

1 **Title:** Assessing the accuracy of California county level COVID-19

2 hospitalization forecasts to inform public policy decision making

3 Lauren A. White*¹, Ryan McCorvie¹, David Crow¹, Seema Jain¹, Tomás M. León¹

4 ¹California Department of Public Health, Richmond, CA, USA

5

6 **Corresponding Author:** *Lauren A. White, **Email:** lauren.white@cdph.ca.gov

7 **Abstract**

8 **Background:** The COVID-19 pandemic has highlighted the role of infectious disease
9 forecasting in informing public policy. However, significant barriers remain for
10 effectively linking infectious disease forecasts to public health decision making,
11 including a lack of model validation. Forecasting model performance and accuracy
12 should be evaluated retrospectively to understand under which conditions models were
13 reliable and could be improved in the future.

14 **Methods:** Using archived forecasts from the California Department of Public Health's
15 [California COVID Assessment Tool \(https://calcat.covid19.ca.gov/cacovidmodels/\)](https://calcat.covid19.ca.gov/cacovidmodels/), we
16 compared how well different forecasting models predicted COVID-19 hospitalization
17 census across California counties and regions during periods of Alpha, Delta, and
18 Omicron variant predominance.

19 **Results:** Based on mean absolute error estimates, forecasting models had variable
20 performance across counties and through time. When accounting for model availability
21 across counties and dates, some individual models performed consistently better than the
22 ensemble model, but model rankings still differed across counties. Local transmission
23 trends, variant prevalence, and county population size were informative predictors for
24 determining which model performed best for a given county based on a random forest
25 classification analysis. Overall, the ensemble model performed worse in less populous
26 counties, in part because of fewer model contributors in these locations.

27 **Conclusions:** Ensemble model predictions could be improved by incorporating
28 geographic heterogeneity in model coverage and performance. Consistency in model
29 reporting and improved model validation can strengthen the role of infectious disease
30 forecasting in real-time public health decision making.

31

32 **Keywords:** infectious disease modeling, forecasting, model evaluation, COVID-19,
33 public health

34 **Background**

35 In public health, forecasting has been used to predict infectious disease dynamics
36 for a variety of diseases including influenza, dengue fever, Ebola virus disease, Zika
37 fever, and most recently COVID-19, which has highlighted the importance of infectious
38 disease modeling to help inform public health decision making (1). Nevertheless,
39 significant barriers remain for effectively linking infectious disease forecasts with public
40 health decision making including a lack of model standardization and validation, and
41 difficulty in successfully communicating model complexity and uncertainty (2).
42 Moreover, public health practitioners may need different outcomes or indicators than
43 what forecast models provide (2,3).

44 In June 2020, as part of the COVID-19 response, the California Department of
45 Public Health's (CDPH) COVID Modeling Team launched the [California Communicable](#)
46 [diseases Assessment Tool \(CalCAT\)](#) to compile available COVID-19 models, mostly
47 from academic groups, to inform policy and public health action (4). CalCAT provides
48 nowcasts (R-effective estimates), forecasts (short-term predictions for hospitalizations,
49 ICU admissions, and deaths), and longer-range scenario models for a variety of COVID
50 indicators at the state, regional, and county scales. Some contributors are national,
51 forecasting for all states, while others focus only on California and may not be publicly
52 available elsewhere (Table 1). The models on CalCAT have been used throughout the
53 COVID-19 pandemic to evaluate current transmission trends and prospective hospital and
54 intensive care unit capacity. This information combined with other evidence and policy
55 considerations has helped to inform the implementation of stay-at-home orders and

56 statewide mask mandates (e.g., reinstating a mask mandate during the emergence of
57 Omicron/BA.1). In addition, models combined with other data streams were used to
58 inform metrics for the Blueprint for a Safer Economy including the nation's first health
59 equity metric and to support planning for vaccine allocation and distribution (5).

60 **Table 1. Constituent models providing county-level hospitalization census predictions that**
 61 **are archived on CalCAT and included in the analysis.**

Model	Forecast update frequency	Forecast horizon	Methods/Approach	Documentation
Columbia	Weekly	Up to 6 weeks	County level metapopulation model	(6)
UCSF, COVID NearTerm	Daily	2-4 weeks	Bootstrap-based method based on an autoregressive model	(7)
UCB LEMMA	Daily	Up to 4 weeks	SEIR compartmental model with parameters fit using case series data of COVID-19 hospital and ICU census, hospital admissions, deaths, cases and seroprevalence	(8)
CDPH Simple Growth	Daily	Up to 4 weeks	Assumes new cases grow exponentially according to the rate given by the latest ensemble R-effective. Assumes a fixed severity and average length of stay to generate hospitalizations	(4)
CalCAT Ensemble	Daily	Up to 4 weeks	The ensemble forecast takes the median of all the forecasts available on a given date and fits a smoothed spline to the trend.	(4)
CA Baseline	Daily	Up to 4 weeks	Retroactive 7-day rolling average mean of past hospitalization values	Methods

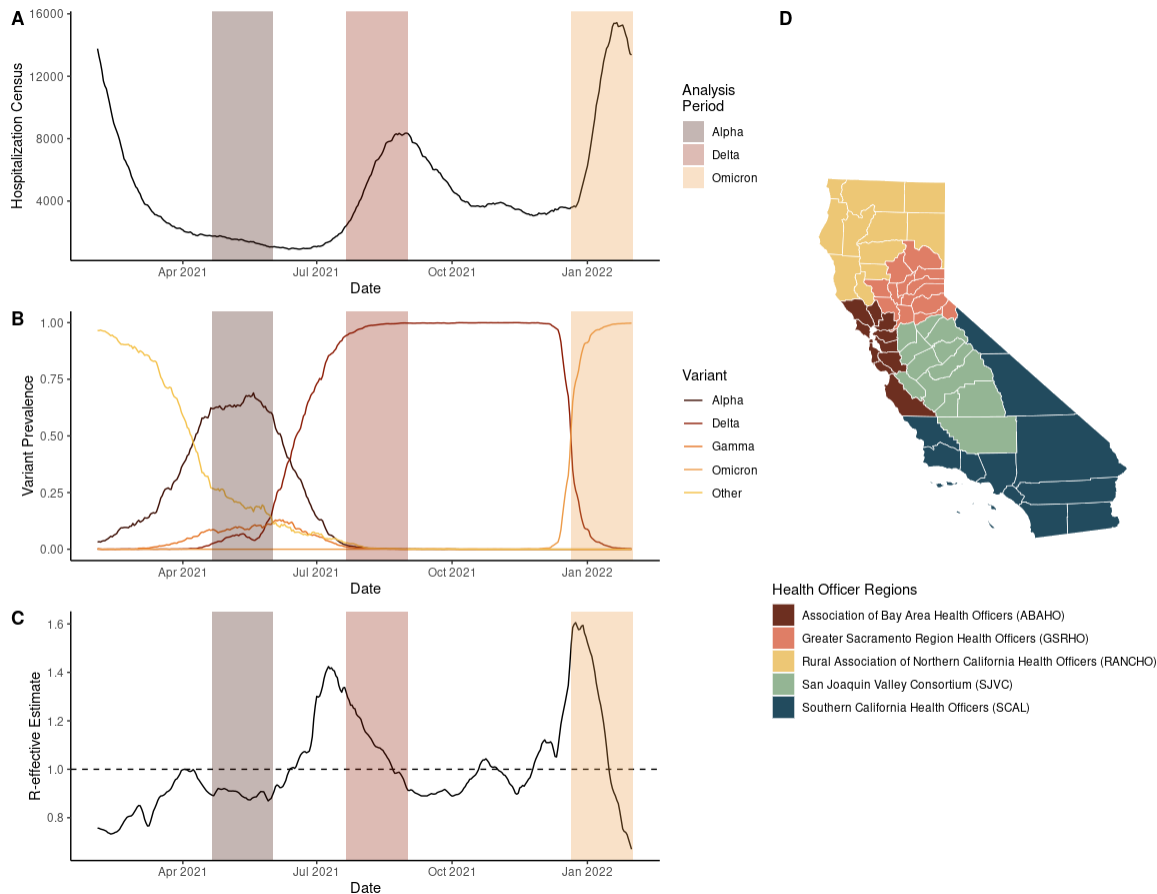
62

63 During the COVID-19 pandemic response, many California local health

64 jurisdictions communicated the importance of forecasts focused on the relevant scale of

