

1 When are predictions useful? A new method for 2 evaluating epidemic forecasts

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

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

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

7 Abstract

8 **Background:** COVID-19 will not be the last pandemic of the 21st century. To
9 better prepare for the next one, it is essential that we make honest appraisals of
10 the utility of different responses to COVID. In this paper we focus specifically
11 on epidemiologic forecasting. Characterizing forecast efficacy over the history of
12 the pandemic is challenging, especially given its significant spatial, temporal, and
13 contextual variability. In this light, we introduce the Weighted Contextual Inter-
14 val Score (WCIS), a new method for retrospective interval forecast evaluation.
15 The WCIS reflects the potential utility of predictions, resulting in a score that
16 is easily comparable across different pandemic scenarios despite remaining intu-
17 itively representative of the in-situ quality of individual forecasts.

18 **Methods:** The central tenet of the WCIS is a direct incorporation of contextual
19 utility into the evaluation. This necessitates a specific characterization of forecast
20 efficacy depending on the use case for predictions, accomplished via defining a
21 utility threshold parameter. In essence, changes in forecast accuracy beyond this
22 threshold do not map to changes in the utility of a prediction. This idea is gener-
23 alized to probabilistic interval-form forecasts, which are the preferred prediction
24 format for epidemiological modeling, as an adaptation of the existing Weighed
25 Interval Score (WIS).

26 **Results:** We apply the WCIS to two different forecasting scenarios. The first
27 assesses the performance of facility-level COVID-19 hospital bed occupancy pre-
28 dictions for the state of Maryland during the Omicron wave, and the second
29 evaluates state-level hospitalization forecasts drawn from the COVID-19 Forecast
30 Hub. We use these applications to demonstrate the parameterization of contex-
31 tual utility, compare the WCIS to the WIS, and explore the utility of the WCIS.

32 **Conclusions:** The WCIS provides a pragmatic utility-based characterization
33 of probabilistic predictions. This method is expressly intended to enable prac-
34 titioners and policymakers who may not have expertise in forecasting but are
35 nevertheless essential partners in epidemic response to use and provide insightful

analysis of predictions. We note that the WCIS is intended specifically for retrospective forecast evaluation and should not be used as a minimized penalty in a competitive context as it lacks statistical propriety.

Keywords: COVID-19, Epidemiology, Public health, Statistics

1 Background

1.1 Introduction

Given the devastating impact of COVID-19, and in the face of future pandemic threats, it is incumbent upon the epidemic forecasting community to deploy prediction tools that provide meaningful and actionable utility to those who need them. An important piece of this effort is candid retrospective evaluation of the utility of forecasting during the COVID-19 pandemic. In this light, we present a new probabilistic forecast evaluation method, the Weighted Contextual Interval Score (WCIS). It is a relative metric that encodes a simple question. How useful could a forecast have been where and when it was made? Unlike other scores, the WCIS is designed specifically as a retrospective way to judge whether or not forecasting could have been useful. It is not intended for real-time model ranking and ensemble construction. Instead, the WCIS is meant for broader pandemic preparedness efforts, for taking an honest look at how helpful forecasts could have been and thus potentially could be in the future. Despite the high spatial and temporal variability of pandemic scenarios, the WCIS evaluates forecasts in a comparable and communicable way by scoring them as a function of their potential utility.

The advent of the COVID-19 pandemic precipitated a massive public health response, including a significant modeling effort [1, 2]. In the United States, this quickly resulted in the formation of the COVID-19 Forecast Hub, a repository for short-term pandemic predictions [3]. Similar to prior collective forecasting efforts focused on seasonal influenza, dengue, and Ebola, the Forecast Hub solicited predictions from a large and diverse group of modelers, synthesizing their submissions into ensemble forecasts of COVID-19 cases, deaths, and hospitalizations. These outputs were provided to the United States Centers for Disease Control and Prevention (CDC) for policy making and dissemination to the public [4–9]. In addition to modeling efforts like the Hub at the regional level, COVID prompted a considerable amount of more granular forecasting, such as predictions for individual medical facilities [10, 11]. Despite this abundance of pandemic modeling, translating short-term epidemiological forecasts into applicable, actionable, and insightful decision-making remains a significant challenge [7, 12–17]. Understanding whether or not a forecast could have been useful requires understanding the conditions in which the forecast was made. It also requires knowledge about the type of decision the forecast would be used to inform. The WCIS was designed around these two requirements. It uses a utility-based normalization scheme to enable intuitive and meaningful comparison of forecast quality despite potentially dissimilar prediction contexts.

76 1.2 Motivation

77 Probabilistic predictions are preferred in many disciplines, including the epidemic
78 forecasting community. Unlike single outcome “point” predictions, probabilistic fore-
79 casts convey the uncertainty of the underlying model. This is particularly important
80 given the difficulty of correctly interpreting a quickly-evolving pandemic [7, 18]. The
81 extant Weighted Interval Score (WIS), an error metric for quantile forecasts that
82 approximates the Continuous Ranked Probability Score, is the primary method used
83 to evaluate Forecast Hub submissions [19, 20]. As summarized by Bracher et al., “the
84 (Weighted Interval) score can be interpreted heuristically as a measure of distance
85 between the predictive distribution and the true observation, where the units are those
86 of the absolute error” [19]. The WIS is an effective metric for real-time prediction
87 scoring, model comparison, and ensemble forecast creation [20]. However, the WIS
88 is limited in its ability to be used for intuitive forecast utility analysis, in particular
89 because the score is scaled on the order of the prediction data [19]. Retrospective pan-
90 demic evaluation involves comparing scenarios of highly different scales. One example
91 of such a comparison would be between regions with large baseline differences in data
92 magnitudes, such as highly vs sparsely populated regions (as in the Forecast Hub).
93 Another situation where scale-related contextualization is essential to consider is the
94 comparison of periods of high vs low epidemic activity (surge vs non-surge). In fact,
95 both of these spatial and temporal scaling challenges are often necessary to consider
96 at the same time (see Additional file 1: Section 1.1 for motivating examples of these
97 issues drawn from state-level pandemic scenarios in the United States).

