

Supplemental Materials

When are predictions useful? A new method for evaluating epidemic forecasts

Maximilian Marshall^{1*}, Felix Parker¹ and Lauren M. Gardner¹

¹Dept. of Civil and Systems Engineering, Johns Hopkins University,
Baltimore, MD, USA.

*Corresponding author(s). E-mail(s): mmarsh29@jhu.edu;

1 Formulation

1.1 Motivating Examples - Spatial and Temporal Variability

From many perspectives, making and disseminating state-level forecasts is a reasonable strategy. States are the intuitive building blocks of the country, carrying their own governments and public health systems. Accurate state-level forecasts therefore have the potential for direct and meaningful application. However, states have enormously variable characteristics, which makes generalizing forecast performance problematic. Population difference in particular is a key factor. For example, California has the highest population of any state in the US (~40 million), and Wyoming the lowest (~0.6 million). For the second week of January 2022, California reported over 850,000 incident cases. During the same week, Wyoming reported just over 6,600 new cases [1]. Note that California reported over 1.4 times more new cases that week than the entire population of Wyoming. However, in terms of incidence percentages, California and Wyoming were actually much closer at that time, with approximately 2% and 1% of the population testing positive, respectively. Intuitively, this is an easy dynamic to recognize when examining individual states separately. Raw epidemic numbers carry different meanings depending on underlying demographic factors (i.e., population size). However, this is problematic for aggregate and comparative analysis of forecast performance. This becomes clear if we apply a standard metric like mean absolute error (MAE) to this scenario with California and Wyoming. (For simplicity we refer to point predictions instead of probabilistic forecasts in the motivating examples in this section, along with corresponding metrics such as the absolute and percent error.

30 However, as indicated above, probabilistic evaluation is susceptible to the same issues
31 [2].) For the week under consideration, predictions from the Forecast Hub’s baseline
32 model yielded a MAE of 27,130 across all US states [3]. For California, a prediction
33 that overshot the truth by this margin would incur a percent error of only about 3%,
34 whereas for Wyoming, such a prediction would miss by over 400%. Unfortunately, spa-
35 tial inconsistency is not the only obstacle. Accounting for temporal context is equally
36 vital and presents its own difficulties.

37 When examining forecast performance for a single region over time, metrics must
38 be interpreted as a function of time-variant data. This necessity is demonstrated
39 trivially by comparing pandemic surges to times of relatively low epidemic activity.
40 The same value of a non-normalized metric like the absolute error carries an entirely
41 different meaning in each of these situations. Consider the Forecast Hub’s baseline
42 model predictions for cases in Maryland. In mid-December 2020, this model missed
43 its three-weeks-ahead target by about 2,000 cases. In mid-May 2021, the same model
44 also missed by about 2,000 cases [3]. Without knowing the context of each prediction,
45 (namely that the first was made during a massive surge and the second was made
46 during a significant lull), one might be forgiven for assuming that the model performed
47 similarly in both scenarios. However, the December forecast only just missed the mark,
48 undershooting by 12% of the true value. Conversely, the May forecast missed by 213%.
49 Note that in this case, percent error has interpretable utility because it normalizes
50 by the true value, a time-varying data source that directly represents the prevailing
51 condition of the pandemic. Unfortunately, percent error is not an ideal solution as it
52 becomes unstable when true values approach zero [2]. This is especially problematic
53 when analyzing death forecasts (for all of 2020 through 2022, almost 15% of US states
54 had less than ten weekly deaths, and over 8% had below five weekly deaths). In this
55 situation, percent error is in fact too sensitive to the exact circumstances. It indicates
56 a relatively large deviation from the truth which, while technically correct, misses the
57 reality of how forecasts are interpreted. Given the larger context of the pandemic,
58 it is unreasonable to characterize a four-death forecast compared to a target value
59 of one (300% error) as a worse prediction than a 400-death forecast compared to an
60 800-death reality (50% error). Like the spatial case, the numerical value of an error
61 metric, absent any temporal contextualization, cannot be relied on to consistently or
62 intuitively reflect forecast performance.

63 **1.2 Derivation of the CIS as a function of the CRE**

64 For simplicity and consistency with extant metrics, we introduce the Contextual
 65 Interval Score (CIS) in the main body of the paper as a scaled, constrained version of
 66 the Interval Score (IS). However, we developed the CIS as a direct, interval-forecast
 67 extension of our point-forecast CRE function. In this section, we demonstrate that the
 68 formulation of the CIS as a function of the IS is equivalent to a different formulation
 69 that directly incorporates the CRE. Then, we explain each of the components of the
 70 equivalent form of the score to help intuit the motivation for the creation of the score.
 71

72 We begin with the form of the CIS introduced in the main body of the paper:

$$CIS_{\alpha}(F, y, \delta) = \min \left\{ \frac{\alpha}{2\delta} IS_{\alpha}(F, y), 1 \right\} \quad (1)$$

73 Taking the right hand side of this equation, we substitute in the expanded form
 74 of the Interval Score (IS):

$$\min \left\{ \frac{\alpha}{2\delta} \left[(u - l) + \frac{2}{\alpha} (l - y) \mathbb{1}\{y < l\} + \frac{2}{\alpha} (y - u) \mathbb{1}\{y > u\} \right], 1 \right\} \quad (2)$$

75 Simplifying:

$$\min \left\{ \frac{\alpha}{2\delta} (u - l) + \frac{l - y}{\delta} \mathbb{1}\{y < l\} + \frac{y - u}{\delta} \mathbb{1}\{y > u\}, 1 \right\} \quad (3)$$

76 Examining the $\frac{l - y}{\delta} \mathbb{1}\{y < l\}$ term, we observe that if this term reaches or exceeds
 77 1, the minimizer operating over the entire equation will restrict the overall output to 1.
 78 Thus applying a “local” minimizer, constraining this term to a maximum of 1, will not
 79 change the overall value of the score. The same logic applies to the the $\frac{y - u}{\delta} \mathbb{1}\{y > u\}$
 80 term. Including these internal minimizers yields the following form of the CIS:

$$\min \left\{ \frac{\alpha}{2\delta} (u - l) + \min \left\{ \frac{l - y}{\delta}, 1 \right\} \mathbb{1}\{y < l\} + \min \left\{ \frac{y - u}{\delta}, 1 \right\} \mathbb{1}\{y > u\}, 1 \right\} \quad (4)$$

81 We can further exploit the indicator functions to include absolute values in the
 82 two minimized terms:

$$\min \left\{ \frac{\alpha}{2\delta} (u - l) + \min \left\{ \frac{|l - y|}{\delta}, 1 \right\} \mathbb{1}\{y < l\} + \min \left\{ \frac{|u - y|}{\delta}, 1 \right\} \mathbb{1}\{y > u\}, 1 \right\} \quad (5)$$