65 decision making (e.g., county- vs. state-level forecasts) because there was significant
66 geographic heterogeneity in COVID-19 outcomes at regional and local levels (9). A
67 better understanding of how forecasting models have captured these geographical
68 heterogeneities could help inform local public health decision making during future
69 COVID-19 waves and enable local health jurisdictions to employ models judiciously
70 given proven past performance. Lessons learned from COVID-19 forecasting efforts can
71 also be applied to future modeling for other diseases including influenza.

72 We retrospectively evaluated archived forecasting predictions from CalCAT for
73 models that consistently provided county-level hospitalization census predictions across a
74 year long period from February 1, 2021 to February 1, 2022 (Table 1). Hereafter, we will
75 use hospital census to refer to the number of patients currently hospitalized with
76 confirmed COVID-19 for a given county and date. To explore the effects of COVID-19
77 variants on model performance within that period, we also compared forecasting model
78 accuracy during three phases of the COVID-19 pandemic at the county and regional level
79 in California (Figure 1 A-C) with different variant predominance: Alpha (April 22- June
80 1, 2021), Delta (June 21 - September 1, 2021), and Omicron (December 21, 2021 -
81 February 1, 2022). These periods also differed in their hospitalization burden (Figure 1A)
82 and epidemic growth rates (Figure 1C).



83

84 **Figure 1. Time courses of (A) California COVID-19 hospitalization census, (B) variant**

85 **prevalence, (C) statewide R-effective estimate, and (D) California health officer regions. The**

86 **period displayed for panels A-C corresponds to the complete period of analysis February 1, 2021-**

87 **February 1, 2022 used for the pairwise tournament and random forest analyses. Shaded regions**

88 **for panels A:C correspond to the dates of analysis for the three variant predominant periods:**

89 **Alpha, Delta, and Omicron.**

90

91

92 **Methods**

93 Multiple methods exist for measuring epidemic forecast accuracy including
94 metrics that evaluate specific point estimates and uncertainty (10). When full predictive
95 estimates are available, metrics like the logarithmic score or continuous ranked
96 probability score (CRPS) provide context for probabilistic models' predictions and
97 uncertainty. When forecasts are provided in quantile or interval formats, the weighted
98 interval score (WIS) is a potential alternative (10). Since not all models incorporated into
99 CalCAT provided full predictive or interval estimates, or did so with different reporting
100 standards, we focused on the median point estimates (50th percentile) from forecasting
101 models for hospital census at the county scale. In addition to these models, we
102 retroactively created a baseline California forecast that projected forward the 7-day
103 rolling mean from the prior week. Each forecast has the following properties: (1) model
104 (m): the organization or group issuing the forecast (Table 1); (2) location (j): the
105 geographic location for which a forecast was issued (in this case, at the county-level); (3)
106 publication date (i): the date that the forecast was displayed on CalCAT; and (4) target
107 end date (k): the future forecast horizon date for which the prediction was made.

108 We utilized mean absolute error (MAE) and relative error at the 7-, 14-, and 21-
109 day forecast horizons to evaluate the accuracy of these point estimates. To better compare
110 across counties with different population sizes, we normalized both error types by the
111 median hospital capacity of each county (14-day horizon results are highlighted in the
112 main text; the remaining forecast horizons are provided in the Supporting Information).
113 From the MAE scores, we computed a standardized ranking score for every forecasted

114 observation relative to other models issuing a prediction for that same publication date
115 and location (11). In addition, we also conducted pairwise tournaments of model
116 performance to control for the frequency of model participation. Finally, using a
117 classification regression approach, we explored which county-level epidemiological and
118 socio-economic covariates could help explain the “winning” model for a given location
119 and date based on the lowest MAE scores for a given forecast horizon.

120 County results are grouped by health officer regions, which are contiguous
121 groupings of 58 counties used for health mandates in California (Figure 1D): Association
122 of Bay Area Health Officers (ABAHO); Greater Sacramento Region Health Officers
123 (GSRHO); Rural Association of Northern California Health Officers (RANCHO); San
124 Joaquin Valley Consortium (SJVC); and Southern California (SCAL). Some counties do
125 not have major hospitals and therefore lack forecasting predictions, actual numbers of
126 hospitalizations, or both. For this reason, Alpine, Sierra, and Sutter Counties are not
127 included in the analyses that follow.

128 **Mean absolute error**

129 The raw mean absolute error (MAE) for each publication date i with associated
130 target end dates k is calculated as: $\frac{1}{N} \sum_{k=i}^{i+N} |x_{i,k} - \hat{x}_k|$ where N is the number of days into
131 the future that the forecast is made, $x_{i,k}$ is the prediction made on publication date i for
132 target end date k and \hat{x}_k is the actual observed value for a given target end date (11). We
133 then standardized the MAE by h , the median non-surge hospital capacity of a given
134 county: $\frac{MAE}{h} \cdot 100$

135 Median hospital capacity was chosen for standardization because the hospital capacity for
136 facilities, and aggregated for counties, changes through time based on staffing and other
137 factors. Note that not all model forecasts were available for all counties or all dates. A
138 model only received an MAE score for a given publication date if it had predictions
139 available for the target end dates of interest (e.g., to receive a 7-day MAE score, a model
140 must have made predictions for 1-7 days ahead of the publication date). Here we used
141 CA-state specific data (12) for post-hoc evaluation, whereas many model teams may be
142 relying on other data sources (e.g., U.S. Department of Health and Human Services) for
143 fitting or calibration.

144 **Relative error**

145 The relative error for each publication date (i) with associated target end dates k is
146 calculated as: $\frac{1}{N} \sum_{k=i}^{i+N} (x_{i,k} - \hat{x}_k)$ where N is the number of days into the future that the
147 forecast is made, $x_{i,k}$ is the prediction made on publication date i for target end date k
148 and \hat{x}_k is the actual observed value for a given target end date (11). We then standardized
149 the relative error by h , the median non-surge hospital capacity of a given county:

$$150 \frac{\text{relative error}}{h} \cdot 100$$

151 Therefore, a positive relative error corresponds to a model overestimating the hospital
152 census, while a negative relative error corresponds to an underestimation.

153 **Standardized ranking score**

154 For each publication date i and location j , we calculated a standardized rank for
155 every available model m based on its associated MAE: $sr_{m,i,j} = 1 - \frac{r_{m,i,j}-1}{n_{i,j}}$ where $r_{m,i,j}$
156 is the ranking of the MAE of model m out of the n other models that made predictions for
157 publication date i and location j (adapted slightly from (11)). Thus, for a given
158 publication date i and location j , the highest possible standardized ranking score for any
159 given model is 1 and the lowest is the inverse of the lowest possible ranking ($1/n_{i,j}$).
160 Models not participating for a given publication date i and location j receive a zero, and
161 thus, are penalized for lack of coverage.

