

1 **Title:** Challenges of COVID-19 Case Forecasting in the US, 2020-2021

2

3 **Authors:** Velma K Lopez<sup>1\*</sup>, Estee Y Cramer<sup>2</sup>, Robert Pagano<sup>3</sup>, John M Drake<sup>4</sup>, Eamon B O’Dea<sup>4</sup>, Madeline Adee  
4<sup>5</sup>, Turgay Ayer<sup>6</sup>, Jagpreet Chhatwal<sup>7</sup>, Ozden O Dalgic<sup>8</sup>, Mary A Ladd<sup>5</sup>, Benjamin P Linas<sup>9</sup>, Peter P Mueller<sup>7</sup>, Jade  
5 Xiao<sup>6</sup>, Johannes Bracher<sup>10</sup>, Alvaro J Castro Rivadeneira<sup>2</sup>, Aaron Gerding<sup>2</sup>, Tilmann Gneiting<sup>11</sup>, Yuxin Huang<sup>2</sup>,  
6 Dasuni Jayawardena<sup>2</sup>, Abdul H Kanji<sup>2</sup>, Khoa Le<sup>2</sup>, Anja Mühlemann<sup>12</sup>, Jarad Niemi<sup>13</sup>, Evan L Ray<sup>2</sup>, Ariane Stark<sup>2</sup>,  
7 Yijin Wang<sup>2</sup>, Nutcha Wattanachit<sup>2</sup>, Martha W Zorn<sup>2</sup>, Sen Pei<sup>14</sup>, Jeffrey Shaman<sup>14</sup>, Teresa K Yamana<sup>14</sup>, Samuel R  
8 Tarasewicz<sup>15</sup>, Daniel J Wilson<sup>15</sup>, Sid Baccam<sup>16</sup>, Heidi Gurung<sup>16</sup>, Steve Stage<sup>16</sup>, Brad Suchoski<sup>16</sup>, Lei Gao<sup>17</sup>, Zhiling  
9 Gu<sup>13</sup>, Myungjin Kim<sup>18</sup>, Xinyi Li<sup>19</sup>, Guannan Wang<sup>20</sup>, Lily Wang<sup>17</sup>, Yueying Wang<sup>21</sup>, Shan Yu<sup>22</sup>, Lauren Gardner<sup>23</sup>,  
10 Sonia Jindal<sup>23</sup>, Maximilian Marshall<sup>23</sup>, Kristen Nixon<sup>23</sup>, Juan Dent<sup>24</sup>, Alison L Hill<sup>23</sup>, Joshua Kaminsky<sup>24</sup>,  
11 Elizabeth C Lee<sup>24</sup>, Joseph C Lemaitre<sup>25</sup>, Justin Lessler<sup>26</sup>, Claire P Smith<sup>24</sup>, Shaun Truelove<sup>24</sup>, Matt Kinsey<sup>27</sup>, Luke  
12 C. Mullany<sup>27</sup>, Kaitlin Rainwater-Lovett<sup>27</sup>, Lauren Shin<sup>27</sup>, Katharine Tallaksen<sup>27</sup>, Shelby Wilson<sup>27</sup>, Dean Karlen<sup>28</sup>,  
13 Lauren Castro<sup>29</sup>, Geoffrey Fairchild<sup>29</sup>, Isaac Michaud<sup>29</sup>, Dave Osthus<sup>29</sup>, Jiang Bian<sup>30</sup>, Wei Cao<sup>30</sup>, Zhifeng Gao<sup>30</sup>,  
14 Juan Lavista Ferres<sup>30</sup>, Chaozhuo Li<sup>30</sup>, Tie-Yan Liu<sup>30</sup>, Xing Xie<sup>30</sup>, Shun Zhang<sup>30</sup>, Shun Zheng<sup>30</sup>, Matteo Chinazzi  
15<sup>31</sup>, Jessica T Davis<sup>31</sup>, Kunpeng Mu<sup>31</sup>, Ana Pastore y Piontti<sup>31</sup>, Alessandro Vespignani<sup>31</sup>, Xinyue Xiong<sup>31</sup>, Robert  
16 Walraven<sup>32</sup>, Jinghui Chen<sup>33</sup>, Quanquan Gu<sup>33</sup>, Lingxiao Wang<sup>33</sup>, Pan Xu<sup>33</sup>, Weitong Zhang<sup>33</sup>, Difan Zou<sup>33</sup>,  
17 Graham Casey Gibson<sup>34</sup>, Daniel Sheldon<sup>2</sup>, Ajitesh Srivastava<sup>35</sup>, Aniruddha Adiga<sup>22</sup>, Benjamin Hurt<sup>22</sup>, Gursharn  
18 Kaur<sup>22</sup>, Bryan Lewis<sup>22</sup>, Madhav Marathe<sup>22</sup>, Akhil Sai Peddireddy<sup>36</sup>, Przemyslaw Porebski<sup>22</sup>, Srinivasan  
19 Venkatramanan<sup>22</sup>, Lijing Wang<sup>37</sup>, Pragati V Prasad<sup>1</sup>, Jo W Walker<sup>1</sup>, Alexander E Webber<sup>1</sup>, Rachel B Slayton<sup>1</sup>,  
20 Matthew Biggerstaff<sup>1</sup>, Nicholas G Reich<sup>2</sup>, Michael Johansson<sup>1</sup>

21

22

23 **Affiliations:**

24 <sup>1</sup> COVID-19 Response, Centers for Disease Control and Prevention, Atlanta, GA, United States of America

25 <sup>2</sup> University of Massachusetts, Amherst, Amherst, MA, United States of America

26 <sup>3</sup> Unaffiliated, Tucson, AZ, United States of America

27 <sup>4</sup> University of Georgia, Athens, GA, United States of America

28 <sup>5</sup> Massachusetts General Hospital, Boston, MA, United States of America

29 <sup>6</sup> Georgia Institute of Technology, Atlanta, GA, United States of America

30 <sup>7</sup> Massachusetts General Hospital, Harvard Medical School, Boston, MA, United States of America

31 <sup>8</sup> Value Analytics Labs, Boston, MA, United States of America

32 <sup>9</sup> Boston University School of Medicine, Boston, MA, United States of America

33 <sup>10</sup> Chair of Econometrics and Statistics, Karlsruhe Institute of Technology, Karlsruhe, Germany

34 <sup>11</sup> Heidelberg Institute for Theoretical Studies, Heidelberg, Germany

35 <sup>12</sup> Institute of Mathematical Statistics and Actuarial Science, University of Bern, Bern, Switzerland

36 <sup>13</sup> Iowa State University, Ames, IA, United States of America

37 <sup>14</sup> Mailman School of Public Health, Columbia University, New York, New York, United States of America

38 <sup>15</sup> Federal Reserve Bank of San Francisco, San Francisco, CA, United States of America

39 <sup>16</sup> IEM, Bel Air, MD, United States of America

40 <sup>16</sup> IEM, Baton Rouge, LA, United States of America

41 <sup>17</sup> George Mason University, Fairfax, VA, United States of America

42 <sup>18</sup> Kyungpook National University, Bukgu, Daegu, Republic of Korea

43 <sup>19</sup> Clemson University, Clemson, SC, United States of America

44 <sup>20</sup> College of William & Mary, Williamsburg, VA, United States of America

45 <sup>21</sup> Amazon, Seattle, WA, United States of America

- 46 <sup>22</sup> University of Virginia, Charlottesville, VA, United States of America
- 47 <sup>23</sup> Johns Hopkins University, Baltimore, MD, United States of America
- 48 <sup>24</sup> Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, United States of America
- 49 <sup>25</sup> École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
- 50 <sup>26</sup> UNC Gillings School of Global Public Health, Chapel Hill, NC, United States of America
- 51 <sup>27</sup> Johns Hopkins University Applied Physics Lab, Baltimore, MD, United States of America
- 52 <sup>28</sup> University of Victoria, Victoria, BC, Canada
- 53 <sup>29</sup> Los Alamos National Laboratory, Los Alamos, NM, United States of America
- 54 <sup>30</sup> Microsoft, Redmond, WA, United States of America
- 55 <sup>31</sup> Laboratory for the Modeling of Biological and Socio-technical Systems, Northeastern University, Boston,  
56 MA, United States of America
- 57 <sup>32</sup> Unaffiliated, Davis, CA, United States of America
- 58 <sup>33</sup> University of California, Los Angeles, Los Angeles, CA, United States of America
- 59 <sup>34</sup> Los Alamos National Lab, Los Alamos, NM, United States of America
- 60 <sup>35</sup> University of Southern California, Los Angeles, CA, United States of America
- 61 <sup>36</sup> Discreet Labs, Raleigh, NC, United States of America
- 62 <sup>37</sup> New Jersey Institute of Technology, Newark, NJ, United States of America
- 63
- 64 *\*Corresponding author: oko8@cdc.gov (VKL)*

## 65 **Abstract**

66 During the COVID-19 pandemic, forecasting COVID-19 trends to support planning and response was a priority for  
67 scientists and decision makers alike. In the United States, COVID-19 forecasting was coordinated by a large  
68 group of universities, companies, and government entities led by the Centers for Disease Control and Prevention  
69 and the US COVID-19 Forecast Hub (<https://covid19forecasthub.org>). We evaluated approximately 9.7 million  
70 forecasts of weekly state-level COVID-19 cases for predictions 1-4 weeks into the future submitted by 24 teams  
71 from August 2020 to December 2021. We assessed coverage of central prediction intervals and weighted  
72 interval scores (WIS), adjusting for missing forecasts relative to a baseline forecast, and used a Gaussian  
73 generalized estimating equation (GEE) model to evaluate differences in skill across epidemic phases that were  
74 defined by the effective reproduction number. Overall, we found high variation in skill across individual models,  
75 with ensemble-based forecasts outperforming other approaches. Forecast skill relative to the baseline was  
76 generally higher for larger jurisdictions (e.g., states compared to counties). Over time, forecasts generally  
77 performed worst in periods of rapid changes in reported cases (either in increasing or decreasing epidemic

78 phases) with 95% prediction interval coverage dropping below 50% during the growth phases of the winter  
79 2020, Delta, and Omicron waves. Ideally, case forecasts could serve as a leading indicator of changes in  
80 transmission dynamics. However, while most COVID-19 case forecasts outperformed a naïve baseline model,  
81 even the most accurate case forecasts were unreliable in key phases. Further research could improve forecasts  
82 of leading indicators, like COVID-19 cases, by leveraging additional real-time data, addressing performance  
83 across phases, improving the characterization of forecast confidence, and ensuring that forecasts were coherent  
84 across spatial scales. In the meantime, it is critical for forecast users to appreciate current limitations and use a  
85 broad set of indicators to inform pandemic-related decision making.

86

## 87 **Author Summary**

88 As SARS-CoV-2 began to spread throughout the world in early 2020, modelers played a critical role in predicting  
89 how the epidemic could take shape. Short-term forecasts of epidemic outcomes (for example, infections, cases,  
90 hospitalizations, or deaths) provided useful information to support pandemic planning, resource allocation, and  
91 intervention . Yet, infectious disease forecasting is still a nascent science, and the reliability of different types of  
92 forecasts is unclear. We retrospectively evaluated COVID-19 case forecasts, which were often unreliable. For  
93 example, forecasts did not anticipate the speed of increase in cases in early winter 2020. This analysis provides  
94 insights on specific problems that could be addressed in future research to improve forecasts and their use.  
95 Identifying the strengths and weaknesses of forecasts is critical to improving forecasting for current and future  
96 public health responses.

97

## 98 **Introduction**

99 Predicting the trajectory of an epidemic to support control and mitigation planning is the primary objective of  
100 infectious disease forecasting. To this end, large-scale, collaborative forecasting efforts across multiple disease

101 systems, such as influenza (1–3), dengue (4), West Nile (5), and Ebola viruses (6), have been integrated into  
102 routine public health workflows and emergency response (7). Researchers in academia, private institutions, and  
103 the United States (US) government built upon these frameworks to incorporate forecasts into the COVID-19  
104 information systems used to inform pandemic response and created the US COVID-19 Forecast Hub. In April  
105 2020, the US Centers for Disease Control and Prevention (CDC) and the COVID-19 Forecast Hub began collecting  
106 COVID-19 death forecasts (8). Compared to death reports, case reports are a leading indicator of SARS-CoV-2  
107 infections, as the time from infection to case report is typically shorter than that between infection and death  
108 report. Hence, information gleaned from case forecasts is potentially more actionable.

109

110 Case forecasts for all US counties (n=3,143), states (n=50), territories (n=5), the District of Columbia (DC), and  
111 the nation as a whole were generated and collected beginning in July 2020, with ensemble forecasts of cases  
112 first posted on a CDC webpage on August 6, 2020 (8,9). Because of their potential utility, case forecasts were  
113 also integrated into US government web pages and situational awareness updates (10). In addition, county-level  
114 case forecasts were used to inform vaccine trial site selection (11) and COVID-19 case forecasts have been cited  
115 as useful for guiding personal risk-based decisions (12). Because these forecasts influence policies and personal  
116 decisions, accuracy and precision of the forecasts is of the utmost importance. Incorrect forecasts can lead to  
117 inappropriate policy implementation and resource allocation, and also to erosion of trust in public health  
118 institutions (13).

