

Scoring epidemiological forecasts on transformed scales

Nikos I. Bosse^{1,2,3,*}, Sam Abbott^{1,2}, Anne Cori⁴,
Edwin van Leeuwen^{1,5}, Johannes Bracher^{6,7,†}, Sebastian Funk^{1,2,3,†}

July 28, 2023

Abstract

Forecast evaluation is essential for the development of predictive epidemic models and can inform their use for public health decision-making. Common scores to evaluate epidemiological forecasts are the Continuous Ranked Probability Score (CRPS) and the Weighted Interval Score (WIS), which can be seen as measures of the absolute distance between the forecast distribution and the observation. However, applying these scores directly to predicted and observed incidence counts may not be the most appropriate due to the exponential nature of epidemic processes and the varying magnitudes of observed values across space and time. In this paper, we argue that transforming counts before applying scores such as the CRPS or WIS can effectively mitigate these difficulties and yield epidemiologically meaningful and easily interpretable results. Using the CRPS on log-transformed values as an example, we list three attractive properties: Firstly, it can be interpreted as a probabilistic version of a relative error. Secondly, it reflects how well models predicted the time-varying epidemic growth rate. And lastly, using arguments on variance-stabilizing transformations, it can be shown that under the assumption of a quadratic mean-variance relationship, the logarithmic transformation leads to expected CRPS values which are independent of the order of magnitude of the predicted quantity. Applying a transformation of $\log(x + 1)$ to data and forecasts from the European COVID-19 Forecast Hub, we find that it changes model rankings regardless of stratification by forecast date, location or target types. Situations in which models missed the beginning of upward swings are more strongly emphasised while failing to predict a downturn following a peak is less severely penalised when scoring transformed forecasts as opposed to untransformed ones. We conclude that appropriate transformations, of which the natural logarithm is only one particularly attractive option, should be considered when assessing the performance of different models in the context of infectious disease incidence.

* Correspondence to nikos.bosse@lshtm.ac.uk, † Contributed equally

¹Department of Infectious Disease Epidemiology, London School of Hygiene & Tropical Medicine, London, United Kingdom

²Centre for the Mathematical Modelling of Infectious Diseases, London, United Kingdom

³NIHR Health Protection Research Unit in Modelling & Health Economics

⁴MRC Centre for Outbreak Analysis and Modelling, Department of Infectious Disease Epidemiology, School of Public Health, Imperial College London, London, United Kingdom

⁵Modelling & Economics Unit and NIHR Health Protection Research Unit in Modelling & Health Economics, UK Health Security Agency, London, United Kingdom

⁶Chair of Statistical Methods and Econometrics, Karlsruhe Institute of Technology, Karlsruhe, Germany

⁷Computational Statistics Group, Heidelberg Institute for Theoretical Studies, Heidelberg, Germany

1 Introduction

Probabilistic forecasts (Held et al., 2017) play an important role in decision-making in epidemiology and public health (Reich et al., 2022), as well as other areas as diverse as economics (Timmermann, 2018) or meteorology (Gneiting and Raftery, 2005). Forecasts based on epidemiological modelling in particular have received widespread attention during the COVID-19 pandemic. Evaluations of forecasts can provide feedback for researchers to improve their models and train ensembles. They moreover help decision-makers distinguish good from bad predictions and choose forecasters and models that are best suited to inform future decisions.

Probabilistic forecasts are usually evaluated using so-called (strictly) proper scoring rules (Gneiting and Raftery, 2007), which return a numerical score as a function of the forecast and the observed data. Proper scoring rules are constructed such that they encourage honest forecasting and cannot be ‘gamed’ or ‘cheated’. Assuming that the forecaster’s actual best judgement corresponds to a predictive distribution F , a proper score is constructed such that if F was the data-generating process, no other distribution G would yield a better expected score. A scoring rule is called *strictly* proper if there is no other distribution that under F achieves the *same* expected score as F , meaning that any deviation from F leads to a worsening of expected scores. Forecasters (anyone or anything that issues a forecast) are thus incentivised to report their true belief F about the future. Common proper scoring rules are the logarithmic or log score (Good, 1952) and the continuous ranked probability score (CRPS, Gneiting and Raftery, 2007). The log score is the predictive log density or probability mass evaluated at the observed value. It is supported by the likelihood principle (Winkler, 1996) and has many desirable theoretical properties; however, the particularly severe penalties it assigns to occasional misguided forecasts make it little robust (Bracher et al., 2021a). Moreover, it is not easily applied to forecasts reported as samples or quantiles, as used in many recent disease forecasting efforts. It is nonetheless occasionally used in epidemiology (see e.g., Held et al. 2017; Johansson et al. 2019), but in recent years the CRPS and the weighted interval score (WIS, Bracher et al., 2021a) have become increasingly popular.

The CRPS measures the distance of the predictive distribution to the observed data as

$$\text{CRPS}(F, y) = \int_{-\infty}^{\infty} (F(x) - \mathbf{1}(x \geq y))^2 dx, \quad (1)$$

where y is the true observed value, F is the cumulative distribution function (CDF) of the predictive distribution, and $\mathbf{1}(\cdot)$ is the indicator function. The CRPS can be understood as a generalisation of the absolute error to predictive distributions, and interpreted on the natural scale of the data. The WIS is an approximation of the CRPS for predictive distributions represented by a set of predictive quantiles and is currently used to assess forecasts in the so-called COVID-19 Forecast Hubs in the US (Cramer et al., 2020, 2021), Europe (Sherratt et al., 2022) and Germany and Poland (Bracher et al., 2021b, 2022), as well as the US *FluSight* project on influenza forecasting (CDC, 2022). The WIS is defined as

$$\text{WIS}(F, y) = \frac{1}{K} \times \sum_{k=1}^K 2 \times [\mathbf{1}(y \leq q_{\tau_k}) - \tau_k] \times (q_{\tau_k} - y), \quad (2)$$

where q_{τ} is the τ quantile of the forecast F , y is the observed outcome and K is the number of (roughly equally spaced) predictive quantiles provided. The WIS can be decomposed into three components, dispersion, underprediction and overprediction, which reflect the spread of the forecast and whether it was centred above or below the observed value. We show an alternative definition based on central prediction intervals in Supplement A.1 which illustrates this decomposition.

The notion of absolute distance encoded by the CRPS and WIS provides a straightforward interpretation, but may not always be the most useful perspective in the context of infectious disease spread. Especially in their early phase, outbreaks are best conceived as exponential processes, characterized by potentially time varying reproduction numbers R_t (Gostic et al., 2020) or epidemic growth rates r_t (Dushoff and Park, 2021). If the true modelling task revolves around estimating and forecasting these quantities, then evaluating forecasts

71 based on the absolute distance between forecasted and observed incidence values penalises underprediction
72 (of the reproduction number or growth rate) less than overprediction by the same amount. For illustration,
73 consider an incidence forecast issued at time 0 and referring to time t that misses the correct average growth
74 rate \bar{r}_t by either $-\epsilon$ or $+\epsilon$. Then the ratio of the resulting absolute errors on the scale of observed incidences
75 y_t is

$$76 \quad \frac{|y_0 \exp[(\bar{r}_t - \epsilon) \times t] - y_0 \exp(\bar{r}_t t)|}{|y_0 \exp[(\bar{r}_t + \epsilon) \times t] - y_0 \exp(\bar{r}_t t)|} = \exp(-\epsilon t) < 1. \quad (3)$$

77 If one is to measure the ability to forecast the underlying infection dynamics, it may thus be more desirable
78 to evaluate errors on the scale of the growth rate directly.

79 Another argument against using notions of absolute distance between predicted and observed incidence
80 values is that forecast consumers may find errors on a relative scale easier to interpret and more useful in
81 order to track predictive performance across targets of different orders of magnitude. Bolin and Wallin (2023)
82 have proposed the scaled CRPS (SCRPS) which is locally scale invariant; however, it does not correspond
83 to a relative error measure and lacks a straightforward interpretation as available for the CRPS.

84 Lastly, it may be considered desirable to give all forecast targets similar weight in an overall performance
85 evaluation. As the CRPS typically scales with the order of magnitude of the quantity to be predicted, this
86 is not the case for the CRPS, which will typically assign higher scores to forecast targets with high expected
87 values (e.g., in large locations or around the peak of an epidemic). Bracher et al. (2021a) have argued
88 that this is a desirable feature, directing attention to situations of particular public health relevance. An
89 evaluation based on absolute errors, however, will assign little weight to other potentially important aspects,
90 such as the ability to correctly predict future upswings while observed numbers are still low.

