WHEN ARE PREDICTIONS USEFUL? A NEW METHOD FOR EVALUATING EPIDEMIC FORECASTS

Maximilian Marshall[1,*], Felix Parker[1], Lauren M Gardner[1]

June 29, 2023

ABSTRACT

We introduce the Weighted Contextual Interval Score (WCIS), a new method for evaluating the performance of short-term interval-form forecasts. The WCIS provides a pragmatic utility-based characterization of probabilistic predictions, developed in response to the challenge of evaluating forecast performances in the turbulent context of the COVID-19 pandemic. Current widely-used scoring techniques generally fall into two groups: those that generate an individually interpretable metric, and those that generate a comparable and aggregable metric. The WCIS harmonizes these attributes, resulting in a normalized score that is nevertheless intuitively representative of the in-situ quality of individual forecasts. This method is expressly intended to enable practitioners and policy-makers who may not have expertise in forecasting but are nevertheless essential partners in epidemic response to use and provide insightful analysis of predictions. In this paper, we detail the methodology of the WCIS and demonstrate its utility in the context of US state-level COVID-19 predictions.

Keywords COVID-19 · Epidemiology · Public health · Statistics

1 Introduction

1.1 Background

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. The Forecast Hub connects academic and industry forecasters to the United States Centers for Disease Control and Prevention (CDC), providing projections of COVID-19 related cases, deaths, and hospitalizations that the CDC uses for policy making and dissemination to the public [3]. Similar to prior collective forecasting efforts focused on seasonal influenza, dengue, and Ebola, the Forecast Hub solicited predictions from a large group of modelers using diverse techniques, synthesizing the submissions into ensemble forecasts that were judged to consistently outperform their component predictions [4, 5, 6, 7, 8, 9]. In this article, we use these ensemble forecasts as test cases for our new metric: the Weighted Contextual Interval Score (WCIS). While the WCIS could easily be applied to other types of forecasting, it was designed with efforts like the COVID-19 Forecast Hub in mind. As a collaboration between modelers, public health practitioners, and government officials, the Hub is representative of efforts that will remain vital given the danger posed by both extant and heretofore unknown epidemic threats [10]. However, using forecasts for consistent real time decision making remains a challenge, a vital component which is translating forecast data into actionable insights [7, 11, 12, 13, 14]. In this light, we present the WCIS as a way to alleviate challenges in this space that arise from comparing, aggregating, and interpreting forecasts made across highly spatially and temporally variable prediction instances. It does so by encoding a simple question: how useful was the prediction where and when it was made?

[1]Department of Civil and Systems Engineering, Johns Hopkins University, Baltimore, MD
[*]Correspondence to Maximilian Marshall: mmarsh29@jhu.edu

Probabilistic predictions are increasingly preferred in many disciplines, including the epidemic forecasting community. Unlike single outcome "point" predictions, probabilistic forecasts convey the uncertainty of the underlying model [15]. This is particularly important given the inherent difficulty of correctly interpreting a quickly-evolving pandemic [7]. Additionally, reporting only point forecasts (thus neglecting to convey any uncertainty) runs the public health risk of disseminating potentially incorrect predictions with an apparently high degree of confidence [16]. In keeping with these currently accepted best practices, the Forecast Hub requires submissions to be reported in quantile form [3]. The Weighted Interval Score (WIS), an error metric for quantile/interval scores that approximates the Continuous Ranked Probability Score, is the primary method used to evaluate Forecast Hub submissions [15, 17]. Note that the mechanics of Weighted Interval Score are necessarily considered in more detail in a subsequent section of this paper, as it is a direct precursor to our new WCIS score. In brief, its functionality is elegantly summarized by Bracher et al.: "the (Weighted Interval) score can be interpreted heuristically as a measure of distance between the predictive distribution and the true observation, where the units are those of the absolute error" [15]. This conveys an important benefit, intuitive interpretability, to the WIS. In spite of its relatively complex formulation, it can easily be understood as a probabilistic analogue of the absolute error. Unfortunately, this means that the Weighted Interval Score also suffers from similar limitations to the absolute error. In particular, it is sensitive (especially for the purposes of comparison and aggregation) to differences in scale. This presents a significant problem when used in a context like the COVID-19 Forecast Hub, where target scale varies significantly in space and time.

1.2 Motivation - Spatial and Temporal Instability

Creation of the WCIS arose from attempts to retrospectively characterize and intuitively communicate the utility of short-term COVID-19 prediction efforts like the Forecast Hub. We found that standard metrics were not able to sufficiently and robustly capture the interaction of predictions with the spatially variable and temporally dynamic reality of the evolving pandemic. Meaningful analysis, given extant scoring methods, always required a substantial expenditure of effort characterizing the on-the-ground reality when and where forecasts were made. In this section, we provide motivating examples of this issue drawn from US state-level Forecast Hub predictions.

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 ($\sim$40 million), and Wyoming the lowest ($\sim$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 [18]. 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. However, as indicated above, probabilistic evaluation is susceptible to the same issues [15].) For the week under consideration, predictions from the Forecast Hub's baseline model yielded a MAE of 27,130 across all US states [19]. For California, a prediction that overshot the truth by this margin would incur a percent error of only about 3%, whereas for Wyoming, such a prediction would miss by over 400%. Admittedly, this is not by itself a particularly vexing problem (normalization by population, for example, would likely suffice in this case). Unfortunately, spatial inconsistency is not the only obstacle. Accounting for temporal context is equally vital and presents its own difficulties.

When examining forecast performance for a single region over time, metrics must be interpreted as a function of time-variant data. This necessity is demonstrated trivially by comparing pandemic surges to times of relatively low epidemic activity. The same value of a non-normalized metric like the absolute error carries an entirely different meaning in each of these situations. Consider the Forecast Hub's baseline model predictions for cases in Maryland. In mid-December 2020, this model missed its three-weeks-ahead target by about 2,000 cases. In mid-May 2021, the same model also missed by about 2,000 cases [19]. Without knowing the context of each prediction, (namely that the first was made during a massive surge and the second was made during a significant lull), one might be forgiven for assuming that the model performed similarly in both scenarios. However, the December forecast only just missed the mark, undershooting by 12% of the true value. Conversely, the May forecast missed by 213%. Note that in this case, percent error has interpretable utility because it normalizes by the true value, a time-varying data source that directly represents the prevailing condition of the pandemic. Unfortunately, percent error is not an ideal solution as it becomes

unstable when true values approach zero [15]. This is especially problematic when analyzing death forecasts (for all of 2020 through 2022, almost 15% of US states had less than ten weekly deaths, and over 8% had below five weekly deaths). In this situation, percent error is in fact too sensitive to the exact circumstances. It indicates a large deviation from the truth which, while technically correct, misses the reality of how forecasts are interpreted. Given the larger context of the pandemic, it is unreasonable to characterize a four-death forecast compared to a target value of one (300% error) as a worse prediction than a 400-death forecast compared to an 800-death reality (50% error). Like the spatial case, if context is not very carefully considered, the numerical value of an error metric can be inconsistent with reality.