162 **Pairwise tournament**

163 To conduct a pairwise tournament, we calculated a relative MAE for each pair of
164 models m and m' : $\theta_{m,m'} = \text{median} \left\{ \frac{MAE(m; i,j,k)}{MAE(m'; i,j,k)} \right\}$ where $\theta_{m,m'}$ is the median of the ratio
165 of the simultaneously available MAE scores for model m to model m' with shared
166 publication dates i , target end dates k , and locations j (13). Importantly, the common
167 locations, publication dates, and observation dates may differ for each pair of models m
168 and m' . This approach varies slightly from some previous examples, as the order of
169 operations is scale then aggregate rather than aggregate then scale (11,14).

170 An overall performance score of a given model, m is then calculated as the
171 geometric mean of all relative MAE scores: $\theta_m = \left(\prod_{m'=1}^M \theta_{m,m'} \right)^{1/M}$ where M is the
172 total number of all models available for comparison. At the county level for counties with
173 smaller hospital capacities, there was a non-trivial probability of certain models achieving

174 an MAE of zero, which leads to relative MAE scores of zero or infinity depending on the
175 order of comparison. To eliminate these irregularities in the pairwise comparisons, we
176 excluded counties with median non-surge hospital capacities ≤ 25 (i.e., Calaveras,
177 Lassen, Mariposa, Modoc, Mono, San Benito, and Trinity).

178 **Random forest classification analysis**

179 To explore whether the model with the lowest MAE score for a given location and
180 observation date could be explained by county-specific epidemiological or
181 socioeconomic factors, we conducted a random forest classification analysis. Random
182 forest analysis is a recursive partitioning method that improves classification accuracy by
183 synthesizing the predictions from many classification trees (15,16). The response variable
184 (i.e., classification label) was the best performing model for a given county and date
185 combination based on the lowest MAE score of the available models. We explored the
186 covariates (i.e., features) of: progressive vaccination coverage at the county level, county-
187 level R-effective, 7-day change in county-level R-effective, variant prevalence at the
188 health officer region level (17), county population size, percent of county residents in
189 poverty (2019), percent unemployment (2020), median income (2019), five-year average
190 percentage completing college (university degree) (2015-2019), and 2013 Rural-Urban
191 Continuum code. All socioeconomic variables were taken from U.S. Department of
192 Agriculture Economic Research Service county-level data sets (18). For pre-processing,
193 data were centered and scaled using the caret package (19). For model training and
194 tuning, 70% of original data was used with K-fold validation (four-fold, repeated four
195 times). The final accuracy of the random forest classification models were 61% with mtry

196 = 7, 66% with mtry = 7, and 68% with mtry = 7 for 7, 14, and 21 day forecast horizons
197 respectively.

198 **Data and code availability**

199 The forecasts and R-effective values analyzed in this paper are available from
200 CalCAT (4). California-specific hospitalization data is available on the California Open
201 Data Portal (12). Because of reporting delays and backfilling, datasets used in the
202 analysis may represent a snapshot of what was available at that point in time. All data and
203 code used for analysis and figure generation is available in the public repository:
204 <https://doi.org/10.5281/zenodo.7851280>. Analyses were performed in R (v 3.6.0) (20).

205 **Results**

206 **Model performance varied across locations and under different periods of variant** 207 **predominance**

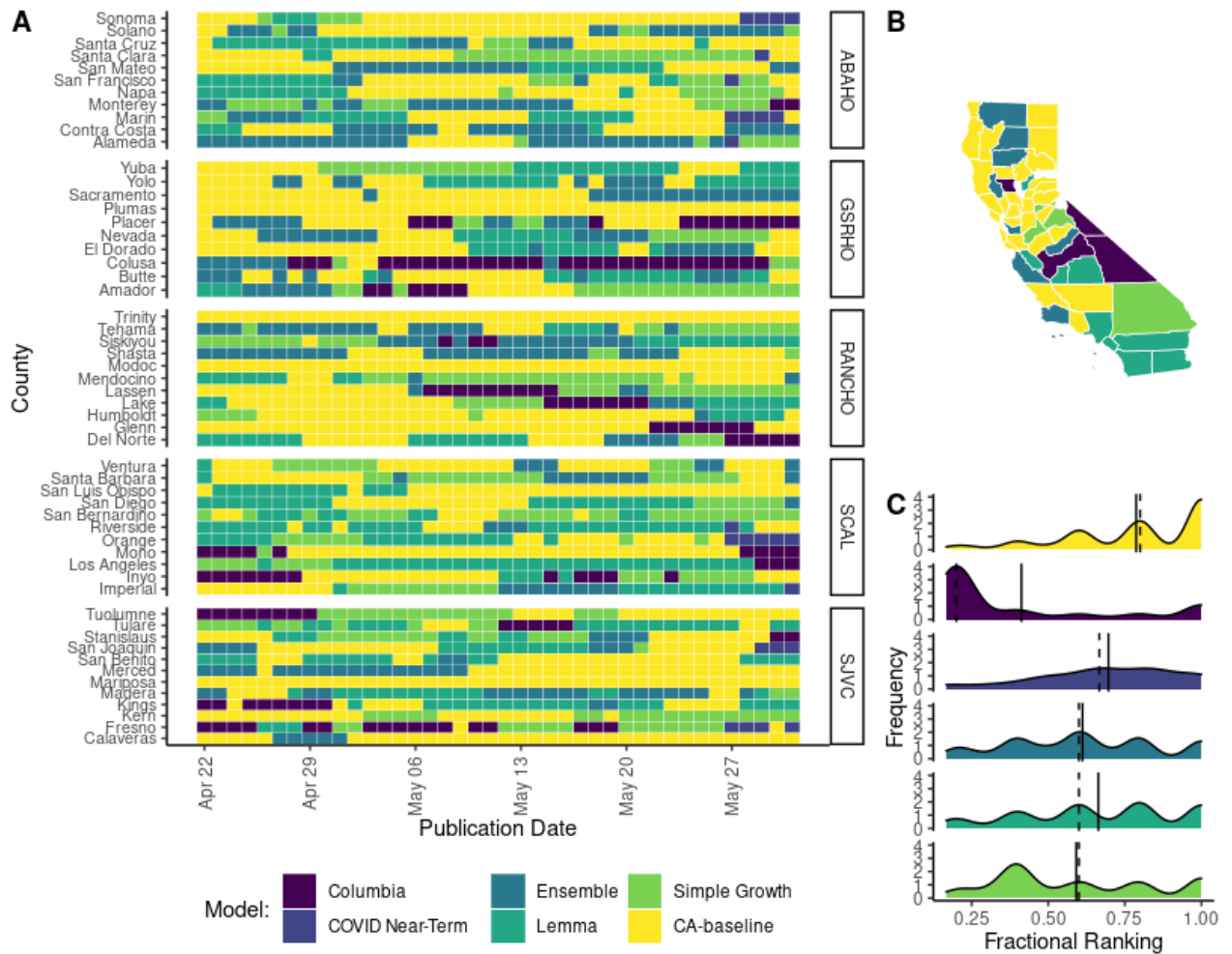
208 Model performance was heterogeneous across counties and during different
209 periods of variant predominance (Figure 2A, 3A, 4A), in part reflecting that the number
210 of models available for a given publication date and location varied through time; fewer
211 models were available during the Omicron variant period and for less populous health
212 officer regions such as RANCHO (Supplementary Figures 3, 11, 15). For example, in
213 Trinity County – one of California’s least populous counties – the Simple Growth model
214 had the lowest 14-day normalized MAE for most forecast publication dates during the
215 Alpha and Omicron predominant periods (Figure 2A, 4A), whereas the Columbia model
216 had the lowest 14-day normalized MAE during the Delta period (Figure 3A). In San
217 Diego County, California’s second most populous county, the LEMMA model had the

218 lowest 14-day normalized MAE during the Alpha period (Figure 2A), and the COVID
219 NearTerm model had the lowest 14-day MAE for the most days during Delta and
220 Omicron periods (Figure 3A & 4A). Overall, the Simple Growth model performed
221 particularly well in the RANCHO region during the Omicron period as demonstrated by a
222 lower 14-day MAE for many counties in that region (Figure 4A). The LEMMA model
223 had the lowest 14-day MAE across many regions during the Omicron period on or after
224 January 13, 2022 (Figure 4A). In general, the range of the relative error distributions
225 increased with longer time horizons and during the Omicron period (Supplementary
226 Figures 1-2). During the Omicron period, most relative error distributions were right
227 skewed with median relative error values less than zero, indicating a tendency for
228 underprediction, but a non-zero probability of sizeable overprediction (Supplementary
229 Figure 1).