98 The WCIS is an adaptation of the WIS that is framed around the two following
99 ideas. First, any meaningful measurement of forecast quality must arise from the con-
100 text into which predictions are disseminated. In other words, a useful forecast improves
101 real-time knowledge and/or decision-making capabilities. The reverse also holds: a
102 forecast is not useful if it is incapable of (or if it provides information detrimental to)
103 gaining real-time information or improving decision-making. Second, for the purposes
104 of enabling the comparison of forecast performances in disparate scenarios without
105 post-processing, a helpful score should be a relative metric. Taken together, these two
106 concepts informed the central idea of the WCIS: that a consistently meaningful score
107 must have endogenous contextualization. In essence, the WCIS normalizes forecast
108 performance as a function of the ability of the forecast to be used in the specific envi-
109 ronment in which it was made. This way, despite potentially occurring in radically
110 different spatial and temporal scenarios, individual evaluations can be meaningfully
111 compared to others.

112 Before moving to its technical basis and formulation in the next sections, it is
113 necessary to address the intended purpose of the WCIS, and similarly, tasks for which
114 it should not be used. The WCIS is not a statistically proper score (see Additional
115 file 1: Section 1.3), which means it should not be used in competitive forecasting
116 contexts. In these situations, such as real-time evaluation of COVID-19 Forecast Hub
117 submissions, scores that are not statistically proper have the potential to be gamed
118 [21]. Again, the WCIS is not designed for and should not be used for such purposes.
119 It was not created to replace the WIS, which functions well for real-time forecast
120 scoring and ensemble generation. Instead, the WCIS is designed to reflect relative

forecast quality using a flexible and contextually specific parameterization of utility. As is demonstrated in the test cases below, this results in a score that is not just intuitively interpreted, but is easy to compare and convey visually. We believe that these attributes are highly important in the context of pandemic preparedness efforts, given the need to more strongly connect the modeling and policy-making spheres of the public health community. Decision-makers need to be able to assess whether or not forecasting has the capacity to positively contribute to pandemic response. We believe that the WCIS enables an intuitive and flexible exploration of this question.

1.3 Review of the Weighted Interval Score

The Weighted Contextual Interval Score (WCIS) builds directly from the Weighted Interval Score (WIS). Bracher et al. [19] provide an excellent explanation of the mechanics of the score and its applications in epidemiology, and we endeavor to use the same symbology whenever possible. For brevity, the entire WIS formulation is not reviewed here, but the key elements (that are also important pieces the WCIS) are necessarily summarized:

$$IS_{\alpha}(F, y) = (u - l) + \frac{2}{\alpha} (l - y) \mathbb{1}\{y < l\} + \frac{2}{\alpha} (y - u) \mathbb{1}\{y > u\} \quad (1)$$

$$WIS_{\alpha\{0:K\}}(F, y) = \frac{1}{K + \frac{1}{2}} \left(w_0 \cdot |y - m| + \sum_{k=1}^K \{w_k \cdot IS_{\alpha_k}(F, y)\} \right) \quad (2)$$

- We assume a submission of K interval forecasts drawn from a predicted distribution F , a probabilistic representation of the target variable. Each of the K forecasts represents a $(1 - \alpha_k)$ prediction interval (PI). These intervals are delineated by their lower and upper bounds l and u , the $\frac{\alpha}{2}$ and $1 - \frac{\alpha}{2}$ quantiles of the predicted distribution, respectively. For example, a 95% interval would be represented by an α_k of 0.05, its lower and upper bounds defined by the 0.025 and 0.975 quantiles of F .
- A predictive median m (point prediction) is submitted, and the true target value y is known.
- For each interval $k \in \{1, 2, \dots, K\}$, an individual Interval Score (IS) is calculated, penalizing both the width/sharpness of the interval: $u - l$, and (if necessary) the amount by which the interval missed the true value: $\frac{2}{\alpha} (l - y) \mathbb{1}\{y < l\} + \frac{2}{\alpha} (y - u) \mathbb{1}\{y > u\}$ [21]. Note that the “miss” component is scaled by the inverse of α , thus narrower prediction intervals are penalized less for missing than are higher confidence submissions.
- The WIS is a weighted average of each of the K Interval Scores and the absolute error of the predictive median, with the weights w_k used for the average corresponding to $\frac{\alpha}{2}$ for each interval.

154 2 Methods

155 2.1 Contextualizing Point Forecasts

156 Although the WCIS (like the WIS) is an interval score, it is framed around a point
157 score that we call the Contextual Relative Error (CRE). The CRE maps the absolute
158 error of a point forecast x to its contextual utility. This is achieved by specifying
159 δ , the utility threshold parameter. (Note that δ is the only parameter in the WCIS
160 formulation that does not already appear in the WIS score).

$$CRE(x, y, \delta) = \min \left\{ \frac{|x - y|}{\delta}, 1 \right\} \quad (3)$$

161 δ is the magnitude of the absolute error above which a forecast loses its utility.
162 The CRE is so named because instead of mapping to the distance between a predicted
163 value and its target like absolute error, it maps to an interval from 0 to 1. A score of 0
164 indicates a forecast with maximum possible utility (with an absolute error of 0), and a
165 score of 1 indicates a useless forecast (with an absolute error of δ or more). See panel
166 (a) of Additional file 1: Fig. S1 for a graphical representation of the CRE. An impor-
167 tant feature to note is the “plateau” of the metric when the absolute error exceeds δ .
168 This might seem problematic, given that beyond the δ threshold the absolute error
169 is capable of increasing without any commensurate increase in the CRE. This is, in
170 fact, the desired behavior of the CRE and warrants a slight re-framing of perspective.
171 Selecting δ requires, when applying the CRE (and the WCIS, as it is a generalization
172 of the CRE from point to interval scores), identification of a practical limit for how
173 a forecast is used or interpreted in a particular context or for a particular purpose.
174 For example, in many scenarios we have a finite capacity to respond to an expected
175 outcome. If the “demand” imparted by an incorrect forecast exceeds that capacity,
176 we are unable to alter our response despite an apparent increase in need. Therefore,
177 an incorrect forecast with an absolute error of 2δ wastes exactly as many resources
178 as a incorrect forecast of magnitude δ , where δ precipitates the maximum allocation
179 in response to the forecast. A different way to interpret δ is as an “absorbable error
180 magnitude.” The test cases later in the paper frame δ this way, wherea decision maker
181 has limited capacity to recover from plans made according to forecasted outcomes. If
182 the forecast is wrong enough that it precipitates an action that cannot be recovered
183 from, such a forecast has met or exceeded the δ threshold.