Consequently, we submit that any definition of forecast quality must arise from the context into which predictions are disseminated. In other words, a useful real-time forecast is capable of improving real-time decision-making. The reverse also holds: a forecast is not useful if it is incapable of (or if it provides information detrimental to) meaningfully informing a decision made using the forecast. This link between forecast utility and in situ decision-making is key. In fact, it is the basis for the core functionality of the WCIS. In essence, the WCIS normalizes forecast performance as a function of the ability of the forecast to be used *in the specific environment in which it was made*. This way, despite (potentially) occurring in radically different spatial and temporal scenarios, individual evaluations can be meaningfully compared to others.

1.3 Review of the Weighted Interval Score

As our choice of name is intended to suggest, the Weighted Contextual Interval Score (WCIS) builds directly from the Weighted Interval Score (WIS), a robust and widely-used evaluation metric for quantile forecasts. If the reader is unfamiliar with the WIS, Bracher et al. [15] provide an excellent explanation of the mechanics of the score and its applications in epidemiology. We encourage familiarity with their formulation and endeavor to use the same symbology in this paper whenever possible. For brevity, the entire WIS formulation is not reviewed here, but the key elements (that are also important pieces of our new WCIS score) are necessarily summarized here:

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

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

- 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 $k \in \{1,2,...,K\}$, 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: $\frac{2}{\alpha}(l-y)\,\mathbb{1}\{y<l\} + \frac{2}{\alpha}(y-u)\,\mathbb{1}\{y>u\}$ [20]. 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.

2 Methods

2.1 Contextual Absolute Error (CAE)

Although the WCIS (like the WIS) is an interval score, its logic is fundamentally rooted in a much simpler point score that we call the Contextual Absolute Error (CAE). In effect, the CAE is a function that maps the absolute error of a point forecast x to its contextual utility. This is achieved by specifying δ, a utility threshold parameter. (Note that δ is the only parameter in the WCIS formulation that does not already appear in the WIS score).

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

What is δ, and how is it chosen? In essence, it is the magnitude of the absolute error above which a forecast loses its ability to be useful. This requires the person applying the CAE (and the WCIS, as the CAE is the foundational part

of the WCIS) to identify some practical limit for how forecasts might be used. The CAE is so named because it is analogous to absolute error, but instead of mapping to the distance between a predicted value and its target, the CAE maps to an interval from 0 to 1. A result of 0 represents a perfect forecast, and a result of 1 represents a useless forecast (a forecast with an absolute error greater than δ). Thus, a decision made based on a forecast with an error beyond this margin might prompt an unrecoverable response, rendering such a forecast useless or misinformative. However, it is important to emphasize that selecting a specific δ creates a judgment of forecast value *for a particular purpose*. Again, we emphasize that forecasts in spatially and temporally heterogeneous scenarios should not be numerically compared unless context is taken into account; context that, in this methodology, is defined by the selection of δ. While the requirement of using a specific δ value might be seen as adding unnecessary complexity to a forecast evaluation metric, such an argument presupposes that a "simpler" score (like WIS or MAE) possesses broad functionality. As we have demonstrated, a score that is not robust to the dynamics of the forecasting landscape does not convey a largely consistent meaning, thus is not broadly functional. The CAE, conversely, conveys an intuitive forecast evaluation precisely because of the selection of an appropriate δ. See panel *(a)* of Figure 1 for a graphical representation of the CAE.

2.2 Weighted Contextual Interval Score (WCIS)

We first consider the case of a single prediction interval, just as the Interval Score (IS) is used as a single-interval constituent of the WIS. We define the single-interval Contextual Interval Score (CIS) as follows:

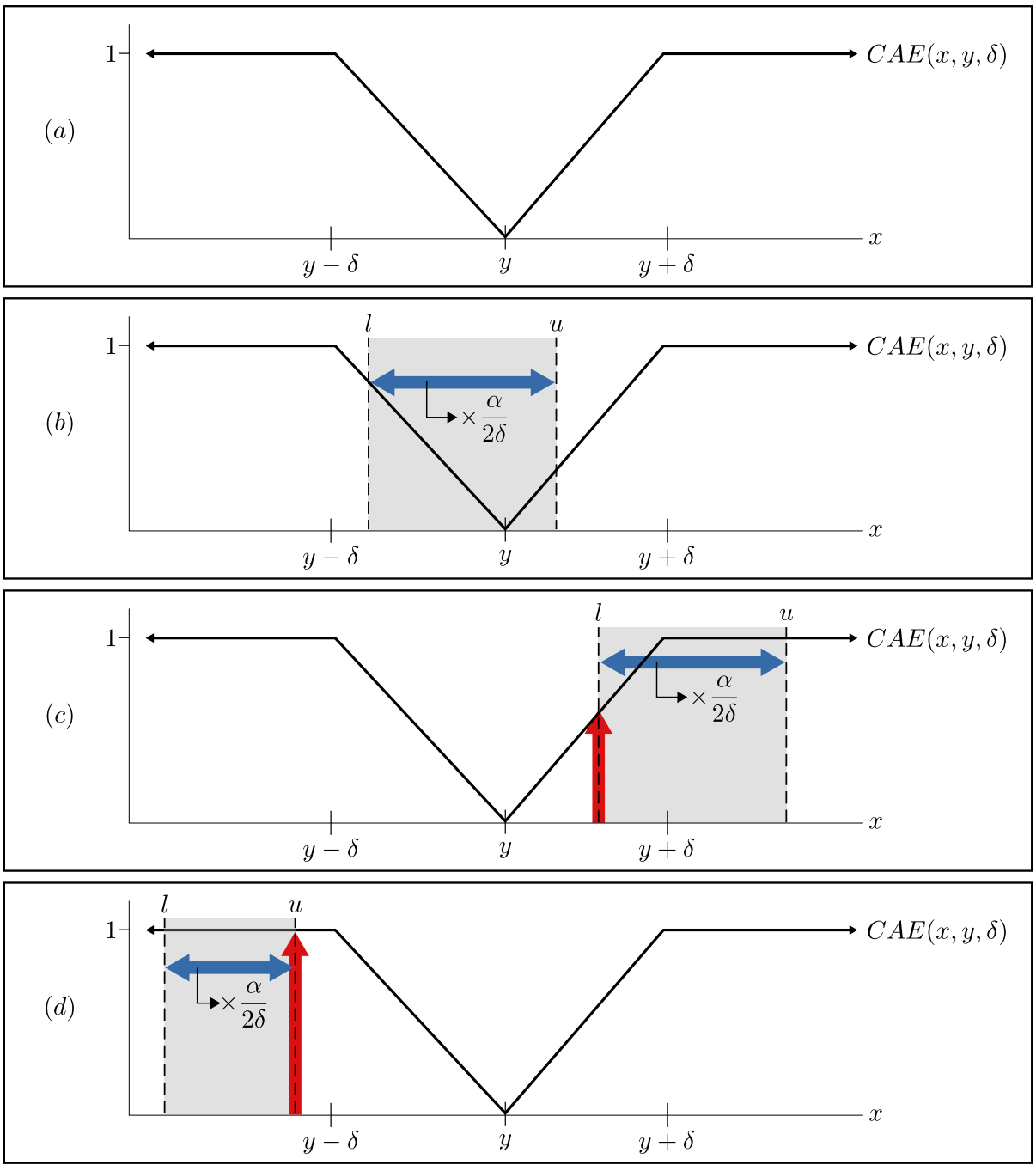
$$CIS_{\alpha}(F, y, \delta) = min\left\{\frac{\alpha}{2\delta}(u-l) + CAE(l, y, \delta)\,\mathbb{1}\{y < l\} + CAE(u, y, \delta)\mathbb{1}\{y > u\},\ 1\right\} \tag{4}$$