91 In many fields, it is common practice to forecast transformed quantities (see e.g. Taylor (1999) in finance,
92 Mayr and Ulbricht (2015) in macroeconomics, Löwe et al. (2014) in hydrology or Fuglstad et al. (2015) in
93 meteorology). While the goal of the transformations is often to improve the accuracy of the predictions,
94 they can also be used to enhance and complement the evaluation process. In this paper, we argue that the
95 aforementioned issues with evaluating epidemic forecasts based on measures of absolute error on the natural
96 scale can be addressed by transforming the forecasts and observations prior to scoring using some strictly
97 monotonic transformation. Strictly monotonic transformations can shift the focus of the evaluation in a
98 way that may be more appropriate for epidemiological forecasts, while guaranteeing that the score remains
99 proper. Many different transformations may be appropriate and useful, depending on the exact context,
100 the desired focus of the evaluation, and specific aspects forecast consumers care most about (see a broader
101 discussion in Section 4).

102 For conceptual clarity and to allow for a more in-depth discussion, we focus mostly on the natural logarithm
103 as a particularly attractive transformation in the context of epidemic phenomena. We refer to this trans-
104 formation as 'log-transformation' and to scores that have been computed from log-transformed forecasts
105 and observations as scores 'on the log scale' (as opposed to scores 'on the natural scale', which involve no
106 transformation). In the theoretical discussion in Section 2, 'log-transformation' and 'log scale' generally refer
107 to a transformation of $\log_e(x)$. For practical applications (Section 3) we also use these terms to describe a
108 transformation of $\log_e(x + a)$ with a small $a > 0$ in order to keep the terminology and notation simple. For
109 a prediction target with strictly positive support, the CRPS after applying a log-transformation is given by

$$110 \quad \text{CRPS}(F_{\log}, \log y) = \int_{-\infty}^{\infty} (F_{\log}(x) - \mathbf{1}(x \geq \log y))^2 dx. \quad (4)$$

111 Here, y is again the observed outcome and F_{\log} is the predictive CDF of the log-transformed outcome, i.e.,

$$112 \quad F_{\log}(x) = F(\exp(x)), \quad (5)$$

113 with F the CDF on the original scale. Instead of a score representing the magnitude of absolute errors,
114 applying a log-transformation prior to the CRPS yields a score which a) measures relative error (see Section
115 2.1), b) provides a measure for how well a forecast captures the exponential growth rate of the target quantity
116 (see Section 2.2) and c) is less dependent on the expected order of magnitude of the quantity to be predicted
117 (see Section 2.3). We therefore argue that such evaluations on the logarithmic scale should complement the
118 prevailing evaluations on the natural scale. Other transformations may likewise be of interest. We briefly
119 explore the square root transformation as an alternative transformation. Our analysis mostly focuses on
120 the CRPS (or WIS) as an evaluation metric for probabilistic forecasts, given its widespread use throughout
121 the COVID-19 pandemic. We note that the logarithmic score has scale invariance properties which imply
122 that score differences between different forecasts are invariant to strictly monotonic transformations (see
123 Lehmann 1950 on corresponding properties of likelihood ratios and Diks et al. 2011). The question of the
124 right scale to evaluate forecasts on does therefore not arise for the log score.

125 The remainder of the article is structured as follows. In Sections 2.1–2.3 we provide some mathematical
126 intuition on applying the log-transformation prior to evaluating the CRPS, highlighting the connections to
127 relative error measures, the epidemic growth rate and variance stabilizing transformations. We then discuss
128 the effect of the log-transformation on forecast rankings (Section 2.4) as well as practical considerations for
129 applying transformations in general and the log-transformation in particular (Section 2.5). To analyse the
130 real-world implications of the log-transformation we use forecasts submitted to the European COVID-19
131 Forecast Hub (European Covid-19 Forecast Hub, 2021; Sherratt et al., 2022, Section 3). Finally, we provide
132 scoring recommendations, discuss alternative transformations that may be useful in different contexts, and
133 suggest further research avenues (Section 4).

134 2 Logarithmic transformation of forecasts and observations

135 2.1 Interpretation as a relative error

136 To illustrate the effect of applying the natural logarithm prior to evaluating forecasts we consider the absolute
137 error, which the CRPS and WIS generalize to probabilistic forecasts. We assume strictly positive support
138 (meaning that no specific handling of zero values is needed), a restriction we will address when applying this
139 transformation in practice. When considering a point forecast \hat{y} for a quantity of interest y , such that

$$140 \quad y = \hat{y} + \varepsilon, \quad (6)$$

141 the absolute error is given by $|\varepsilon|$. When taking the logarithm of the forecast and the observation first, thus
142 considering

$$143 \quad \log y = \log \hat{y} + \varepsilon^*, \quad (7)$$

144 the resulting absolute error $|\varepsilon^*|$ can be interpreted as an approximation of various common relative error
145 measures. Using that $\log(a) \approx a - 1$ if a is close to 1, we get

$$146 \quad |\varepsilon^*| = |\log \hat{y} - \log y| = \left| \log \left(\frac{\hat{y}}{y} \right) \right| \stackrel{\text{if } \hat{y} \approx y}{\approx} \left| \frac{\hat{y}}{y} - 1 \right| = \left| \frac{\hat{y} - y}{y} \right|. \quad (8)$$

147 The absolute error after log transforming is thus an approximation of the *absolute percentage error* (APE,
148 Gneiting, 2011) as long as forecast and observation are close. As we assumed that $\hat{y} \approx y$, we can also
149 interpret it as an approximation of the *relative error* (RE, Gneiting, 2011)

$$150 \quad \left| \frac{\hat{y} - y}{\hat{y}} \right| \quad (9)$$

151 and the *symmetric absolute percentage error* (SAPE; see e.g., Flores 1986)

$$152 \quad \left| \frac{\hat{y} - y}{y/2 + \hat{y}/2} \right|. \quad (10)$$

153 As Figure 1 shows, the alignment with the SAPE is in fact the closest and holds quite well even if predicted
 154 and observed value differ by a factor of two or three. Generalising to probabilistic forecasts, the CRPS
 155 applied to log-transformed forecasts and outcomes can thus be seen as a probabilistic counterpart to the
 156 symmetric absolute percentage error, which offers an appealing intuitive interpretation.

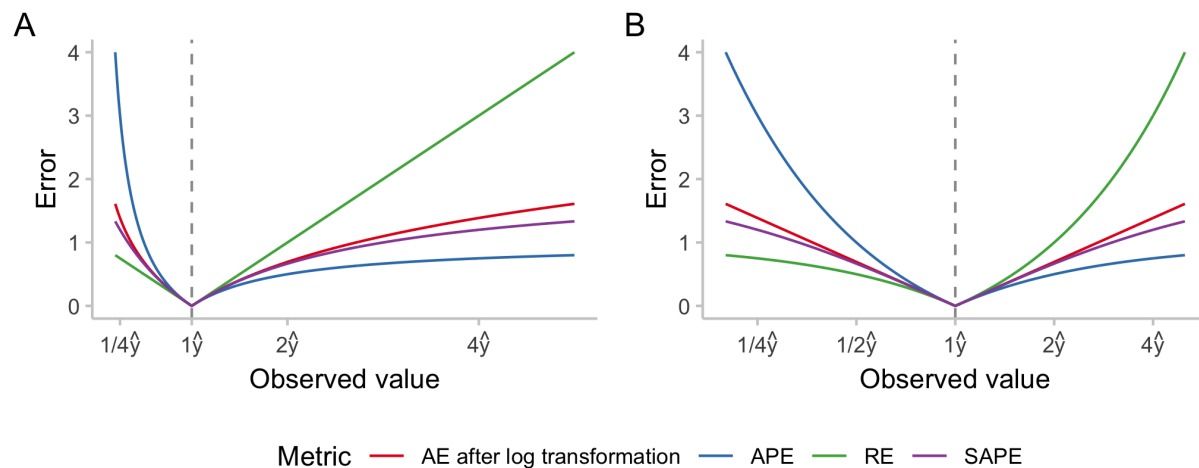


Figure 1: Numerical comparison of different measures of relative error: absolute percentage error (APE), relative error (RE), symmetric absolute percentage error (SAPE) and the absolute error applied to log-transformed predictions and observations. We denote the predicted value by \hat{y} and display errors as a function of the ratio of observed and predicted value. A: x-axis shown on a linear scale. B: x-axis shown on a logarithmic scale.

157 2.2 Interpretation as scoring the exponential growth rate

158 Another interpretation for the log-transform is possible if the generative process is framed as exponential
 159 with a time-varying growth rate $r(t)$ (see, e.g., Wallinga and Lipsitch, 2007), i.e.

$$160 \quad \frac{d}{dt}y(t) = r(t)y(t) \quad (11)$$

161 which is solved by

$$162 \quad y(t) = y_0 \exp\left(\int_0^t r(t')dt'\right) = y_0 \exp(\bar{r}_t t) \quad (12)$$

163 where y_0 is an initial data point and \bar{r}_t is the mean of the growth rate between the initial time point 0 and
 164 time t .