184 Note that δ is both a normalizer and a limit. Thus a forecast with an absolute
185 error greater than δ is not at all useful, and a forecast with an absolute error less
186 than δ is evaluated as a ratio of δ . This gives the CRE (and the WCIS) the ability
187 to provide information about both forecast quality and how frequently forecasts are
188 useful, which, as demonstrated later, is helpful for intuitive analysis and performance
189 visualization.

190 2.2 Contextualizing Interval-Form Forecasts

191 We begin by introducing the Contextual Interval Score (CIS). The CIS is both a
192 probabilistic extension of the Contextual Relative Error, and a contextualized version

of the Interval Score. Like the CRE, it maps a forecast’s error to the δ -parameterized utility space, and like the IS, it generates a score for a single interval forecast. (In fact, the CIS can be equivalently formulated in two different ways, based on either the IS or the CRE. For brevity, we use the IS-based formulation here but, particularly if more intuition about the score is desired, we suggest referencing Section 1.2 in Additional file 1 for the explanation of the CRE-based formulation.)

$$CIS_{\alpha}(F, y, \delta) = \min \left\{ \frac{\alpha}{2\delta} IS_{\alpha}(F, y), 1 \right\} \quad (4)$$

The WCIS is the simple average of the CIS across all α -intervals and the CRE of the predictive median m :

$$WCIS_{\alpha\{0:K\}}(F, y, \delta) = \frac{1}{K+1} \left(CRE(m, y, \delta) + \sum_{k=1}^K CIS_{\alpha_k}(F, y, \delta) \right) \quad (5)$$

Note that we still retain the descriptor “Weighted” in the WCIS title even though there are no weights directly included in its formulation, whereas each component of the WIS is multiplied by $\frac{\alpha}{2}$. However, in our formulation, the same weights are effectively applied directly to the individual constituent CIS scores. Instead of the “miss” components of the score being multiplied by $\frac{2}{\alpha}$, the “width” term is scaled by $\frac{\alpha}{2}$. Thus when the average is taken to create the WCIS the scaling effect is the same as the WIS, but the weights are applied in this way because it preserves the interpretability of the individual single-interval CIS components as described above. Another notable difference is the WCIS uses $K+1$ for the denominator of the average (unlike $K+\frac{1}{2}$ in the WIS) because like the single-interval components, the predictive median component of the score has a maximum penalty of 1. This, and the bound on each CIS term, means the WCIS also takes values only on the interval from 0 to 1. Note the natural equivalence between the WCIS for interval forecasts and the CRE for point forecasts, which mirrors that between the WIS and the absolute error. In both cases, the interval scoring method preserves the behavior and intuitive interpretation of the corresponding point forecast technique.

3 Results

The WCIS is expressly intended to be a flexible scoring method and as such there are many possible and highly variable ways to apply it. We use this Results section to present two demonstrative use cases. Both scenarios evaluate COVID-19 hospitalization forecasts, but each works at a different scale and uses a necessarily different δ formulation. The first scenario applies the WCIS to results from a multi-facility-level forecasting model. We use this first application primarily to develop the intuition for the δ selection process. We show via a direct demonstration how δ can be chosen to represent contextually specific utility as a function of time-varying data, and explore how the choices made during this parameterization map onto the output of the WCIS.

227 Since this section focuses more on the WCIS formulation and less on interpreting the
 228 real-world applicability of the predictions, we use forecasts from a model developed
 229 in-house. Conversely, the second test case evaluates four weeks ahead predictions from
 230 the COVID-19 Forecast Hub’s ensemble model, examining hospitalization forecasts
 231 from May 2021 to May 2022 [3]. This period includes both the Delta and Omicron
 232 variant waves and allows for a larger exploration of the utility and communicability of
 233 the WCIS. Data for these analyses are sourced from the COVID-19 Reported Patient
 234 Impact and Hospital Capacity by Facility dataset for the first section and from the
 235 Forecast Hub’s repository for the second [22, 23].

236 3.1 Facility-level Analysis (First Test Case)

237 As introduced above, our first test case evaluates a facility-level hospitalization model.
 238 More specifically, the model forecasts daily COVID-19 bed occupancy, for each indi-
 239 vidual hospital in Maryland, from one to twenty-one days out, from July 2021 to July
 240 2022. Because our δ selection reflects capacity management within the three-week
 241 forecast window, we only use hospitals listed as “short-term” type (this excludes long-
 242 term and pediatric facilities) and for relevance only include facilities that had at least
 243 ten COVID-19 patients at some point during the time range specified. The partic-
 244 ular time range used was chosen because contextualization is vital when comparing
 245 and contrasting scenarios with highly different levels of pandemic activity, and July
 246 2021 to July 2022 includes the Omicron wave in Maryland. This scenario and facility
 247 selection yields 42 hospitals with an overall capacity range of 30 beds at the smallest
 248 facility to 919 beds at the largest facility.

249 The model used is a Time Series Dense Encoder, using the prior ninety days for
 250 each hospital at each time point to predict the following twenty-one days [24]. For a
 251 complete model formulation see Additional file 1: Section 2.1 but in brief, this model
 252 type was selected because it is a state-of-the-art general-purpose time series forecaster
 253 that is efficient to train and flexible across different covariates, prediction horizons,
 254 output types, and loss functions. We note that the purpose of this test case is to
 255 explore and explain the formulation and application of the WCIS. Thus, we developed
 256 this relatively basic model in order to apply the WCIS to a facility-level scenario, not
 257 to refine a specific method for forecasting hospitalizations. The predictions from this
 258 section are not necessarily indicative of those performed in real time. Because the data
 259 used for training and scoring this model may contain retrospective corrections of errors
 260 that were present in the real time data, it has the potential for higher performance
 261 when compared to an equivalent in-situ forecaster.