Each term in the CIS is analogous to a term in the IS. We begin with the "width" term: $\frac{\alpha}{2\delta}(u-l)$. This term is built upon the logic that because $y-\delta$ to $y+\delta$ represents the upper and lower limits of forecast utility, a prediction interval that spans this entire distance should incur an unweighted penalty of 1. In other words, if a point forecast at or past the "plateau" of the CAE curve incurs a penalty of 1, an unweighted interval forecast that spans this region should get the same score. However, the α-weight is included to distinguish between different prediction intervals. For example, let us compare two intervals that have identical bounds but different α values: 0.05 (95% prediction interval) and 0.9 (10% prediction interval). In this case, the 95% interval should be treated less harshly that the 10% interval, because we expect higher-confidence forecasts to span larger ranges. Next, we examine the "miss" term of the CIS: $(CAE(l, y, \delta))\,\mathbb{1}(y < l) + (CAE(u, y, \delta))\,\mathbb{1}(y > u)$. It is essentially performing the same function as the "miss" term of the IS, but instead of expressing the magnitude of the miss in terms of *distance*, the CIS term is expressed in terms of *utility*. This component of the score can be seen in panels *(c)* and *(d)* of Figure 1 as the vertical arrows. In sum, the CIS is a single-interval analogue of the point-forecast CAE. Regardless of interval width, if a probabilistic forecast is entirely outside the useful region, a value of 1 is returned (panel *(d)* in Figure 1). Additionally, like the IS, the CIS naturally collapses to only its "miss" term when applied to a point forecast.

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

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

Note that we still retain the descriptor "Weighted" in the WCIS title despite the fact that 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 CAE 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.



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

Figure 1: Illustration of different scoring possibilities. Panel *(a)* shows only the Contextual Absolute Error (CAE) point score (Equation 3), with the others displaying different realizations of the Contextual Interval Score (CIS, Equation 4). Blue arrows represent the width penalty term (note that they are scaled by $\frac{\alpha}{2\delta}$). Red arrows indicate the miss term of the CIS. Observe that because the miss term is not scaled, any forecast that entirely misses the $y - \delta$ to $y + \delta$ region, regardless of width, will incur the maximum penalty of 1. For clarity, each of the panels refers to a single-interval evaluation. The full Weighed Contextual Interval Score (WCIS) is composed of an average across multiple α intervals.

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 test cases. Each uses four weeks ahead predictions from the Forecast Hub's ensemble model, and each conveys an essential aspect of the value of the WCIS [3]. The first is a close look at the performance of incident case forecasts for California and Maryland during the Delta variant wave of 2021. It demonstrates the normalization effect of the δ-parameterization and how this contributes to the contextual robustness of the WCIS. Next, we examine hospitalization predictions from May 2021 to May 2022. This period includes both the Delta and Omicron variant waves and allows for a larger exploration of the utility and communicability of the WCIS. Target data for these analyses are sourced directly from the Forecast Hub's repository of truth data [19]. Original sources for these data are the JHU CSSE for cases and mortality, and the US Department of Health & Human Services for hospitalizations [18, 21]. Note that we use a rolling, centered 7-day mean to smooth the target data and minimize the effects of uneven real-time reporting.

3.1 Metric Comparison (First Test Case)

In Section 1.2 we developed intuition regarding the challenges posed by forecast scenarios with high spatial and temporal variability and demonstrated that given these challenges, extant evaluation methods can be inconsistently meaningful. In this section, we examine how the WCIS fares in this context as compared to other evaluation strategies. We use the Forecast Hub ensemble model's four week ahead incident case forecasts for California and Maryland as a demonstrative test case, as shown in Figure 2.

The utility threshold δ chosen for this test case is an expanding mean of historical incidence values weighted to emphasize recent epidemic activity (see Appendix A1 for the detailed formulation). This is intended to reflect the public's intuitive understanding of the evolving state of the pandemic, accounting for humans' natural recency bias while ensuring that the institutional memory of dynamics further in the past is still accounted for. Weighing recent data more heavily makes the WCIS penalty harsher following extended periods of low incident cases and reduces the penalty following times of significantly higher activity. This is intended to reflect complacency in COVID-19 management/resource allocation and human behavior, based on the following characterization. First, extended periods of low epidemic intensity trigger riskier behavior (such as returning to crowded indoor bar/restaurant settings) and policy changes (such as the lifting of mask mandates) that would make early warning of a surge much more beneficial. Second, forecast inaccuracy is less detrimental following periods of very high activity since there are more likely to be higher levels of population immunity and risk mitigation behavior in response. We emphasize again that δ definitions characterize specifically defined representations of utility. The parameterization chosen here is not intended to provide an assessment of forecast quality outside the utility scenario posited by the assumptions given above. However, it demonstrates an important capability of utility threshold selection: δ can be a defined as a dynamic function of data that changes in time and space. Since contextually meaningful forecast utility varies significantly over these same dimensions, a broadly applicable and interpretable score must be able to account for this instability.

First, we compare the WCIS to the WIS. As previously stated, the WIS provides an intuitive evaluation on a *per-state* basis but can fail for the purposes of comparing forecast quality across different scales. The WIS plot in Figure 2 shows this problem, as Maryland's WIS curve (which has significant temporal variation relative to itself) is dwarfed by California's. The WCIS, on the other hand, shows normalized curves that allow for intuitive comparison of the two states' forecast performances. Note that California's WCIS and WIS curves are similarly shaped (Maryland's are as well, but the similarity is harder to see due to scaling). In fact, on a per-state basis, we do not expect or desire the WCIS and WIS to be radically different, except when predictions significantly deviate from the δ-bounded region. (This general correlation, except for more "extreme" errors, can be clearly seen in the comparison scatter plots available in Appendix A2.) Next, we examine the WIS per capita as compared to the WCIS. Again, we observe the curves are relatively similarly shaped. The important differences here are not ones of scale, but of intuition. Per capita evaluation can be a helpful normalization, but it lacks the intuitive meaning of the WCIS. Additionally, population is (relatively) constant in time. While the WCIS normalization parameter δ in this test case is also relatively consistent, thus yielding a similar shape to the per capita curve, δ is not constrained to be so. Thus, it allows for a much more dynamic and meaningful evaluation. Finally, we compare the WCIS to the WIS PE, which is a probabilistic analogue to the percent error (the quotient of the WIS and the target value scaled by 100). Like the other scores, we observe that the WIS PE offers some periods of intuitive, meaningful evaluation. but fails for Maryland in mid-June when target values are low and the score is artificially inflated. Because of its utility-based normalization scheme, the WCIS does not have this problem.