165 If a forecast $\hat{y}(t)$ for the value of the time series at time t is issued at time 0 based on the data point y_0 then
 166 the absolute error after log transformation is

$$167 \quad \begin{aligned} \epsilon^* &= |\log[\hat{y}(t)] - \log[y(t)]| \\ &= |\log[y_0 \exp(\hat{r}_t t)] - \log[y_0 \exp(\bar{r}_t t)]| \\ &= t |\hat{r}_t - \bar{r}_t| \end{aligned} \quad (13)$$

168 where \bar{r}_t is the true mean growth rate and \hat{r}_t is the forecast mean growth rate. We thus evaluate the error
 169 in the mean exponential growth rate, scaled by the length of the time period considered. Again generalising
 170 this to the CRPS and WIS implies a probabilistic evaluation of forecasts of the epidemic growth rate.
 171

172 2.3 Interpretation as a variance-stabilising transformation

173 When evaluating models across sets of forecasting tasks, it may be desirable for each target to have a similar
174 impact on the overall results. This could be motivated by the assumption that forecasts from different
175 geographical units and time periods provide similar amounts of information about how well a forecaster
176 performs. One would then like the resulting scores to be independent of the order of magnitude of the target
177 to predict. CRPS values on the natural scale, however, typically scale with the order of magnitude of the
178 quantity to be predicted. Average scores are then dominated by the results achieved for targets with high
179 expected outcomes in a way that does not necessarily reflect the underlying predictive ability well.

180 If the predictive distribution for the quantity Y equals the true data-generating process F (an ideal forecast),
181 the expected CRPS is given by (Gneiting and Raftery, 2007)

$$182 \quad \mathbb{E}[\text{CRPS}(F, y)] = 0.5 \times \mathbb{E}|Y - Y'|, \quad (14)$$

183 where Y and Y' are independent samples from F . This corresponds to half the *mean absolute difference*,
184 which is a measure of dispersion. If F is well-approximated by a normal distribution $N(\mu, \sigma^2)$, the approxi-
185 mation

$$186 \quad \mathbb{E}_F[\text{CRPS}(F, y)] \approx \frac{\sigma}{\sqrt{\pi}} \quad (15)$$

187 can be used. This means that the expected CRPS scales roughly with the standard deviation, which in
188 turn typically increases with the mean in epidemiological forecasting. In order to make the expected CRPS
189 independent of the expected outcome, a *variance-stabilising transformation* (VST, Bartlett, 1936; Dunn and
190 Smyth, 2018) can be employed. The choice of this transformation depends on the mean-variance relationship
191 of the underlying process.

192 If the mean-variance relationship of the data-generating distribution is quadratic with $\sigma^2 = c \times \mu^2$, the
193 natural logarithm can serve as the VST. Denoting by F_{\log} the predictive distribution for $\log(Y)$, we can use
194 the delta method (a first-order Taylor approximation, see e.g., Dunn and Smyth 2018), to show that

$$195 \quad \mathbb{E}_F[\text{CRPS}\{F_{\log}, \log(y)\}] \approx \frac{\sigma/\mu}{\sqrt{\pi}} = \frac{\sqrt{c}}{\sqrt{\pi}}. \quad (16)$$

196 As σ and μ are linked through the quadratic mean-variance relationship (or linear mean-standard deviation
197 relationship, $\sigma = \sqrt{c} \times \mu$), the expected CRPS thus stays constant regardless of the expected value of the
198 data-generating distribution μ . The assumption of a quadratic mean-variance relationship is closely linked
199 to the aspects discussed in Sections 2.1 and 2.2. It implies that relative errors have constant variance and
200 can thus be meaningfully compared across different targets. Also, it arises naturally if we assume that our
201 capacity to predict the epidemic growth rate does not depend on the expected outcome, i.e. does not depend
202 on the current phase of the epidemic or the order of magnitude of current observations.

203 If the mean-variance relationship is linear with $\sigma^2 = c \times \mu$, as with a Poisson-distributed variable, the square
204 root is known to be a VST (Dunn and Smyth, 2018). Denoting by $F_{\sqrt{\cdot}}$ the predictive distribution for \sqrt{Y} ,
205 the delta method can again be used to show that

$$206 \quad \mathbb{E}_F[\text{CRPS}\{F_{\sqrt{\cdot}}, \sqrt{y}\}] \approx \frac{\sigma/\sqrt{\mu}}{2\sqrt{\pi}} = \frac{\sqrt{c}}{2\sqrt{\pi}}. \quad (17)$$

207 We note that while standard in the derivation of variance-stabilizing transformations, the application of the
208 delta method in equations (16) and (17) requires the probability mass of F to be tightly distributed. If this
209 is not the case, the approximation and thus the variance stabilization may be less accurate.

210 To strengthen our intuition on how transforming outcomes prior to applying the CRPS shifts the emphasis
211 between targets with high and low expected outcomes, Figure 2 shows the expected CRPS of ideal forecasters

212 under different mean-variance relationships and transformations. We consider a Poisson distribution where
 213 $\sigma^2 = \mu$, a negative binomial distribution with size parameter $\theta = 10$ and thus $\sigma^2 = \mu + \mu^2/10$, and a
 214 truncated normal distribution with practically constant variance. We see that when applying the CRPS
 215 on the natural scale, the expected CRPS grows monotonically as the variance of the predictive distribution
 216 (which is equal to the data-generating distribution for the ideal forecaster) increases. The expected CRPS
 217 is constant only for the distribution with constant variance, and grows in μ for the other two. When applying a
 218 log-transformation first, the expected CRPS is almost independent of μ for the negative binomial distribution
 219 and large μ , while smaller targets have higher expected CRPS in case of the Poisson distribution and the
 220 normal distribution with constant variance. When applying a square-root-transformation, the expected
 221 CRPS is independent of the mean for the Poisson-distribution, but not for the other two (with a positive
 222 relationship in the normal case and a negative one for the negative binomial). As can be seen in Figures 2
 223 and SI.3, the approximations presented in equations (16) and (17) work quite well for our simulated example.

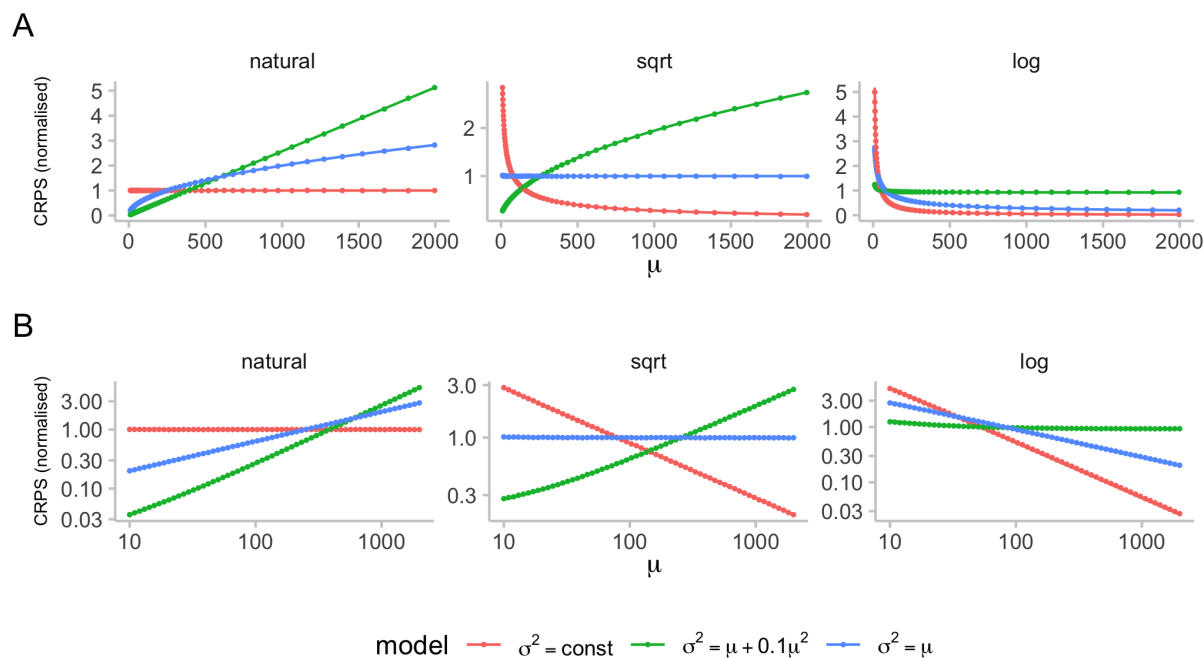


Figure 2: Expected CRPS scores as a function of the mean and variance of the forecast quantity. We computed expected CRPS values for three different distributions, assuming an ideal forecaster with predictive distribution equal to the true underlying (data-generating) distribution. These expected CRPS values were computed for different predictive means based on 10,000 samples each and are represented by dots. Solid lines show the corresponding approximations of the expected CRPS from equations (16) and (17). Figure SI.3 shows the quality of the approximation in more detail. The first distribution (red) is a truncated normal distribution with constant variance (we chose $\sigma = 1$ in order to only obtain positive samples). The second (green) is a negative binomial distribution with variance $\theta = 10$ and variance $\sigma^2 = \mu + 0.1\mu^2$. The third (blue) is a Poisson distribution with $\sigma^2 = \mu$. To make the scores for the different distributions comparable, scores were normalised to one, meaning that the mean score for every distribution (red, green, blue) is one. A: Normalised expected CRPS for ideal forecasts with increasing means for three distribution with different relationships between mean and variance. Expected CRPS was computed on the natural scale (left), after applying a square-root transformation (middle), and after adding one and applying a log-transformation to the data (right). B: A but with x and y axes on the log scale.