262 The δ -parameterization used for this analysis is intended to characterize the capac-
 263 ity of each facility to absorb an incorrect allocation of COVID-19 bed space based on a
 264 flawed forecast. We assume that capacity allocations are made at forecast time, under
 265 the in-situ assumption that forecasts perfectly reflect future outcomes. Thus, the δ
 266 value represents an achievable capacity correction during the time interval separating
 267 the making of the forecast and the realization of its true target value. For example,
 268 the δ value for a seven-day-ahead forecast for each facility is the amount of COVID-
 269 19 beds that each individual hospital can add or take away over a week. Specifically,

270 this δ is determined as follows. The daily capacity change for each hospital is calcu-
 271 lated as the mean of all single day, non-zero capacity changes over the entire available
 272 time series for each facility. δ for a particular forecast is then set as the product of
 273 the forecast horizon and the facility-specific daily change capacity. This means that
 274 the further out a forecast is, the larger (and thus more forgiving) the delta value is,
 275 based on the idea that the more time a facility has to respond to a poor allocation of
 276 resources, the greater the magnitude of the response can be. Please note that the par-
 277 ticular formulation chosen here is not intended to provide an assessment of forecast
 278 quality outside the utility scenario posited by the assumptions given above. How-
 279 ever, it demonstrates an important capability of utility threshold selection: δ can be a
 280 defined as a dynamic function of data that can change in time and space. Since con-
 281 textually meaningful forecast utility varies significantly over these same dimensions,
 282 a broadly applicable and interpretable score must be similarly adaptable.

283 Using Figures 1 and 2, we are able to interpret some important aspects of how
 284 this selection of δ maps onto the scoring of our facility-level model. First, consider
 285 the relationship between the breadth of the confidence intervals and the δ region in
 286 Figure 1, which visualizes a single facility. The larger prediction intervals for the four-
 287 teen day-ahead forecasts indicate less model certainty than those of the two day-ahead
 288 predictions and, all else equal, would yield a worse score. However, δ is significantly
 289 higher for the fourteen day-ahead scenario, given the assumption that facilities have
 290 more time to adapt to inaccurate forecasts over longer horizons. This results in gen-
 291 erally better performance for the fourteen-day model. However, there remain in the
 292 fourteen-day scenario several forecasts that still receive a high penalty despite the
 293 more forgiving δ parameterization. Note that these instances tend to occur when the
 294 forecast median approaches or exceeds the utility threshold. Moving to Figure 2, we
 295 can see that these trends are also visible in the aggregate performance across all 42
 296 facilities. Comparing the WIS to the WCIS over these instances reveals a relatively
 297 linear relationship in the more forgiving scenarios, i.e. non-wave with a larger delta.
 298 During the wave, when absolute performance was broadly worse (as evidenced by the
 299 WIS values), the δ -limit was reached significantly more often. We also draw attention
 300 to the differences in marginal distributions that are visible in the scatter plot column
 301 of Figure 2. The scaling and limiting action of the WCIS distributes performances sig-
 302 nificantly more evenly than the WIS (see Additional file 1: Section 2.2 for plots with
 303 these marginal distributions included).

304 In general, we are able to observe that given a contextually relevant δ choice, the
 305 score is able to simultaneously convey an intuitive sense of both relative quality and
 306 the overall frequency of useful forecasts, as shown in the histograms of Figure 2.

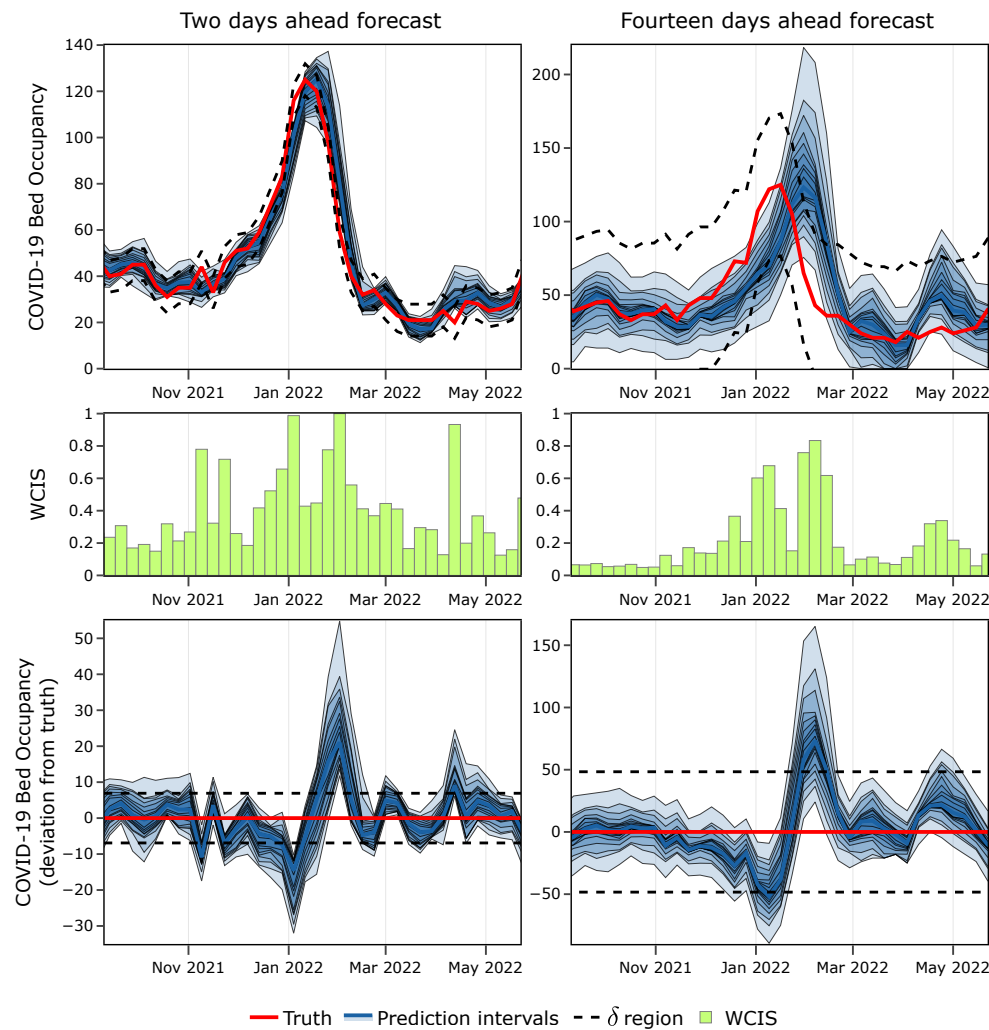


Fig. 1 Illustrated here are facility-level forecasts over two prediction horizons for one hospital: the University of Maryland Medical Center. The top and bottom rows both show the same forecasts, truth data, and δ (utility threshold) region. The top row displays these values normally, whereas the bottom row shows how far each value deviates from the truth. The middle row displays the WCIS, aligned with the data in the other rows. (Note that the facility-level analysis includes more prediction intervals and more dates than are shown in this figure, the extent of both displayed here are reduced for clarity.)