119

120 As part of routine use of the case forecasts in the COVID-19 response, real-time evaluation was conducted. One  
121 of the performance metrics included in the evaluation was the 95% prediction interval (PI) coverage, an estimate  
122 of the frequency at which the interval captures the eventually observed data. The 95% PI of a reliable forecast  
123 should capture eventually reported cases 95% of the time. However, the real-time evaluation indicated that case  
124 forecasts were not always reliable, with much lower 95% PI coverage than expected (14). For example, in  
125 November 2020 as the 2020-2021 winter wave began, the 95% PI coverage for all states and territories was less

126 than 50% for even the shortest, 1-week ahead forecasts from the ensemble – generally the most reliable  
127 forecast. Repeated periods of low coverage during subsequent surges led CDC to stop posting COVID-19 case  
128 forecasts in December 2021. Though these forecasts showed poor performance, there are opportunities to  
129 develop more precise and reliable future predictions.

130

131 Evaluation of forecast performance provides an opportunity not only to assess prediction skill for the purposes  
132 of improving forecasts, but also to assess the reliability of the forecasts and foster transparency between  
133 forecast users and creators. While evaluation is recommended in forecasting research guidelines (i.e., EPFORGE  
134 2020 (15), a systematic review of COVID-19 models showed that half of published models did not include  
135 probabilistic predictions and that approximately one-fourth of published models did not include performance  
136 evaluations (16). We have previously evaluated forecast performance of cumulative (17) and incident (18)  
137 COVID-19 deaths submitted to the COVID-19 Forecast Hub. Given that an ensemble of submitted models  
138 provided consistently accurate probabilistic forecasts at different scales in both evaluations, here we apply  
139 similar methods to assess the prediction skill of the COVID-19 case forecasters, with particular interest in the  
140 COVIDhub ensemble model (that is, a model that combine predictions from forecasts submitted to the Forecast  
141 Hub). Specifically, we analyze prediction interval coverage and other aspects of nearly 10 million individual  
142 forecasts collected by the COVID-19 Forecast Hub for US jurisdictions between July 2020 and December 2021,  
143 the full period over which COVID-19 case forecasts were published by the CDC. We analyze relative forecast  
144 performance across spatial scales and phases of the pandemic to identify limitations and opportunities for  
145 future improvement of case forecasts.

146

## 147 **Results**

### 148 **Summary of Included Team Forecasts**

149 A total of 14,960,171 forecasts were submitted by 67 teams throughout the analysis period (see Supporting  
150 Information [S] 1 for submission patterns over time). Because forecasts were submitted at multiple geographic  
151 scales, we stratified analyses for 1) national forecasts, 2) state (including all 50 states), territory (US Virgin  
152 Islands and Puerto Rico), and DC forecasts), 3) county level forecasts (include all 3,143 counties and county  
153 equivalentents), split into five equal sized groups based on county population size.

154

155 We first evaluated forecasts for inclusion criteria based on numbers of locations, horizons, and time periods  
156 forecast with the same model. Briefly, teams were included if they submitted the full range of required  
157 quantiles, included at least 50 of states/territories/DC or 75% of counties, and produced forecasts at least four  
158 weeks into the future for at least 50% of the time points in the study period. At the national level, 22 sets of  
159 team forecasts met these criteria (5,136 forecasts across dates and forecast horizons), 23 sets of team forecasts  
160 met the state/territory level criteria (280,132 forecasts across jurisdictions, dates, and forecast horizons), and 15  
161 sets of team forecasts met the county-level criteria (9,415,460 forecasts across counties, dates, and forecast  
162 horizons). Overall, 64.8% of all submitted forecasts were included in the analysis (9,700,728 forecasts). Of the  
163 included forecasts, 11 sets of team forecasts met the inclusion criteria for analyzing submissions across all  
164 geospatial scales (8,125,220 forecasts for specific locations, date and forecast horizon).

165

166 Each team included in the analysis submitted forecasts that were generated from unique model structures, data  
167 inputs, and assumptions (S1). Two naïve models (the COVIDhub-baseline and CEID-Walk) and four ensemble  
168 models (the COVIDhub-4\_week\_ensemble, the COVIDhub-trained\_ensemble, LNQ-ens1, and UVA-Ensemble),  
169 which combined multiple forecasts into one, were included in the 26 models evaluated (see S1 Table 1.1). The  
170 COVIDhub-baseline model projects the number of reported cases in the most recent week as the median  
171 predicted value for the next 4 weeks. CEID-Walk is a random walk model with a simple method for removing

172 outliers. A total of seven models included data on COVID-19 hospitalizations, 12 models incorporated  
173 demographic data, and seven models used mobility data. Of the 26 evaluated models, three assumed that social  
174 distancing and other behavioral patterns changed during the prediction period.

175

176 The evaluation period consisted of 1-4 week ahead forecasts submitted in the 73 weeks from July 28, 2020  
177 through December 21, 2021. Multiple phases of the US epidemic were included: the late summer 2020 increase  
178 in several locations, a large late-fall/early-winter surge in 2020/2021, the rise and fall of the Delta variant in the  
179 summer and fall of 2021, and the early phase of the Omicron variant's dominance in winter 2021 (Figure 1A).  
180 Performance of the national ensemble forecasts varied over this period (Figure 1B). For some forecasts, the  
181 median predictions were close to the cases eventually reported, and most reported numbers fell within the 95%  
182 PIs. However, forecasts made at other times, such as January 2021 or December 2021, diverged widely from the  
183 reported data. At those times, the forecasts missed substantial decreases and increases, respectively, with  
184 reported cases falling within the 95% prediction interval for only 1-week ahead forecasts.

185

186 **Figure 1. Weekly incident reported COVID-19 cases per 100K population, nationally (in black) and per**  
187 **state/territory/DC (in gray), over time in panel A. Panel B shows a subset of COVIDhub-4\_week\_ensemble**  
188 **forecasts (in green) over time, with the median predictions represented as lines and points and the 95%**  
189 **prediction intervals in bands. Reported incident cases (counts per week) are shown in gray. In both plots, the**  
190 **black, dashed vertical line shows the date that public communication of the case forecasts was paused.**

191

## 192 **Aggregate performance**

193 We evaluated aggregate forecast performance with two metrics: Weighted Interval Score (WIS), a proper score  
194 considering both precision and accuracy, and prediction interval coverage, an indicator of forecast uncertainty.

195 Lower WIS values reflect forecasts with probability mass closer to observed values. We assessed scaled pairwise

196 WIS relative to the baseline model (referred to throughout as relative WIS, or rWIS) for national and  
197 state/territory/DC forecasts (Figure 2). A rWIS less than one indicates performance that is better than the  
198 baseline model.

199

200 **Figure 2: Percent of weeks with complete submissions for all sets of team forecasts, scaled, pairwise relative**  
201 **Weighted Interval Score (rWIS; see *Methods* for description), observed 95% prediction interval coverage, by**  
202 **geographical scale of submitted forecasts. Teams are sorted by increasing state/territory/DC rWIS values.**

203

204 Overall, seven of 22 team's forecast models outperformed the COVIDhub-baseline model at the  
205 state/territory/DC level (i.e., had rWIS values less than 1.0), and 11 outperformed the baseline model at the  
206 national level. Six of these teams outperformed the baseline model at both scales: LNQ-ens1, COVIDhub-  
207 4\_week\_ensemble, USC-SI\_kJalpha, LANL-GrowthRate, Microsoft-DeepSTIA, and CU-select.

208

209 PI coverage at the 95% level should be close to 95% for well calibrated forecasts. However, it was lower for most  
210 sets of team forecasts, with only one (LNQ-ens1) having coverage of at least 90% at all scales, while others were  
211 as low as 23%. PI coverage at 50% and 80% levels were also well below nominal levels for most sets of team  
212 forecasts, including the COVIDhub-4\_week\_ensemble (Figure 3). For the 50% prediction interval, no sets of team  
213 forecasts had coverage better than 36% at any scale. Only two sets of team forecasts had better coverage than  
214 the COVIDhub-4\_week\_ensemble for the geographic scales in which they submitted forecasts: LNQ-ens1 (all  
215 scales) and JHU\_UNC\_GAS-StatMechPool (state/territory/DC and large county levels).

216

217 **Figure 3: Expected and observed coverage rates for central 50%, 80% and 95% prediction intervals aggregated**  
218 **over time and horizon for national forecasts (panel A), state/territory/DC forecasts (panel B), the largest**  
219 **county forecasts (panel C). The dashed line represents optimal expected-coverage. Team forecasts that had**



220 **closer to nominal coverage than the COVIDhub-4\_week\_ensemble model at all three coverage levels are**  
221 **labeled on the right side of the plots.**

222

223 Forecast skill also showed distinct patterns across jurisdictional scales, with rWIS decreasing for larger  
224 jurisdiction scales (e.g., national vs. state/territory) or population sizes (e.g., larger counties vs. smaller counties,  
225 Figure 4) for most sets of team forecasts. In contrast to this general trend, for three sets of team forecasts, that  
226 pattern was inverted, one team had no distinct pattern, and the COVIDhub-4\_week\_ensemble had markedly  
227 consistent rWIS across all scales. Consistent with the aggregate findings, both LNQ-ens1 and COVIDhub-  
228 4\_week\_ensemble had rWIS lower than 1.0 at all scales, while LANL-GrowthRate had rWIS greater than 1.0 for  
229 smaller counties.

230

231 **Figure 4: Scaled, pairwise relative Weighted Interval Score (rWIS) (see *Methods* for description) by spatial**  
232 **scale for sets of team forecasts that submitted forecasts for the US nation, states/territories/DC, and all US**  
233 **counties. WIS is averaged across all horizons. The COVIDhub-baseline model has, by definition, a rWIS of 1**  
234 **(horizontal dashed line). Teams are ordered by increasing state/territory/DC rWIS with the most accurate**  
235 **model on the left. Points for each team are staggered horizontally to show overlapping WIS values.**

236

### 237 **Performance across jurisdictions**

238 There was additional variability in forecast skill between jurisdictions. Only two team forecasts (LNQ-ens1 and  
239 COVIDhub-4\_week\_ensemble) performed as well as or better than the baseline for all included states and  
240 territories (Figure 5). Variation was higher between team forecasts than between specific jurisdictions, but the  
241 baseline model tended to outperform more models in some jurisdictions (e.g., the baseline was better in  
242 Colorado, Kansas, Puerto Rico) than in others (e.g., the baseline was worse in Mississippi, South Carolina, West  
243 Virginia).

244

245 **Figure 5: Scaled, pairwise relative Weighted Interval Score (rWIS; see *Methods* for description) by location for**  
246 **national and state/territory/DC forecasts, averaged across all horizons through the entire analysis period.**  
247 **National estimates are displayed first, followed by jurisdictions in alphabetical order. Team forecasts are**  
248 **ordered by increasing average state/territory/DC rWIS.**

249

## 250 **Performance over time**

251 While rWIS varied between team forecasts and jurisdictions, it varied even more over time (Figure 6). For  
252 example, all models had relatively high WIS in December 2020-January 2021 and low WIS in June 2021.  
253 Prediction interval coverage also varied between teams and over time, with most team forecasts exhibiting  
254 times of low coverage. Across most time points, the baseline model outperformed many team forecasts,  
255 including the COVIDhub-4\_week\_ensemble, though the ensemble more often outperformed the baseline in  
256 both metrics at the national, state/territory, and large county scales. Increased WIS and decreased prediction  
257 interval coverage generally occurred with increasing case counts, such as in the fall of 2020 and summer of  
258 2021. The worst performance was in the early Omicron wave in the winter of 2021. For the last set of ensemble  
259 forecasts posted by CDC in December 2021 ([https://www.cdc.gov/coronavirus/2019-](https://www.cdc.gov/coronavirus/2019-ncov/science/forecasting/forecasts-cases.html)  
260 [ncov/science/forecasting/forecasts-cases.html](https://www.cdc.gov/coronavirus/2019-ncov/science/forecasting/forecasts-cases.html)), the WIS reached the highest level ever for all scales and the  
261 reported case numbers were outside the 95% prediction interval for most locations at every forecast horizon.

262

263 **Figure 6: Forecast accuracy over time, aggregated by geographic units, forecast horizon, and prediction date.**  
264 **Panels A-C show average Weighted Interval Score (WIS); panels D-F show 95% prediction interval coverage.**  
265 **The black, dashed vertical line in all panels shows the date that public communication of the case forecasts**  
266 **was paused. The black, dashed horizontal line in panels D-F shows nominal 95% interval coverage. National**

267 **level forecasts are presented in A and D, state/territory/DC forecasts in B and E and large county level**  
268 **forecasts in C and F.**

269

270 To further investigate these temporal patterns in performance, we first classified each forecast week as  
271 *increasing, peak, decreasing, or nadir* based on the estimated time-varying reproduction number for that given  
272 week and jurisdiction. We then fitted Gaussian generalized estimating equations (GEE) models for each set of  
273 team forecasts, using a normalized, log transformed WIS value per forecast time and location as the model  
274 outcome. The regression models were adjusted for each prediction horizon and included a natural spline with  
275 two degrees of freedom for the time/state reported case counts to adjust for intrinsic increases in WIS due to  
276 higher values in reported cases (see S6). In agreement with the aggregated results (Figure 2), we found that the  
277 expected WIS at the mean number of case counts across all jurisdictions was lower than the baseline for the  
278 better performing models (6 team forecasts and the ensemble) and higher than the baseline for others (8 team  
279 forecasts).