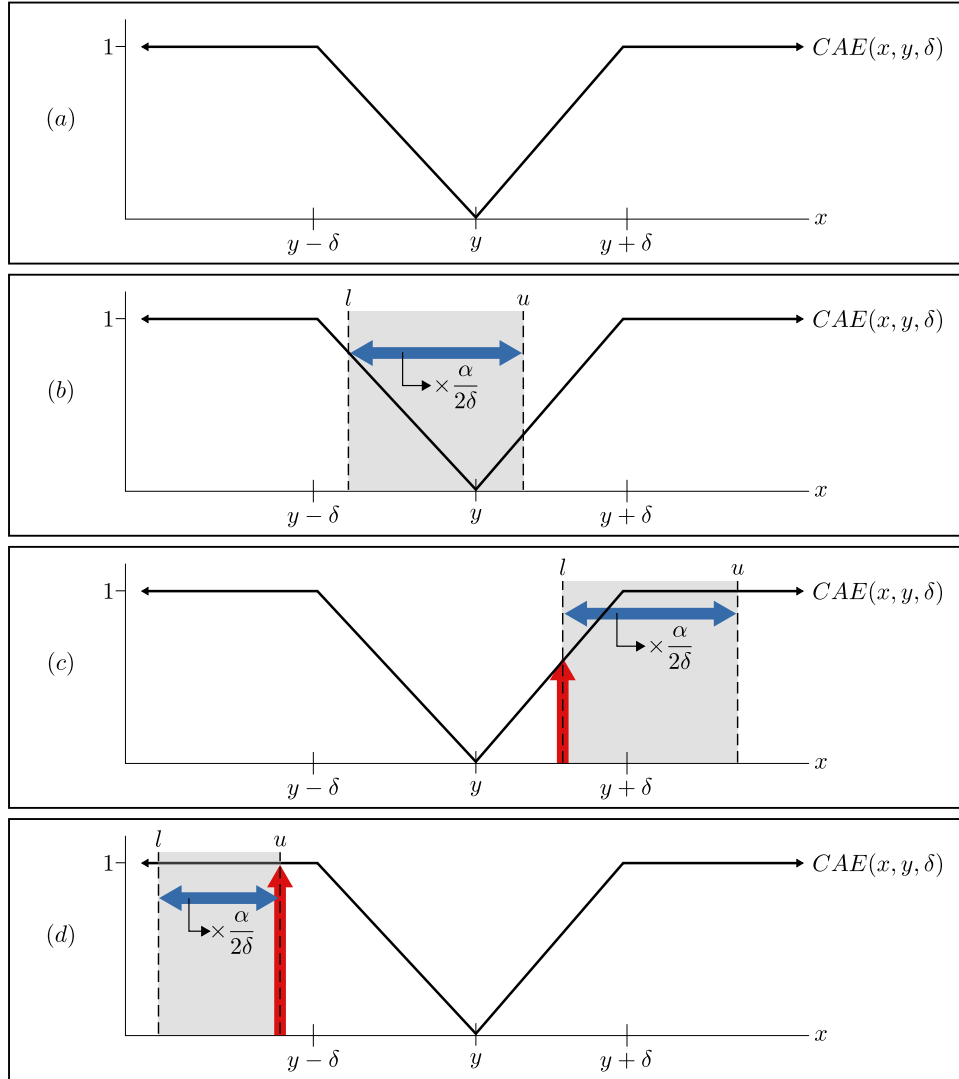
83 This equation now directly includes the formulation of the CRE. Thus, we can
 84 substitute the CRE in to show the complete alternate formulation of the CIS:

$$CIS_{\alpha}(F, y, \delta) = \min \left\{ \begin{array}{c} \frac{\alpha}{2\delta} (u - l) + CRE(l, y, \delta) \mathbb{1}\{y < l\} + CRE(u, y, \delta) \mathbb{1}\{y > u\} \\ 1 \end{array} \right\} \quad (6)$$

85 Each term in the CIS is analogous to a term in the IS. We begin with the “width”
 86 term: $\frac{\alpha}{2\delta} (u - l)$. Because $y - \delta$ to $y + \delta$ represents the upper and lower limits of forecast
 87 utility, a prediction interval that spans this entire distance should incur an unweighted
 88 penalty of 1. In other words, if a point forecast at or past the “plateau” of the CRE
 89 curve incurs a penalty of 1, an unweighted interval forecast that spans this region
 90 should get the same score. However, the α -weight is included to distinguish between
 91 different prediction intervals. Consider two intervals that have identical bounds but
 92 different α values: 0.05 (95% prediction interval) and 0.9 (10% prediction interval).
 93 In this case, the 95% interval should be treated less harshly than the 10% interval,
 94 because we expect higher-confidence forecasts to span larger ranges. Next, we examine
 95 the “miss” term of the CIS: $(CRE(l, y, \delta)) \mathbb{1}(y < l) + (CRE(u, y, \delta)) \mathbb{1}(y > u)$. It is
 96 essentially performing the same function as the “miss” term of the IS, but instead of
 97 expressing the magnitude of the miss in terms of distance, the CIS term is expressed
 98 in terms of utility. This component of the score can be seen in panels (c) and (d) of
 99 Additional file 1: Fig. S1 as the vertical arrows. In sum, the CIS is a single-interval
 100 analogue of the point-forecast CRE. Regardless of interval width, if a probabilistic
 101 forecast is entirely outside the useful region, a value of 1 is returned (panel (d) in Fig.
 102 S1). Like the IS, the CIS naturally collapses to only its “miss” term when applied to
 103 a point forecast.

104 1.3 Visualization of the CRE and CIS

105 The following figure, S1, provides a visualization of the CRE and the three different
 106 ways the CIS can arise, depending on the relative positions of the prediction inter-
 107 val bounds and the true value. Panel (a) shows only the Contextual Relative Error
 108 (CRE) point score (Equation 3 in the main text), with the others displaying differ-
 109 ent realizations of the Contextual Interval Score (CIS, Equation 4 in the main text).
 110 Blue arrows represent the width penalty term (note that they are scaled by $\frac{\alpha}{2\delta}$). Red
 111 arrows indicate the miss term of the CIS. Observe that because the miss term is not
 112 scaled, any forecast that entirely misses the $y - \delta$ to $y + \delta$ region, regardless of width,
 113 will incur the maximum penalty of 1. For clarity, each of the panels refers to a single-
 114 interval evaluation. The full Weighed Contextual Interval Score (WCIS) is composed
 115 of an average across multiple α intervals.



$$CIS_{\alpha}(F, y, \delta) = \min \left\{ \frac{\alpha}{2\delta} \left[(u - l) + CAE(l, y, \delta) \mathbb{1}\{y < l\} + CAE(u, y, \delta) \mathbb{1}\{y > u\} \right], 1 \right\}$$

Fig. S1 Demonstration of the CRE (Panel a) and the three different calculation modes that the CIS can take (Panels b,c,d).

116 **1.4 Empirical Impropriety Demonstration**

117 In this test case, we select an arbitrary distribution to represent the output of a
 118 forecasting model. This distribution functions as the source of the modeler’s “good-
 119 faith” predictions, i.e. a proper score will incentivize submission of forecasts that
 120 are legitimately representative of this distribution. Next, we select an arbitrary δ -
 121 parameterization and interval represented by α . “Good-faith” predictions are drawn
 122 as the $\frac{\alpha}{2}$ and $1 - \frac{\alpha}{2}$ quantiles of the distribution. Next, we iterate over the domain of
 123 the distribution and take the expected score for each feasible interval pair. If any pair
 124 has a score that is lower in expectation than the “good-faith” interval, then the score
 125 is not proper.

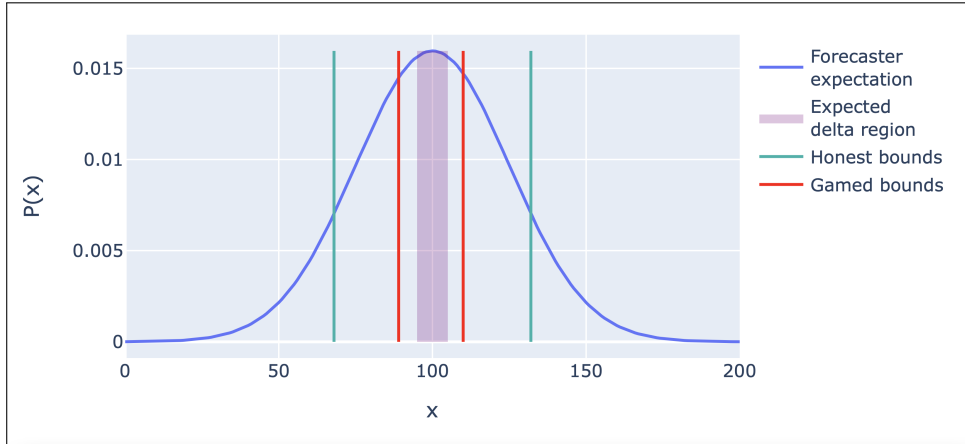


Fig. S2 Empirical demonstration of the impropriety of the Contextual Interval Score (CIS). Note the difference between the “honest” and “gamed” interval bounds, indicating an incentive to deviate from submitting good-faith statistical realizations in a effort to minimize the expected penalty. The parameterization here is a normal distribution defined with $\mu = 100$ and $\sigma = 25$. δ and α were selected to be 5 and 0.2, respectively.

126 As is clearly demonstrated by figures S2 and S3, the CIS and therefore the multi-
 127 interval WCIS is not a statistically proper interval score. However, we propose that a
 128 score with the desired features of the WCIS is inherently improper. The foundation
 129 of the WCIS is the notion of a specific and *constrained* region around the target value
 130 wherein predictions are applicable, represented by the V-shaped CRE function. This
 131 means that from a gaming/error minimization perspective, the WCIS could encourage
 132 probabilistic forecasts that are affected by the size of the δ -region [4]. Similar to prior
 133 forecasting efforts when improper metrics were used, propriety is sacrificed in exchange
 134 for other, desirable properties of the score [5–7]. Additionally, ongoing work by Bosse
 135 et al. indicates that applying monotonic transformations like the natural logarithm to
 136 target data can help to alleviate the domination of higher-activity forecasting scenarios
 137 for model comparison and aggregation while retaining propriety [8].