230 The sum of the standardized rank score ($\sum sr_{m,i,j}$) in each county, j , rewards both
231 performance (model accuracy) and frequent participation (model coverage). During the
232 Alpha period, the LEMMA model had the highest score in 20/55 counties, and the
233 Ensemble model was a close second with the highest score in 17/55 counties (Figure 2B).
234 During the Delta period, the Ensemble model had the highest score in 21/55 counties
235 (Figure 3B). During the Omicron period, the Ensemble model had the highest score in
236 22/55 counties, and the Simple Growth model was a close second with 20/55 counties
237 (Figure 4B).

238 The density distributions of standardized rank ($sr_{m,i,j}$) allow for comparison of
239 model performance while controlling for frequency of model participation (Figures 2C,

240 3C, 4C). Although COVID NearTerm did not have the highest sum of the standardized
241 rank score in any counties, it had the highest median standardized rank score during the
242 Alpha and Delta periods (Figures 2C, 3C). The LEMMA model had the highest median
243 standardized. rank score during the Omicron period (Figure 4C). The same pattern of
244 ranking was present for 7-day MAE (Supplementary Figures 4-6). For 21-day MAE, the
245 COVID NearTerm model had the highest median standardized rank score during the
246 Alpha and Omicron periods (Supplementary Figures 16 & 18), while the LEMMA model
247 had the highest median rank scores during the Delta period (Supplementary Figure 17).



248

249 **Figure 2. Forecasting accuracy results at the county level during the Alpha wave in**

250 **California as measured by mean absolute error (MAE). (A) Heat map of the best daily**

251 **performing model for a given prediction date as measured by 14-day MAE. Each cell in the heat**

252 **map corresponds to a normalized MAE calculated for the day that a model forecast was**

253 **published. Counties are grouped into panels by California health officer regions. (B) A summary**

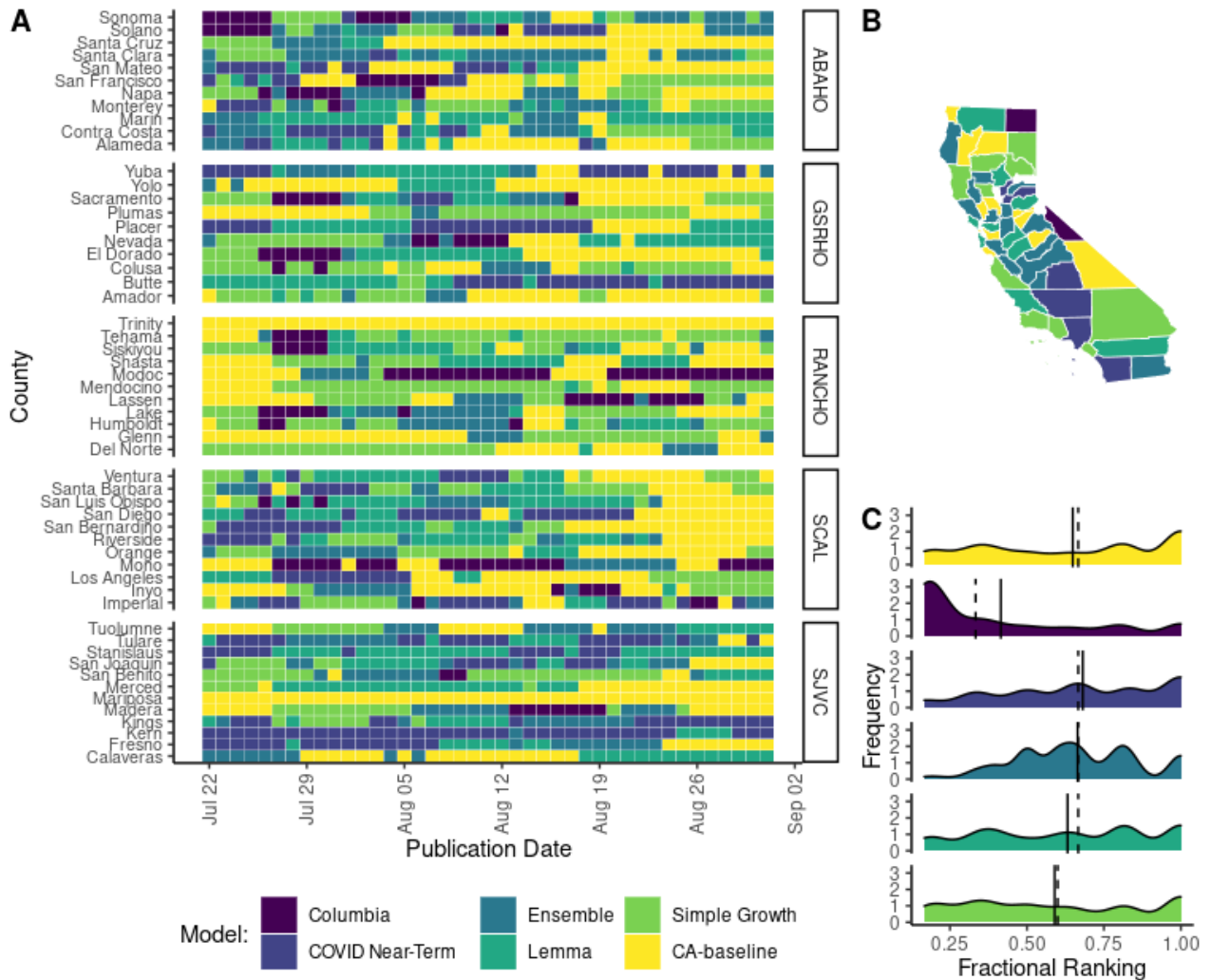
254 **map of California where the color of the county corresponds to the model with the highest sum of**

255 **the standardized rank score for that period ($\sum sr_{m,i,j}$). Note that by using the summation of the**

256 **standardized ranking score, models are penalized for lack of participation. (C) A density**

257 **distribution of the standardized rank score ($sr_{m,i,j}$) that depicts the median (dashed) and mean**