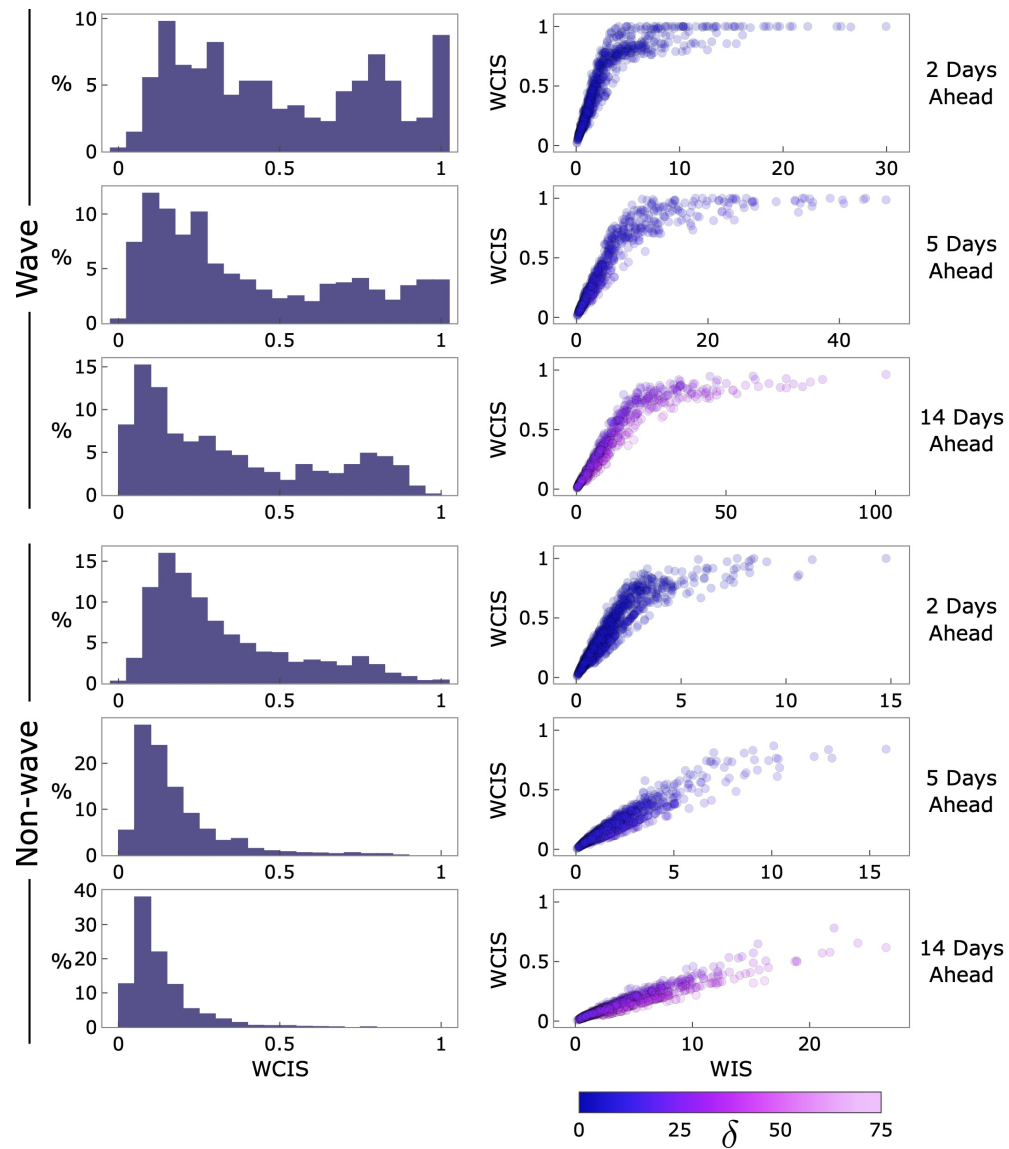


Fig. 2 Results in this figure are generated from all 42 hospitals, for all prediction dates in the facility-level model. The top three rows are from forecasts during the Omicron wave, and the bottom three from before and after the wave. We define the wave as lasting from November 14 2021 through May 15 2022, as illustrated in Additional file 1: Fig. S4.

307 3.2 State-level Analysis (Second Test Case)

308 For this test case, we apply the WCIS to real-world predictions drawn from the Fore-
 309 cast Hub, asking how much contextual utility hospitalization forecasts provided at
 310 the state level from May 2021 to May 2022 [3]. (Note that Forecast Hub hospitaliza-
 311 tion predictions were performed at daily resolution, but for the sake of visualizing a
 312 longer-term analysis we aggregate to and evaluate at weekly totals.)

313 The WCIS always requires a specific interpretation of the use-case for forecasts
 314 in the selection of the utility threshold δ . Similar to Section 3.1, we choose to assess
 315 hospitalization predictions as a function of potential capacity changes. However, we
 316 assume a different decision-making scenario for hospital capacity at the state level
 317 than for its facility level counterpart. Due to the disaggregate decision-making appa-
 318 ratus across statewide hospitals and the inherent institutional inertia that must be
 319 overcome for larger scale change, we take a more conservative approach to estimating
 320 the absorbable error magnitude. Specifically, δ is the 0.9 quantile of the prior devia-
 321 tions in each state's hospital bed capacity over the prediction horizon of the forecast.
 322 We assume prior bed capacity deviations are indicative of a state's capacity to make
 323 changes, and that it is more difficult to make changes over a shorter timeline. Thus, any
 324 deviation over a shorter-term horizon can also occur for longer term horizons, but not
 325 the reverse. For example, when examining one week ahead predictions, only historical
 326 capacity changes over the course of a single week are considered. For four weeks ahead
 327 predictions, capacity changes for one, two, three, and four weeks ahead are considered.
 328 Finally, the 0.9 quantile is selected as the threshold under the assumption that states
 329 are not necessarily able to repeat their largest historical deviations, but can approach
 330 them. To be clear, this choice of δ is a heuristic for the amount of resource alloca-
 331 tion, staffing changes, and other matters that hospitals might practically accomplish
 332 in response to an assumed change in pandemic dynamics. It is intended to demon-
 333 strate the WCIS given a reasonable, data-driven parameterization of forecast utility.
 334 Namely, a response predicated on a forecast outside the δ -range as defined here would
 335 require corrective action of a magnitude that could not be reasonably expected over
 336 such a forecast's prediction horizon.