280

281 Forecasts skill also varied across epidemic phases (Figure 7B). Compared to the baseline model across all phases,  
282 overall skill for most models was better in nadir and peak phases and worse in increasing and decreasing phases.  
283 LNQ-ens1 and the COVIDhub ensemble outperformed the baseline model in all epidemic phases between  
284 August 1, 2020 and January 15, 2022, while several other team models outperformed the baseline in some  
285 phases.

286

287 **Figure 7. Estimated marginal mean Weighted Interval Score (WIS) and 95% confidence intervals for mean**  
288 **cases from team-specific GEE models for all 51 jurisdictions (Panel A). The 95% confidence intervals for the**  
289 **COVIDhub-baseline model are shown in dashed red vertical lines. Panel B presents each team's estimated**  
290 **marginal mean WIS per phase, scaled to the COVIDhub-baseline model's estimated marginal mean WIS for all**  
291 **epidemic phases. Teams with higher estimated marginal mean WIS values (i.e., greater than 1.0) are**

292 **presented in shades of orange while teams with lower estimated marginal mean WIS (i.e., less than 1.0) are**  
293 **shown in shades of green. Forecasts for a team in a particular phase are marked with an asterisk (\*) if the 80%**  
294 **confidence interval of the expected WIS outcome (normalized and on the log scale) was estimated by a model**  
295 **to be lower than the expected WIS of the COVIDhub-baseline model for all phases.**

296

297 To examine whether our results were affected by reporting anomalies, we also conducted sensitivity analyses  
298 for data revisions, when data were revised at a later date, and for outlier data points, when reported cases were  
299 outside of weekly expected ranges (see S2). We first identified weeks in which revised case counts as of April 2,  
300 2022 differed from the case counts initially reported for that week, excluded them from the dataset, and reran  
301 the GEE models. With this partial dataset, the results were essentially unchanged. Next, we identified outliers as  
302 reported case counts outside of the expected range by at least two of the three following algorithms: a rolling  
303 median, a seasonal trend decomposition, and a seasonal trend decomposition without a seasonality term, each  
304 method over a 21-day window. Approximately 3% of weeks (686 of 27,489 total week-location combinations in  
305 the analysis period) had at least one day of reported cases identified as an outlier. We then excluded the weeks  
306 with outliers and the week following an outlier and reran the GEE models. This sensitivity analysis had  
307 comparable results to the models with the full data (see S2 Figure 2.3, Panel A.).

308

## 309 **Discussion**

310 We evaluated performance of 9.7 million COVID-19 case forecasts at multiple geospatial scales in the US over  
311 approximately a year and a half. Real-time analyses and those presented here revealed important limitations in  
312 these forecasts. Forecast prediction intervals were largely over-confident, that is, prediction interval coverage  
313 was lower than the nominal value, particularly when case numbers were changing rapidly and forecasts could  
314 have been most useful. Few team forecasts outperformed a relatively simple and minimally informative baseline  
315 model. Forecast skill degraded for smaller geographic scales where forecasts could potentially be most useful.

316 Forecast skill was also lowest when case counts were changing the most, in phases of increasing or decreasing  
317 transmission. These limitations of case forecasts indicate key areas for improvement and important reasons to  
318 use case forecasts with caution.

319

320 Several technical challenges for forecasts were evident in these analyses. First, cases are a relatively early  
321 indicator of transmission, with no clear leading signal in traditional public health surveillance data (e.g., unlike  
322 for death forecasts, where case counts themselves can provide information for predicting future deaths). While  
323 non-traditional data sources may provide a useful predecessor to changing population case counts, the evidence  
324 from previous work is unclear. For example, internet searches, medical claims, and online surveys have been  
325 used to modestly improve case forecast accuracy relative to models without those data (19). Estimating case  
326 counts using both wastewater and clinical surveillance data has shown mixed results (20–23). Additional  
327 integration of temporal dynamics could also be helpful. The case forecasts analyzed here were developed and  
328 evaluated based on the date when cases were reported, not when individuals were infected, became ill, sought  
329 care, or were tested. Additional detail on those dates could enable models to better capture the current  
330 dynamics using nowcasting approaches giving earlier signals of change.

331

332 Second, and likely related to the challenge of cases being an early indicator, the models had substantial variation  
333 in skill between epidemic phases. In general, forecast skill was worst for the increasing phase followed by the  
334 decreasing phase. In many of these periods of low performance (e.g., the 2020-2021 winter, Delta, and Omicron  
335 waves), the COVIDhub ensemble predicted possible or probable increases or decreases, but not at the rate that  
336 actually occurred. This effect may be even stronger than our results show as they rely on a comparison to the  
337 baseline which, by definition, does not predict change. While epidemic phase is unknown in real time, it too can  
338 be estimated, and these results and others suggest that accounting for epidemic phase when making predictions  
339 could improve the forecast skill of ensemble models (24,25). Additional data, as discussed above, or model  
340 components associated with distinct phases could also help improve predictive capabilities. Seasonal changes in

341 transmission biology and human behavior, emergence of variants, and changing mitigation behavior all  
342 contribute to transmission dynamics. While some forecasting models incorporate seasonality and variants,  
343 integration of human behavior to characterize the link between behavior and transmission has lagged (13,26–  
344 28). Ensemble approaches offer another opportunity to mitigate phase-specific differences. Team modeling skill  
345 across phases was highly heterogeneous, but two ensemble approaches were better than the baseline in all  
346 phases.

347

348 Another challenge across most forecasts was overconfidence, a pattern seen with other infectious disease  
349 forecasts (4,18). The baseline model predicted a flat trend, yet it outperformed many sets of team forecasts in  
350 the increasing and decreasing phase only because its predictions had high uncertainty around that flat trend.  
351 The COVIDhub ensemble performance, on the other hand, benefitted by combining uncertainty across multiple  
352 models, yet, like the constituent models, also exhibited overconfidence. The temporal and phase-specific  
353 analyses suggest that it is, during rapid increases and decreases, that model overconfidence is most pronounced.  
354 Previous infectious disease forecasting work has shown that ensembles tend to have wider prediction intervals  
355 that are more likely to capture the eventually reported outcome and thus reduce overconfidence compared to  
356 their constituent models (4,18). Wider prediction intervals, reflecting increased uncertainty, can mediate some  
357 impacts of overconfidence. However, forecasts would be most useful if they were both reliable and informative -  
358 that is, if they could accurately capture the uncertainty, while also providing more precise estimates, rather than  
359 merely increased uncertainty (29,30).

360

361 Finally, while forecasts would be most actionable at local scales, performance was generally worse for smaller  
362 than larger jurisdictions. Other infectious disease forecasting systems have found better forecast skill at smaller  
363 geographic scales, likely because local transmission dynamics (e.g., a county) are a better predictor of local than  
364 aggregate transmission (e.g., a state) (31). We compared WIS across scales by comparison to the baseline model  
365 to adjust for missing forecasts and for WIS scaling relative to the magnitude of observed outcomes. After those

366 adjustments, population size had a clear association with forecast , likely reflecting the relative role of stochastic  
367 dynamics. For better local forecasts, models may need to explicitly account for stochasticity. Forecasts could  
368 also be improved by better leveraging spatial information, such as dynamics in neighboring counties or nearest  
369 urban centers. Local forecasts remain a key public health need, as local forecasts are more likely to reflect local  
370 conditions and motivate local mitigation action.

371

372 Overall, these findings, as well as the real-time evaluations, indicated that COVID-19 case forecasts were not  
373 reliable as a single indicator for pandemic response of a novel pathogen. Similar to other forecasting studies, we  
374 found that the ensemble was among the most reliable forecasts (3,4,18,32), outperformed only by LNQ-ens1  
375 across the metrics evaluated here. Thus, while the overall best forecasts had poor performance at key times,  
376 other forecasts were often even worse at these same time points. Weighted (or trained) ensembles offer  
377 another potential avenue for improvement (33–35), but the version implemented here did not outperform the  
378 simple, median ensemble, likely reflecting limited historical data (36) and variation in team forecast submissions  
379 (37,38).

380

381 While COVID-19 deaths are a more lagging indicator of infections than case reports, and so may be less useful as  
382 an input to public health decision making, forecasts of deaths have generally been more reliable (18). Similarly,  
383 COVID-19 hospitalization forecasts in France have also shown high forecast skill (39). Better performing US death  
384 and French hospitalization forecasts share one factor in common: models generally used local case reports as an  
385 input to inform their forecasts. While public health decision making should not rely on case forecasts alone, they  
386 may still be helpful in the context of other important indicators, such as the case, hospitalizations, and death  
387 reports. Nowcasts of reports and real-time estimates of the effective reproductive number can also provide  
388 insight on current dynamics (40–43). Together, a suite of indicators is more informative for outbreak response  
389 than a leading indicator alone.

390

391 The analysis presented here includes important findings about real-time applied forecasting in an emerging  
392 pandemic to inform pandemic response rather than to address specific research aims of improving predictions.  
393 Several factors limit the strength of our findings and ability to understand underlying mechanisms of predictive  
394 performance. Notably, we compared the forecasts to a changing record of reported cases. Throughout the  
395 COVID-19 outbreak, cases have been reported with jurisdiction- and time-varying delays and have been revised  
396 over time, resulting in varying forecast targets. In addition, the definition of a reported COVID-19 case also  
397 changed over time and varied between states. These changes were a result of many factors, including laboratory  
398 capacity and implementation of home-based testing, and may have affected forecast skill in other ways. Our  
399 sensitivity analyses found no qualitative differences in our main findings when we excluded forecasts for time  
400 points with revised data or when we excluded outlier data points. Nevertheless, forecasting teams were greatly  
401 impacted by the evolving landscape of COVID-19 case surveillance. More timely and consistent reports likely  
402 would improve both the process of making forecasts and forecast skill.

403

404 The overall goal of the COVID-19 Forecast Hub was to provide forecasts in near real-time for decision making.  
405 While the collaborative efforts of the Hub achieved this goal despite a changing epidemic landscape,  
406 nevertheless, the open nature of COVID-19 forecasting also limits understanding the drivers of forecast  
407 performance. Many teams participated at different times, some intermittently, and provided varied and limited  
408 descriptions of their forecast methods. While we were able to adjust our evaluation for differences in in varying  
409 submissions, we are unable to assess the underlying impact of modeling approaches on performance since we  
410 do not have the granular details on forecast methods and how they evolved over time for all team forecasts. For  
411 example, the LNQ-ens1, which outperformed all other forecasts by most metrics, only submitted forecasts for  
412 approximately two thirds of the analysis period and stopped in June 2021 (prior to the Delta wave). The model is  
413 described as a combination of three machine learning models, leveraging other embedded models and datasets,  
414 with weights that “are chosen by hand each week based on performance in the previous week” (see LNQ-ens1  
415 metadata, <https://github.com/reichlab/covid19-forecast->



416 [hub/blob/b12f916abc859bf59ea584b64f53afc2982042fd/data-processed/LNQ-ens1/metadata-LNQ-ens1.txt](https://hub.blob/b12f916abc859bf59ea584b64f53afc2982042fd/data-processed/LNQ-ens1/metadata-LNQ-ens1.txt), at  
417 (44)). The ensemble approach used in the LNQ-ens1 model building likely contributed to the overall  
418 performance. However, several other ensemble models had lower performance than the LNQ-ens1 model; we  
419 are unable to assess whether LNQ-ens1 performance gains were due to a particular component model or  
420 dataset, the hand weighting procedure, or something else. The brief descriptions submitted to the COVID-19  
421 Forecast Hub, such as for the LNQ-ens1, must include a summary of the methods used and may indicate a  
422 variety of unique features such as input data, parameters, model fitting, etc. (44). However, the level of detail  
423 provided in these descriptions varies between teams, and we do not have enough information to determine  
424 which aspects of individual models were important determinants of forecast performance. To elucidate  
425 associations between modeling approaches and forecast skill, additional research is needed. Future work to  
426 support improved forecasting will require assessing the impact of specific features (e.g., through ablation  
427 analyses) using retrospective, stable data systems and retrospective evaluation of the full forecasting process  
428 (e.g., from data wrangling to final forecast production).

429

430 Infectious disease forecasting continues to present many challenges and opportunities for improving outbreak  
431 response. Forecasts should be leading indicators of future activity and, while the COVID-19 case ensemble  
432 forecasts were good leading indicators at many points in time; they were unreliable, especially during periods of  
433 rapid change. Case data were integrated in COVID-19 mortality forecasts, which proved to be more reliable,  
434 likely in part due to reported cases being leading indicators of reported deaths (18,45). However, because  
435 deaths are a lagging indicator, death forecasts are less useful for short-term outbreak responses. Evaluation of  
436 the case forecasts provided insight on limitations of early forecasts and research avenues for improving them.  
437 These insights and the real-time forecasts provided by this effort were the product of large-scale collaboration  
438 between researchers and public health responders to confront the COVID-19 pandemic. Learning from and  
439 improving forecasting for COVID-19, other infectious diseases, and future pandemics will benefit from  
440 continuing and expanding these collaborative efforts.