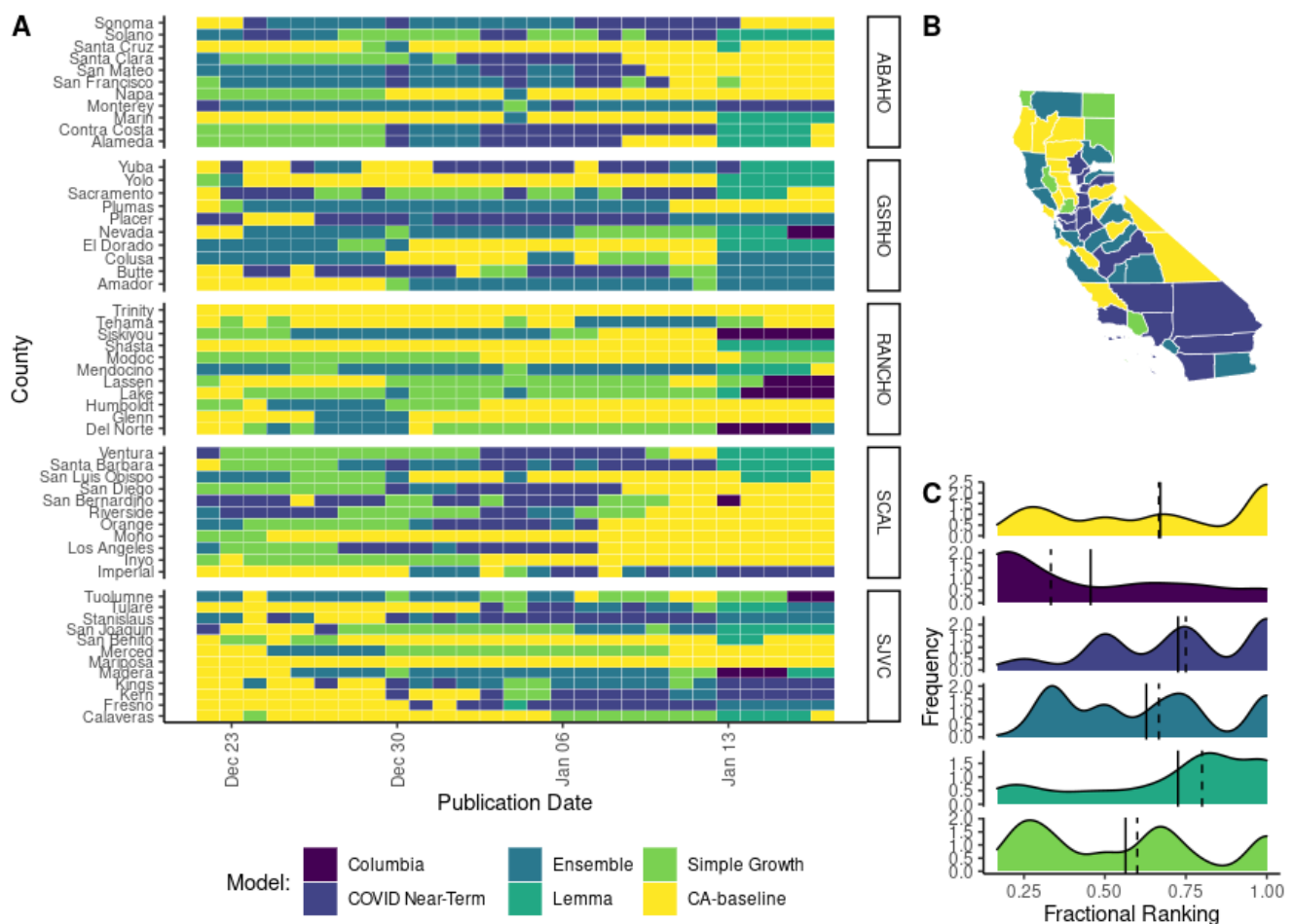
258 (solid) as vertical lines for each model distribution. A standardized rank score of one indicates
 259 that a model came in first relative to other participating models for a given date and location,
 260 values closer to zero indicate that a model had a lower ranking compared to other participating
 261 models, and a value of zero corresponds to no participation.



262

263 **Figure 3. Forecasting accuracy results at the county level during the Delta wave in**
 264 **California as measured by mean absolute error (MAE). (A)** Heat map of the best daily
 265 performing model for a given prediction date as measured by 14-day MAE. Each cell in the heat
 266 map corresponds to a standardized MAE calculated for the day that a model forecast was

267 published. Counties are grouped into panels by California health officer regions. **(B)** A summary
 268 map of California where the color of the county corresponds to the model with the highest sum of
 269 the standardized rank score for that period ($\sum sr_{m,i,j}$). Note that by using the summation of the
 270 standardized ranking score models are penalized for lack of participation. **(C)** A density
 271 distribution of the standardized rank score ($sr_{m,i,j}$) that depicts the median (dashed) and mean
 272 (solid) as vertical lines for each model distribution. A standardized rank score of one indicates
 273 that a model came in first relative to other participating models for a given date and location,
 274 values closer to zero indicate that a model had a lower ranking compared to other participating
 275 models, and a value of zero corresponds to no participation.
 276



277

278 **Figure 4. Forecasting accuracy results at the county level during the Omicron wave in**
279 **California as measured by mean absolute error (MAE).** (A) Heat map of the best daily
280 performing model for a given prediction date as measured by 14-day MAE. Each cell in the heat
281 map corresponds to a standardized MAE calculated for the day that a model forecast was
282 published. Counties are grouped into panels by California health officer regions. (B) A summary
283 map of California where the color of the county corresponds to the model with the highest sum of
284 the standardized rank score for that period ($\sum sr_{m,i,j}$). Note that by using the summation of the
285 standardized ranking score models are penalized for lack of participation. (C) A density
286 distribution of the standardized rank score ($sr_{m,i,j}$) that depicts the median (dashed) and mean
287 (solid) as vertical lines for each model distribution. A standardized rank score of one indicates
288 that a model came in first relative to other participating models for a given date and location,
289 values closer to zero indicate that a model had a lower ranking compared to other participating
290 models, and a value of zero corresponds to no participation.