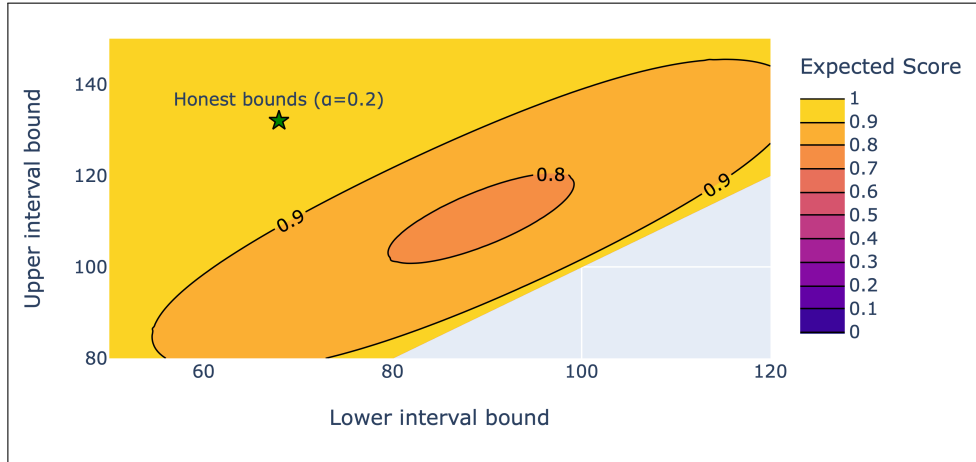


Fig. S3 Contour plot showing the expected Contextual Interval Score (CIS) value for a set of different prediction interval bounds. The distribution and score parameterization used here is the same as in Figure S2 above. Note that the “honest” bounds do not exist at the minimum expected score, and selecting bounds that do minimize the expectation results in the “gamed” bounds (approximately 89 and 110) that are seen in Figure S2.

138 2 Facility-Level Analysis

139 2.1 Facility-Level Model Formulation

140 **Data:** We obtained facility-level data on COVID-19 hospitalizations from the
 141 COVID-19 Reported Patient Impact and Hospital Capacity by Facility dataset,
 142 collected at various times by the CDC, HHS, and CDC again. This dataset includes
 143 metrics related to COVID-19 hospitalizations, hospital occupancy, and capacity [9].
 144 For this analysis, we focused only on the time series of COVID-19 bed occupancy.
 145 The raw data was provided at a weekly resolution for each hospital. To enable more
 146 granular modeling, we performed temporal disaggregation to obtain daily resolution
 147 data. We assumed that the weekly trends at each hospital followed the same pattern
 148 as the aggregated state-level trends, which were available at a daily resolution. For
 149 each week, we normalized the daily state-level values to sum to 1, then multiplied
 150 the normalized values by each hospital’s weekly totals to impute daily hospital-level
 151 values. Any remaining missing values were imputed using local regression smoothing
 152 [10]. We selected 42 hospitals in Maryland for this analysis. Hospitals were included
 153 if they were classified as short-term acute care hospitals and if they had at least 10
 154 COVID-19 patients at some point between July 2021 and July 2022. We chose to focus
 155 on a single state because modeling and analyzing all U.S. hospitals was not practical.
 156

157 **Model:** To forecast future COVID-19 hospitalizations, we used the Time Series
 158 Dense Encoder (TiDE) model, a deep neural network architecture that has achieved
 159 state-of-the-art performance on general time-series forecasting tasks [11]. TiDE uses
 160 a simple but flexible encoder-decoder structure that can incorporate covariates and
 161 accommodate various prediction horizons, output distributions, and loss functions.

162 We configured the TiDE model with 4 encoder layers, 4 decoder layers, a decoder
 163 output dimension of 32, hidden size of 128, past temporal width of 4, future temporal
 164 width of 4, dropout probability of 0.1, and layer normalization. The model used
 165 the previous 90 days as context to predict hospitalizations for the next 21 days.
 166 Rather than making sequential autoregressive predictions, the model predicted all 21
 167 days at once. To obtain probabilistic forecasts, we used quantile regression, with the
 168 model directly outputting predictions for the 0.01, 0.025, 0.05, 0.1, 0.15, 0.2, 0.25,
 169 0.3, 0.35, 0.4, 0.45, 0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95, 0.975, and 0.99
 170 quantiles. The input features were the target variable (total COVID-19 census) and
 171 time covariates including year, day of year, day of week, and days since July 1, 2021.
 172 Predictions were generated for each Monday in the time span included, simulating a
 173 weekly system like the Forecast Hub’s [3].

174

175 **Training and Calibration:** We trained a separate model for each hospital and
 176 prediction date using an expanding window of training data starting from August 1,
 177 2020. Models were trained for 100 epochs to minimize the pinball loss. The model
 178 predictions were post-processed in two steps to improve calibration. First, we applied
 179 a non-negativity constraint, thresholding all predictions to be at least 0 since neg-
 180 ative patient counts are impossible. Second, we applied the conformalized quantile
 181 regression (CQR) method [12]. CQR adjusts the predicted quantiles based on the
 182 model’s historical quantile errors to achieve better coverage. We did not hold out a
 183 separate calibration dataset, instead using the training data for the CQR calibration.

184

185 **Implementation:** We implemented the models in Python using the darts time-series
 186 library, while the data processing was done in Julia [13]. Model training took approxi-
 187 mately 10 seconds per hospital and prediction window using an NVIDIA 4070Ti GPU.
 188 β

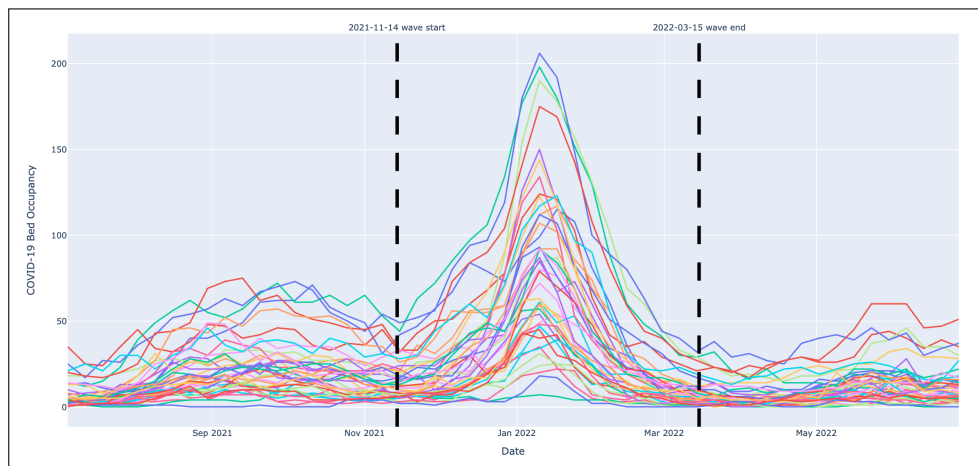


Fig. S4 Delineation of the segments of the facility-level analysis that are in and out of the Omicron wave for the purposes of our analysis. Each line represents one of the 42 Maryland facilities predicted for.