441

## 442 **Methods**

443 The US COVID-19 Forecast Hub (46) is a consortium of researchers that develop and share forecasts of COVID-19  
444 reported cases, hospitalizations, and deaths with the goal of leveraging information from individual models that  
445 predict the near-term burden of COVID-19 in the United States. Teams that submitted models to the US COVID-  
446 19 Forecast Hub used a wide variety of methodology and data (S1, Table S1). Beyond serving as a repository for  
447 forecasts, submitted data were also aggregated by scientists at the COVID-19 Forecast Hub to generate two  
448 models that we included in this analysis: the COVIDhub-4\_week\_ensemble and the COVIDhub-  
449 trained\_ensemble. Since the beginning of the COVID-19 Forecast Hub, the quantile predictions from each week's  
450 submitted models were used as input data for the COVIDhub-4\_week\_ensemble. Ensemble aggregation  
451 methods evolved over time; for this analysis period, the ensemble forecast was calculated as the median across  
452 forecasts from all models at each quantile level. Additionally, beginning on February 1, 2021, the COVID-19  
453 Forecast Hub also generated a weighted ensemble (COVIDhub-trained\_ensemble). Models were selected for  
454 weighted ensemble inclusion based on their past performance over various window period and given a weight  
455 prior to aggregation. The methodology evolved over time and details are available on the model's metadata file  
456 on the COVID-19 Forecast Hub GitHub repository (see *Data and code availability and reporting guidelines*).

457

458 The COVID-19 Forecast Hub, and death forecasts submitted to the Hub have been described in detail elsewhere  
459 (8,17,18). The Hub's incident COVID-19 case forecasts, which were first solicited in July 2020, have similar  
460 submission requirements to the death forecasts. Important differences include an expanded geographical scale  
461 (national; state, territory, and DC; and county levels) and reduced number of required quantiles in the  
462 probability distribution (7 quantiles in total: 0.025, 0.10, 0.25, 0.50, 0.75, 0.90, and 0.975). Predictions for weekly  
463 incident COVID-19 cases can be submitted for up to 8 weeks in the future, although our analysis only includes  
464 predictions made for 1-4 weeks into the future.

465

466 We evaluated submitted forecasts between July 28, 2020, and December 21, 2021 (2020 epi week [EW] 31 –  
467 2021 EW 51), which encompasses 73 weeks. Because forecasts were submitted at multiple geographic scales,  
468 we conducted separate analyses for 1) national forecasts, 2) state, territory, and DC forecasts, 3) county level  
469 forecasts, and 4) sets of team forecasts for all three geographic scales. When appropriate, we compared forecast  
470 performance to that of a naïve model, created by the COVID-19 Forecast Hub, the COVIDhub-baseline. The  
471 COVIDhub-baseline model, created each week, was designed to be a neutral model to provide a simple  
472 reference point of comparison for all models. This baseline model forecasts a predictive median incidence equal  
473 to the number of reported cases in the most recent week, with uncertainty based on the empirical distribution  
474 of previous differences between the median and observed values (18).

475

#### 476 **Inclusion criteria**

477 Teams were included in the evaluation when they submitted forecasts with a complete set of quantiles for each  
478 1- through 4-week ahead target predictions. Additionally, teams must have met the following inclusion criteria:

- 479 1. had predictions for at least 50 locations (states, territories, or DC) for the state, territory, and DC level  
480 analyses; and for at least 75% of counties included in each population size quantile per submission week  
481 for the county-level analyses;
- 482 2. had submissions for at least 50% of the weeks included in the analysis period per location forecasted.

483

484 Teams meeting these inclusion criteria, and their submissions over time, are depicted in S1, Figure S1.

485

#### 486 **Ground Truth**

487 Forecasts were evaluated against the reported COVID-19 case reports collated by the Johns Hopkins Center for  
488 Systems Science and Engineering (CSSE) (47). To calculate weekly incident reported cases, we subtracted the

489 cumulative count for each Saturday from the cumulative count for the next Saturday, such that each incident  
490 weekly count reflects the number of cases reported from Sunday through Saturday in a given week. We  
491 aggregated reported counts from smaller geographic units into their larger unit. For example, counts in a given  
492 state are the aggregate of the county level reported counts and national counts are the sum of all states,  
493 territories, and DC.

494

495 CSSE reports data in real-time. Thus, data may be revised if the reporting health system makes public updates to  
496 their surveillance data. At times, such revisions may result in negative daily counts or in increases to case counts  
497 if the date of cases is shifted from one day to another or the definition of a reportable case is changed. We  
498 examined the percent change between the first reported cases in each state, DC, and territory per date relative  
499 to the counts in the surveillance file from April 2, 2022. We also assessed the influence of revised data on the  
500 final model outcomes (see S2) and the presence of negative case counts in the timeseries. Less than 1 percent of  
501 time points in the analysis period had negative daily case counts in the largest US counties. Negative counts  
502 were observed at the state/territory level only twice: in Missouri during the week of April 17, 2021, and Virgin  
503 Islands during the week ending October 10, 2020. The state of Florida reported 0 cases on November 27, 2021.  
504 We excluded all weeks and locations with negative counts as well as the week with 0 incidence in Florida in our  
505 primary analyses.

506

507 Additionally, we also examined whether a reported case count was an outlier in the case trend for each state.  
508 Anomalies in case data trends have not been uncommon throughout the pandemic, as reporting entities have  
509 uploaded large batches of surveillance data on a single day. To assess whether cases were outside of the  
510 expected range of reported cases over time, we applied three outlier detection algorithms, each with a 21-day  
511 window: a rolling median, a seasonal trend decomposition, and a seasonal trend decomposition without a  
512 seasonality term. We then classified a given count as an outlier if it was detected as such by at least two of the  
513 three algorithms. Using these data, we ran several sensitivity analyses to assess the likely impact of anomalous

514 data points on model performance. Sensitivity analyses examining the robustness of our findings to reporting  
515 anomalies are presented in S2.

516

517 Additional information about the CSSE data, and revisions to the dataset, is publicly available on a GitHub  
518 repository:

519 [https://github.com/CSSEGISandData/COVID-19/tree/master/csse\\_covid\\_19\\_data](https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data).

520

### 521 **Forecast locations**

522 Forecasts for incident cases were submitted for the national level, 50 states, 5 territories (American Samoa,  
523 Guam, the Northern Mariana Islands, Puerto Rico, and the US Virgin Islands), the DC, and 3,142 US counties. We  
524 excluded two counties in Alaska because they were not forecasted by most sets of team forecasts (Federal  
525 Information Processing Standard code 02063 and 02066 were excluded). Because fewer teams submitted  
526 forecasts for American Samoa, Guam, the Northern Mariana Islands, we excluded these territories from the  
527 analysis. Some teams treated DC as both a county and a jurisdiction, so we excluded DC from the county  
528 forecasts. In addition, because county population size and transmission are correlated and case counts and  
529 forecast performance are also correlated, we grouped counties into 5 quantiles based on their population sizes,  
530 with cut points at 8,908; 18,662; 36,742; and 93,230 people; most analyses used forecasts from the quantile  
531 with the largest population size (n=628). We hypothesized that small counties would be more likely to have  
532 better forecast accuracy because they had zero or very few reported cases. We thus chose to stratify counties by  
533 size to minimize any bias from aggregation. Performance results for most county forecasts are presented in S3.

534

### 535 **Defining epidemic phases**

536 For every state and DC, we independently classified each forecast week based on the estimated time-varying  
537 reproduction number ( $R_t$ ) for that given week. State-level  $R_t$  estimates were obtained from

538 <https://github.com/epiforecasts/epiforecasts.github.io> (48). We extracted the  $R_t$  estimate for the Wednesday of  
539 each week from all available files. Because  $R_t$  estimates were updated on a rolling basis in near real time, there  
540 were multiple estimates generated for the same date; we calculated the median estimated  $R_t$  per date for the  
541 upper and lower 90% credible interval and the median value (August 1, 2020 – January 15, 2022, or 2020 EW 31  
542 – 2022 EW 2, reflecting 77 weeks in total). Each forecast week was then classified into one of the following  
543 categories based on the  $R_t$  estimates: *increasing*, *peak*, *decreasing*, *nadir*.

544  
545 *Increasing* and *decreasing* phases reflect weeks in which  $R_t$  had a 90% probability of being greater than or less  
546 than 1.0, respectively. There were several periods of rapid transmission in certain jurisdictions where  $R_t$  dipped  
547 above/below the 1.0 threshold but did not remain on an upward or downward trajectory. Thus, we classified  
548 weeks between two increasing phases as *increasing* and weeks between two decreasing periods as *decreasing*.  
549 Weeks between increasing and decreasing phases were classified as *peaks*, whereas *nadirs* were defined as  
550 periods between decreasing and increasing phases. Periods at the beginning or the end of an analysis period  
551 were classified as a continuation of whichever phase preceded or followed them. Graphical depictions of  $R_t$  are  
552 provided in S4 and show general concordance between  $R_t$  and reported cases.

553

## 554 **Evaluation methodology**

555 We evaluated probabilistic forecast accuracy using two different metrics, empirical prediction interval coverage  
556 rates and weighted interval scores (WIS) (49). Coverage was calculated by determining the frequency with which  
557 the prediction interval contained the eventually observed outcome for the 50%, 80% and 95% intervals. WIS  
558 reflects a weighted estimate of sharpness (i.e., the range of the predicted interval) and calibration (i.e., precision  
559 or error) across the three prediction intervals and the median prediction, with higher WIS and indicating lower  
560 forecast skill. Importantly, WIS is highly correlated with the magnitude of observed and forecasted values. We  
561 used mean absolute WIS to assess forecast accuracy over time and mean relative WIS (rWIS) to assess forecast  
562 accuracy over space. Relative WIS was estimated by calculating the geometric mean of WIS across all sets of

563 team forecasts and scaling that value to the WIS of a naïve model, the COVID-hub baseline. This approach eases  
564 interpretation, where values greater than 1.0 reflected worse accuracy than the baseline model and values  
565 below 1.0 reflected better model performance. Additionally, the pairwise relative comparison helps account for  
566 missing forecasts. Both coverage and WIS have been described in detail elsewhere (18,49). Horizon specific  
567 results for national, state/territory/DC, and large counties are presented in S5.

568

569 To assess the association between WIS and epidemic phase for each team, we fitted separate Gaussian  
570 generalized estimating equation (GEE) models per team (equation 1) with an independent working correlation  
571 structure at the state-level. This structure assumes that observations are not correlated over time in a state  
572 (denoted as  $l$  in the equations below). Cases and weighted interval scores were log transformed and then  
573 standardized (subtracting the mean and dividing by the standard deviation) prior to fitting the model, as this  
574 transformation yielded more computationally and numerically stable estimates. We define those resulting  
575 variables as  $stdWIS$  and  $stdCases$ . The expected value for a standardized WIS for time ( $t$ ) and location ( $l$ ), with  
576 forecasts from a given team's model, is as follows:

577

$$578 \quad \log (stdWIS_{t,l,h}) = \beta_0 + \alpha_{p[t,l]} + \gamma_h + ns(\log (stdCases_{t,l,h})) + \epsilon_{t,l,h} \quad (1)$$

579

580 Where  $p[t, l]$  is an index that reflects the phase of each time ( $t$ ) and location ( $l$ ), ( $h$ ) is the horizon of the  
581 forecast in weeks, and  $ns(\cdot)$  represents a natural spline with two degrees of freedom. Using a regression model  
582 allows us to summarize patterns of overall average performance between teams while accounting for high  
583 correlation and variation in the scores. Comparisons of  $rWIS$ , in contrast, do not allow for formal inference on  
584 the differences in performance between teams. Prior to applying this regression model structure, our model  
585 building approach included exploratory analysis of several structures appropriate for longitudinal analysis. We  
586 examined model residuals, influential observations, goodness of fit metrics, and the impact of changing the  
587 functional form of the variables included in the model.

588

589 The inclusion of reported cases in models permitted flexible adjustment for the wide range in cases between  
590 and within jurisdictions, which led to a wide range of possible WIS values, as WIS values tend to be higher when  
591 counts are higher. Expected WIS values were computed by first obtaining a marginal mean from the GEE model  
592 and then undoing the transformations by exponentiating and un-standardizing the marginal mean. This was  
593 done separately for each team for all phases and for each team and each phase individually (see S6 for  
594 estimated team-specific marginal mean WIS relative to reported case counts). Additionally, we calculated  
595 whether the 80% confidence interval (based on Gaussian distributional assumptions) for each team's expected  
596 WIS outcome (on the log-scale and normalized, as described above) was less than the baseline model for all  
597 phases (i.e., the marginal mean WIS for the baseline model).