2.4 Effects on model rankings

Rankings between different forecasters based on the CRPS may change when making use of a transformation, both in terms of aggregate and individual scores. We illustrate this in Figure 3 with two forecasters, A and B, issuing two different distributions with different dispersion. When showing the obtained CRPS as a function of the observed value, it can be seen that the ranking between the two forecasters may change when scoring the forecast on the logarithmic, rather than the natural scale. In particular, on the natural scale, forecaster A, who issues a more uncertain distribution, receives a better score than forecaster B for observed values far away from the centre of the respective predictive distribution. On the log scale, however, forecaster A receives a lower score for large observed values, being more heavily penalised for assigning large probability to small values (which, in relative terms, are far away from the actual observation). We note that the chosen example involving a geometric forecast distribution is somewhat constructed; as shown in Section 3.4 and Figure 8A, rankings between models in practice stay quite stable for a single forecast.

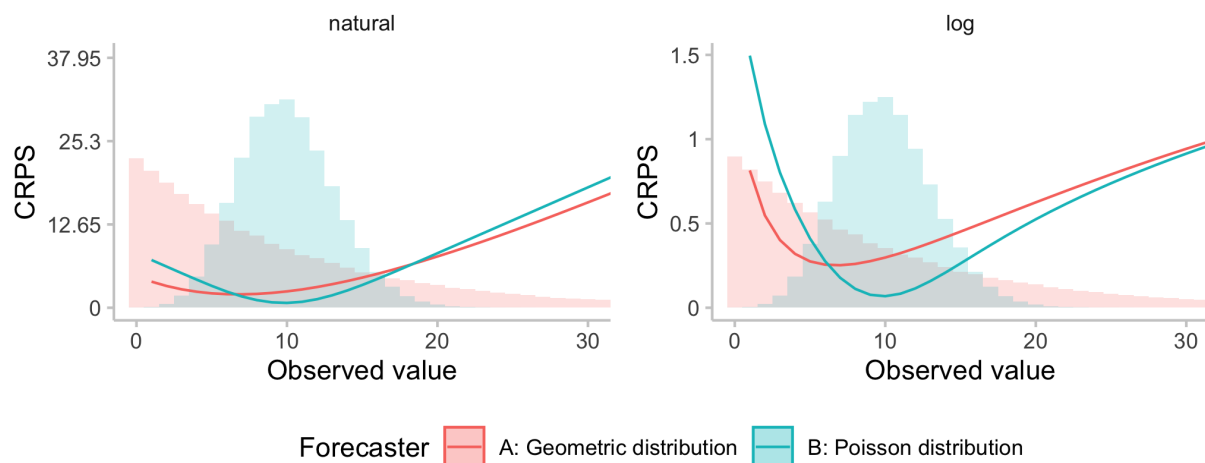


Figure 3: Illustration of the effect of the log-transformation of the ranking for a single forecast. Shown are CRPS (or WIS, respectively) values as a function of the observed value for two forecasters. Model A issues a geometric distribution (a negative binomial distribution with size parameter $\theta = 1$) with mean $\mu = 10$ and variance $\sigma^2 = \mu + \mu^2 = 110$, while Model B issues a Poisson distribution with mean and variance equal to 10. Zeroes in this illustrative example were handled by adding one before applying the natural logarithm.

Overall model rankings would be expected to differ more when scores are averaged across multiple forecasts or targets. The change in rankings of aggregate scores usually is mainly driven by the order of magnitude of scores for different forecast targets across time, location and target type and less so by the kind of changes in model rankings for single forecasts discussed above (see Figure 8 for a practical example). Large observations will dominate average CRPS values when evaluation is done on the natural scale, but much less so after log transformation. Depending on how different models perform across targets of different orders of magnitude, rankings in terms of average scores may change when applying a transformation.

2.5 Practical considerations and other transformations

In practice, one issue with the log transform is that it is not readily applicable to negative or zero values, which need to be removed or otherwise handled. One common approach to this end is to add a small positive quantity, such as $a = 1$, to all observations and predictions before taking the logarithm (Bellégo et al., 2022). This still represents a strictly monotonic transformation, but the choice of a does influence scores and

248 rankings (measures of relative errors shrink the larger the chosen value a). As a rule of thumb, if if $x > 5a$,
249 the difference between $\log(x + a)$ and $\log(x)$ is small, and it becomes negligible if $x > 50a$. Choosing a
250 suitable offset a thus balances two competing concerns: on the one hand, choosing a small a makes sure that
251 the transformation is as close to a natural logarithm as possible and scores can be interpreted as outlined
252 in the previous sections. On the other hand, choosing a larger a can help stabilise scores for forecasts and
253 observations close to zero, avoiding giving excessive weight to forecasts of small quantities. For increasing
254 a , less relative weight is given to smaller forecast targets. For very large values of a , $\log(x + a)$ is roughly
255 linear in x , so that using a very large a implies similar relative weighting as applying no transformation at
256 all. In practice, a user could explore the effect of different values of a graphically and choose a such that
257 the relative weightings of times and regions with high and low incidence correspond to their preferences (see
258 Figure 6 in our example application, Section 3).

259 A related issue occurs when the predictive distribution has a large probability mass on zero (or on very small
260 values), as this can translate into an excessively wide forecast in relative terms. In our applied example this
261 is illustrated in Figure SI.7. In such instances, the dispersion component of the WIS is inflated for scores
262 obtained after applying the natural logarithm because forecasts contained zero in its prediction intervals.
263 To deal with this issue one could choose to use a higher a value when applying a transformation $\log(x + a)$,
264 for example $a = 10$ instead of the $a = 1$ that we chose to use.

265 A natural question is which other transformations could be applied and whether resulting scores remain
266 (strictly) proper. In principle, any transformation function can be applied simultaneously to forecasts and
267 observations as long as the definition of the transformation is independent of the forecasts and any quantities
268 unknown at the time of forecasting, including the observed value. This simply corresponds to a re-definition
269 of the forecasting target. However, applying non-invertible transformations leads to a loss in information
270 conveyed by forecasts, which we consider undesirable. The resulting score will be proper, but it may not be
271 strictly proper anymore (as forecasts differing from the forecaster's true belief on the original scale may be
272 identical on the transformed scale). When using the CRPS or the WIS, it seems most appropriate to use
273 only strictly monotonic transformations such as the natural logarithm or the square root as otherwise the
274 encoded notion of distance may become meaningless.

275 Some other strictly monotonic transformations that can be applied are scaling by the population size or
276 scaling by past observations. The latter, as discussed in Section 4, is similar to applying a log-transformation,
277 but corresponds to evaluating a forecast of multiplicative, rather than exponential growth rates. The arising
278 issue of dividing by zero can again be solved by adding a small offset a . Scaling a forecast by the later
279 observed value (as opposed to scaling by past observations) is generally not permissible as it can result in
280 improper scores (see Lerch et al. 2015 on the closely related topic of weighting scores with a function of the
281 observed value). Similarly, scaling forecasts and observations by a function of the predictive distribution
282 (like the predictive mean) may lead to improper scores; however, we are unaware of existing theoretical
283 arguments on this.

284 When applying a transformation, the order of the operations matters, and applying a transformation after
285 scores have been computed generally does not guarantee that the score remains proper. In the case of log
286 transforms, taking the logarithm of the CRPS values, rather than scoring the log-transformed forecasts and
287 data, results in an improper score. We illustrate this point using simulated data in Figure SI.1, where it
288 can be seen that in the example overconfident models perform best in terms of the log WIS. We note that
289 strictly speaking, re-scaling average scores by the average score of a baseline model or average scores across
290 different models to obtain skill scores likewise leads to improper scores (Gneiting and Raftery, 2007). The
291 application of such skill scores, however, is established practice and considered largely unproblematic.

292 We note that in the practical evaluation of operational forecasting systems several additional challenges
293 arise, which we do not study in detail. These concern e.g., the removal of outlying observations and forecasts
294 and the handling of missing forecasts. The solutions we employed in practice are provided in Section 3.1.

