences (R35GM119582) and the US CDC  
159 (1U01IP001122). The content is solely the responsibility of the authors and does not necessarily represent  
160 the official views of CSTE, CDC, NIGMS, or the National Institutes of Health.

## 161 Biographical Sketch

162 Dr. Spencer J. Fox is an Assistant Professor at the University of Georgia in the Department of  
163 Epidemiology & Biostatistics and the Institute of Bioinformatics. His research interests include statistical  
164 modeling of emerging infectious diseases and outbreak forecasting.

## 165 Conflicts of Interest

166 The authors declare no conflicts of interest.  
167  
168  
169  
170  
171



172

173

174