598

#### 599 **Data and code availability and reporting guidelines**

600 The forecasts from models used in this paper are available from the COVID-19 Forecast Hub GitHub repository  
601 (<https://github.com/reichlab/covid19-forecast-hub>) (8) and the Zoltar forecast archive  
602 (<https://zoltardata.com/project/44>) (50). The code used to generate all figures and tables in the manuscript is  
603 available in a public repository  
604 (<https://github.com/cdcepi/Evaluation-of-case-forecasts-submitted-to-COVID19-Forecast-Hub>). All analyses  
605 were conducted using the R language for statistical computing (v 4.0.3) (51), and the following packages were  
606 used for the main analyses: *scoringutils* (52), *covidhubUtils* (53), *geepack* (54). Additionally, we included the  
607 EPIFORGE 2020 reporting guideline checklist in S7 to indicate each page in this evaluation that corresponds to  
608 each specific recommendation (15).

609

610 This activity was reviewed by CDC and was conducted consistent with applicable federal law and CDC policy.



611 **CDC disclaimer:** The findings and conclusions in this report are those of the authors and do not necessarily  
612 represent the official position of the Centers for Disease Control and Prevention.

## 613 **References**

- 614 1. Biggerstaff M, Johansson M, Alper D, Brooks LC, Chakraborty P, Farrow DC, et al. Results from the second  
615 year of a collaborative effort to forecast influenza seasons in the United States. *Epidemics*. 2018 Sep  
616 1;24:26–33.
- 617 2. McGowan CJ, Biggerstaff M, Johansson M, Apfeldorf KM, Ben-Nun M, Brooks L, et al. Collaborative efforts to  
618 forecast seasonal influenza in the United States, 2015–2016. *Sci Rep* 2019 91. 2019 Jan 24;9(1):1–13.
- 619 3. Reich NG, Brooks LC, Fox SJ, Kandula S, McGowan CJ, Moore E, et al. A collaborative multiyear, multimodel  
620 assessment of seasonal influenza forecasting in the United States. *Proc Natl Acad Sci U S A*. 2019 Feb  
621 19;116(8):3146–54.
- 622 4. Johansson MA, Apfeldorf KM, Dobson S, Devita J, Buczak AL, Baugher B, et al. An open challenge to advance  
623 probabilistic forecasting for dengue epidemics. *Proc Natl Acad Sci*. 2019 Nov 26;116(48):24268–74.
- 624 5. Holcomb KM, Barker CM, Keyel Wadsworth Center Matteo Marcantonio AC, Childs ML, Gorris ME, Hamins-  
625 Puértolas M, et al. Evaluation of an open forecasting challenge to assess skill of West Nile virus  
626 neuroinvasive disease prediction. 2022 Aug 26 [cited 2022 Dec 16]; Available from:  
627 <https://www.researchsquare.com>
- 628 6. Viboud C, Sun K, Gaffey R, Ajelli M, Fumanelli L, Merler S, et al. The RAPIDD ebola forecasting challenge:  
629 Synthesis and lessons learnt. *Epidemics*. 2018 Mar 1;22:13–21.
- 630 7. Lutz CS, Huynh MP, Schroeder M, Anyatonwu S, Dahlgren FS, Danyluk G, et al. Applying infectious disease  
631 forecasting to public health: A path forward using influenza forecasting examples. *BMC Public Health*. 2019  
632 Dec 10;19(1):1–12.
- 633 8. Cramer EY, Huang Y, Wang Y, Ray EL, Cornell M, Bracher J, et al. The United States COVID-19 Forecast Hub  
634 dataset. *Sci Data*. 2022 Aug 1;9(1):462.

- 635 9. Centers for Disease Control and Prevention. COVID-19 Forecasts: Cases | CDC [Internet]. [cited 2022 Dec  
636 16]. Available from: <https://www.cdc.gov/coronavirus/2019-ncov/science/forecasting/forecasts-cases.html>
- 637 10. Biggerstaff M, Slayton RB, Johansson MA, Butler JC. Improving Pandemic Response: Employing  
638 Mathematical Modeling to Confront Coronavirus Disease 2019. *Clin Infect Dis*. 2022 Mar 9;74(5):913–7.
- 639 11. Bertsimas D, Boussioux L, Cory-Wright R, Delarue A, Digalakis V, Jacquillat A, et al. From predictions to  
640 prescriptions: A data-driven response to COVID-19. *Health Care Manag Sci*. 2021 Jun 1;24(2):253–72.
- 641 12. Padilla L, Hosseinpour H, Fygenson R, Howell J, Chunara R, Bertini E. Impact of COVID-19 forecast  
642 visualizations on pandemic risk perceptions. *Sci Rep* 2022 121. 2022 Feb 7;12(1):1–14.
- 643 13. Eksin C, Paarporn K, Weitz JS. Systematic biases in disease forecasting – The role of behavior change.  
644 *Epidemics*. 2019 Jun 1;27:96–105.
- 645 14. Reich NG, Tibshirani RJ, Ray EL, Rosenfeld R. On the predictability of COVID-19 - International Institute of  
646 Forecasters [Internet]. [cited 2022 Dec 16]. Available from: [https://forecasters.org/blog/2021/09/28/on-](https://forecasters.org/blog/2021/09/28/on-the-predictability-of-covid-19/)  
647 [the-predictability-of-covid-19/](https://forecasters.org/blog/2021/09/28/on-the-predictability-of-covid-19/)
- 648 15. Pollett S, Johansson MA, Reich NG, Brett-Major D, Del Valle SY, Venkatramanan S, et al. Recommended  
649 reporting items for epidemic forecasting and prediction research: The EPIFORGE 2020 guidelines. *PLOS Med*.  
650 2021 Oct 1;18(10):e1003793.
- 651 16. Nixon K, Jindal S, Parker F, Reich NG, Ghobadi K, Lee EC, et al. An evaluation of prospective COVID-19  
652 modelling studies in the USA: from data to science translation. *Lancet Digit Health*. 2022 Oct 1;4(10):e738–  
653 47.
- 654 17. Ray EL, Wattanachit N, Niemi J, Kanji AH, House K, Cramer EY, et al. Ensemble Forecasts of Coronavirus  
655 Disease 2019 (COVID-19) in the U.S. *medRxiv*. 2020 Aug 22;2020.08.19.20177493.
- 656 18. Cramer EY, Ray EL, Lopez VK, Bracher J, Brennen A, Castro Rivadeneira AJ, et al. Evaluation of individual and  
657 ensemble probabilistic forecasts of COVID-19 mortality in the United States. *Proc Natl Acad Sci*. 2022 Apr  
658 12;119(15):e2113561119.
- 659 19. McDonald DJ, Bien J, Green A, Hu AJ, DeFries N, Hyun S, et al. Can auxiliary indicators improve COVID-19

- 660 forecasting and hotspot prediction? *Proc Natl Acad Sci.* 2021 Dec 21;118(51):e2111453118.
- 661 20. McMahan CS, Self S, Rennert L, Kalbaugh C, Kriebel D, Graves D, et al. COVID-19 wastewater epidemiology:  
662 a model to estimate infected populations. *Lancet Planet Health.* 2021 Dec 1;5(12):e874–81.
- 663 21. Nourbakhsh S, Fazil A, Li M, Mangat CS, Peterson SW, Daigle J, et al. A wastewater-based epidemic model  
664 for SARS-CoV-2 with application to three Canadian cities. *Epidemics.* 2022 Jun 1;39:100560.
- 665 22. Proverbio D, Kemp F, Magni S, Ogorzaly L, Cauchie HM, Gonçalves J, et al. Model-based assessment of  
666 COVID-19 epidemic dynamics by wastewater analysis. *Sci Total Environ.* 2022 Jun 25;827:154235.
- 667 23. Cao Y, Francis R. On forecasting the community-level COVID-19 cases from the concentration of SARS-CoV-2  
668 in wastewater. *Sci Total Environ.* 2021 Sep 10;786:147451.
- 669 24. Adiga A, Kaur G, Wang L, Hurt B, Porebski P, Venkatramanan S, et al. Phase-Informed Bayesian Ensemble  
670 Models Improve Performance of Covid-19 Forecasts. In *Thirty-Fifth Annual Conference on Innovative  
671 Applications of Artificial Intelligence*; 2023.
- 672 25. Adiga A, Kaur G, Hurt B, Wang L, Porebski P, Venkatramanan S, et al. Enhancing COVID-19 Ensemble  
673 Forecasting Model Performance Using Auxiliary Data Sources. In *IEEE International Conference on Big Data*;  
674 2022.
- 675 26. Funk S, Salathé M, Jansen V a a. Modelling the influence of human behaviour on the spread of infectious  
676 diseases: a review. *J R Soc Interface R Soc.* 2010;7(50):1247–56.
- 677 27. Moran KR, Fairchild G, Generous N, Hickmann K, Osthus D, Priedhorsky R, et al. Epidemic Forecasting is  
678 Messier Than Weather Forecasting: The Role of Human Behavior and Internet Data Streams in Epidemic  
679 Forecast. *J Infect Dis.* 2016 Dec 1;214(suppl\_4):S404–8.
- 680 28. Buckee C, Noor A, Sattenspiel L. Thinking clearly about social aspects of infectious disease transmission. *Nat*  
681 2021 5957866. 2021 Jun 30;595(7866):205–13.
- 682 29. Bodner K, Fortin MJ, Molnár PK. Making predictive modelling ART: accurate, reliable, and transparent.  
683 *Ecosphere.* 2020;11(6):e03160.
- 684 30. Yanvi I, Foster DP. Graininess of Judgment Under Uncertainty: An Accuracy–Informativeness Trade-Off. *J Exp*

- 685 Psychol Gen. 1995 Dec;124(4):24–432.
- 686 31. Reis J, Yamana T, Kandula S, Shaman J. Superensemble forecast of respiratory syncytial virus outbreaks at  
687 national, regional, and state levels in the United States. *Epidemics*. 2019 Mar 1;26:1–8.
- 688 32. Oidtman RJ, Omodei E, Kraemer MUG, Castañeda-Orjuela CA, Cruz-Rivera E, Misnaza-Castrillón S, et al.  
689 Trade-offs between individual and ensemble forecasts of an emerging infectious disease. *Nat Commun*.  
690 2021 Sep 10;12(1):5379.
- 691 33. Brooks LC, Farrow DC, Hyun S, Tibshirani RJ, Rosenfeld R. Nonmechanistic forecasts of seasonal influenza  
692 with iterative one-week-ahead distributions. *PLOS Comput Biol*. 2018 Jun 1;14(6):e1006134.
- 693 34. Ray EL, Reich NG. Prediction of infectious disease epidemics via weighted density ensembles. *PLOS Comput*  
694 *Biol*. 2018 Feb 1;14(2):e1005910.
- 695 35. McAndrew T, Reich NG. Adaptively stacking ensembles for influenza forecasting. *Stat Med*. 2021 Dec  
696 30;40(30):6931–52.
- 697 36. Smith J, Wallis KF. A simple explanation of the forecast combination puzzle. *Oxf Bull Econ Stat*. 2009;71(3).
- 698 37. Ray EL, Brooks LC, Bien J, Biggerstaff M, Bosse NI, Bracher J, et al. Comparing trained and untrained  
699 probabilistic ensemble forecasts of COVID-19 cases and deaths in the United States. *Int J Forecast*. 2022 Jul  
700 1;
- 701 38. Taylor JW, Taylor KS. Combining probabilistic forecasts of COVID-19 mortality in the United States. *Eur J*  
702 *Oper Res*. 2023;304(1).
- 703 39. Paireau J, Andronico A, Hozé N, Layan M, Crépey P, Roumagnac A, et al. An ensemble model based on early  
704 predictors to forecast COVID-19 health care demand in France. *Proc Natl Acad Sci U S A [Internet]*. 2022 May  
705 3 [cited 2022 Dec 16];119(18). Available from: <https://doi.org/10.1073/pnas.2103302119>
- 706 40. Morozova O, Li ZR, Crawford FW. One year of modeling and forecasting COVID-19 transmission to support  
707 policymakers in Connecticut. *Sci Rep*. 123AD;11:20271.
- 708 41. Greene SK, McGough SF, Culp GM, Graf LE, Lipsitch M, Menzies NA, et al. Nowcasting for Real-Time COVID-  
709 19 Tracking in New York City: An Evaluation Using Reportable Disease Data From Early in the Pandemic.