295 3 Empirical example: the European Forecast Hub

296 3.1 Setting

297 As an empirical comparison of evaluating forecasts on the natural and on the log scale, we use forecasts from
298 the European Forecast Hub (European Covid-19 Forecast Hub, 2021; Sherratt et al., 2022). The European
299 COVID-19 Forecast Hub is one of several COVID-19 Forecast Hubs (Cramer et al., 2021; Bracher et al.,
300 2021b) which have been systematically collecting, aggregating and evaluating forecasts of several COVID-19
301 targets created by different teams every week. Forecasts are made one to four weeks ahead into the future
302 and follow a quantile-based format with a set of 23 quantiles (0.01, 0.025, 0.05, ..., 0.5, ...0.95, 0.975, 0.99).

303 The forecasts used for the purpose of this illustration are forecasts submitted between the 8th of March
304 2021 and the 5th of December 2022 for reported cases and deaths from COVID-19. Target dates range from
305 the 13th of March 2021 to the 10th of December 2022, for a total of 92 weeks. See Sherratt et al. (2022)
306 for a more thorough description of the data. We filtered all forecasts submitted to the Hub to only include
307 the seven models which have submitted forecasts for both deaths and cases for 4 horizons in 32 locations
308 on at least 46 forecast dates (see Figure SI.4). We removed all observations marked as data anomalies by
309 the European Forecast Hub (Sherratt et al., 2022) as well as all remaining negative observed values. These
310 anomalies made up a relevant fraction of all observations. On average across locations, 12.1 out of 92 (13.2%)
311 observations were removed for cases and 12.4 out of 92 (13.5%) for deaths. Figure SI.5 displays the number
312 of anomalies removed for each location. In addition, we filtered out a small number of erroneous forecasts
313 that were in extremely poor agreement with the observed data, as defined by any of the conditions listed in
314 Table SI.2. Figure SI.6 shows the percentage of forecasts removed for each model. Those few (less than 0.2%
315 of forecasts for each model) erroneous outlier forecasts had excessive influence on average scores and relative
316 skill scores in a way that was not representative of normal model behaviour. We removed them here in order
317 to better illustrate the effects of the log-transformation on scores that one would expect in a well-behaved
318 scenario. In a regular forecast evaluation such erroneous forecasts should usually not be removed and would
319 count towards overall model scores.

320 All predictive quantiles were truncated at 0. We applied the log-transformation after adding a constant
321 $a = 1$ to all predictions and observed values. The choice of $a = 1$ in part reflects convention, but also
322 represents a suitable choice as it avoids giving excessive weight to forecasts close to zero, while at the same
323 time ensuring that scores for observations > 5 can be interpreted reasonably. The analysis was conducted
324 in R (R Core Team, 2022), using the `scoringutils` package (Bosse et al., 2022) for forecast evaluation. All
325 code is available on GitHub (<https://github.com/epiforecasts/transformation-forecast-evaluation>). Where
326 not otherwise stated, we report results for a two-week-ahead forecast horizon.

327 In addition to the WIS we use pairwise comparisons (Cramer et al., 2021) to evaluate the relative performance
328 of models across countries in the presence of missing forecasts. In the first step, score ratios are computed
329 for all pairs of models by taking the set of overlapping forecasts between the two models and dividing the
330 score of one model by the score achieved by the other model. The relative skill for a given model compared
331 to others is then obtained by taking the geometric mean of all score ratios which involve that model. Low
332 values are better, and the “average” model receives a relative skill score of 1.

333 3.2 Illustration and qualitative observations

334 When comparing examples of forecasts on the natural scale with those on the log scale (see Figures 4, SI.7,
335 SI.8) a few interesting patterns emerge. Missing the peak, i.e. predicting increasing numbers while actual
336 observations are already falling, tends to contribute a lot to overall scores on the natural scale (see forecasts
337 during the peak in May 2022 in Figure 4A, B). On the log scale, these have less of an influence, as errors are

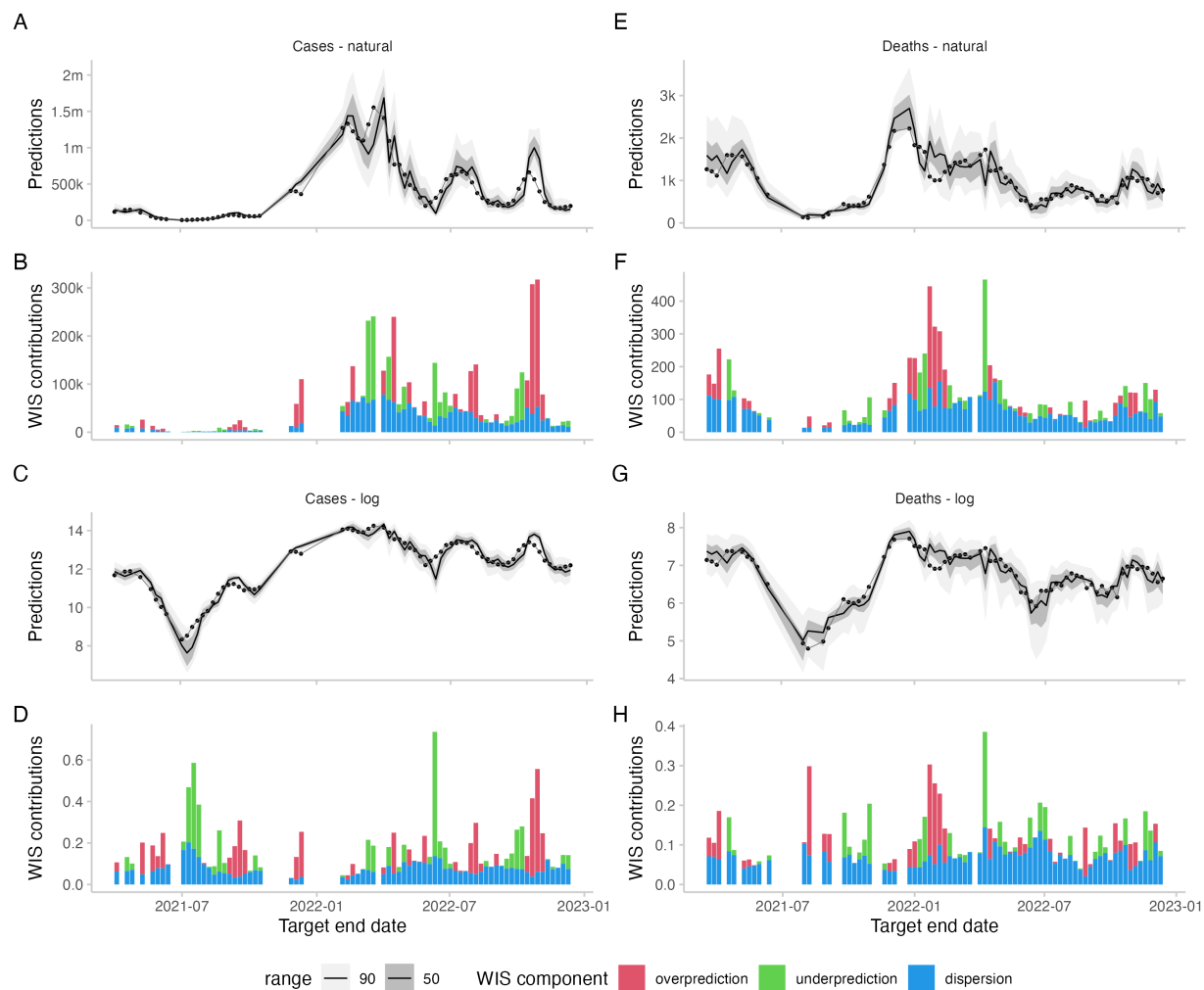


Figure 4: Forecasts and scores for two-week-ahead predictions from the EuroCOVIDhub-ensemble made in Germany. Missing values are due to data anomalies that were removed (see section 3.1). A, E: 50% and 90% prediction intervals and observed values for cases and deaths on the natural scale. B, F: Corresponding scores. C, G: Forecasts and observations on the log scale. D, H: Corresponding scores.

338 smaller in relative terms (see 4C, D). Conversely, failure to predict an upswing while numbers are still low,
 339 is less severely penalised on the natural scale (see forecasts in July 2021 and to a lesser extent in July 2022
 340 in Figure 4 A, B), as overall absolute errors are low. On the log scale, missing lower inflection points tends
 341 to lead to more severe penalties (see Figure 4C, D). One can also observe that on the natural scale, scores
 342 tend to track the overall level of the target quantity (compare for example forecasts for March-July with
 343 forecasts for September-October in Figure 4E, F). On the log scale, scores do not exhibit this behaviour and
 344 rather increase whenever forecasts are far away from the truth in relative terms, regardless of the overall
 345 level of observations.