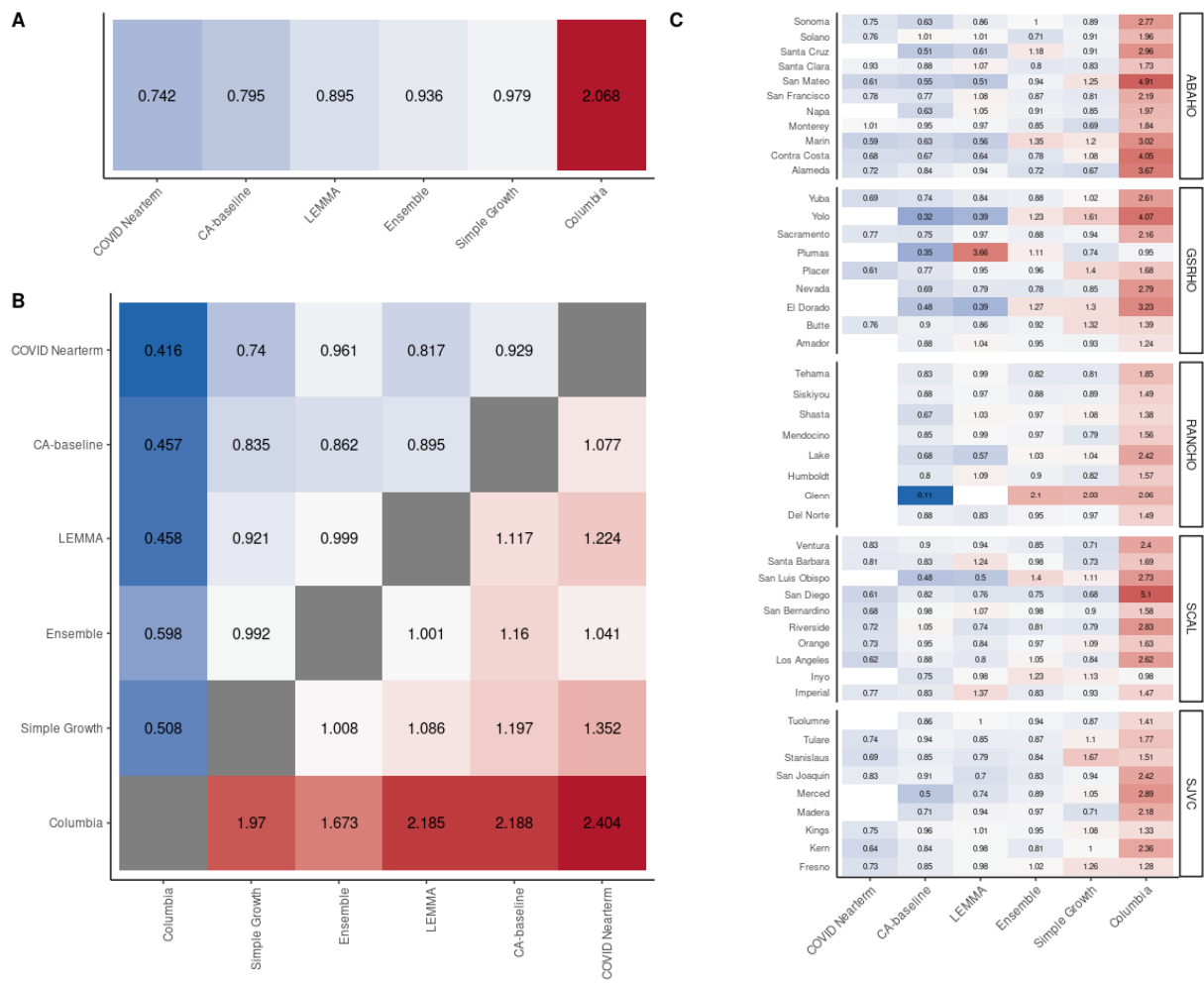
291

292 **When controlling for participation, some models outperformed the ensemble, but**
293 **pairwise model rankings varied across counties**

294 When matching across all locations and all observation dates, two models—
295 COVID NearTerm and LEMMA—performed better in pairwise comparisons relative to
296 the Ensemble model for 14-day MAE (Figure 5A & B). However, pairwise rankings were
297 quite variable when disaggregated by county and also highlighted the differences in
298 coverage and availability across locations for different models (Figure 5C). For example,
299 although the Simple Growth model came fourth in the overall pairwise ranking (Figure
300 5A), it came first in twelve individual counties. Similarly, the Columbia model came last

301 in the overall pairwise ranking (Figure 5A) and generally performed worse than average
 302 ($\theta_m > 1$), but performed better than average ($\theta_m < 1$) in Plumas and Inyo counties
 303 (Figure 5C).

304 Overall pairwise rankings were robust to forecast horizon length for the complete
 305 analysis period (Supplementary Figures 5A & 17A). However, overall pairwise rankings
 306 were more unstable during specific periods of variant predominance, particularly for
 307 shorter forecast horizons (Supplementary Figures 6-8, 10-12, 18-20) and for county-
 308 specific rankings (Supplementary Figures 7C-9C, 12C-14C, 19C-21C).



309

310 **Figure 5. Pairwise tournament median rankings of models for the whole analysis period for**
311 **14-day MAE. (A)** Overall median rankings (θ_m) across all locations and observation dates. **(B)**
312 Median pairwise rankings ($\theta_{m,m'}$) comparing each model m relative to every other model m' .
313 The grid is symmetrical, so the ratio of model m : model m' is the inverse score of the ratio of
314 model m' : model m . **(C)** Overall median rankings for all available observation dates
315 disaggregated by county.

316 **Epidemiological traits, county population size, and variant traits best predicted**
317 **forecast “winners”**

318 For the entire analysis period (February 1, 2021-February 1, 2022), time-varying
319 vaccine coverage at the county-level, local transmission dynamics (R-effective and 7-day
320 change in R-effective), county population size, and regional proportion of variants, were
321 most important in predicting which model had the lowest MAE for a given county on a
322 given publication date (Supplementary Figure 23). Other static socio-economic variables
323 like income, percent unemployment, percent of residents with a university degree, and
324 percent of residents in poverty were less important for predicting model outcomes. These
325 variable importance rankings were robust to the forecast horizon used for MAE
326 calculations (Supplementary Figure 23).

327 **Less populated counties have ensemble predictions with higher median MAE and**
328 **more variable MAE**

329 When comparing 14-day MAE normalized by hospital capacity, counties with
330 smaller population sizes typically had a higher median MAE score and more variable

331 MAE distributions compared to more populous counties (Supplementary Figure 24B).
332 Based on a linear regression, the logarithmic of the normalized MAE score was
333 negatively correlated with county population size (coefficient estimate: $1.7 \cdot 10^{-7}$; p-
334 value $< 2 \cdot 10^{-16}$). This relationship held true regardless of the forecast horizon used for
335 MAE calculations (Supplementary Figure 24 A & C).

336

337 **Discussion**

338 **Ensemble model could be improved by incorporating geographic heterogeneity in** 339 **model coverage and performance**

340 Echoing other analyses of COVID-19 forecast performance that have described a
341 large variation in model accuracy by location (11,21), forecasting models performed
342 differentially across California counties and regions and for different periods of variant
343 predominance during the COVID pandemic (Figures 2-4, 5C, Supplementary Figures 4-
344 6, 16-18). Moreover, location-specific features like local transmission dynamics or
345 county population size helped explain model performance (Supplementary Figure 24).
346 This geographic variation in model performance points to the importance of location-
347 specific model evaluation in order for local health jurisdictions to best employ forecasts
348 for public health decision making.

349 In general, combining multiple models into ensembles allows for better
350 performance (11,22–25). However, in this case, COVID NearTerm and LEMMA
351 consistently outperformed the Ensemble model when controlling for frequency of

352 participation (Figures 2-4C, Figure 5), although pairwise ranking scores remained
353 variable at the county level (Figure 5C). The higher performance of individual models
354 over the Ensemble model combined with the variability in performance at the county-
355 level suggests that the Ensemble model does not have to be applied uniformly across all
356 locations; public health decision making could benefit from model selection and
357 ensemble weighting that reflects location-specific past performance as well as local
358 transmission trends (26).

359 **Lower forecast coverage in less populated counties weakens evidence-based decision**
360 **making**

361 One interesting question from a public health decision making context is whether
362 model coverage (i.e., frequent issuing of forecasts across all potential locations) and
363 model accuracy should be weighed equally when establishing the criteria for a “winning”
364 forecast. In this analysis, there was typically a mismatch between raw model performance
365 based on availability as measured by the sum of standardized ranking (Figures 2-4) and
366 model performance when controlling for participation via pairwise tournaments (Figure
367 5). In part, this disagreement reflects that not all models provided estimates for all
368 counties, especially for less populous regions or counties (Supplementary Figures 1, 9,
369 13). For example, the COVID NearTerm model ranked first in the pairwise ranking but
370 provided no coverage for any counties in the less populous RANCHO region (Figure 5C).
371 In contrast, the Ensemble model came first in the majority of counties during the Delta
372 and Omicron periods as measured by sum of the standardized rank score for that period
373 ($\sum sr_{m,i,j}$) (Figures 3B, 4B), but was generally outperformed in pairwise ranking

374 evaluations both overall and for individual counties (Figure 5). Although the Ensemble
375 model in less populous counties exhibited a higher median normalized MAE and a more
376 variable normalized MAE regardless of the forecast horizon (Supplementary Figure 22),
377 this observation may be a direct result of calculating MAE from median point estimates
378 rather than accounting for forecast uncertainty, since stochastic effects likely contribute
379 more significantly to the forecast predictions for counties with smaller population sizes.