- 710 JMIR Public Health Surveill 2021;7(1):e25538. 2021 Jan 15;7(1):e25538.
- 711 42. covidestim: COVID-19 nowcasting [Internet]. [cited 2023 Jan 9]. Available from: <https://covidestim.org/>
- 712 43. Abbott S, Sherratt K, Bevan J, Gibbs H, Hellewell J, Munday J, et al. Temporal variation in transmission during  
713 the COVID-19 outbreak [Internet]. Vol. 2020 [cited 2023 Jan 9]. Available from:  
714 <https://epiforecasts.io/covid/>
- 715 44. reichlab/covid19-forecast-hub [Internet]. The Reich Lab at UMass-Amherst; 2022 [cited 2023 Jan 10].  
716 Available from: <https://github.com/reichlab/covid19-forecast-hub>
- 717 45. Bracher J, Wolfram D, Deuschel J, Görden K, Ketterer JL, Ullrich A, et al. National and subnational short-  
718 term forecasting of COVID-19 in Germany and Poland during early 2021. *Commun Med* 2022 21. 2022 Oct  
719 31;2(1):1–17.
- 720 46. COVID 19 forecast hub [Internet]. [cited 2022 Dec 16]. Available from: <https://covid19forecasthub.org/>
- 721 47. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect*  
722 *Dis.* 2020 May 1;20(5):533–4.
- 723 48. Abbott S, Bennett C, Hickson J, Allen J, Sherratt K, Funk S. National and Subnational Estimates of the Covid  
724 19 Reproduction Number (R) for the United States of America Based on Test Results. *Harvard Dataverse*,  
725 V292. 2020.
- 726 49. Bracher J, Ray EL, Gneiting T, Reich NG. Evaluating epidemic forecasts in an interval format. *PLoS Comput*  
727 *Biol.* 2021 Feb 1;17(2).
- 728 50. Reich NG, Cornell M, Ray EL, House K, Le K. The Zoltar forecast archive, a tool to standardize and store  
729 interdisciplinary prediction research. *Sci Data.* 2021 Feb 11;8(1):59.
- 730 51. R Core Team. R: A language and environment for statistical computing. [Internet]. Vienna, Austria: R  
731 Foundation for Statistical Computing; 2022. Available from: <https://www.R-project.org/>.
- 732 52. Bosse NI, Gruson H, Cori A, van Leeuwen E, Funk S, Abbott S. Evaluating Forecasts with scoringutils in R.  
733 2022 May 14 [cited 2022 Dec 16]; Available from: <https://arxiv.org/abs/2205.07090v1>
- 734 53. Wang S, Ray EL, Reich NG, Shah A. Tools for working with COVID-19 Forecast Hub data: a brief tour of the

- 735 `covidHubUtils` R package [Internet]. [cited 2022 Dec 16]. Available from:  
736 <http://reichlab.io/covidHubUtils/articles/covidHubUtils-overview.html>
- 737 54. Højsgaard S, Halekoh U, Yan J. The R Package geePack for Generalized Estimating Equations. *J Stat Softw.*  
738 2006;15:1–11.
- 739 55. Linas BP, Xiao J, Dalgic OO, Mueller PP, Adey M, Aaron A, et al. Projecting COVID-19 Mortality as States  
740 Relax Nonpharmacologic Interventions. *JAMA Health Forum.* 2022 Apr 1;3(4):e220760.
- 741 56. Sen P, Yamana TK, Kandula S, Galanti M, Shaman J. Burden and characteristics of COVID-19 in the United  
742 States during 2020. *Nature.* 2021;598(7880).
- 743 57. Suchoski B, Stage S, Gurung H, Baccam P. GPU Accelerated Parallel Processing for Large-Scale Monte Carlo  
744 Analysis: COVID-19 Parameter Estimation and New Case Forecasting. *Front Appl Math Stat.* 2022;8.
- 745 58. Wang Y, Kim M, Yu S, Li X, Wang G, Wang L. Nonparametric estimation and inference for spatiotemporal  
746 epidemic models. *J Nonparametric Stat.* 2022;34(3).
- 747 59. Lemaitre JC, Grantz KH, Kaminsky J, Meredith HR, Truelove SA, Lauer SA, et al. A scenario modeling pipeline  
748 for COVID-19 emergency planning. *Sci Rep* 2021 111. 2021 Apr 6;11(1):1–13.
- 749 60. Karlen D. Characterizing the spread of CoVID-19. 2020 Jul 14 [cited 2023 Mar 10]; Available from:  
750 <https://arxiv.org/abs/2007.07156v1>
- 751 61. Castro L, Fairchild G, Michaud I, Osthus D. COFFEE: COVID-19 Forecasts using Fast Evaluations and  
752 Estimation. 2021 Oct 4 [cited 2023 Mar 10]; Available from: <https://arxiv.org/abs/2110.01546v1>
- 753 62. Zheng S, Gao Z, Cao W, Bian J, Liu TY. HierST: A Unified Hierarchical Spatialoral Framework for COVID-19  
754 Trend Forecasting. *Int Conf Inf Knowl Manag Proc.* 2021 Oct 26;4383–92.
- 755 63. Davis JT, Chinazzi M, Perra N, Mu K, Pastore y Piontti A, Ajelli M, et al. Cryptic transmission of SARS-CoV-2  
756 and the first COVID-19 wave. *Nat* 2021 6007887. 2021 Oct 25;600(7887):127–32.
- 757 64. Srivastava A. The Variations of SIKJalpha Model for COVID-19 Forecasting and Scenario Projections. 2022 Jul  
758 6 [cited 2023 Mar 10]; Available from: <https://arxiv.org/abs/2207.02919v1>
- 759 65. Adiga A, Wang L, Hurt B, Peddireddy A, Porebski P, Venkatramanan S, et al. All Models Are Useful: Bayesian

760 Ensembling for Robust High Resolution COVID-19 Forecasting. Proc ACM SIGKDD Int Conf Knowl Discov Data  
761 Min. 2021 Aug 14;2505–13.

762

## 763 **Supporting information captions**

### 764 **Supporting Information 1: Team submissions, methods, and data**

765

766 **SI Figure 1.1. Forecasts submitted over time at the national, state-territory-DC level in panel A and at the**  
767 **country scale in Panel B. The number of forecasted locations submitted each week nationally or at the state,**  
768 **territory and DC level is included, while the country level forecast submissions shows the percent of counties**  
769 **per quantile that were submitted each week. Sets of team forecasts meeting the inclusion criteria for this**  
770 **main analysis are labeled with an asterisk (\*).**

771

772 **S1 Table 1.1. List of models evaluated, including sources for case, hospitalization, death, demographic, and**  
773 **mobility data when used as inputs for the given model. We evaluated 26 models contributed by 24 teams. The**  
774 **COVIDhub team submitted three models including the baseline model and the ensemble model. A brief**  
775 **description is included for each model, with a reference where available. The last column indicates whether**  
776 **the model made assumptions about how and whether social distancing measures were assumed to change**  
777 **during the period for which forecasts were made.**

### 778 **Supporting Information 2: Revision and outlier sensitivity analyses**

779

780 **S2 Figure 2.1. To assess the influence of data revisions on our evaluation of forecast skill, we compared daily**  
781 **differences in cumulative reported cases during the week they were first reported to reported case counts for**

782 **the same week in the complete data as of April 2, 2022. In total 721 weeks had at least one day with a revised**  
783 **case count (17% of all weeks, n=4,241 weeks) and revisions occurred in 43 of 51 jurisdictions. These**  
784 **jurisdiction specific plots compare cases reported as of the date in the subtitle (in red) compared to cases**  
785 **reported as of April 2, 2022 (in black).**

786

787 **S2 Figure 2.2. After identifying weeks with revised case counts, we then excluded them from the dataset and**  
788 **reran the GEE models and estimated the marginal mean Weighted Interval Score (WIS). Panel A shows the**  
789 **estimated marginal mean WIS and 95% confidence intervals for mean cases from team-specific GEE models**  
790 **for all 48 jurisdictions from this sensitivity analysis. The 95% confidence intervals for the COVIDhub-baseline**  
791 **model are shown in dashed red vertical lines. Panel B presents each team’s estimated marginal mean WIS per**  
792 **phase, scaled to the COVIDhub-baseline model’s estimated marginal mean WIS for all epidemic phases, using**  
793 **the dataset with excluded week. Teams with higher estimated marginal mean WIS values (i.e., greater than**  
794 **1.0) are presented in shades of orange while teams with lower estimated marginal mean WIS (i.e., less than**  
795 **1.0) are shown in shades of green. Team forecasts are denoted with an asterisk (\*) if the 80% confidence**  
796 **interval of the expected WIS outcome (normalized and on the log scale) was estimated by a model to be lower**  
797 **than the expected WIS of the COVIDhub-baseline model for all phases.**

798

799 **S2 Figure 2.3. Outliers were defined as non-revised reported case counts that were outside of the expected**  
800 **range by at least two of the three algorithms: a rolling median, a seasonal trend decomposition, and a**  
801 **seasonal trend decomposition without a seasonality term. Each method used a 21-day window.**  
802 **Approximately three percent of weeks (686 of 27,489 total weeks in the analysis period) had at least one day**  
803 **of reported cases identified as an outlier.**

804



805 **Supporting Information 3: Incident COVID-19 case forecasts were submitted for all US counties. The**  
806 **plots shown here depicted average, scaled pairwise Weighted Interval Score (WIS; see *Methods* for**  
807 **description), 95% coverage, and submissions (S3 Figure 3.1), average 50%, 80% and 95% coverage for**  
808 **eligible submitted forecasts (S3 Figure 3.2), and average WIS and 95% coverage over time (S3 Figure**  
809 **3.2). Each figure shows spatial disaggregated results, with increasing population size and quantile**  
810 **numbers. For example, counties with the smallest population are grouped in Quantile 1 and the**  
811 **largest population sizes are grouped in Quantile 5. The following teams are included in these figures:**  
812 **CEID-Walk, LNQ-ens1, Microsoft\_DeepSTIA, COVIDhub-4\_week\_ensemble, COVIDhub-**  
813 **trained\_ensemble, COVIDhub-baseline, CU-select, FAIR-NRAR, FRBSF\_Wilson-Econometric,**  
814 **IowasStateLW-STEM, JHU\_IDD-CovidSP, JHU\_CSSE-DECOM, JHUAPL-Bucky, LANL-GrowthRate, LNQ-**  
815 **esn1, UVA-Ensemble.**

816  
817 **S3 Figure 3.1. Percent of weeks with complete submissions for all sets of team forecasts, scaled, pairwise**  
818 **relative Weighted Interval Score (rWIS), 95% coverage, and by geographical scale of submitted forecasts.**  
819 **Teams are sorted by increasing rWIS values.**

820  
821 **S3 Figure 3.2. Expected and observed coverage rates aggregated over time and horizon for county forecasts.**  
822 **The dashed line represents optimal expected-coverage. Team forecasts that outperformed the COVIDhub-**  
823 **4\_week\_ensemble model at all coverage levels are labeled on the right hand side of the plots.**

824  
825 **S3 Figure 3.3. Mean Weighted Interval Score (WIS) over time, aggregated by geographic units and forecast**  
826 **horizon in A and 95% coverage over time, aggregated by geographic units and forecast horizon in B. The black,**  
827 **dashed vertical line in all panels shows the date that public communication of the case forecasts was paused.**  
828 **The black, dashed horizontal line in panels B show nominal 95% interval coverage**

829

830 **Supporting Information 4: Estimated time-varying reproduction number and epidemic phase**  
831 **classifications. For each state, the top panel shows the median  $R_t$  and median upper and lower 90%**  
832 **credible interval over time in red. The bottom panel shows reported case counts over time. Both**  
833 **plots have vertical bands representing the epidemic phase of each forecast week: *increasing, peak,***  
834 ***decreasing, nadir.***

835

836 **Supporting Information 5: Each location specific forecast submitted to the COVID19 Forecast Hub**  
837 **included at least 4 weeks of future predictions. Here, we present disaggregated 1 and 4 week ahead**  
838 **predictions of model performance for each team model that submitted national and**  
839 **state/territory/DC forecasts and were included in the main analyses. Specific plots include the**  
840 **average 50%, 80% and 95% coverage for eligible submitted forecasts (S5 Figure 5.1), average**  
841 **absolute Weighted Interval Score (WIS) and 95% coverage over time (S5 Figure 5.2), and scaled,**  
842 **pairwise rWIS by location (S5 Figure 5.3)**

843

844 **S5 Figure 5.1. Expected and observed coverage rates aggregated for 1 and 4 week ahead forecasts over time**  
845 **for national forecasts in A, state/territory/DC forecasts in B, the largest country forecasts in C. The dashed line**  
846 **represents optimal expected-coverage. Teams that outperformed the COVIDhub-4\_week\_ensemble model at**  
847 **all coverage levels are labeled on the right-hand side of the plots.**

848

849 **S5 Figure 5.2. Mean Weighted Interval Score (WIS) over time for 1 and 4 week ahead forecasts, aggregated by**  
850 **geographic units, and 95% coverage over time for 1 and 4 week ahead forecasts, aggregated by geographic**

851 units. The black, dashed vertical line in all panels shows the date that public communication of the case  
852 forecasts was paused. The black, dashed horizontal line in panels D, E, and F show nominal 95% interval  
853 coverage. Teams that submitted national forecasts are presented in A. and D., state/territory/DC forecasts  
854 presented in B. and E., and teams that submitted large county level forecasts are presented in C. and F.