Table S1 Hospitals Included in Facility-Level Model

Hospital ID	Facility Name
210001	Meritus Medical Center
210002	University of Maryland Medical Center
210003	University of Maryland Prince George's Hospital Center
210004	Holy Cross Hospital
210005	Frederick Health Hospital
210006	University of Maryland Harford Memorial Hospital
210008	Mercy Medical Center
210009	The Johns Hopkins Hospital
210011	Saint Agnes Hospital
210012	Sinai Hospital of Baltimore
210015	MedStar Franklin Square Medical Center
210016	Adventist Healthcare White Oak Medical Center
210017	Garrett County Memorial Hospital
210018	MedStar Montgomery Medical Center
210019	TidalHealth Peninsula Regional, Inc.
210022	Suburban Hospital
210023	Anne Arundel Medical Center
210024	MedStar Union Memorial Hospital
210027	U.P.M.C. Western Maryland
210028	MedStar Saint Mary's Hospital
210029	Johns Hopkins Bayview Medical Center
210032	Union Hospital of Cecil County
210033	Carroll Hospital Center
210034	MedStar Harbor Hospital
210035	University of Maryland Charles Regional Medical Center
210037	University of Maryland Shore Medical Center at Easton
210038	University of Maryland Medical Center Midtown Campus
210039	CalvertHealth Medical Center
210040	Northwest Hospital Center
210043	University of Maryland Baltimore Washington Medical Center
210044	Greater Baltimore Medical Center
210048	Howard County General Hospital
210049	University of Maryland Upper Chesapeake Medical Center
210051	Doctors Community Hospital
210056	MedStar Good Samaritan Hospital
210057	Adventist Healthcare Shady Grove Medical Center
210060	Adventist Healthcare Fort Washington Medical Center
210061	Atlantic General Hospital
210062	MedStar Southern Maryland Hospital Center
210063	University of Maryland St. Joseph Medical Center
210064	Levindale Hebrew Geriatric Center And Hospital
210065	Holy Cross Germantown Hospital

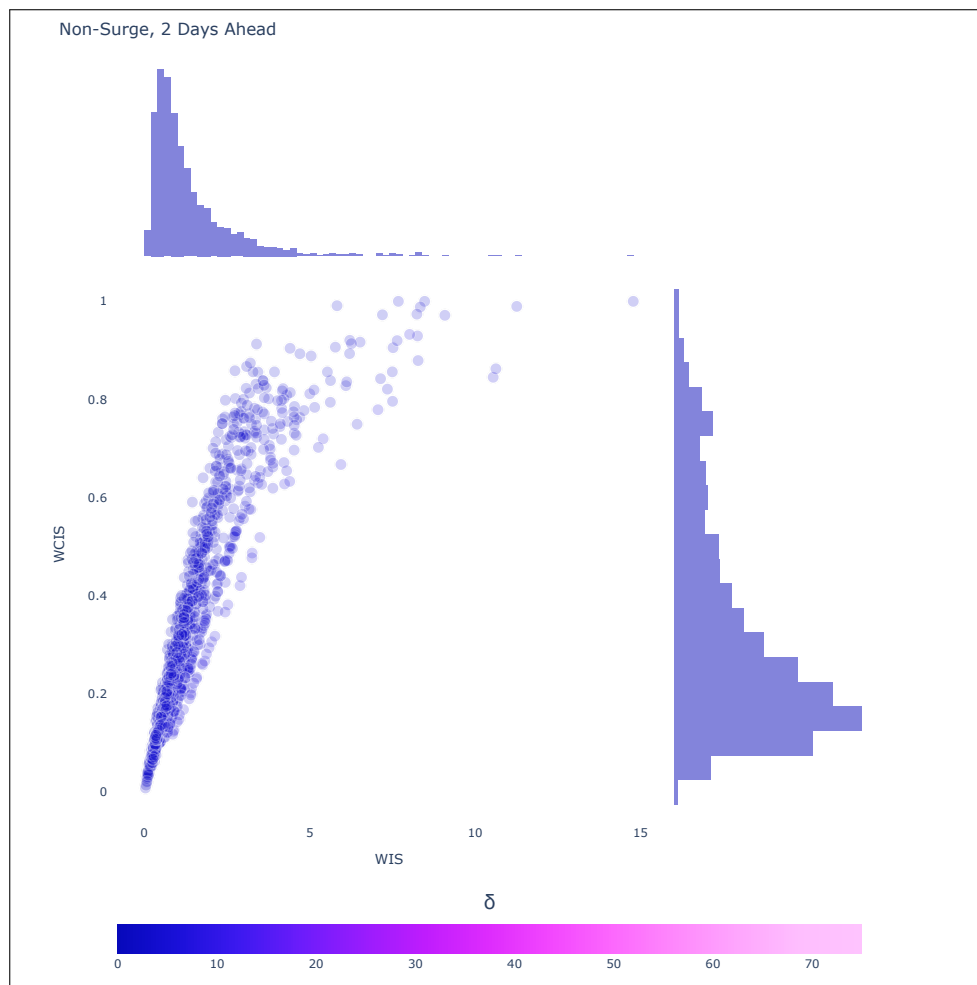


Fig. S5 WIS vs WCIS values for all 42 facilities, for 2-day-ahead forecasts, for all prediction dates outside of the Omicron surge.

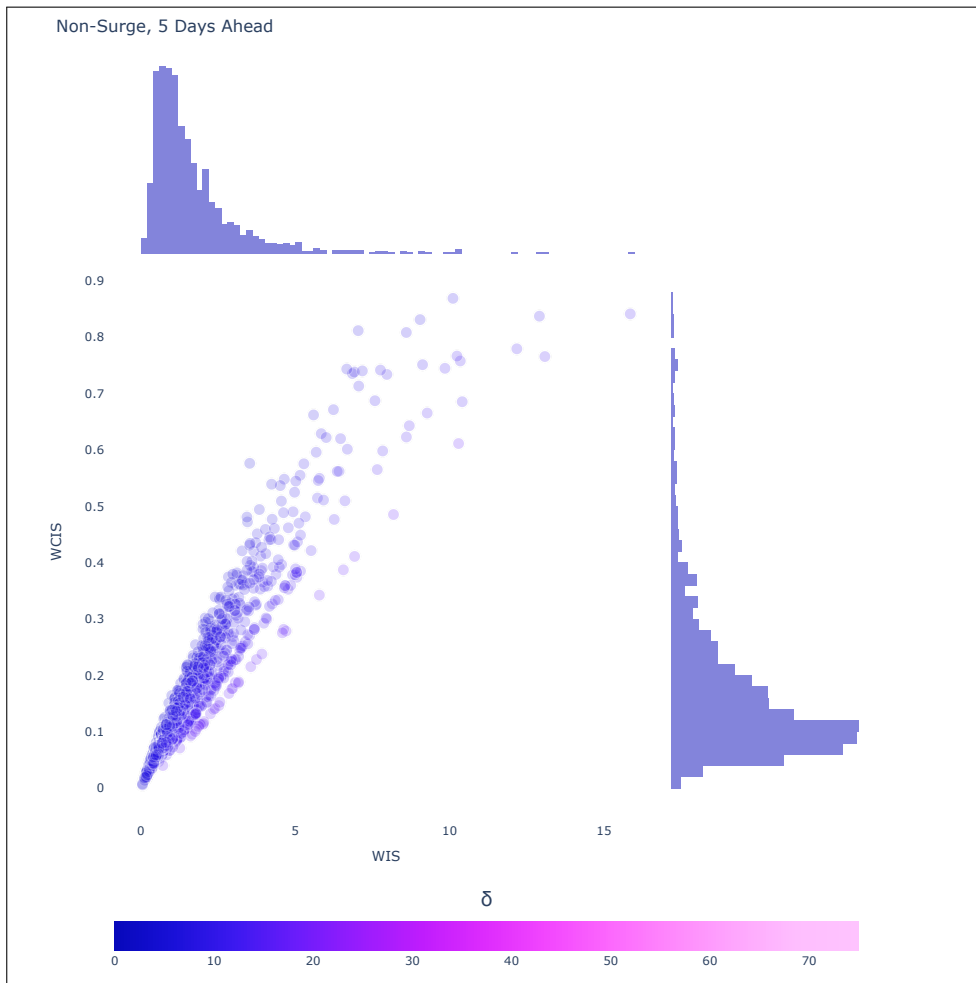


Fig. S6 WIS vs WCIS values for all 42 facilities, for 5-day-ahead forecasts, for all prediction dates outside of the Omicron surge.