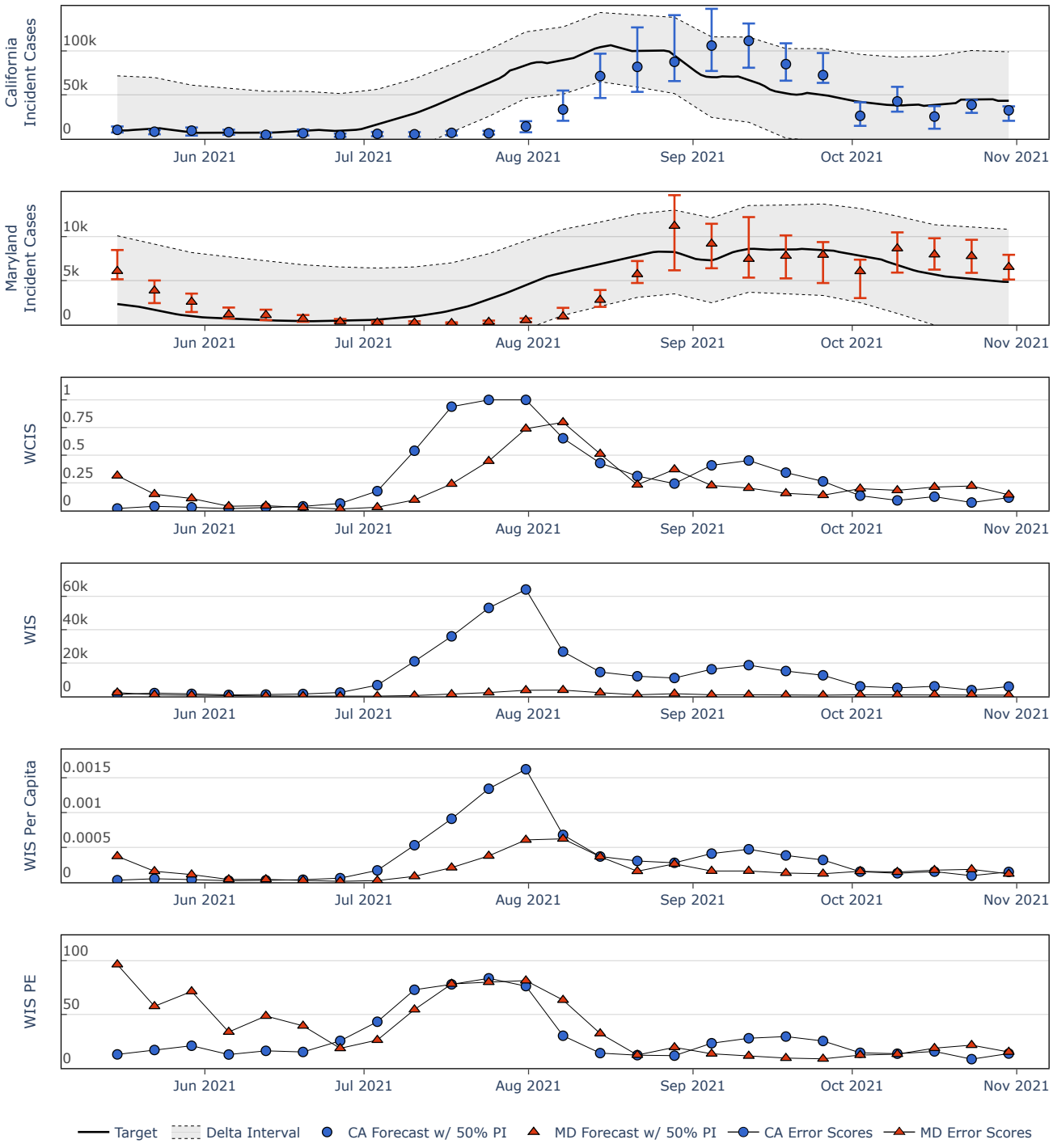


Figure 2: Comparison of WCIS with other metrics for California and Maryland. The top two plots are weekly incident COVID-19 case forecasts for California and Maryland performed by the Forecast Hub's ensemble model, predicting four weeks ahead. The four final plots are the performance for each state according to different evaluation metrics, aligned temporally with the forecasts. (Note that the "WIS PE" plot is the quotient of the WIS and the target value scaled by 100, which can be interpreted heuristically as a probabilistic percent error. Additionally, COVIDhub forecasts (both point and quantile) have a lower bound of zero, so error bars that may appear to extend below the range of the figure are actually constrained to zero. Finally, both the WCIS and the various WIS-derived scores are calculated using *all* submitted prediction intervals (which for case forecasts were 50%, 80%, and 95%), not just the 50% interval shown. Only one interval is shown here for visual clarity.)

3.2 Aggregation and Comparison (Second Test Case)

The prior section examines the in-situ benefits of the WCIS as compared to other probabilistic evaluation metrics. Here, we demonstrate why those benefits are helpful in a higher level, more general, multi-period and multi-location analysis. For this test case we choose a different target variable: state-level weekly COVID-19 hospitalizations. However, in this test case we are not concerned with specific states over a short time period. Instead, we ask whether hospitalization forecasts were helpful for *all* states over the course of a full year. (Note that Forecast Hub hospitalization predictions were performed at daily resolution, but for the sake of visualizing a longer-term analysis we aggregate to and evaluate at weekly totals.)

As ever, using the WCIS requires a specific interpretation of forecast efficacy in the selection of the utility threshold δ. In this case, we choose to assess hospitalization predictions as a function of potential hospital capacity changes. The utility threshold chosen for this is a heuristic for the amount of resource allocation, staffing changes, and other matters that hospitals might practically accomplish in response to an assumed change in pandemic dynamics at the state level. Specifically, δ is the 0.9 quantile of the historical deviations in each state's hospital bed capacity over the prediction horizon of the forecast. Our assumption is that the historical bed capacity deviations are generally indicative of a state's capacity to make changes. Additionally, we assume that it is more difficult to make changes over a shorter timeline. Thus, any deviation over a shorter-term horizon can also occur for longer term horizons, but the reverse is not true. For example, when examining one week ahead predictions, only historical capacity changes over the course of a single week are considered. For four weeks ahead predictions, capacity changes for one, two, three, and four weeks ahead are considered. Finally, the 0.9 quantile is selected as the threshold under the assumption that states are not necessarily able to repeat their larger historical deviations, but can approach them. To be clear, this choice of δ is a simple approximation of state level hospitalization prediction utility intended to enable a demonstration of the WCIS, not to conclusively determine the quality of hospitalization forecasts. Facility or city level hospitalization forecasts, for which the use of much more specific capacity management data might be available, would likely warrant entirely different selections of the utility threshold. However, given the loss of granularity inherent to state level aggregation, we contend that a response predicated on a forecast outside the δ-range as defined here would ask a state to make changes beyond what could be reasonably expected over the forecast's prediction horizon.