380 While maximizing model accuracy is important, a forecast cannot add value if it
381 is not available for decision making. As county-level contributors are lost to attrition,
382 ensemble estimates may further decrease in accuracy or may not be possible in these less
383 populous counties. Policy and public health decision makers should evaluate what
384 investments or innovations in modeling are needed to improve results for underserved
385 counties with lower forecast coverage. In addition, decision makers could seek to
386 incentivize the best-performing models to serve smaller counties that neither have the
387 resources to do this work in-house nor have academic partners readily available.

388 The lack of coverage in smaller counties also points to the inherent complexities
389 of interpreting in hospitalization burden—since hospitalizations are typically recorded via
390 hospital location rather than patient residency (27). As others have suggested, forecasting
391 at the geographic unit of hospital referral networks could be another solution to low
392 model coverage in less populous counties (27).

393 **Continuity of contributors, forecast structure, and documentation helps real-time**
394 **public health decision making and post-hoc analysis**

395 The overall continuity of forecasting contributions has proved challenging for
396 post-hoc evaluation. Although CalCAT has had roughly ten unique forecast contributors
397 through time, many of these groups have ceased contributing as the COVID-19 landscape
398 has increased in complexity (e.g., emerging variants, prior immunity, boosters). Although
399 less relevant to forecasting hospitalizations, changes in case ascertainment and testing
400 practices make retrospective analyses more challenging. Interruptions to forecast
401 continuity can also limit post-hoc evaluation. For example, some modeling groups paused
402 forecasts in order to reset or recalibrate for new variants like Omicron.

403 While initiatives like the COVID-19 Forecast Hub have worked to standardize
404 forecast output and reporting (11), one additional challenge for this analysis was that the
405 reporting across external forecast contributors differed. For example, across three of the
406 externally contributed forecasts they all produced interval estimates at different cutoff
407 points: COVID Nearterm (10, 20, 30, 40, 50, 60, 70, 80, and 90 percentiles), Columbia
408 (2.5, 25, 50, 75 and 97.5 percentiles), and LEMMA (5, 50, 95 percentiles). This
409 discrepancy precluded the use of more robust measures like CRPS or WIS and means that
410 that our results are much more sensitive to the median point estimates (10). Changing
411 repository structures, file nomenclature, and data formatting can also disrupt the
412 archiving process necessary for ensemble generation and subsequent post-hoc review.
413 This analysis is a snapshot of what was available on CalCAT—and therefore to the
414 general public and public health decision makers—and may not entirely reflect what

415 model contributors would intend to be their contributing forecast at all times. The
416 CalCAT team updated data and results iteratively as often as possible, but not all model
417 changes were announced. As the COVID-19 pandemic necessitated rapid changes in data
418 reporting and data infrastructure, other information technology issues may have
419 introduced unintended errors.

420 The current classification regression analysis in this manuscript does not include
421 model-specific traits. In order to truly evaluate whether underlying model traits and
422 assumptions help to explain performance for specific locations through time, it would be
423 necessary to have a larger number of forecasting contributors and consistent metadata on
424 both the changes in model construction and the timing of those changes. Therefore,
425 another potential area of documentation might include not just existing model
426 assumptions and structure but how those characteristics have changed over time. This
427 analysis may be easier to do at a state or national scale, where more model contributors
428 are available, and reporting is better standardized through initiatives like the Forecast and
429 Scenario Hubs.

430 Reporting and communicating infectious disease forecasting results, with all their
431 inherent uncertainty and complexity, remain areas for improvement and growth for public
432 health departments and their academic and industry collaborators to support evidence-
433 based public health policy planning and decision making. Importantly, forecasting
434 models may also serve as a communication tool to influence behavior change by the
435 general public. One phenomenon not explored in this analysis is the potential for
436 forecasts to alter human behavior, and subsequently model accuracy.

437 **Conclusions**

438 Major progress in infectious disease forecasting has been made during the
439 COVID-19 pandemic, while ongoing challenges, such as those around data and
440 communication, have persisted. We retrospectively investigated hospitalization census
441 forecast model performance at the county level in the state of California. Model
442 performance and ranking varied through space and time and by metric, highlighting the
443 difficulty of making blanket recommendations for which models to use for individual
444 counties, including an ensemble approach. Calibrating based on past model performance
445 may help improve ensemble forecast generation, and counties may benefit by considering
446 which individual model contributors have historically served them the best. Going
447 forward, closer collaboration between forecasters, researchers, and policymakers may
448 create positive feedback loops that inform the ongoing COVID-19 response and other
449 future public health action.

450

451 **List of abbreviations**

- 452 • Association of Bay Area Health Officers (ABAHO)
- 453 • [California Communicable diseases Assessment Tool \(CalCAT\)](#)
- 454 • Greater Sacramento Region Health Officers (GSRHO)
- 455 • Mean absolute error (MAE)
- 456 • Rural Association of Northern California Health Officers (RANCHO)
- 457 • San Joaquin Valley Consortium (SJVC)
- 458 • Southern California (SCAL)

459

460 **Declarations**

461 **Ethics approval and consent to participate**

462 Not applicable

463

464 **Consent for publication**

465 Not applicable

466

467 **Availability of data and materials**

468 The forecasts and R-effective values analyzed in this paper are available from CalCAT
469 (4). California-specific hospitalization data is available on the California Open Data
470 Portal (12). Because of reporting delays and backfilling, datasets used in the analysis may
471 represent a snapshot of what was available at that point in time. All data and code used
472 for analysis and figure generation is available in the public repository:
473 https://github.com/whit1951/CA_COVID_Forecasting_Accuracy.

474

475 **Competing interests**

476 The authors declare that they have no competing interests.

477

478 **Funding**

479 This work was supported by the California Department of Public Health. The findings
480 and conclusions in this article are those of the authors and do not necessarily represent
481 the views or opinions of the California Department of Public Health or the California
482 Health and Human Services Agency. This work was funded by Centers for Disease

483 Control and Prevention, Epidemiology and Laboratory Capacity for Infectious Diseases,
484 Cooperative Agreement Number 6 NU50CK000539.

485

486 **Authors' contributions**

487 LW and TL designed research and wrote the paper. LW performed the research and
488 analyzed data. RM and DC contributed new analytic tools. RM, DC, and SJ
489 revised/edited the manuscript. SJ supervised the project. All authors read and approved
490 the final manuscript.

491

492 **Acknowledgments**

493 The authors thank the CMU Delphi Group including Ryan Tibshurani and Daniel
494 J. McDonald, MIDAS members, and the COVID-19 Forecasting Hub for discussion and
495 feedback. The authors also thank Californian local health jurisdictions and members of
496 the CDPH Modeling and Advanced Analytics team including Chris Hoover, Mugdha
497 Thakur, Natalie Linton, Phoebe Lu, Sindhu Ravuri, and Sophie Zhu for conversations and
498 insights that improved these analyses.