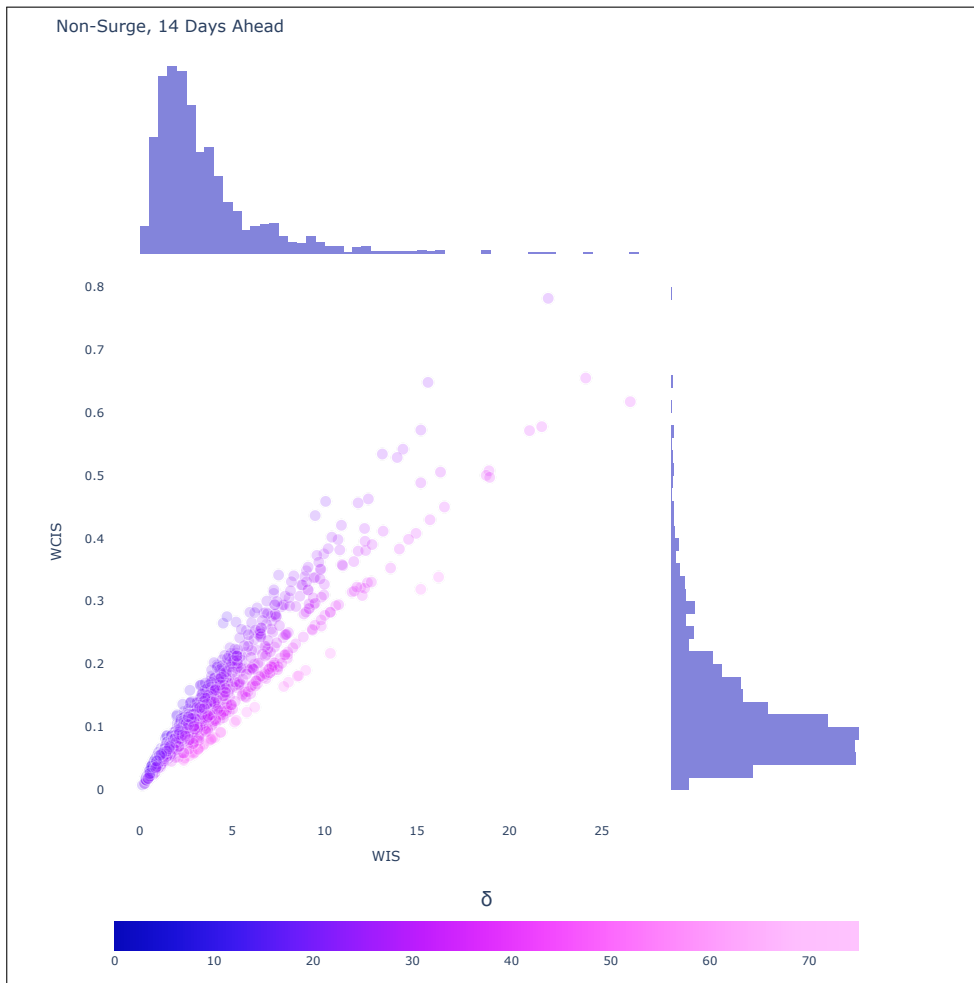


Fig. S7 WIS vs WCIS values for all 42 facilities, for 14-day-ahead forecasts, for all prediction dates outside of the Omicron surge.

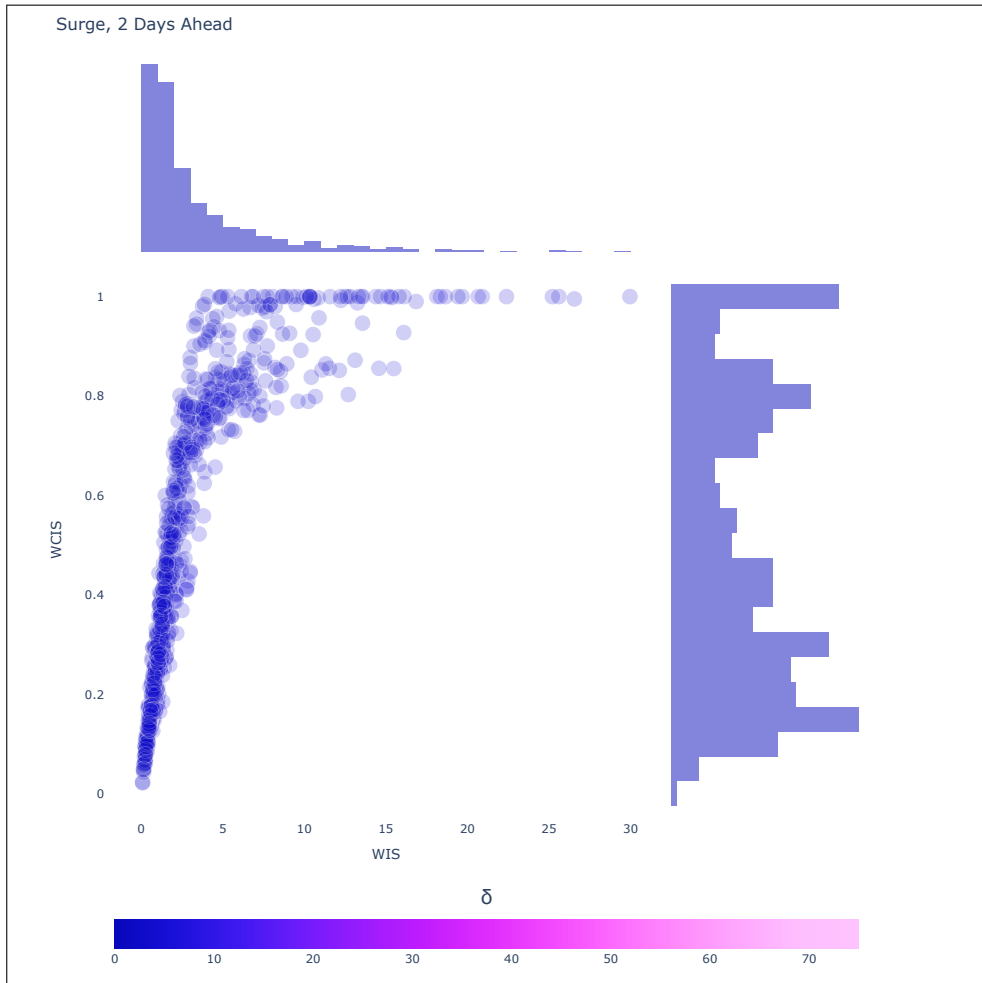


Fig. S8 WIS vs WCIS values for all 42 facilities, for 2-day-ahead forecasts, for all prediction dates within the Omicron surge.

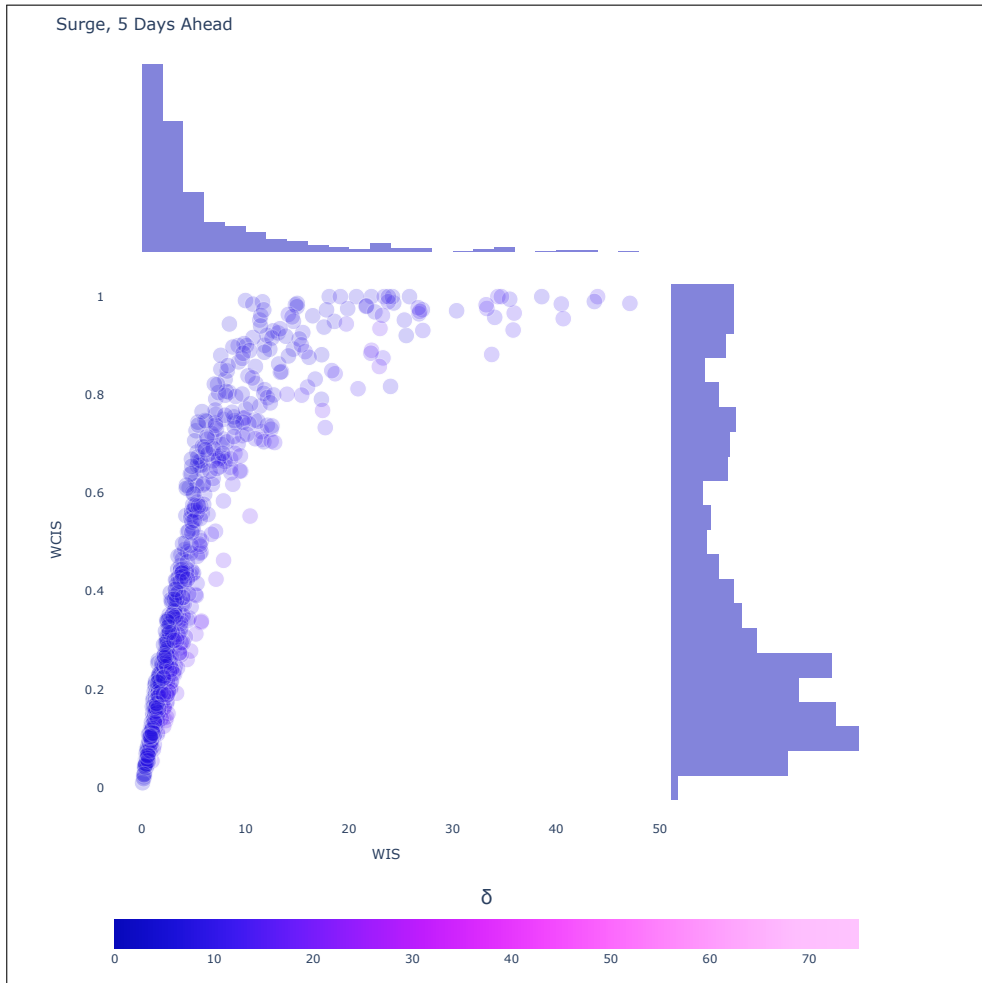


Fig. S9 WIS vs WCIS values for all 42 facilities, for 5-day-ahead forecasts, for all prediction dates within the Omicron surge.