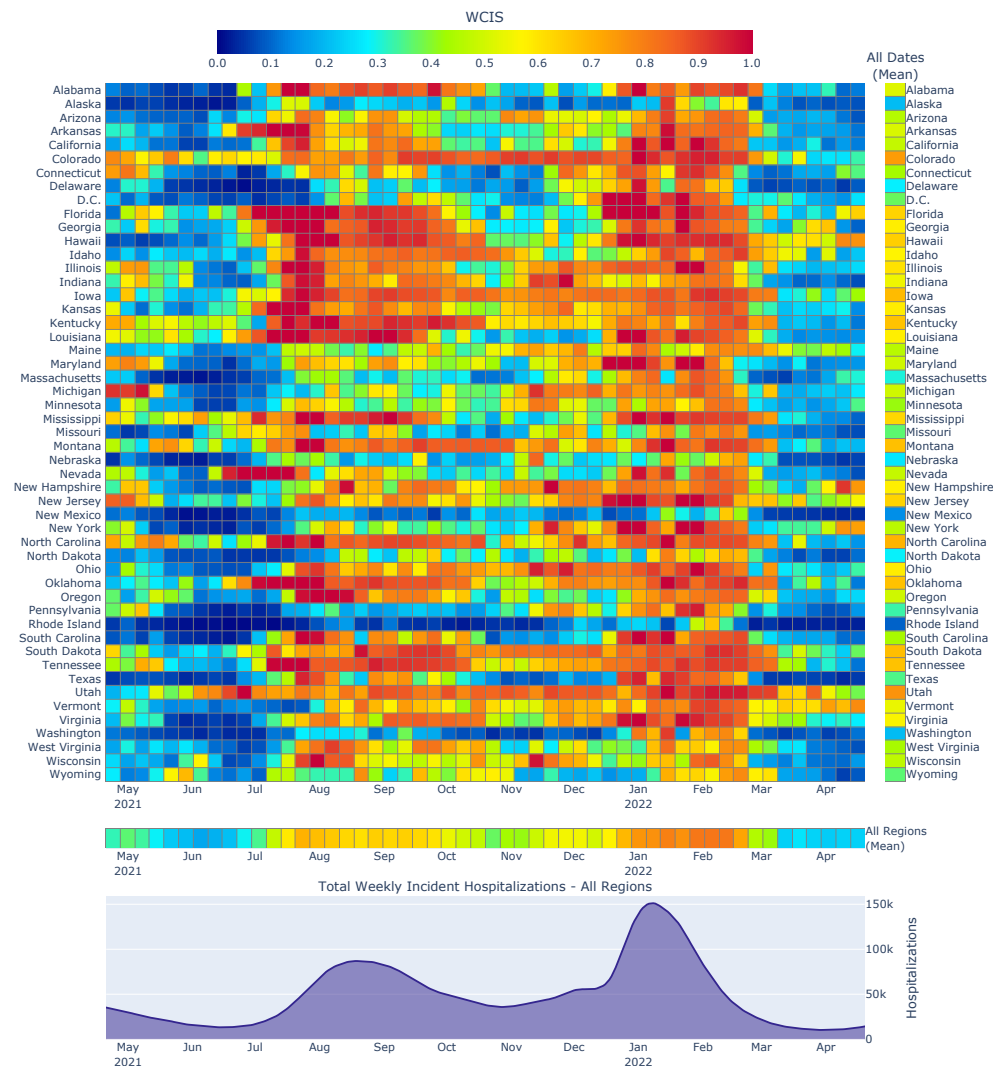


Fig. 3 Heatmap of the WCIS for 4 week ahead hospitalization forecasts, performed by the Forecast Hub's ensemble model. The central and largest grid shows the most granular results: region- and time-specific performance. On the right and lower sides of the grid are average performances over time and space, respectively. The shaded line plot at the bottom of the figure is the target hospitalization variable aggregated across all regions. Note that its domain is aligned exactly with those of the time-dependent heatmaps above, to provide insight into the trends of the overall pandemic alongside the more granular information in the heatmaps. (See Additional file 1: Section 3 for heatmaps over differing prediction horizons).

WCIS performance results for four weeks ahead state-level hospitalization predictions are demonstrated in Figure 3. Since the WCIS was designed primarily as a way to meaningfully evaluate and compare forecasts in disparate contexts, we can easily use it to observe several important aspects of hospitalization forecasting performance. For example, during surges and declines, forecast utility decreases substantially. We can intuit that this is a consistent trend across different locations both by directly observing the large central grid and by examining the lower, spatially averaged array of the figure. In contrast, if we examine the right-side, temporally averaged array, we observe that there is less variability in space than there is in time. Thus, by making an up-front determination about what constitutes a useful prediction (performing the δ -parameterization), we are capable of making, displaying, and intuitively evaluating forecasts. This allows, given a well-informed choice of δ , for meaningful overall analysis without needing to repeatedly delve into the specific circumstances during which each forecast was made. Without contextual normalization, conveying informative and comparable performance would be much more challenging. This capability, demonstrated by the ease of interpreting Figure 3, is the overall aim for our creation of the WCIS. It permits substantive and easily interpretable performance evaluation.

4 Discussion

The WCIS is framed around our belief that a useful forecast contributes meaningful and/or actionable information given uncertain future outcomes. Determining whether or not forecasts accomplish this necessitates an explicit definition of utility. This brings up an important philosophical difference between the WCIS and other techniques. The WCIS formulation, centered around a user-defined utility threshold δ , arises from our assertion that there will never be a one-size-fits-all solution for assessing and comparing short-term forecast quality. One must always consider prediction context and purpose lest standard metrics tell a misleading story. Additionally, different forecast use-cases yield different judgments of predictions. The helpfulness of a model that predicts rainfall, for example, will be judged very differently by a user deciding whether or not to bring an umbrella on a walk as compared to a user deciding whether or not to issue regional flood warnings. An incorrect forecast of light rain with a realization of heavy rain is good enough for the first user but may be catastrophic for the second. Again, forecast purpose is essential to consider. The WCIS ensures this by building a definition of forecast utility directly into the formulation of the score.