How does the WCIS help assess forecast efficacy in this multi-region, yearlong analysis? Since it was designed primarily as a way to meaningfully aggregate and compare forecasts in disparate contexts, the results for each state over the entire time period of interest can be easily and intuitively displayed. One way of doing so is demonstrated in Figure 3, in which we can easily observe several important aspects of hospitalization forecasting performance. For example, during times of rapidly changing pandemic activity (surges and declines), the utility of forecasts 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 spatially averaged (lower) array of the figure. In contrast, if we examine the temporally averaged (right-side) array, we observe that the there is less variability in average quality 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 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 a substantive and easily interpretable performance evaluation.

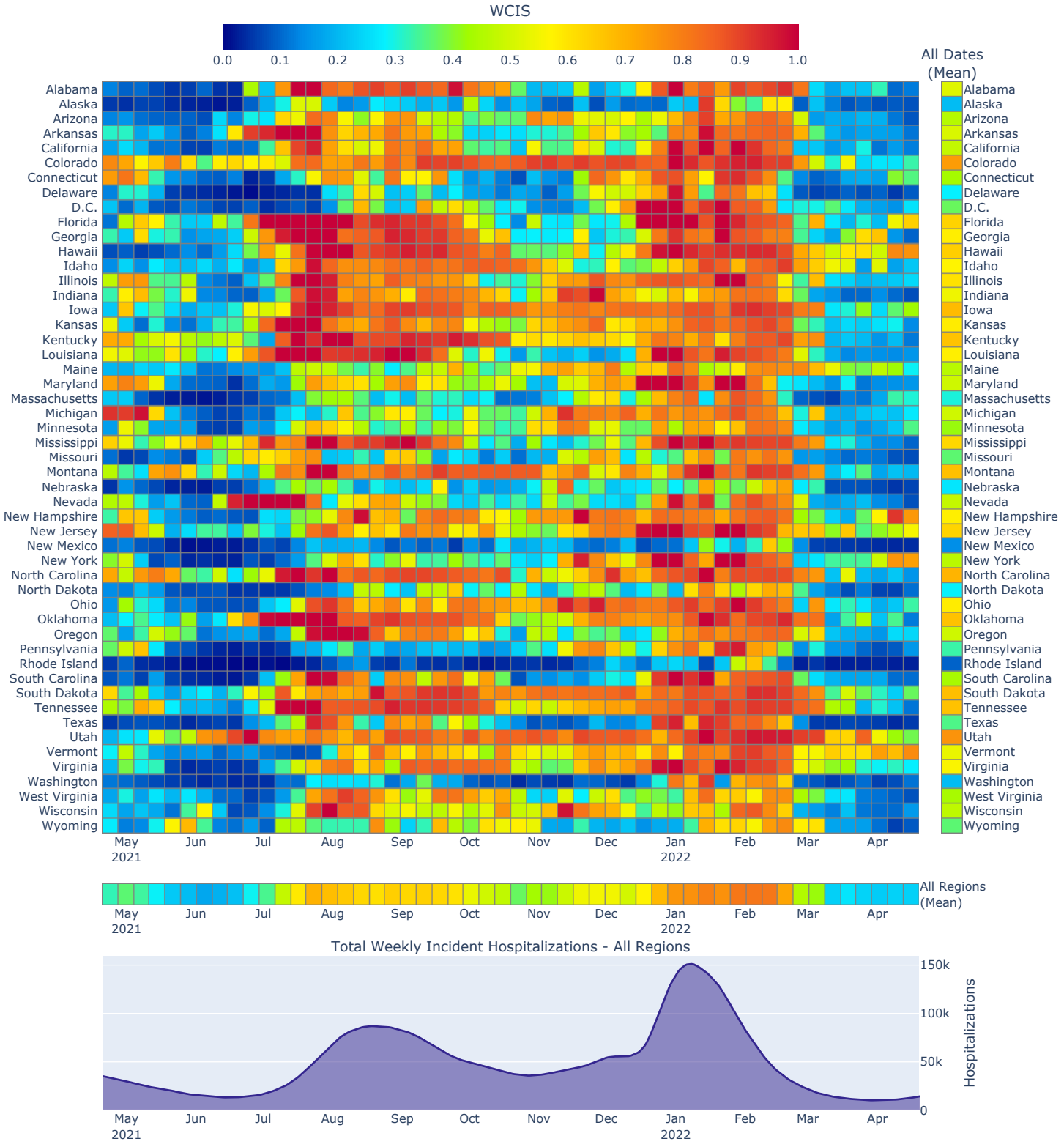


Figure 3: Heatmap of the WCIS for 4 week ahead hospitalization forecasts, performed by the Forecast Hub's ensemble model The central and largest grid shows the most granular results: region- and time-specific performance. On the right and lower sides of the grid are average performances over time and space, respectively. The shaded line plot at the bottom of the figure is the target 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 Appendices A3 and B3 for heatmaps of other target variables and for hospitalizations over differing prediction horizons).

4 Discussion

4.1 Contextually Relevant Retrospective Evaluation

What is the purpose of using a framework like the Forecast Hub to provide real-time epidemiology predictions? We contend that at its most fundamental, the goal must be to add utility. A useful prediction provides meaningful and actionable information for someone making a decision subject to uncertain future pandemic outcomes. Determining whether or not forecasts accomplish this necessitates an explicit definition of utility, which brings up an important philosophical difference between the WCIS and other techniques. The WCIS formulation, centered around a user-defined utility threshold δ, is based on our assertion that there will *never* be a one-size-fits-all solution for assessing and aggregating short-term forecast quality. One must always consider prediction context lest standard metrics tell a misleading story. Additionally, different decision-making mechanisms 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, forecasts use is essential to consider. The WCIS ensures this by requiring a direct definition of the utility threshold.

In this light, we summarize the contribution of the WCIS. In brief, it is a probabilistic forecast evaluation metric that is intuitively interpretable (like the WIS), easily comparable (like other normalized metrics), and robust to numerical problems (unlike other normalized metrics). Real-world use cases for epidemic predictions 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 forecast evaluation 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, aggregated, and compared in a technically meaningful and intuitively understandable way.