346 Across the dataset, the average number of observed cases and deaths varied considerably by location and
 347 target type (see Figure 5A and B). On the natural scale, scores show a pattern quite similar to the ob-
 348 servations across targets (see Figure 5D) and locations (see Figure 5C). On the log scale, scores were more
 349 evenly distributed between targets (see Figure 5D) and locations (see Figure 5C). Both on the natural scale
 350 as well on the log scale, scores increased considerably with increasing forecast horizon (see Figure 5E). This

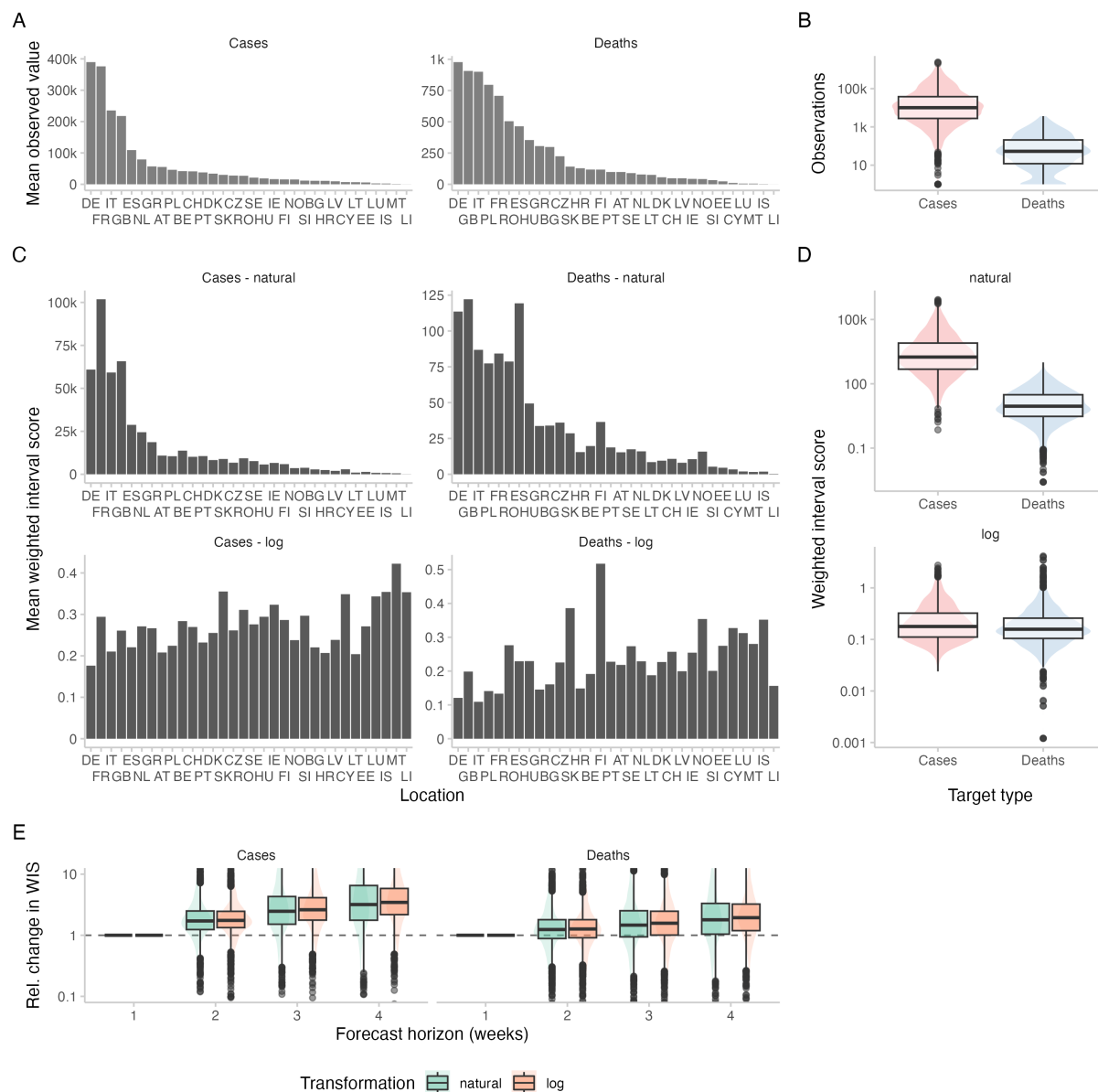


Figure 5: Observations and scores across locations and forecast horizons for the European COVID-19 Forecast Hub data. Locations are sorted according to the mean observed value in that location. A: Average (across all time points) of observed cases and deaths for different locations. B: Corresponding boxplot (y-axis on log-scale) of all cases and deaths. C: Scores for two-week-ahead forecasts from the EuroCOVIDhub-ensemble (averaged across all forecast dates) for different locations, evaluated on the natural scale as well as after transforming counts by adding one and applying the natural logarithm. D: Corresponding boxplots of all individual scores of the EuroCOVIDhub-ensemble for two-week-ahead predictions. E: Boxplots for the relative change of scores for the EuroCOVIDhub-ensemble across forecast horizons. For any given forecast date and location, forecasts were made for four different forecast horizons, resulting in four scores. All scores were divided by the score for forecast horizon one. To enhance interpretability, the range of visible relative changes in scores (relative to horizon = 1) was restricted to [0.1, 10].

351 reflects the increasing difficulty of forecasts further into the future and, for the log scale, corresponds with

our expectations from Section 2.2.

To assess the impact of the choice of offset value a we extend the display from Figure 5C by results obtained under different specifications. Results are shown in Figure 6, where for completeness we also added the square root transformation. As discussed in Section 2.5, smaller values of a increase the relative weight of smaller locations in the overall evaluation. In the most extreme considered case $a = 0.001$, the smallest locations in fact receive the largest weight both for deaths and cases. For very large values (see the third row of Figure 6), the relative weights strongly resemble those of the evaluation on the natural scale. We recommend using displays of this type to get an intuition for the role different locations may play for overall evaluation results.

3.3 Regression analysis to determine the variance-stabilizing transformation

As argued in Section 2.3, the mean-variance, or mean-CRPS, relationship determines which transformation can serve as a VST. We can analyse this relationship empirically by running a regression that explains the WIS (which approximates the CRPS) as a function of the central estimate of the predictive distribution. We ran the regression

$$\log[\text{WIS}(F, y)] = \alpha + \beta \times \log[\text{median}(F)], \quad (18)$$

where the predictive distribution F and the observation y are on the natural scale. This is equivalent to

$$\text{WIS}(F, y) = \exp(\alpha) \times \text{median}(F)^\beta, \quad (19)$$

meaning that we estimate a polynomial relationship between the predictive median and achieved WIS. Note that we are using predictive medians rather than means as only the former are available in the European COVID-19 Forecast Hub. As (under the simplifying assumption of normality; see Section 2.3) the WIS/CRPS of an ideal forecaster scales with the standard deviation, a value of $\beta = 1$ would imply a quadratic median-variance relationship; the natural logarithm could then serve as a VST. A value of $\beta = 0.5$ would imply a linear median-variance relationship, suggesting the square root as a VST. We applied the regression to case and death forecasts, stratified for one through four-week-ahead forecasts. Results are provided in Table 1. It can be seen that the estimates of β always take a value somewhat below 1, implying a slightly sub-quadratic mean-variance relationship. The logarithmic transformation should thus approximately stabilize the variance (and WIS), possibly leading to somewhat higher scores for smaller forecast targets. The square-root transformation, on the other hand, can be expected to still lead to higher WIS values for targets of higher orders of magnitude.

To check the relationship after the transformation, we ran the regressions

$$\text{WIS}(F_{\log}, \log y) = \alpha_{\log} + \beta_{\log} \cdot \log(\text{median}(F)), \quad (20)$$

where F_{\log} is the predictive distribution for $\log(y)$, and

$$\text{WIS}(F_{\sqrt{\cdot}}, \sqrt{y}) = \alpha_{\sqrt{\cdot}} + \beta_{\sqrt{\cdot}} \cdot \sqrt{\text{median}(F)}, \quad (21)$$

where $F_{\sqrt{\cdot}}$ is the predictive distribution on the square-root scale. A value of $\beta_{\log} = 0$ (or $\beta_{\sqrt{\cdot}} = 0$, respectively) would imply that scores are linearly independent of the median prediction after the transformation. A value smaller (larger) than 0 would imply that smaller (larger) targets lead to higher scores. As can be seen from Table 1, the results indeed indicate that small targets lead to larger average WIS when using the log transform ($\beta_{\log} < 0$), while the opposite is true for the square-root transform ($\beta_{\sqrt{\cdot}} > 0$). The results of the three regressions are also displayed in Figure 7. In this empirical example, the log transformation thus helps (albeit not perfectly), to stabilise WIS values, and it does so more successfully than the square-root transformation. As can be seen from Figure 7, the expected WIS scores for case targets with medians of 10 and 100,000 differ by more than a factor of ten for the square root transformation, but only a factor of around 2 for the logarithm.