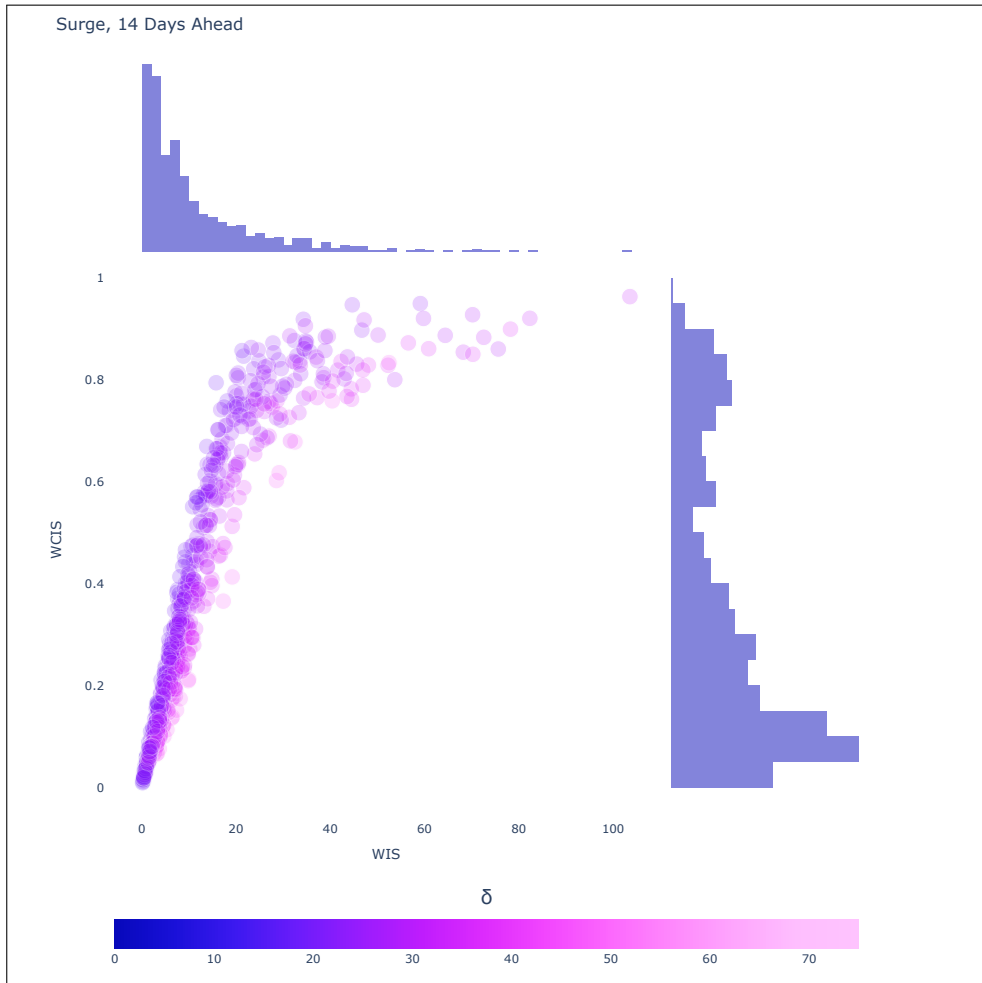


Fig. S10 WIS vs WCIS values for all 42 facilities, for 14-day-ahead forecasts, for all prediction dates within the Omicron surge.

190 **3 COVIDhub Ensemble Hospitalization Forecasts**
191 **(Second Test Case)**

192 Included here are heatmaps of the WCIS vs the WIS for hospitalization fore-
193 casts for each prediction horizon (one, two, three, and four weeks ahead) from
194 the Forecast Hub’s ensemble model. The δ used for the hospitalization analysis is
195 detailed in full in section 3.2 of the main text of the paper. We note here that
196 the column used to generate the delta values is “inpatient_beds” in the COVID-19
197 Reported Patient Impact and Hospital Capacity by Facility dataset (archive link:
198 <https://healthdata.gov/d/j4ip-wfsv>).

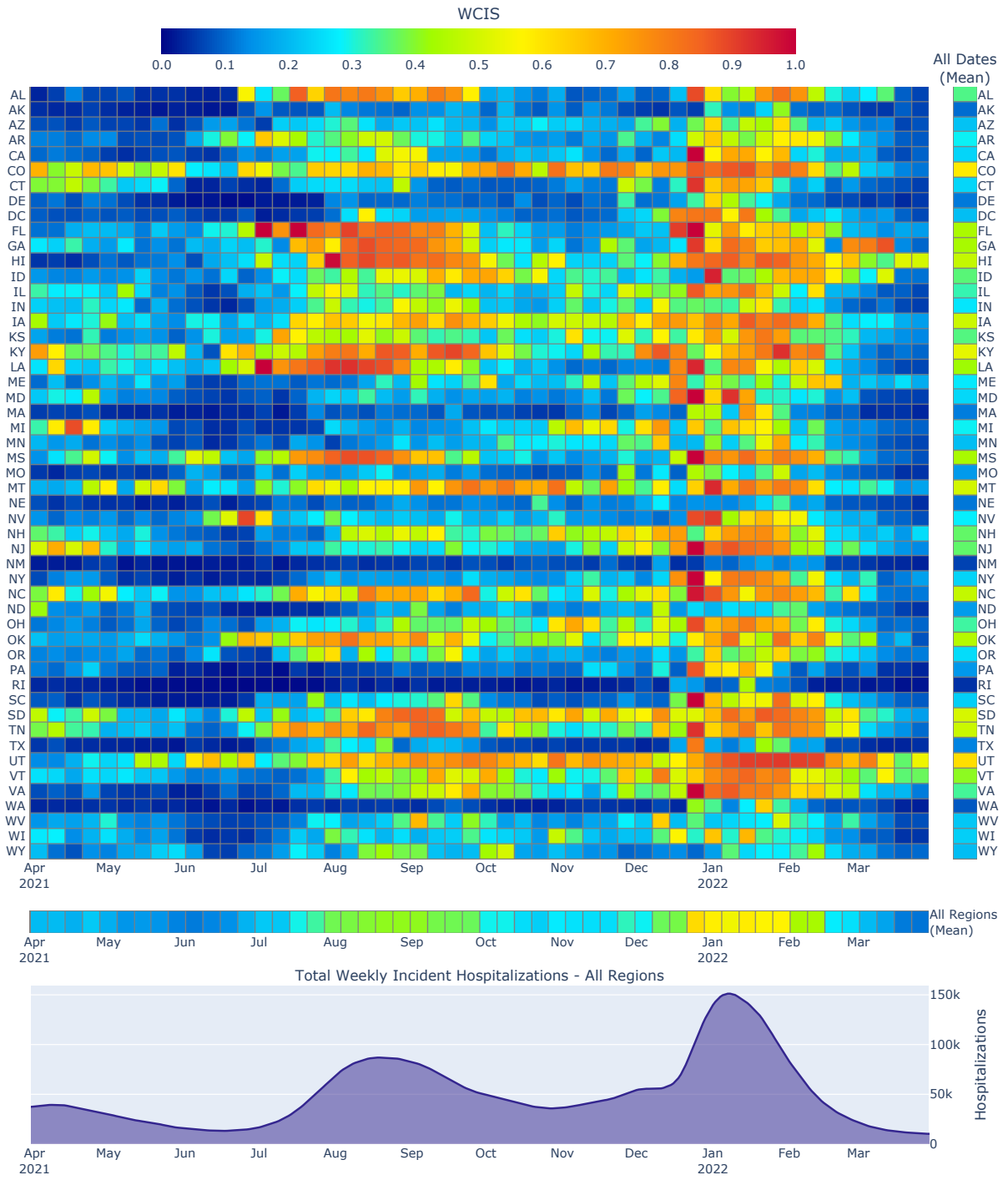


Fig. S11 Heatmap of the WCIS for 1 week ahead hospitalization forecasts, performed by the Forecast Hub's ensemble model.