4.2 On Propriety and the Application of the WCIS

In the context of a collaborative and influential effort like the COVID-19 Forecast Hub, predictions are constantly evaluated, ranked, and potentially published widely. This naturally results in pressure for participants to maximize the performance of their submissions. Using a proper score function for forecast evaluation helps to ensure honest reporting. Briefly, if a participant attempts to game the system by submitting forecasts calculated to maximize their performance (instead of just reporting what they honestly believe is most likely to occur), a proper score makes sure that *in expectation*, no submission can perform better than their true belief [20]. Using proper scores makes sure that forecasts disseminated by collaborative efforts like this represent the best-faith projections of modelers. The WIS is a proper score, and is the primary metric used by the Forecast Hub for evaluating probabilistic forecasts [17].

The WCIS is not a statistically proper interval score. However, we propose that a score with the desired features of the WCIS is inherently improper. The foundation of the WCIS is the notion of a specific and *constrained* region around the target value wherein predictions are applicable, represented by the V-shaped CAE function. This means that from a gaming/error minimization perspective, the WCIS could encourage probabilistic forecasts that are affected by the size of the δ-region (see Appendix C1 for an empirical demonstration of this) [22]. Similar to prior forecasting efforts when improper metrics were used, propriety is sacrificed in exchange for other, desirable properties of the score [23, 24, 25]. Additionally, ongoing work by Bosse et al. indicates that applying monotonic transformations like the natural logarithm to target data can help to alleviate the domination of higher-activity forecasting scenarios for model comparison and aggregation while retaining propriety [26]. Therefore, we expressly do not recommend the WCIS for real-time forecast ranking or ensemble creation, because such use introduces options for forecast hedging and could encourage dishonest reporting. Instead, we propose that the WCIS is best suited for retrospective use to answer specific questions about forecast utility.

4.3 Conclusion

The WCIS builds on the strengths of the Weighted Interval Score while adding advantageous capabilities for retrospective evaluation of pandemic forecasts. The central tenet of the WCIS is the δ-parameterization, which impels 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, compare, and aggregate 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.

Our goal is to enable and encourage honest and contextually specific discourse about the utility of short-term epidemic predictions. The design of the WCIS reflects this desire. 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.

References

[1] Serge P. J. M. Horbach. Pandemic publishing: Medical journals strongly speed up their publication process for COVID-19. *Quantitative Science Studies*, 1(3):1056–1067, August 2020.

[2] Nicholas Fraser, Liam Brierley, Gautam Dey, Jessica K. Polka, Máté Pálfy, Federico Nanni, and Jonathon Alexis Coates. The evolving role of preprints in the dissemination of COVID-19 research and their impact on the science communication landscape. *PLOS Biology*, 19(4):e3000959, April 2021. Publisher: Public Library of Science.

[3] Estee Y. Cramer, Yuxin Huang, Yijin Wang, Evan L. Ray, Matthew Cornell, Johannes Bracher, Andrea Brennen, Alvaro J. Castro Rivadeneira, Aaron Gerding, Katie House, Dasuni Jayawardena, Abdul Hannan Kanji, Ayush Khandelwal, Khoa Le, Vidhi Mody, Vrushti Mody, Jarad Niemi, Ariane Stark, Apurv Shah, Nutcha Wattanchit, Martha W. Zorn, and Nicholas G. Reich. The United States COVID-19 Forecast Hub dataset. *Scientific Data*, 9(1):462, August 2022. Number: 1 Publisher: Nature Publishing Group.

[4] Craig J. McGowan, Matthew Biggerstaff, Michael Johansson, Karyn M. Apfeldorf, Michal Ben-Nun, Logan Brooks, Matteo Convertino, Madhav Erraguntla, David C. Farrow, John Freeze, Saurav Ghosh, Sangwon Hyun, Sasikiran Kandula, Joceline Lega, Yang Liu, Nicholas Michaud, Haruka Morita, Jarad Niemi, Naren Ramakrishnan, Evan L. Ray, Nicholas G. Reich, Pete Riley, Jeffrey Shaman, Ryan Tibshirani, Alessandro Vespignani, Qian Zhang, and Carrie Reed. Collaborative efforts to forecast seasonal influenza in the United States, 2015–2016. *Scientific Reports*, 9(1):683, January 2019. Number: 1 Publisher: Nature Publishing Group.

[5] Michael A. Johansson, Karyn M. Apfeldorf, Scott Dobson, Jason Devita, Anna L. Buczak, Benjamin Baugher, Linda J. Moniz, Thomas Bagley, Steven M. Babin, Erhan Guven, Teresa K. Yamana, Jeffrey Shaman, Terry Moschou, Nick Lothian, Aaron Lane, Grant Osborne, Gao Jiang, Logan C. Brooks, David C. Farrow, Sangwon Hyun, Ryan J. Tibshirani, Roni Rosenfeld, Justin Lessler, Nicholas G. Reich, Derek A. T. Cummings, Stephen A. Lauer, Sean M. Moore, Hannah E. Clapham, Rachel Lowe, Trevor C. Bailey, Markel García-Díez, Marilia Sá Carvalho, Xavier Rodó, Tridip Sardar, Richard Paul, Evan L. Ray, Krzysztof Sakrejda, Alexandria C. Brown, Xi Meng, Osonde Osoba, Raffaele Vardavas, David Manheim, Melinda Moore, Dhananjai M. Rao, Travis C. Porco, Sarah Ackley, Fengchen Liu, Lee Worden, Matteo Convertino, Yang Liu, Abraham Reddy, Eloy Ortiz, Jorge Rivero, Humberto Brito, Alicia Juarrero, Leah R. Johnson, Robert B. Gramacy, Jeremy M. Cohen, Erin A. Mordecai, Courtney C. Murdock, Jason R. Rohr, Sadie J. Ryan, Anna M. Stewart-Ibarra, Daniel P. Weikel, Antarpreet Jutla, Rakibul Khan, Marissa Poultney, Rita R. Colwell, Brenda Rivera-García, Christopher M. Barker, Jesse E. Bell, Matthew Biggerstaff, David Swerdlow, Luis Mier-y Teran-Romero, Brett M. Forshey, Juli Trtanj, Jason Asher, Matt Clay, Harold S. Margolis, Andrew M. Hebbeler, Dylan George, and Jean-Paul Chretien. An open challenge to advance probabilistic forecasting for dengue epidemics. *Proceedings of the National Academy of Sciences*, 116(48):24268–24274, November 2019. Publisher: Proceedings of the National Academy of Sciences.

[6] Cécile Viboud, Kaiyuan Sun, Robert Gaffey, Marco Ajelli, Laura Fumanelli, Stefano Merler, Qian Zhang, Gerardo Chowell, Lone Simonsen, and Alessandro Vespignani. The RAPIDD ebola forecasting challenge: Synthesis and lessons learnt. *Epidemics*, 22:13–21, March 2018.

[7] Nicholas G. Reich, Justin Lessler, Sebastian Funk, Cecile Viboud, Alessandro Vespignani, Ryan J. Tibshirani, Katriona Shea, Melanie Schienle, Michael C. Runge, Roni Rosenfeld, Evan L. Ray, Rene Niehus, Helen C. Johnson, Michael A. Johansson, Harry Hochheiser, Lauren Gardner, Johannes Bracher, Rebecca K. Borchering, and Matthew Biggerstaff. Collaborative Hubs: Making the Most of Predictive Epidemic Modeling. *American Journal of Public Health*, 112(6):839–842, June 2022. Publisher: American Public Health Association.