499

500

501

502 **References**

- 503 1. Bertozzi AL, Franco E, Mohler G, Short MB, Sledge D. The challenges of modeling
504 and forecasting the spread of COVID-19. *Proc Natl Acad Sci*. 2020 Jul
505 21;117(29):16732–8.
- 506 2. Lutz CS, Huynh MP, Schroeder M, Anyatonwu S, Dahlgren FS, Danyluk G, et al.
507 Applying infectious disease forecasting to public health: a path forward using
508 influenza forecasting examples. *BMC Public Health*. 2019 Dec;19(1):1659.
- 509 3. Doms C, Kramer SC, Shaman J. Assessing the Use of Influenza Forecasts and
510 Epidemiological Modeling in Public Health Decision Making in the United States.
511 *Sci Rep*. 2018 Dec;8(1):12406.
- 512 4. California Department of Public Health. California COVID Assessment Tool
513 [Internet]. 2022. Available from: <https://calcat.covid19.gov/cacovidmodels/>
- 514 5. California Department of Public Health. Blueprint for a Safer Economy [Internet].
515 2021 [cited 2023 Mar 6]. Available from:
516 [https://www.cdph.ca.gov/Programs/CID/DCDC/Pages/COVID-](https://www.cdph.ca.gov/Programs/CID/DCDC/Pages/COVID-19/COVID19CountyMonitoringOverview.aspx)
517 [19/COVID19CountyMonitoringOverview.aspx](https://www.cdph.ca.gov/Programs/CID/DCDC/Pages/COVID-19/COVID19CountyMonitoringOverview.aspx)
- 518 6. Pei S, Shaman J. Initial Simulation of SARS-CoV2 Spread and Intervention Effects
519 in the Continental US [Internet]. *Epidemiology*; 2020 [cited 2022 Mar 22]. Available
520 from: <http://medrxiv.org/lookup/doi/10.1101/2020.03.21.20040303>
- 521 7. Olshen AB, Garcia A, Kapphahn KI, Weng Y, Wesson PD, Rutherford GW, et al.
522 COVIDNearTerm: A Simple Method to Forecast COVID-19 Hospitalizations
523 [Internet]. *Infectious Diseases (except HIV/AIDS)*; 2021 Oct [cited 2022 Mar 22].
524 Available from: <http://medrxiv.org/lookup/doi/10.1101/2021.10.08.21264785>
- 525 8. Schwab J, Peterson M. Local Epidemic Modeling for Management and Action
526 (LEMMA) [Internet]. 2021. Available from:
527 <https://localepi.github.io/LEMMA/index.html>
- 528 9. California State Government. California’s commitment to health equity [Internet].
529 California for All. 2022 [cited 2022 May 3]. Available from:
530 <https://covid19.ca.gov/equity/>
- 531 10. Bracher J, Ray EL, Gneiting T, Reich NG. Evaluating epidemic forecasts in an
532 interval format. Pitzer VE, editor. *PLOS Comput Biol*. 2021 Feb 12;17(2):e1008618.
- 533 11. Cramer EY, Ray EL, Lopez VK, Bracher J, Brennen A, Castro Rivadeneira AJ, et al.
534 Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality
535 in the United States. *Proc Natl Acad Sci U S A*. 2022 Apr 12;119(15):e2113561119.

- 536 12. California Department of Public Health. COVID-19 Hospital Data [Internet].
537 California Open Data Portal. Available from: <https://data.ca.gov/group/covid-19>
- 538 13. Green A, Hu A, Jahja M, Ventura V, Wasserman L, Tibshirani R, et al. CMU Delphi
539 Covid-19 Forecasts [Internet]. 2021. Available from: [https://github.com/cmu-](https://github.com/cmu-delphi/covid-19-forecast/tree/develop#delphi-forecasting-efforts)
540 [delphi/covid-19-forecast/tree/develop#delphi-forecasting-efforts](https://github.com/cmu-delphi/covid-19-forecast/tree/develop#delphi-forecasting-efforts)
- 541 14. Bracher J. Evaluating probabilistic COVID19 forecasts under partial missingness: A
542 pairwise comparison approach [Internet]. COVID-19 Forecast Hub; 2020 Oct 27.
543 Available from: [https://covid19forecasthub.org/talks/2020-10-27-](https://covid19forecasthub.org/talks/2020-10-27-Bracher_Pairwise_Comparisons.pdf)
544 [Bracher_Pairwise_Comparisons.pdf](https://covid19forecasthub.org/talks/2020-10-27-Bracher_Pairwise_Comparisons.pdf)
- 545 15. Breiman L. Random Forests. *Mach Learn*. 2001 Oct 1;45(1):5–32.
- 546 16. Cutler DR, Edwards Jr. TC, Beard KH, Cutler A, Hess KT, Gibson J, et al. Random
547 Forests for Classification in Ecology. *Ecology*. 2007;88(11):2783–92.
- 548 17. California Department of Public Health. Variants [Internet].
549 <https://covid19.ca.gov/variants/>. Available from: <https://covid19.ca.gov/variants/>
- 550 18. U.S. Department of Agriculture Economic Research Service. County-level Data Sets
551 [Internet]. Data Products. 2021 [cited 2022 Mar 23]. Available from:
552 ers.usda.gov/data-products/county-level-data-sets/
- 553 19. Kuhn M. caret: Classification and Regression Training [Internet]. 2021. Available
554 from: <https://github.com/topepo/caret/>
- 555 20. R Core Team. R: A language and environment for statistical computing [Internet].
556 Vienna, Austria: R Foundation for Statistical Computing; 2021. Available from:
557 <https://www.R-project.org/>
- 558 21. Reich NG, Tibshirani RJ, Ray EL, Rosenfeld R. On the predictability of COVID-19
559 [Internet]. 2021 [cited 2022 Jun 3]. Available from:
560 <https://forecasters.org/blog/2021/09/28/on-the-predictability-of-covid-19/>
- 561 22. Reich NG, McGowan CJ, Yamana TK, Tushar A, Ray EL, Osthus D, et al. Accuracy
562 of real-time multi-model ensemble forecasts for seasonal influenza in the U.S. Pitzer
563 VE, editor. *PLOS Comput Biol*. 2019 Nov 22;15(11):e1007486.
- 564 23. Johansson MA, Apfeldorf KM, Dobson S, Devita J, Buczak AL, Baugher B, et al. An
565 open challenge to advance probabilistic forecasting for dengue epidemics. *Proc Natl*
566 *Acad Sci*. 2019 Nov 26;116(48):24268–74.
- 567 24. Viboud C, Sun K, Gaffey R, Ajelli M, Fumanelli L, Merler S, et al. The RAPIDD
568 ebola forecasting challenge: Synthesis and lessons learnt. *Epidemics*. 2018
569 Mar;22:13–21.

- 570 25. The Influenza Forecasting Working Group, McGowan CJ, Biggerstaff M, Johansson
571 M, Apfeldorf KM, Ben-Nun M, et al. Collaborative efforts to forecast seasonal
572 influenza in the United States, 2015–2016. *Sci Rep.* 2019 Dec;9(1):683.
- 573 26. Ray EL, Reich NG. Prediction of infectious disease epidemics via weighted density
574 ensembles. Viboud C, editor. *PLOS Comput Biol.* 2018 Feb 20;14(2):e1005910.
- 575 27. Rosenfeld R, Tibshirani RJ. Epidemic tracking and forecasting: Lessons learned from
576 a tumultuous year. *Proc Natl Acad Sci U S A.* 2021 Dec 21;118(51):e2111456118.
- 577

Table 1. Constituent models providing county-level hospitalization census predictions that are archived on CalCAT and included in the analysis.

Model	Forecast update frequency	Forecast horizon	Methods/Approach	Documentation
Columbia	Weekly	Up to 6 weeks	County level metapopulation model	(5)
UCSF, COVID NearTerm	Daily	2-4 weeks	Bootstrap-based method based on an autoregressive model	(6)
UCB LEMMA	Daily	Up to 4 weeks	SEIR compartmental model with parameters fit using case series data of COVID-19 hospital and ICU census, hospital admissions, deaths, cases and seroprevalence	(7)
CDPH Simple Growth	Daily	Up to 4 weeks	Assumes new cases grow exponentially according to the rate given by the latest ensemble R-effective. Assumes a fixed severity and average length of stay to generate hospitalizations	(4)
CalCAT Ensemble	Daily	Up to 4 weeks	The ensemble forecast takes the median of all the forecasts available on a given date and fits a smoothed spline to the trend.	(4)
CA Baseline	Daily	Up to 4 weeks	Retroactive 7-day rolling average mean of past hospitalization values	Methods