The core of the WCIS is the combined normalization and thresholding imposed by the δ parameterization, which incorporates a vital aspect of real-world forecast utility. Namely, past a certain point, changes in a prediction's absolute error do not equate to changes in outcomes predicated on that forecast. Even when one forecast is more accurate than another, if both are beyond the utility horizon then the "better" one is not actually more useful, just arbitrarily closer to the truth. This idea is the basis for the plateaued CRE point scoring function, which in turn is the basis for the WCIS. While a metric that does not always increase the penalty as forecast accuracy diminishes may seem counterintuitive, we believe that for characterizing contextual

utility, a score with a limited scope of relevance is actually more intuitive than a score that gets arbitrarily worse (or better) no matter how far away it is from being helpful. The WCIS builds on the Weighted Interval Score, adding the δ -parameterization to impel users to directly characterize contextual utility. Judging predictions in this way allows for a powerful and effective normalization of the error, making the WCIS easy to interpret and compare across heterogeneous forecasting scenarios. Importantly, this robust efficacy exists *only for each individual definition of utility*. We belabor this point because it is inherent to our overall assertion about forecast interpretability: that a specific use case is necessary to meaningfully evaluate prediction quality. Without an explicit link to how forecasts are used, there is no way to consistently and meaningfully evaluate them over variable spatial and temporal conditions. Other evaluation metrics are in essence arbitrary until they are contextualized, whereas the WCIS builds this contextualization directly into the formulation of the score.

5 Conclusions

Determining the future role of pandemic forecasting, as well as identifying areas of forecasting that need improvement, must at some point include the translation of modeling results to policy and decision makers. The WCIS is expressly intended to function well in this process, allowing for intuitive characterization of forecast utility that can be easily communicated to an audience with less technical expertise. Figure 3 demonstrates this directly. Without effective contextual normalization, generating such a display would be challenging given large differences in error magnitude, likely requiring a transformation (such as log-scaling) that limits interpretability. Instead, the WCIS allows for a direct, clearly defined interpretation of forecast utility to be displayed and compared in a technically meaningful and intuitively understandable way.

We created the WCIS to enable and encourage honest and contextually specific discourse about the utility of short-term epidemic predictions. It incorporates prediction uncertainty, keeps the technical definition of utility as simple as possible, and generates an intuitively interpretable and comparable numerical output. Our intent is to allow for people without specific technical experience to be able to interact with and evaluate probabilistic forecasting in a meaningful way. As the public health community learns from COVID-19 and prepares for future challenges, explicit analysis of the utility of historical predictions is essential. We hope the WCIS will help with effective and meaningful communication between modelers and practitioners in this effort.

Declarations

- **Ethics approval and consent to participate:** Not applicable. Ethics approval was not sought or required for this study. All human health data used are population-level COVID-19 outcomes sourced from publicly available online repositories.
- **Consent for publication:** Not applicable.
- **Availability of data and materials:** Code and processed data for both the facility- and state-level analyses are accessible from a publicly available GitHub

repository [<https://github.com/cpt-diabetes/wcis>]. Forecast and ground truth data used for our state-level analysis are available from the COVID-19 Forecast Hub repository [<https://doi.org/10.5281/zenodo.6301718>]. The original source for ground truth hospitalization data is the COVID-19 Reported Patient Impact and Hospital Capacity by Facility repository [<https://healthdata.gov/d/j4ip-wfsv>].

- **Competing interests:** The authors declare that they have no competing interests
- **Funding:** This study was supported by the United States National Science Foundation (NSF) under grant no. 2108526.
- **Authors' contributions:** MM conceived of the study, performed the analysis, and wrote the manuscript. MM and FP developed the methodology. LMG supervised the project and provided essential methodological guidance. All authors read, edited, and approved the final manuscript.
- **Acknowledgements:** Not applicable.

References

- [1] Horbach SPJM. Pandemic publishing: Medical journals strongly speed up their publication process for COVID-19. *Quantitative Science Studies*. 2020 Aug;1(3):1056–1067. <https://doi.org/10.1162/qss.a.00076>.
- [2] Fraser N, Brierley L, Dey G, Polka JK, Pálffy M, Nanni F, et al. The evolving role of preprints in the dissemination of COVID-19 research and their impact on the science communication landscape. *PLOS Biology*. 2021 Apr;19(4):e3000959. Publisher: Public Library of Science. <https://doi.org/10.1371/journal.pbio.3000959>.
- [3] Cramer EY, Huang Y, Wang Y, Ray EL, Cornell M, Bracher J, et al. The United States COVID-19 Forecast Hub dataset. *Scientific Data*. 2022 Aug;9(1):462. Number: 1 Publisher: Nature Publishing Group. <https://doi.org/10.1038/s41597-022-01517-w>.
- [4] McGowan CJ, Biggerstaff M, Johansson M, Apfeldorf KM, Ben-Nun M, Brooks L, et al. Collaborative efforts to forecast seasonal influenza in the United States, 2015–2016. *Scientific Reports*. 2019 Jan;9(1):683. Number: 1 Publisher: Nature Publishing Group. <https://doi.org/10.1038/s41598-018-36361-9>.
- [5] Johansson MA, Apfeldorf KM, Dobson S, Devita J, Buczak AL, Baugher B, et al. An open challenge to advance probabilistic forecasting for dengue epidemics. *Proceedings of the National Academy of Sciences*. 2019 Nov;116(48):24268–24274. Publisher: Proceedings of the National Academy of Sciences. <https://doi.org/10.1073/pnas.1909865116>.
- [6] Viboud C, Sun K, Gaffey R, Ajelli M, Fumanelli L, Merler S, et al. The RAPIDD ebola forecasting challenge: Synthesis and lessons learnt. *Epidemics*. 2018 Mar;22:13–21. <https://doi.org/10.1016/j.epidem.2017.08.002>.