855

856

857 **S5 Figure 5.3. Scaled, pairwise relative Weighted Interval Score (rWIS; see *Methods* for description) for all**  
858 **teams that submitted national and state/territory/DC forecasts by location for 1 and 4 week ahead horizon.**  
859 **National estimates are displayed first, followed by jurisdictions in alphabetical order. Teams are displayed by**  
860 **decreasing average rWIS across all forecast horizons and locations.**

861

862 **Supporting Information 6: Each team model's estimated marginal mean Weighted Interval Score**  
863 **(WIS) over range of reported case counts per epidemic phase. Marginal mean WIS was estimated**  
864 **from GEE model results and reflect values across the 95% confidence interval of mean reported**  
865 **cases. Case counts differ per team model as each team forecasted a different set of locations over a**  
866 **different range of possible dates.**

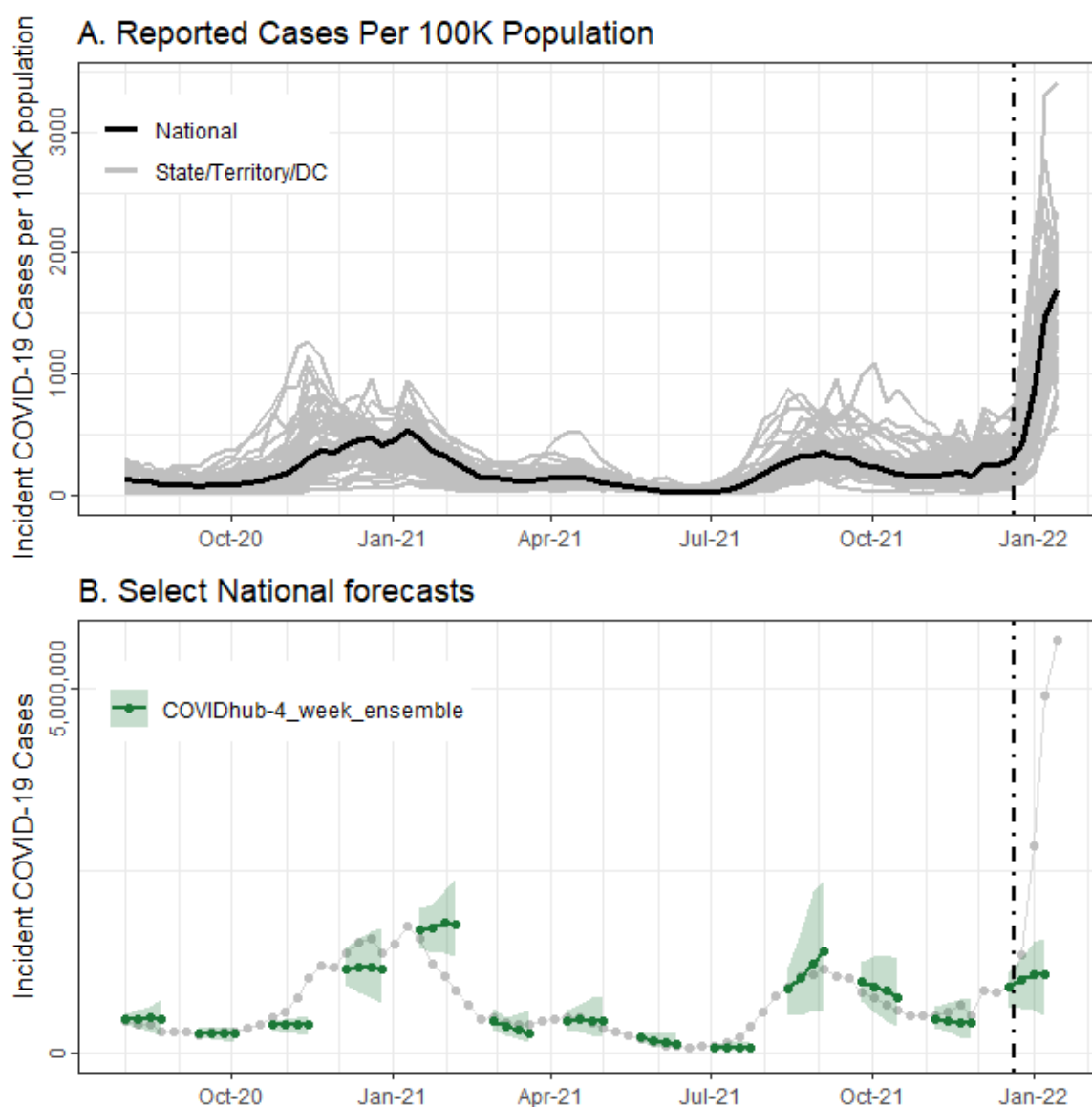
867

868 **Supporting Information 7: EPIFORGE 2020 guidelines outline 19 recommended reporting items for**  
869 **epidemic forecasting and prediction research (15). These items are included in the checklist below,**  
870 **which also include the page number where each item is described or presented within this**  
871 **evaluation.**

## 872 Main Text Figures

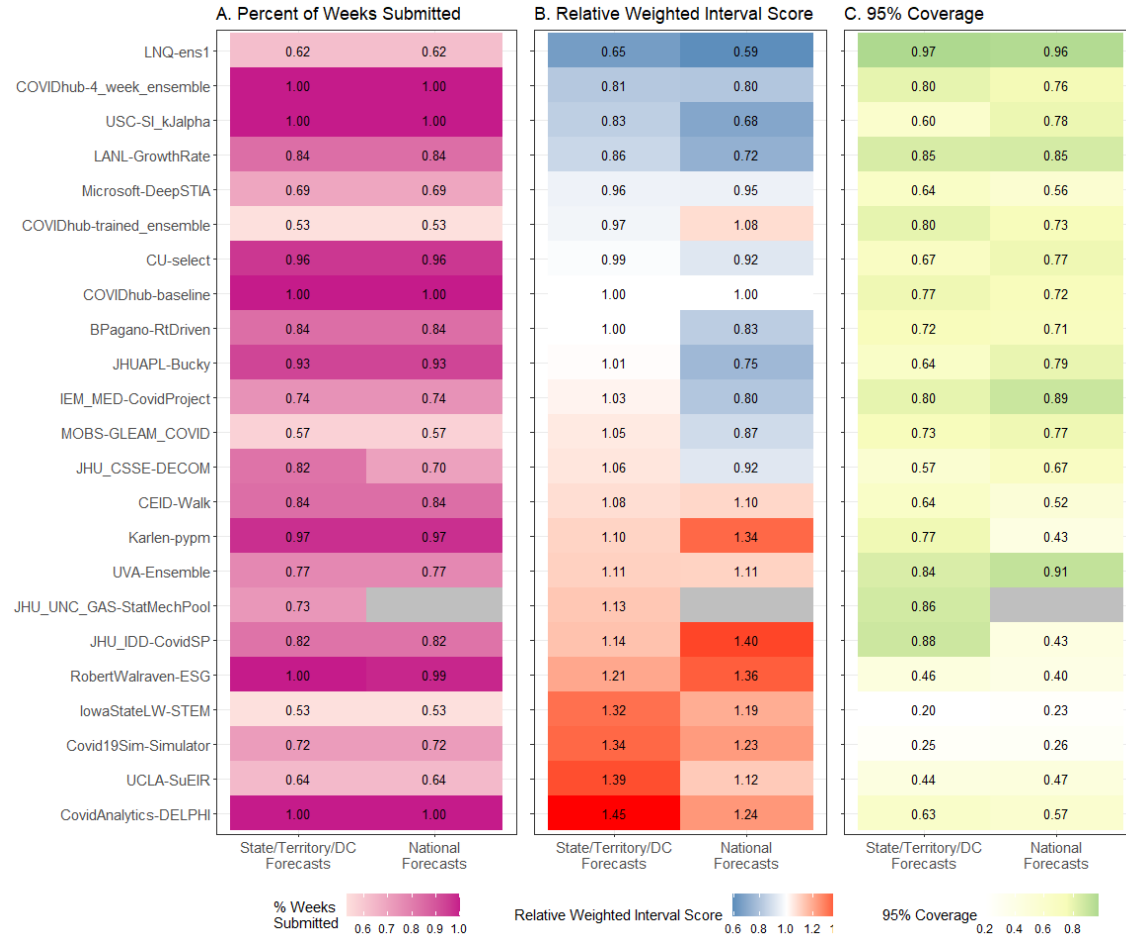
873

874 **Figure 1.** Weekly incident reported COVID-19 cases per 100K population, nationally (in black) and per  
875 state/territory/DC (in gray), over time in panel A. Panel B shows a subset of COVIDhub-4\_week\_ensemble  
876 forecasts (in green) over time, with the median predictions represented as lines and points and the 95%  
877 prediction intervals in bands. Reported incident cases (counts per week) are shown in gray. In both plots, the  
878 black, dashed vertical line shows the date that public communication of the case forecasts was paused.

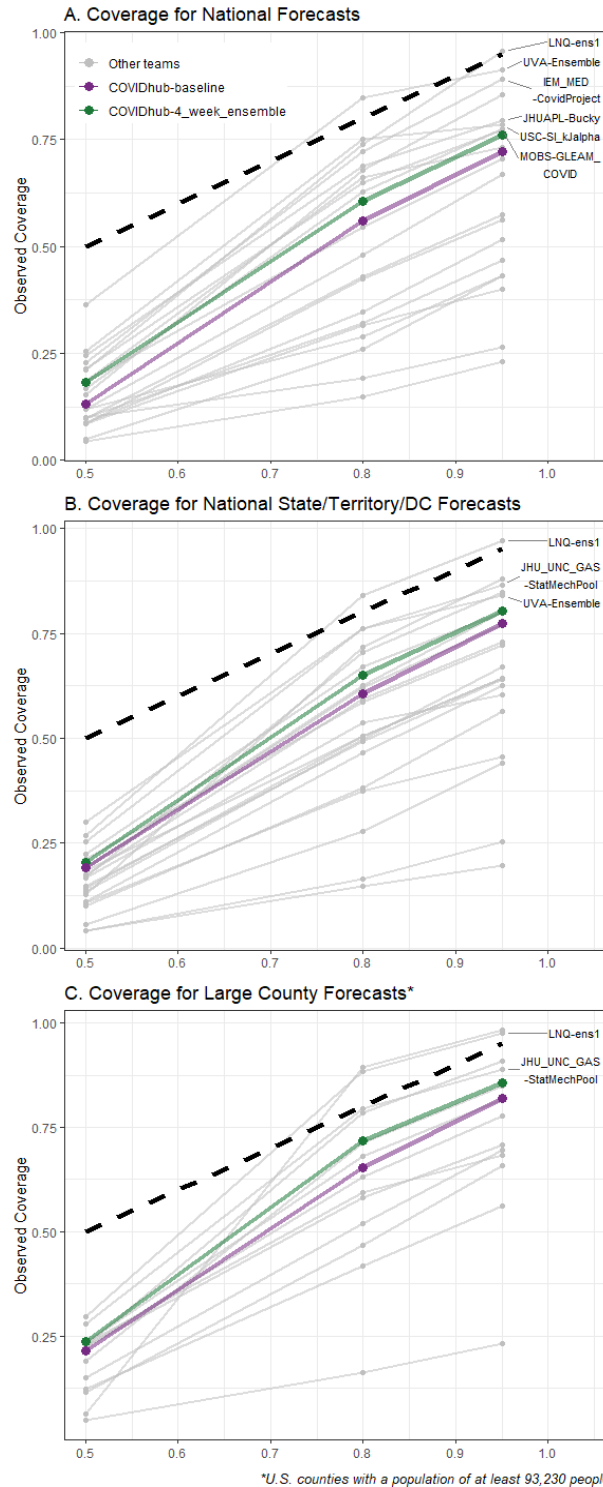


879

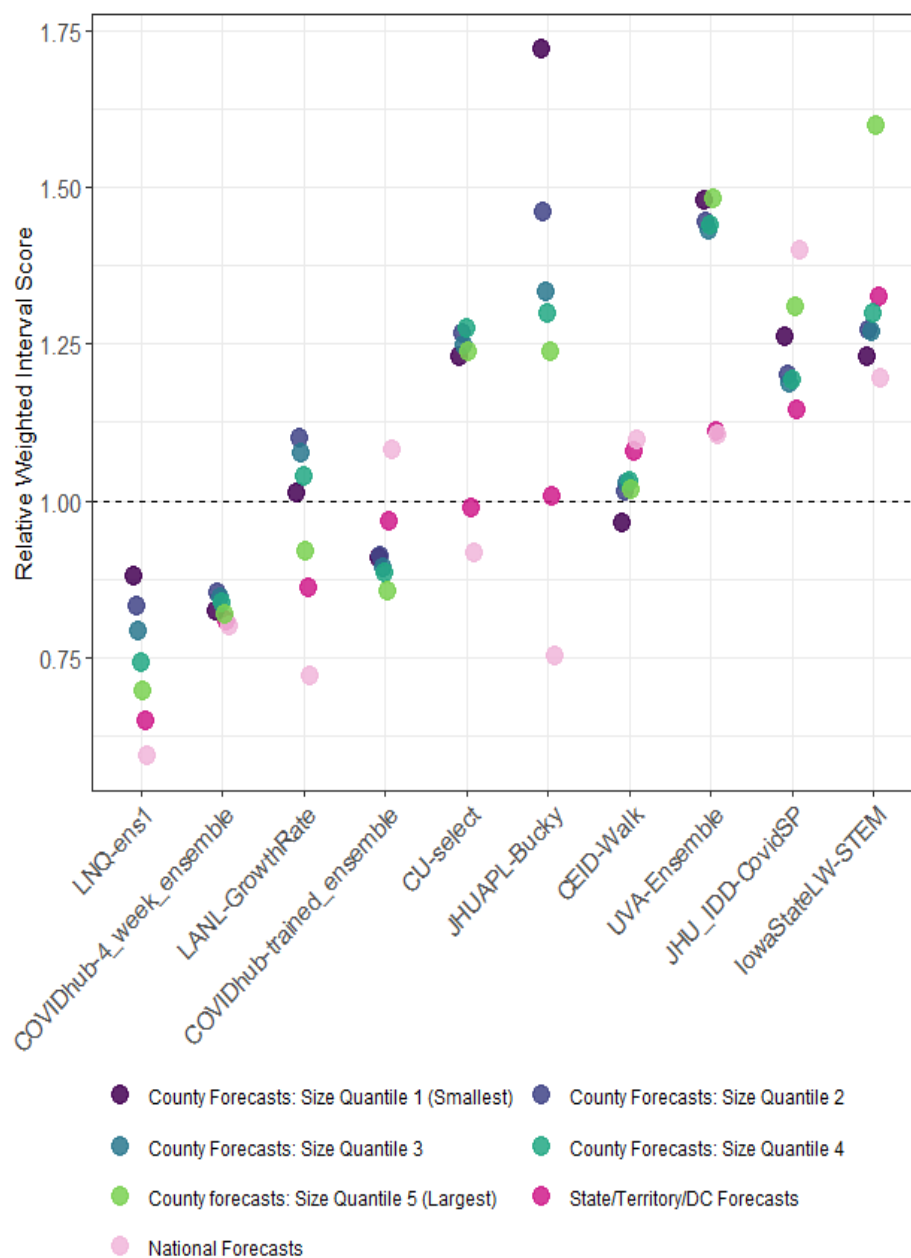
**Figure 2:** Percent of weeks with complete submissions for all sets of team forecasts, scaled, pairwise relative Weighted Interval Score (rWIS; see *Methods* for description), observed 95% prediction interval coverage, by geographical scale of submitted forecasts. Teams are sorted by increasing state/territory/DC rWIS values.