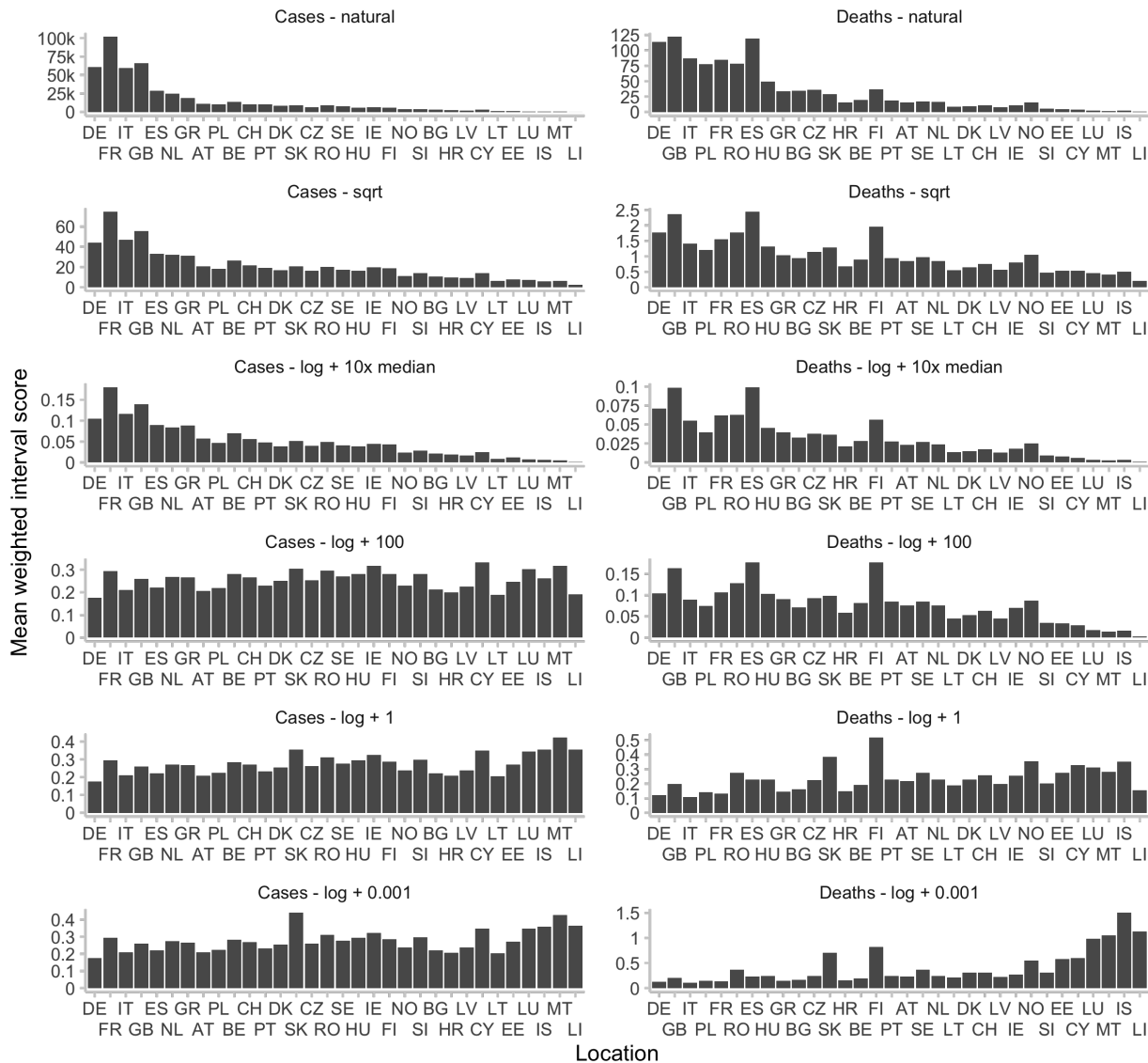


Figure 6: Mean WIS in different locations for different transformations applied before scoring. Locations are sorted according to the mean observed value in that location. Shown are scores for two-week-ahead forecasts of the EuroCOVIDhub-ensemble. On the natural scale (with no transformation prior to applying the WIS), scores correlate strongly with the average number of observed values in a given location. The same is true for scores obtained after applying a square-root transformation, or after applying a log-transformation with a large offset a . For illustrative purposes, a was chosen to be 101630 for cases and 530 for deaths, 10 times the respective median observed value. For large values of a , $\log(x + a)$ grows roughly linearly in x , meaning that we expect to observe the same patterns as in the case with no transformation. For decreasing values of a , we give more relative weight to scores in small locations.

393 3.4 Impact of logarithmic transformation on model rankings

394 For *individual* forecasts, rankings between models for single forecasts are mostly preserved, with differences
 395 increasing across forecast horizons (see Figure 8A). While rankings between forecasters remain similar for
 396 a single forecast, this is not true anymore when looking at rankings obtained after averaging scores across

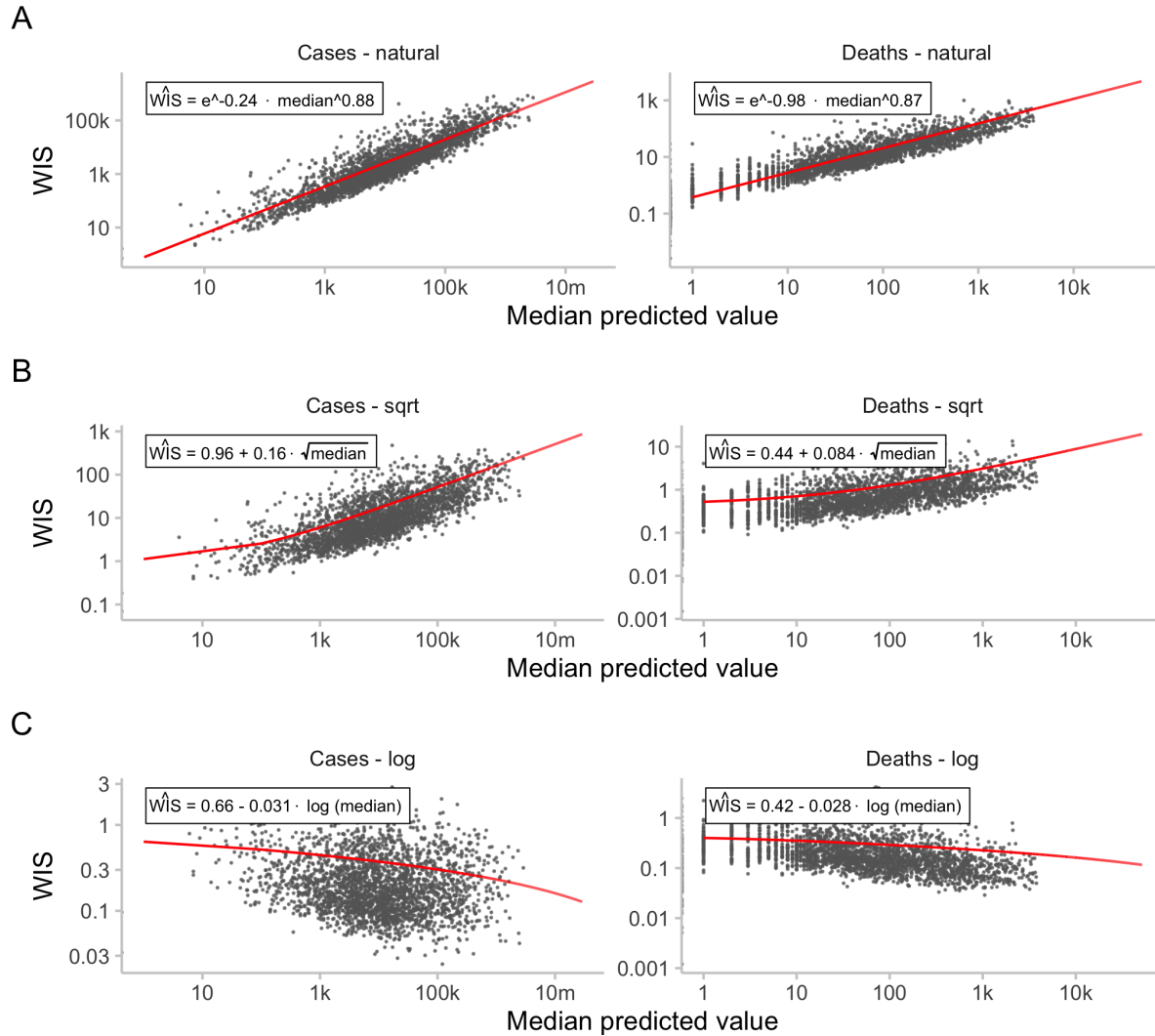


Figure 7: Relationship between median forecasts and scores. Black dots represent WIS values for two-week ahead predictions of the EuroCOVIDhub-ensemble. Shown in red are the regression lines discussed in Section 3.3 shown in Table 1. A: WIS for two-week-ahead predictions of the EuroCOVIDhub-ensemble against median predicted values. B: Same as A, with scores obtained after applying a square-root-transformation to the data. C: Same as A, with scores obtained after applying a log-transformation to the data.