[8] Evan L. Ray, Logan C. Brooks, Jacob Bien, Matthew Biggerstaff, Nikos I. Bosse, Johannes Bracher, Estee Y. Cramer, Sebastian Funk, Aaron Gerding, Michael A. Johansson, Aaron Rumack, Yijin Wang, Martha Zorn,

Ryan J. Tibshirani, and Nicholas G. Reich. Comparing trained and untrained probabilistic ensemble forecasts of COVID-19 cases and deaths in the United States. *International Journal of Forecasting*, July 2022.

[9] Nicholas G. Reich, Craig J. McGowan, Teresa K. Yamana, Abhinav Tushar, Evan L. Ray, Dave Osthus, Sasikiran Kandula, Logan C. Brooks, Willow Crawford-Crudell, Graham Casey Gibson, Evan Moore, Rebecca Silva, Matthew Biggerstaff, Michael A. Johansson, Roni Rosenfeld, and Jeffrey Shaman. Accuracy of real-time multimodel ensemble forecasts for seasonal influenza in the U.S. *PLOS Computational Biology*, 15(11):e1007486, November 2019. Publisher: Public Library of Science.

[10] Nicholas G. Reich and Evan L. Ray. Collaborative modeling key to improving outbreak response. *Proceedings of the National Academy of Sciences*, 119(14):e2200703119, April 2022. Publisher: Proceedings of the National Academy of Sciences.

[11] Colin Doms, Sarah C. Kramer, and Jeffrey Shaman. Assessing the Use of Influenza Forecasts and Epidemiological Modeling in Public Health Decision Making in the United States. *Scientific Reports*, 8(1):12406, August 2018. Number: 1 Publisher: Nature Publishing Group.

[12] Nicholas G. Reich, Yijin Wang, Meagan Burns, Rosa Ergas, Estee Y. Cramer, and Evan L. Ray. Assessing the utility of COVID-19 case reports as a leading indicator for hospitalization forecasting in the United States, March 2023. Pages: 2023.03.08.23286582.

[13] Kristen Nixon, Sonia Jindal, Felix Parker, Maximilian Marshall, Nicholas G. Reich, Kimia Ghobadi, Elizabeth C. Lee, Shaun Truelove, and Lauren Gardner. Real-time COVID-19 forecasting: challenges and opportunities of model performance and translation. *The Lancet Digital Health*, 4(10):e699–e701, October 2022. Publisher: Elsevier.

[14] Chelsea S. Lutz, Mimi P. Huynh, Monica Schroeder, Sophia Anyatonwu, F. Scott Dahlgren, Gregory Danyluk, Danielle Fernandez, Sharon K. Greene, Nodar Kipshidze, Leann Liu, Osaro Mgbere, Lisa A. McHugh, Jennifer F. Myers, Alan Siniscalchi, Amy D. Sullivan, Nicole West, Michael A. Johansson, and Matthew Biggerstaff. Applying infectious disease forecasting to public health: a path forward using influenza forecasting examples. *BMC Public Health*, 19(1):1659, December 2019.

[15] Johannes Bracher, Evan L. Ray, Tilmann Gneiting, and Nicholas G. Reich. Evaluating epidemic forecasts in an interval format. *PLOS Computational Biology*, 17(2):e1008618, February 2021. Publisher: Public Library of Science.

[16] Kristen Nixon, Sonia Jindal, Felix Parker, Nicholas G Reich, Kimia Ghobadi, Elizabeth C Lee, Shaun Truelove, and Lauren Gardner. An evaluation of prospective COVID-19 modelling studies in the USA: from data to science translation. *The Lancet Digital Health*, 4(10):e738–e747, October 2022.

[17] Estee Y. Cramer, Evan L. Ray, Velma K. Lopez, Johannes Bracher, Andrea Brennen, Alvaro J. Castro Rivadeneira, Aaron Gerding, Tilmann Gneiting, Katie H. House, Yuxin Huang, Dasuni Jayawardena, Abdul H. Kanji, Ayush Khandelwal, Khoa Le, Anja Mühlemann, Jarad Niemi, Apurv Shah, Ariane Stark, Yijin Wang, Nutcha Wattanachit, Martha W. Zorn, Youyang Gu, Sansiddh Jain, Nayana Bannur, Ayush Deva, Mihir Kulkarni, Srujana Merugu, Alpan Raval, Siddhant Shingi, Avtansh Tiwari, Jerome White, Neil F. Abernethy, Spencer Woody, Maytal Dahan, Spencer Fox, Kelly Gaither, Michael Lachmann, Lauren Ancel Meyers, James G. Scott, Mauricio Tec, Ajitesh Srivastava, Glover E. George, Jeffrey C. Cegan, Ian D. Dettwiller, William P. England, Matthew W. Farthing, Robert H. Hunter, Brandon Lafferty, Igor Linkov, Michael L. Mayo, Matthew D. Parno, Michael A. Rowland, Benjamin D. Trump, Yanli Zhang-James, Samuel Chen, Stephen V. Faraone, Jonathan Hess, Christopher P. Morley, Asif Salekin, Dongliang Wang, Sabrina M. Corsetti, Thomas M. Baer, Marisa C. Eisenberg, Karl Falb, Yitao Huang, Emily T. Martin, Ella McCauley, Robert L. Myers, Tom Schwarz, Daniel Sheldon, Graham Casey Gibson, Rose Yu, Liyao Gao, Yian Ma, Dongxia Wu, Xifeng Yan, Xiaoyong Jin, Yu-Xiang Wang, YangQuan Chen, Lihong Guo, Yanting Zhao, Quanquan Gu, Jinghui Chen, Lingxiao Wang, Pan Xu, Weitong Zhang, Difan Zou, Hannah Biegel, Joceline Lega, Steve McConnell, V. P. Nagraj, Stephanie L. Guertin, Christopher Hulme-Lowe, Stephen D. Turner, Yunfeng Shi, Xuegang Ban, Robert Walraven, Qi-Jun Hong, Stanley Kong, Axel van de Walle, James A. Turtle, Michal Ben-Nun, Steven Riley, Pete Riley, Ugur Koyluoglu, David DesRoches, Pedro Forli, Bruce Hamory, Christina Kyriakides, Helen Leis, John Milliken, Michael Moloney, James Morgan, Ninad Nirgudkar, Gokce Ozcan, Noah Piwonka, Matt Ravi, Chris Schrader, Elizabeth Shakhnovich, Daniel Siegel, Ryan Spatz, Chris Stiefeling, Barrie Wilkinson, Alexander Wong, Sean Cavany, Guido España, Sean Moore, Rachel Oidtman, Alex Perkins, David Kraus, Andrea Kraus, Zhifeng Gao, Jiang Bian, Wei Cao, Juan Lavista Ferres, Chaozhuo Li, Tie-Yan Liu, Xing Xie, Shun Zhang, Shun Zheng, Alessandro Vespignani, Matteo Chinazzi, Jessica T. Davis, Kunpeng Mu, Ana Pastore y Piontti, Xinyue Xiong, Andrew Zheng, Jackie Baek, Vivek Farias, Andreea Georgescu, Retsef Levi, Deeksha Sinha, Joshua Wilde, Georgia Perakis, Mohammed Amine Bennouna, David Nze-Ndong, Divya Singhvi, Ioannis Spantidakis, Leann Thayaparan, Asterios Tsiourvas, Arnab Sarker, Ali Jadbabaie, Devavrat Shah, Nicolas Della Penna, Leo A. Celi, Saketh Sundar, Russ Wolfinger, Dave Osthus, Lauren