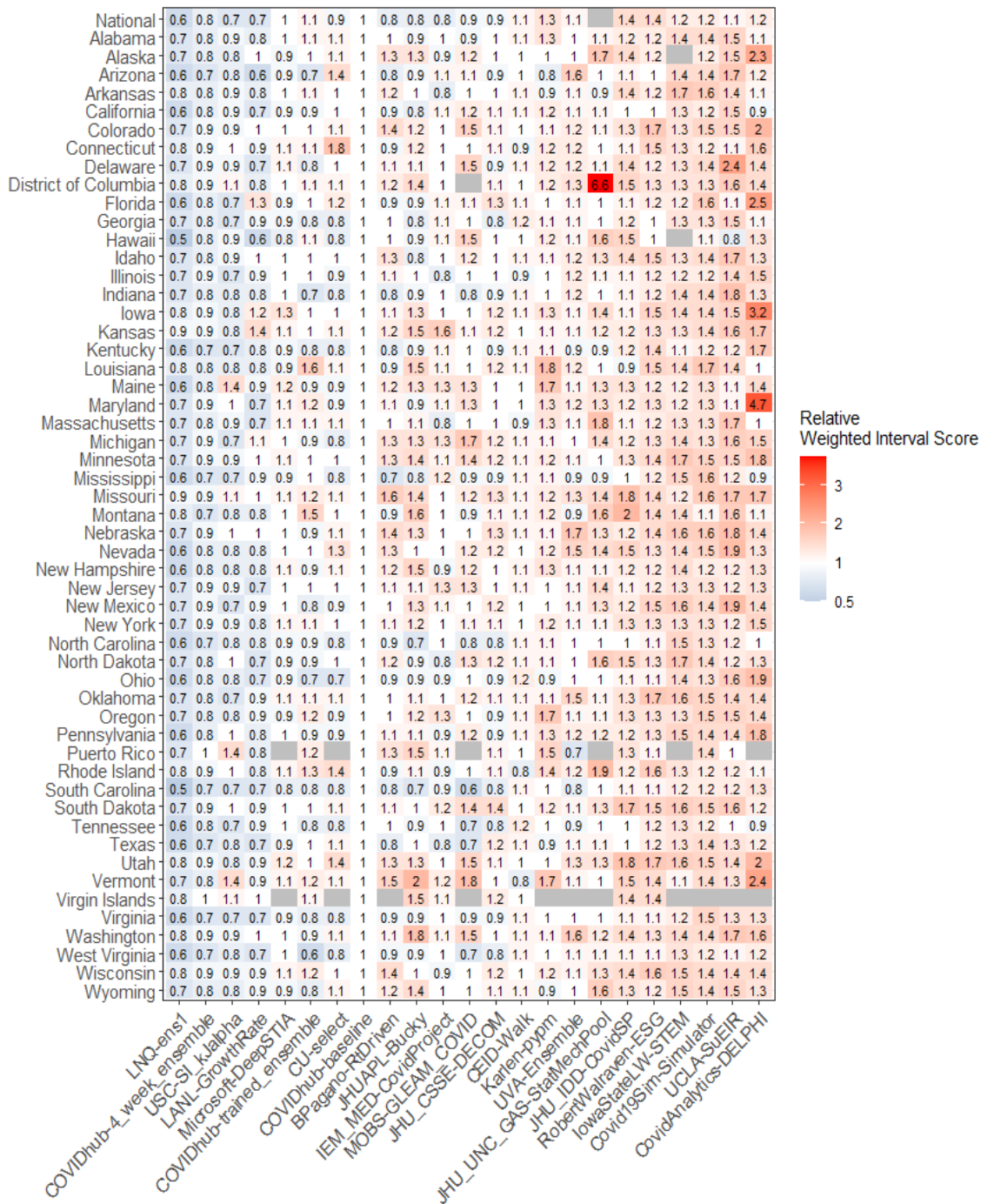
34 **Figure 3:** Expected and observed coverage rates for central 50%, 80% and 95% prediction intervals aggregated over time  
35 and horizon for national forecasts (panel A), state/territory/DC forecasts (panel B), the largest county forecasts (panel  
36 C). The dashed line represents optimal expected-coverage. Team forecasts that had closer to nominal coverage than the  
37 COVIDhub-4\_week\_ensemble model at all three coverage levels are labeled on the right side of the plots.



**Figure 4:** Scaled, pairwise relative Weighted Interval Score (rWIS) (see *Methods* for description) by spatial scale for sets of team forecasts that submitted forecasts for the US nation, states/territories/DC, and all US counties. WIS is averaged across all horizons. The COVIDhub-baseline model has, by definition, a rWIS of 1 (horizontal dashed line). Teams are ordered by increasing state/territory/DC rWIS with the most accurate model on the left. Points for each team are staggered horizontally to show overlapping WIS values.

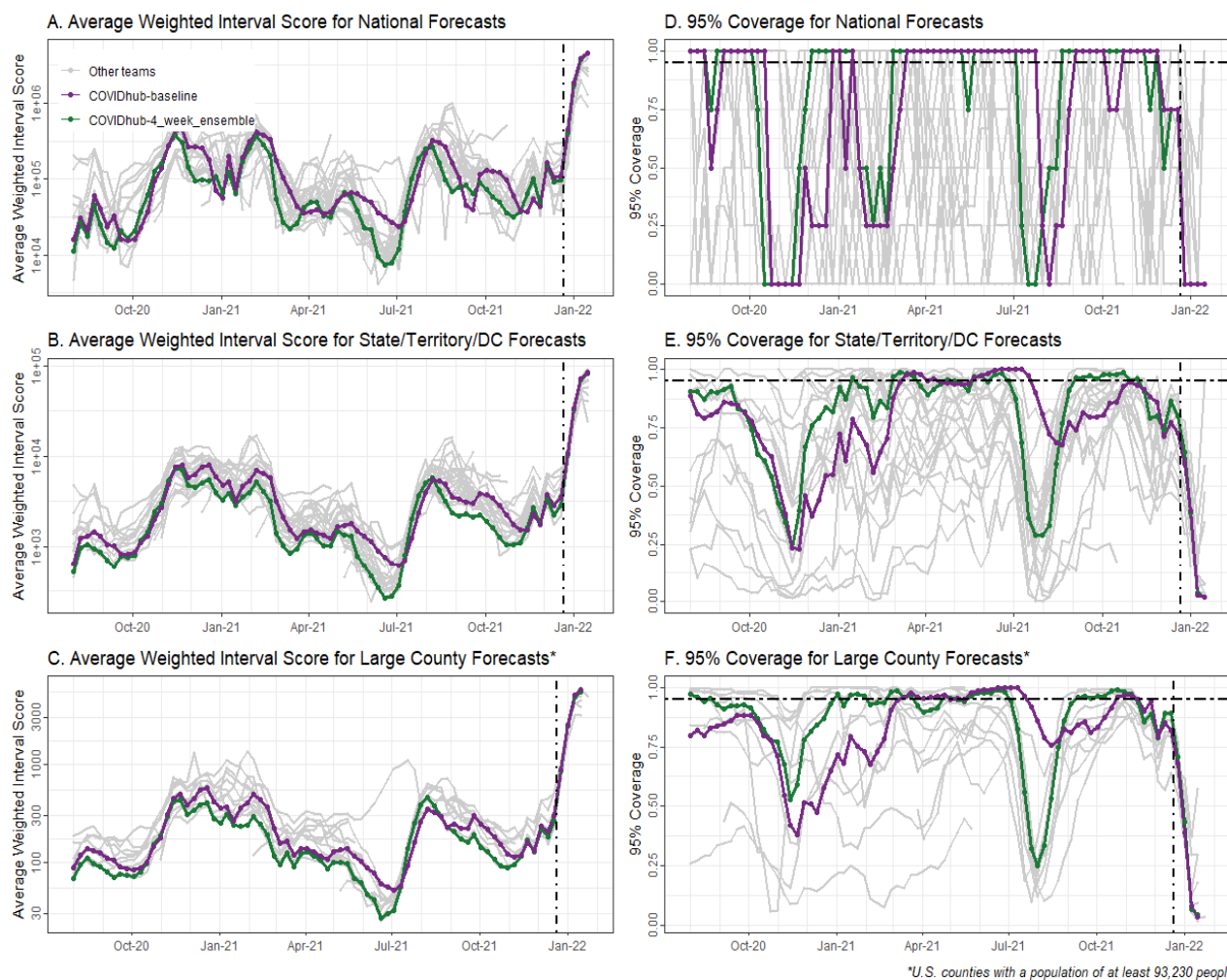


37 **Figure 5:** Scaled, pairwise relative Weighted Interval Score (rWIS; see *Methods* for description) by location for national  
 38 and state/territory/DC forecasts, averaged across all horizons through the entire analysis period. National estimates are  
 39 displayed first, followed by jurisdictions in alphabetical order. Team forecasts are ordered by increasing average  
 40 state/territory/DC rWIS.





**Figure 6:** Forecast accuracy over time, aggregated by geographic units, forecast horizon, and prediction date. Panels A-C show average Weighted Interval Score (WIS); panels D-F show 95% prediction interval coverage. The black, dashed vertical line in all panels shows the date that public communication of the case forecasts was paused. The black, dashed horizontal line in panels D-F shows nominal 95% interval coverage. National level forecasts are presented in A and D, state/territory/DC forecasts in B and E and large county level forecasts in C and F.



Panels A, D, B and E include: LNQ-ens1, Microsoft\_DeepSTIA, COVIDhub-4\_week\_ensemble, USC-SI\_kJalpha, CU-select, LANL-GrowthRate, JHU\_CSSE-DECOM, COVIDhub-trained\_ensemble, COVIDhub-baseline, Karlen-pypm, BPagano-RtDriven, JHUAPL-Bucky, UVA-Ensemble, IEM\_MED-CovidProject, CEID-Walk, Covid19Sim-Simulator, IowasStateLW-STEM, UCLA-SuEIR, JHU\_IDD-CovidSP, RobertWalraven-ESG, MOBS-GLEAM\_COVID, and CovidAnalytics-DELPHI.

Panels C and F include: LNQ-ens1, COVIDhub-4\_week\_ensemble, CU-select, LANL-GrowthRate, COVIDhub-trained\_ensemble, COVIDhub-baseline, JHUAPL-Bucky, UVA-Ensemble, CEID-Walk, JHU\_UNC\_GAS-StatMechPool, IowasStateLW-STEM, JHU\_IDD-CovidSP, UMass-MechBayes, FAIR-NRAR, FRBSF\_Wilson-Econometric.

08

09

10

11 **Figure 7.** Estimated marginal mean Weighted Interval Score (WIS) and 95% confidence intervals for mean cases from  
 12 team-specific GEE models for all 51 jurisdictions (Panel A). The 95% confidence intervals for the COVIDhub-baseline  
 13 model are shown in dashed red vertical lines. Panel B presents each team's estimated marginal mean WIS per phase,  
 14 scaled to the COVIDhub-baseline model's estimated marginal mean WIS for all epidemic phases. Teams with higher  
 15 estimated marginal mean WIS values (i.e., greater than 1.0) are presented in shades of orange while teams with lower  
 16 estimated marginal mean WIS (i.e., less than 1.0) are shown in shades of green. Forecasts for a team in a particular  
 17 phase are marked with an asterisk (\*) if the 80% confidence interval of the expected WIS outcome (normalized and on  
 18 the log scale) was estimated by a model to be lower than the expected WIS of the COVIDhub-baseline model for all  
 19 phases.