- 458 [7] Reich NG, Ray EL. Collaborative modeling key to improving outbreak response.
459 Proceedings of the National Academy of Sciences. 2022 Apr;119(14):e2200703119.
460 Publisher: Proceedings of the National Academy of Sciences. <https://doi.org/10.1073/pnas.2200703119>.
461
- 462 [8] Ray EL, Brooks LC, Bien J, Biggerstaff M, Bosse NI, Bracher J, et al. Comparing
463 trained and untrained probabilistic ensemble forecasts of COVID-19 cases and
464 deaths in the United States. International Journal of Forecasting. 2022 Jul;<https://doi.org/10.1016/j.ijforecast.2022.06.005>.
465
- 466 [9] Reich NG, McGowan CJ, Yamana TK, Tushar A, Ray EL, Osthus D, et al.
467 Accuracy of real-time multi-model ensemble forecasts for seasonal influenza in
468 the U.S. PLOS Computational Biology. 2019 Nov;15(11):e1007486. Publisher:
469 Public Library of Science. <https://doi.org/10.1371/journal.pcbi.1007486>.
- 470 [10] Weissman GE, Crane-Droesch A, Chivers C, Luong T, Hanish A, Levy MZ,
471 et al. Locally Informed Simulation to Predict Hospital Capacity Needs During
472 the COVID-19 Pandemic. Annals of Internal Medicine. 2020 Jul;173(1):21–28.
473 <https://doi.org/10.7326/M20-1260>.
- 474 [11] Kociurzynski R, D'Ambrosio A, Papathanassopoulos A, Bürkin F, Hertweck S,
475 Eichel VM, et al. Forecasting Local Hospital Bed Demand for COVID-19 Using
476 on-Request Simulations. Scientific Reports. 2023 Dec;13(1):21321. <https://doi.org/10.1038/s41598-023-48601-8>.
477
- 478 [12] Doms C, Kramer SC, Shaman J. Assessing the Use of Influenza Forecasts and
479 Epidemiological Modeling in Public Health Decision Making in the United States.
480 Scientific Reports. 2018 Aug;8(1):12406. Number: 1 Publisher: Nature Publishing
481 Group. <https://doi.org/10.1038/s41598-018-30378-w>.
- 482 [13] Reich NG, Wang Y, Burns M, Ergas R, Cramer EY, Ray EL.: Assessing the utility
483 of COVID-19 case reports as a leading indicator for hospitalization forecasting
484 in the United States. medRxiv. Pages: 2023.03.08.23286582. Available from:
485 <https://www.medrxiv.org/content/10.1101/2023.03.08.23286582v1>.
- 486 [14] Nixon K, Jindal S, Parker F, Marshall M, Reich NG, Ghobadi K, et al. Real-time
487 COVID-19 forecasting: challenges and opportunities of model performance and
488 translation. The Lancet Digital Health. 2022 Oct;4(10):e699–e701. Publisher:
489 Elsevier. [https://doi.org/10.1016/S2589-7500\(22\)00167-4](https://doi.org/10.1016/S2589-7500(22)00167-4).
- 490 [15] Lutz CS, Huynh MP, Schroeder M, Anyatonwu S, Dahlgren FS, Danyluk G,
491 et al. Applying infectious disease forecasting to public health: a path forward
492 using influenza forecasting examples. BMC Public Health. 2019 Dec;19(1):1659.
493 <https://doi.org/10.1186/s12889-019-7966-8>.
- 494 [16] Guerrier C, McDonnell C, Magoc T, Fishe JN, Harle CA. Understanding Health
495 Care Administrators' Data and Information Needs for Decision Making during

- the COVID-19 Pandemic: A Qualitative Study at an Academic Health System. MDM Policy & Practice. 2022 Jan;7(1):23814683221089844. Publisher: SAGE Publications Inc. <https://doi.org/10.1177/23814683221089844>.
- [17] Lee TH, Do B, Dantzing L, Holmes J, Chyba M, Hankins S, et al. Mitigation Planning and Policies Informed by COVID-19 Modeling: A Framework and Case Study of the State of Hawaii. International Journal of Environmental Research and Public Health. 2022 Jan;19(10):6119. Number: 10 Publisher: Multidisciplinary Digital Publishing Institute. <https://doi.org/10.3390/ijerph19106119>.
- [18] Nixon K, Jindal S, Parker F, Reich NG, Ghobadi K, Lee EC, et al. An evaluation of prospective COVID-19 modelling studies in the USA: from data to science translation. The Lancet Digital Health. 2022 Oct;4(10):e738–e747. [https://doi.org/10.1016/S2589-7500\(22\)00148-0](https://doi.org/10.1016/S2589-7500(22)00148-0).
- [19] Bracher J, Ray EL, Gneiting T, Reich NG. Evaluating epidemic forecasts in an interval format. PLOS Computational Biology. 2021 Feb;17(2):e1008618. Publisher: Public Library of Science. <https://doi.org/10.1371/journal.pcbi.1008618>.
- [20] Cramer EY, Ray EL, Lopez VK, Bracher J, Brennen A, Castro Rivadeneira AJ, et al. Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the United States. Proceedings of the National Academy of Sciences. 2022 Apr;119(15):e2113561119. Publisher: Proceedings of the National Academy of Sciences. <https://doi.org/10.1073/pnas.2113561119>.
- [21] Gneiting T, Raftery AE. Strictly Proper Scoring Rules, Prediction, and Estimation. Journal of the American Statistical Association. 2007 Mar;102(477):359–378. <https://doi.org/10.1198/016214506000001437>.
- [22] gov H.: COVID-19 Reported Patient Impact and Hospital Capacity by Facility. United States Department of Health & Human Services. Available from: <https://healthdata.gov/d/j4ip-wfsv>.
- [23] Cramer E, Wang SY, Reich NG, Hannan A, Niemi J, Ray E, et al.: reichlab/covid19-forecast-hub: release for Zenodo 20220227. Zenodo. Available from: <https://doi.org/10.5281/zenodo.6301718>.
- [24] Das A, Kong W, Leach A, Mathur S, Sen R, Yu R.: Long-term Forecasting with TiDE: Time-series Dense Encoder. arXiv. ArXiv:2304.08424 [cs, stat]. Available from: <http://arxiv.org/abs/2304.08424>.

Additional Files

- “Additional file 1.pdf” - This document contains the Supplemental Materials for this article. These include sections that provide more detail on and/or motivating examples for the formulation of the score, the impropriety analysis, and the facility-level model formulation. It also includes figures comparing the WIS and the WCIS

533 performance of the facility-level model for different scenarios, and state-level WCIS
534 hospitalization heatmaps for all 4 standard Forecast Hub prediction horizons (1, 2,
535 3, and 4 weeks ahead).