Castro, Geoffrey Fairchild, Isaac Michaud, Dean Karlen, Matt Kinsey, Luke C. Mullany, Kaitlin Rainwater-Lovett, Lauren Shin, Katharine Tallaksen, Shelby Wilson, Elizabeth C. Lee, Juan Dent, Kyra H. Grantz, Alison L. Hill, Joshua Kaminsky, Kathryn Kaminsky, Lindsay T. Keegan, Stephen A. Lauer, Joseph C. Lemaitre, Justin Lessler, Hannah R. Meredith, Javier Perez-Saez, Sam Shah, Claire P. Smith, Shaun A. Truelove, Josh Wills, Maximilian Marshall, Lauren Gardner, Kristen Nixon, John C. Burant, Lily Wang, Lei Gao, Zhiling Gu, Myungjin Kim, Xinyi Li, Guannan Wang, Yueying Wang, Shan Yu, Robert C. Reiner, Ryan Barber, Emmanuela Gakidou, Simon I. Hay, Steve Lim, Chris Murray, David Pigott, Heidi L. Gurung, Prasith Baccam, Steven A. Stage, Bradley T. Suchoski, B. Aditya Prakash, Bijaya Adhikari, Jiaming Cui, Alexander Rodríguez, Anika Tabassum, Jiajia Xie, Pinar Keskinocak, John Asplund, Arden Baxter, Buse Eylul Oruc, Nicoleta Serban, Sercan O. Arik, Mike Dusenberry, Arkady Epshteyn, Elli Kanal, Long T. Le, Chun-Liang Li, Tomas Pfister, Dario Sava, Rajarishi Sinha, Thomas Tsai, Nate Yoder, Jinsung Yoon, Leyou Zhang, Sam Abbott, Nikos I. Bosse, Sebastian Funk, Joel Hellewell, Sophie R. Meakin, Katharine Sherratt, Mingyuan Zhou, Rahi Kalantari, Teresa K. Yamana, Sen Pei, Jeffrey Shaman, Michael L. Li, Dimitris Bertsimas, Omar Skali Lami, Saksham Soni, Hamza Tazi Bouardi, Turgay Ayer, Madeline Adee, Jagpreet Chhatwal, Ozden O. Dalgic, Mary A. Ladd, Benjamin P. Linas, Peter Mueller, Jade Xiao, Yuanjia Wang, Qinxia Wang, Shanghong Xie, Donglin Zeng, Alden Green, Jacob Bien, Logan Brooks, Addison J. Hu, Maria Jahja, Daniel McDonald, Balasubramanian Narasimhan, Collin Politsch, Samyak Rajanala, Aaron Rumack, Noah Simon, Ryan J. Tibshirani, Rob Tibshirani, Valerie Ventura, Larry Wasserman, Eamon B. O'Dea, John M. Drake, Robert Pagano, Quoc T. Tran, Lam Si Tung Ho, Huong Huynh, Jo W. Walker, Rachel B. Slayton, Michael A. Johansson, Matthew Biggerstaff, and Nicholas G. Reich. Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the United States. *Proceedings of the National Academy of Sciences*, 119(15):e2113561119, April 2022. Publisher: Proceedings of the National Academy of Sciences.

[18] Ensheng Dong, Hongru Du, and Lauren Gardner. An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases*, 20(5):533–534, May 2020. Publisher: Elsevier.

[19] Estee Cramer, Serena Yijin Wang, Nicholas G. Reich, Abdul Hannan, Jarad Niemi, Evan Ray, Katie House, Yuxin David Huang, Ariane Stark, Robert Walraven, aniruddhadiga, Shanghong Xie, Dean Karlen, Michael Lingzhi Li, rjpagano, Youyang Gu, zyt9lsb, Aaron Gerding, Xinyue X, Lauren Castro, mzorn-58, Frost Tianjian Xu, stevemcconnell, Graham Gibson, leyouz, Matt Le, Steve Horstman, Hannah Biegel, and EpiDeep. reichlab/covid19-forecast-hub: release for Zenodo 20220227, February 2022.

[20] Tilmann Gneiting and Adrian E Raftery. Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, 102(477):359–378, March 2007.

[21] HealthData.gov. COVID-19 Reported Patient Impact and Hospital Capacity by Facility, December 2020.

[22] Tilmann Gneiting and Roopesh Ranjan. Comparing Density Forecasts Using Threshold- and Quantile-Weighted Scoring Rules. *Journal of Business & Economic Statistics*, 29(3):411–422, July 2011. Publisher: Taylor & Francis.

[23] Nicholas G. Reich, Logan C. Brooks, Spencer J. Fox, Sasikiran Kandula, Craig J. McGowan, Evan Moore, Dave Osthus, Evan L. Ray, Abhinav Tushar, Teresa K. Yamana, Matthew Biggerstaff, Michael A. Johansson, Roni Rosenfeld, and Jeffrey Shaman. A collaborative multiyear, multimodel assessment of seasonal influenza forecasting in the United States. *Proceedings of the National Academy of Sciences*, 116(8):3146–3154, February 2019. Publisher: Proceedings of the National Academy of Sciences.

[24] Johannes Bracher. On the multibin logarithmic score used in the FluSight competitions. *Proceedings of the National Academy of Sciences*, 116(42):20809–20810, October 2019. Publisher: Proceedings of the National Academy of Sciences.

[25] Nicholas G. Reich, Dave Osthus, Evan L. Ray, Teresa K. Yamana, Matthew Biggerstaff, Michael A. Johansson, Roni Rosenfeld, and Jeffrey Shaman. Reply to Bracher: Scoring probabilistic forecasts to maximize public health interpretability. *Proceedings of the National Academy of Sciences*, 116(42):20811–20812, October 2019. Publisher: Proceedings of the National Academy of Sciences.

[26] Nikos I. Bosse, Sam Abbott, Anne Cori, Edwin van Leeuwen, Johannes Bracher, and Sebastian Funk. Transformation of forecasts for evaluating predictive performance in an epidemiological context, January 2023. ISSN: 2328-4722 Pages: 2023.01.23.23284722.