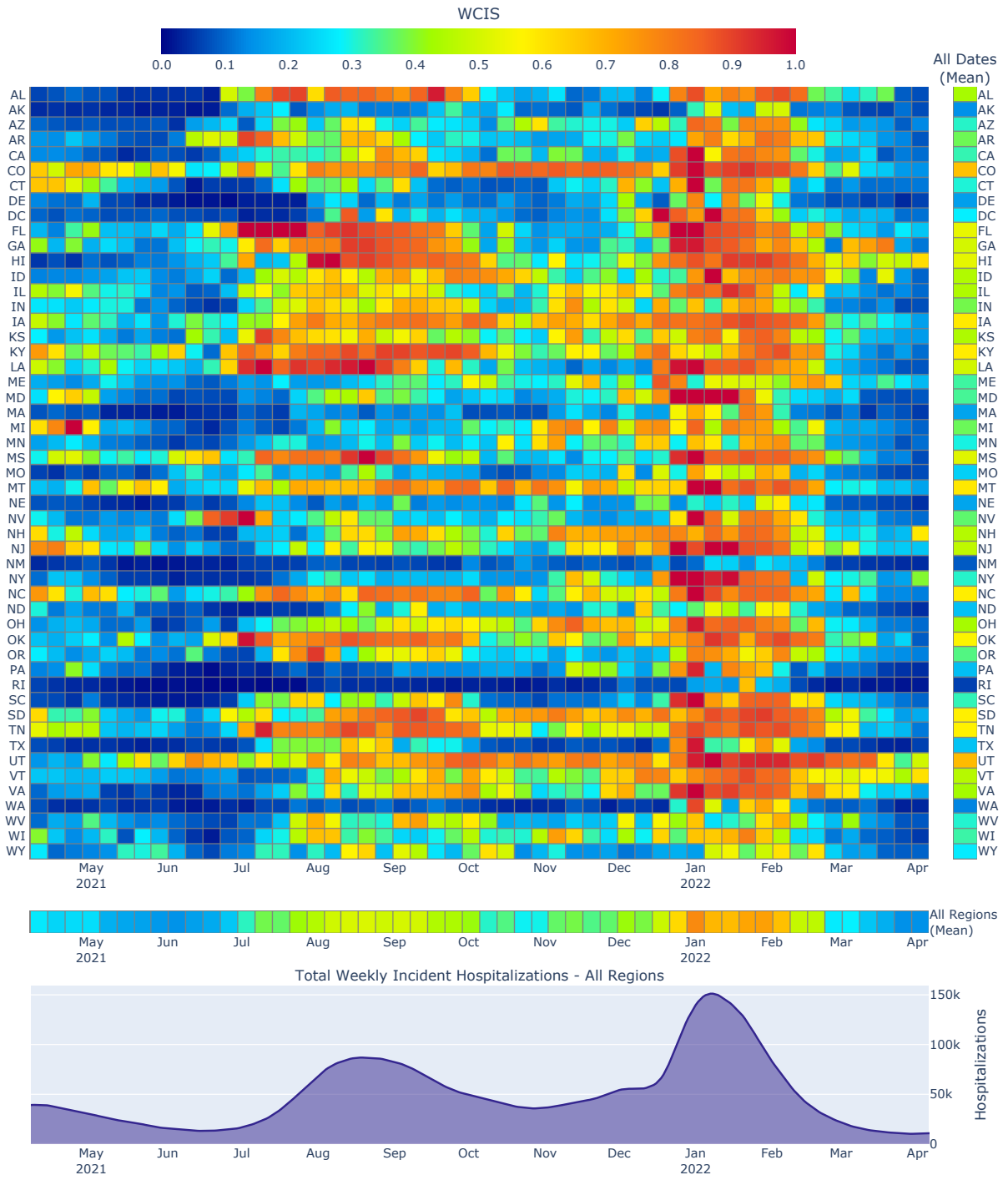


Fig. S12 Heatmap of the WCIS for 2 week ahead hospitalization forecasts, performed by the Forecast Hub's ensemble model.

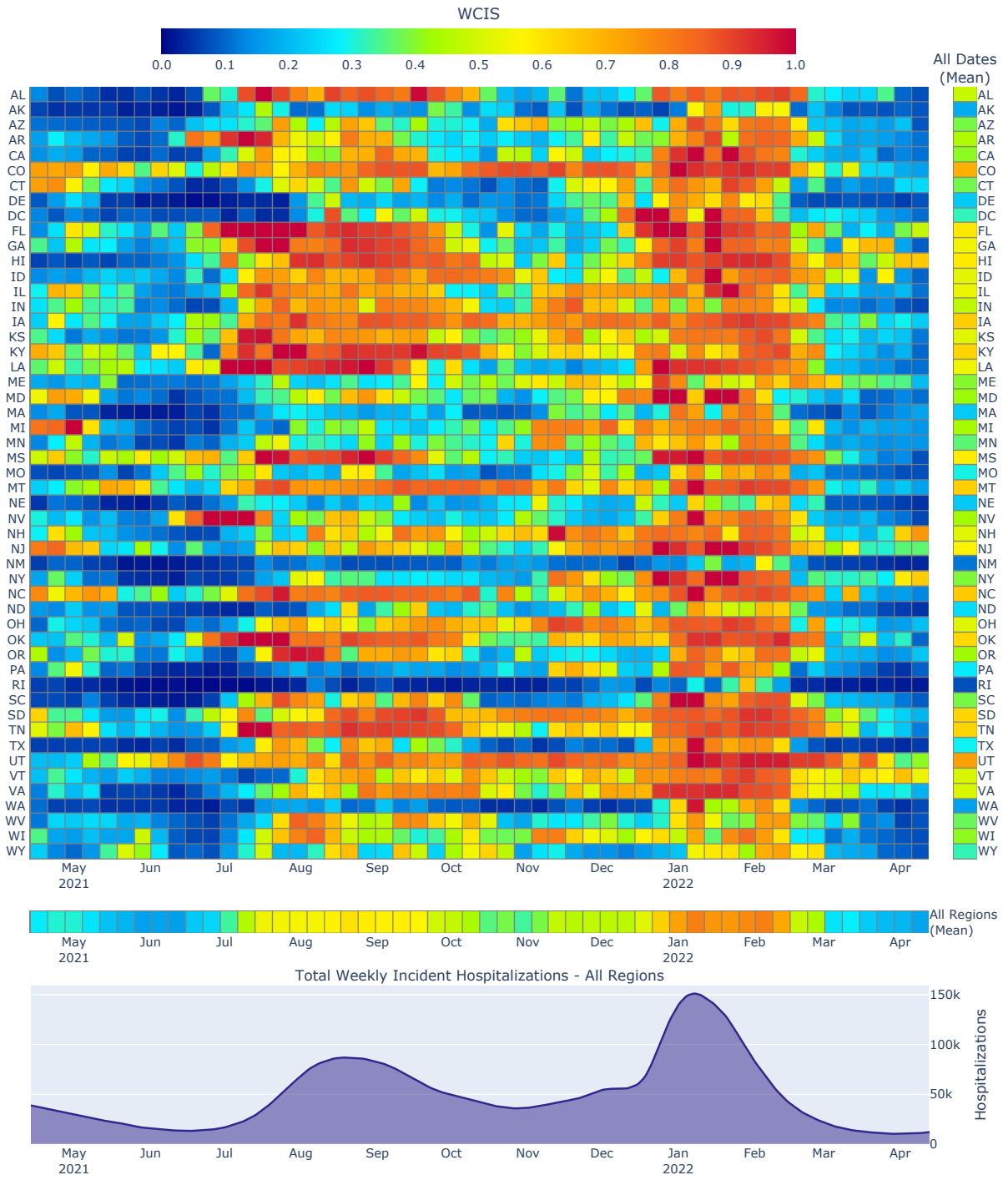


Fig. S13 Heatmap of the WCIS for 3 week ahead hospitalization forecasts, performed by the Forecast Hub's ensemble model.

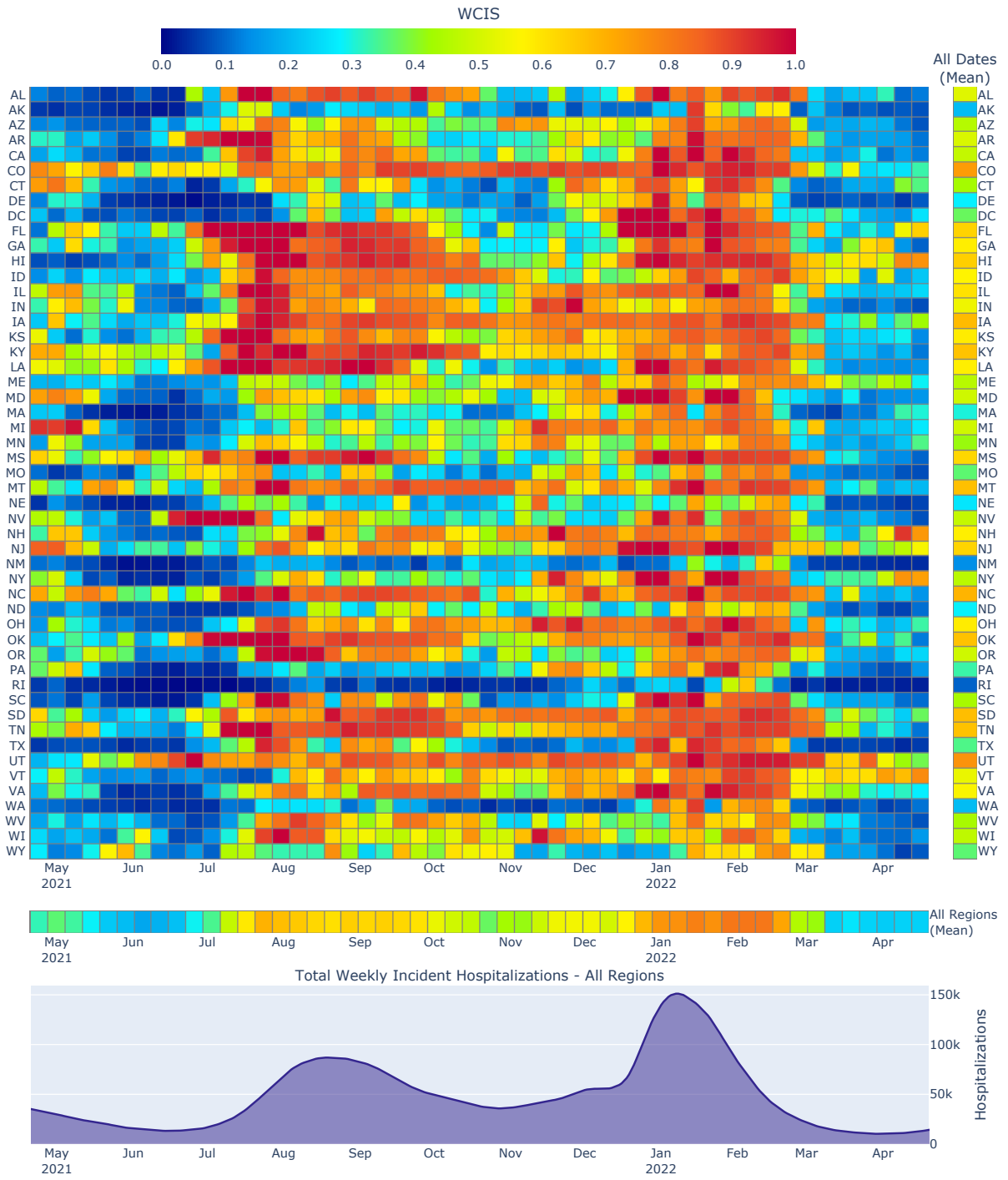


Fig. S14 Heatmap of the WCIS for 4 week ahead hospitalization forecasts, performed by the Forecast Hub's ensemble model.

199 **References**

- 200 [1] Dong E, Du H, Gardner L. An interactive web-based dashboard to track
201 COVID-19 in real time. *The Lancet Infectious Diseases*. 2020 May;20(5):533–534.
202 Publisher: Elsevier. [https://doi.org/10.1016/S1473-3099\(20\)30120-1](https://doi.org/10.1016/S1473-3099(20)30120-1).
- 203 [2] Bracher J, Ray EL, Gneiting T, Reich NG. Evaluating epidemic forecasts in an
204 interval format. *PLOS Computational Biology*. 2021 Feb;17(2):e1008618. Pub-
205 lisher: Public Library of Science. <https://doi.org/10.1371/journal.pcbi.1008618>.
- 206 [3] Cramer EY, Huang Y, Wang Y, Ray EL, Cornell M, Bracher J, et al.
207 The United States COVID-19 Forecast Hub dataset. *Scientific Data*. 2022
208 Aug;9(1):462. Number: 1 Publisher: Nature Publishing Group. [https://doi.org/](https://doi.org/10.1038/s41597-022-01517-w)
209 [10.1038/s41597-022-01517-w](https://doi.org/10.1038/s41597-022-01517-w).
- 210 [4] Gneiting T, Ranjan R. Comparing Density Forecasts Using Threshold- and
211 Quantile-Weighted Scoring Rules. *Journal of Business & Economic Statistics*.
212 2011 Jul;29(3):411–422. Publisher: Taylor & Francis. [https://doi.org/10.1198/](https://doi.org/10.1198/jbes.2010.08110)
213 [jbes.2010.08110](https://doi.org/10.1198/jbes.2010.08110).
- 214 [5] Reich NG, Brooks LC, Fox SJ, Kandula S, McGowan CJ, Moore E, et al. A
215 collaborative multiyear, multimodel assessment of seasonal influenza forecast-
216 ing in the United States. *Proceedings of the National Academy of Sciences*.
217 2019 Feb;116(8):3146–3154. Publisher: Proceedings of the National Academy of
218 Sciences. <https://doi.org/10.1073/pnas.1812594116>.
- 219 [6] Bracher J. On the multibin logarithmic score used in the FluSight competitions.
220 *Proceedings of the National Academy of Sciences*. 2019 Oct;116(42):20809–20810.
221 Publisher: Proceedings of the National Academy of Sciences. [https://doi.org/10.](https://doi.org/10.1073/pnas.1912147116)
222 [1073/pnas.1912147116](https://doi.org/10.1073/pnas.1912147116).
- 223 [7] Reich NG, Osthus D, Ray EL, Yamana TK, Biggerstaff M, Johansson MA,
224 et al. Reply to Bracher: Scoring probabilistic forecasts to maximize public
225 health interpretability. *Proceedings of the National Academy of Sciences*. 2019
226 Oct;116(42):20811–20812. Publisher: Proceedings of the National Academy of
227 Sciences. <https://doi.org/10.1073/pnas.1912694116>.
- 228 [8] Bosse NI, Abbott S, Cori A, Leeuwen Ev, Bracher J, Funk S.: Transformation
229 of forecasts for evaluating predictive performance in an epidemiological context.
230 medRxiv. ISSN: 2328-4722 Pages: 2023.01.23.23284722. Available from: [https:](https://www.medrxiv.org/content/10.1101/2023.01.23.23284722v1)
231 [//www.medrxiv.org/content/10.1101/2023.01.23.23284722v1](https://www.medrxiv.org/content/10.1101/2023.01.23.23284722v1).
- 232 [9] gov H.: COVID-19 Reported Patient Impact and Hospital Capacity by Facility.
233 United States Department of Health & Human Services. Available from: [https:](https://healthdata.gov/d/j4ip-wfsv)
234 [//healthdata.gov/d/j4ip-wfsv](https://healthdata.gov/d/j4ip-wfsv).

- 235 [10] Cleveland WS, Grosse E. Computational methods for local regression. *Statistics*
236 and Computing. 1991 Sep;1(1):47–62. <https://doi.org/10.1007/BF01890836>.
- 237 [11] Das A, Kong W, Leach A, Mathur S, Sen R, Yu R.: Long-term Forecasting with
238 TiDE: Time-series Dense Encoder. arXiv. ArXiv:2304.08424 [cs, stat]. Available
239 from: <http://arxiv.org/abs/2304.08424>.
- 240 [12] Romano Y, Patterson E, Candès EJ.: Conformalized Quantile Regression. arXiv.
241 ArXiv:1905.03222 [stat]. Available from: <http://arxiv.org/abs/1905.03222>.
- 242 [13] Herzen J, Lassig F, Piazzetta SG, Neuer T, Tafti L, Raille G, et al. Darts: User-
243 Friendly Modern Machine Learning for Time Series. *Journal of Machine Learning*
244 *Research*. 2022;23(124):1–6.