397 multiple forecasts made at different times or in different locations. As discussed earlier, scores on the
 398 natural and on the log scale penalise errors very differently, e.g. when looking at performance during peaks
 399 or troughs. When evaluating performance *averaged across* different forecasts and forecast targets, relative
 400 skill scores of the models therefore change considerably (Figure 8B). The correlation between relative skill
 401 scores also decreases noticeably with increasing forecast horizon.

402 Figure 9 shows the changes in the ranking between different forecasting models. Encouragingly for the
 403 European Forecast Hub, the Hub ensemble, which is the forecast the organisers suggest forecast consumers
 404 make use of, remains the top model across scoring schemes. For cases, the ILM-EKF model and the Forecast
 405 Hub baseline model exhibit the largest change in relative skill scores. For the ILM-EKF model the relative
 406 proportion of the score that is due to overprediction is reduced when applying a log-transformation before

Horizon	Target	α	β	$\alpha_{\sqrt{\cdot}}$	$\beta_{\sqrt{\cdot}}$	α_{\log}	β_{\log}
1	Cases	-0.862	0.876	0.790	0.087	0.433	-0.024
2	Cases	-0.243	0.877	0.959	0.162	0.660	-0.031
3	Cases	0.372	0.855	1.109	0.238	0.882	-0.037
4	Cases	0.816	0.837	1.645	0.296	1.009	-0.036
1	Deaths	-1.146	0.832	0.457	0.048	0.376	-0.035
2	Deaths	-0.981	0.867	0.443	0.084	0.416	-0.028
3	Deaths	-0.807	0.885	0.349	0.131	0.453	-0.019
4	Deaths	-0.602	0.891	0.125	0.194	0.501	-0.011

Table 1: Coefficients of three regressions for the effect of the magnitude of the median forecast on expected scores. The first regression was $\log[\text{WIS}(F, y)] = \alpha + \beta \times \log[\text{median}(F)]$, where F is the predictive distribution and y the observed value. The second one was $\text{WIS}(F_{\log}, \log y) = \alpha_{\log} + \beta_{\log} \cdot \log(\text{median}(F))$, where F_{\log} is the predictive distribution for $\log y$. The third one was $\text{WIS}(F_{\sqrt{\cdot}}, \sqrt{y}) = \alpha_{\sqrt{\cdot}} + \beta_{\sqrt{\cdot}} \cdot \sqrt{\text{median}(F)}$, where $F_{\sqrt{\cdot}}$ is the predictive distribution for \sqrt{y} .

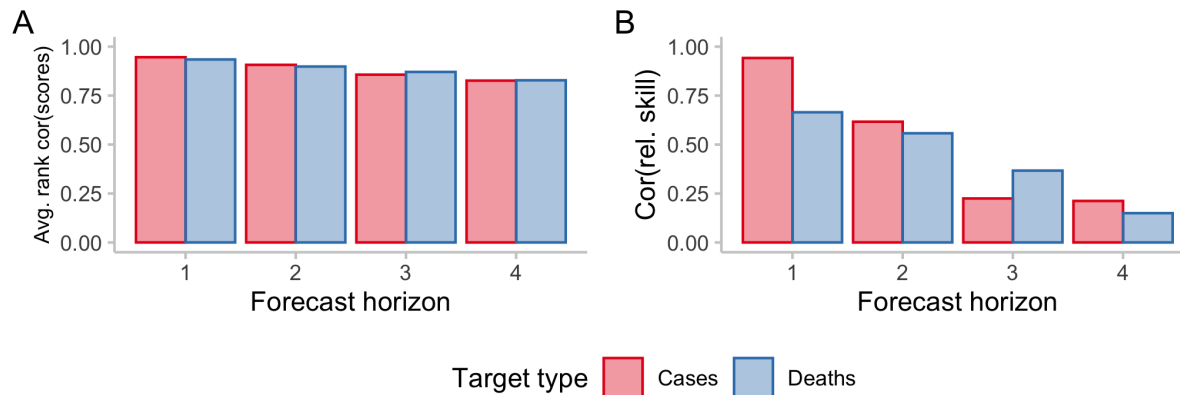


Figure 8: Correlations of rankings on the natural and logarithmic scale. A: Average Spearman rank correlation of scores for individual forecasts. For every individual target (defined by a combination of forecast date, target type, horizon, location), one score was obtained per model. Then, for every forecast target, the Spearman rank correlation was computed between scores on the natural scale and on the log scale for all the models that had made a forecast for that specific target. These individual rank correlations were then averaged across locations and time and are displayed stratified by horizon and target types, representing average accordance of model ranks for a single forecast target on the natural and on the log scale. B: Correlation between relative skill scores. For every forecast horizon and target type, a separate relative skill score was computed per model using pairwise comparisons, which is a measure of performance of a model relative to the others for a given horizon and target type that accounts for missing values. The plot shows the correlation between the relative skill scores on the natural vs. on the log scale, representing accordance of overall model performance as judged by scores on the natural and on the log scale.

407 scoring (see Figure 9E. Instances where the model has overshoot are penalised less heavily on the log scale,
 408 leading to an overall better score. For the Forecast Hub baseline model, the fact that it often puts relevant
 409 probability mass on zero (see Figure SI.7), leads to worse scores after applying log-transformation due to
 410 large dispersion penalties. For deaths, the baseline model seems to get similarly penalised for its in relative
 411 terms highly dispersed forecasts. The performance of other models changes as well, but patterns are less
 412 discernible on this aggregate level.

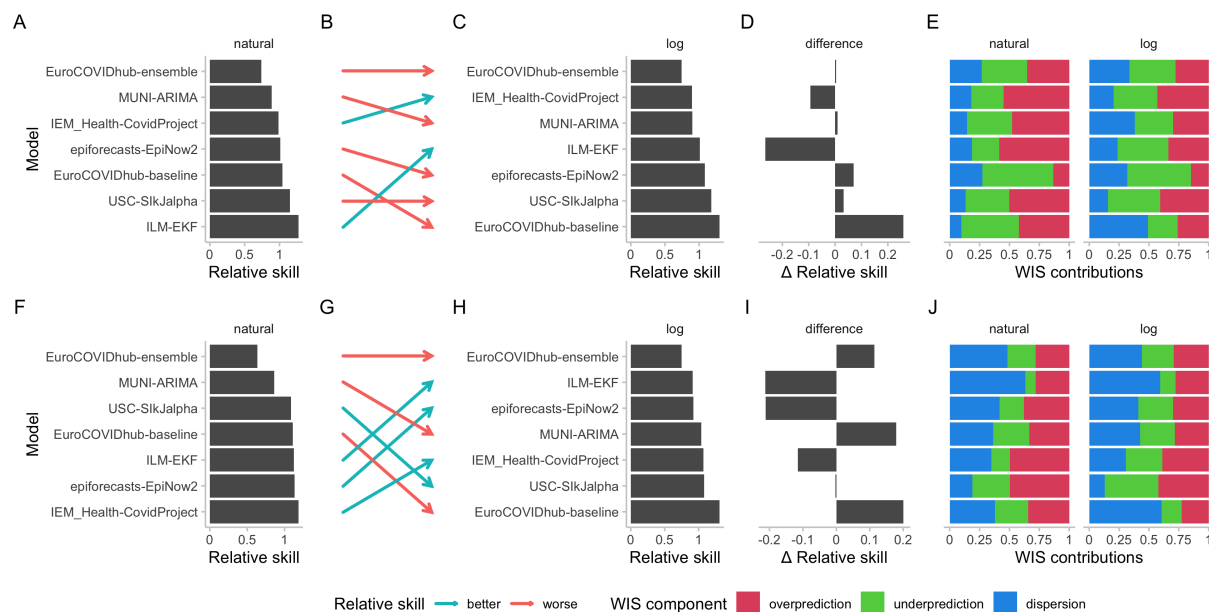


Figure 9: Changes in model ratings as measured by relative skill for two-week-ahead predictions for cases (top row) and deaths (bottom row). A: Relative skill scores for case forecasts from different models submitted to the European COVID-19 Forecast Hub computed on the natural scale. B: Change in rankings as determined by relative skill scores when moving from an evaluation on the natural scale to one on the logarithmic scale. Red arrows indicate that the relative skill score deteriorated when moving from the natural to the log scale, green arrows indicate they improved. C: Relative skill scores based on scores on the log scale. D: Difference in relative skill scores computed on the natural and on the logarithmic scale, ordered as in C. E: Relative contributions of the different WIS components (overprediction, underprediction, and dispersion) to overall model scores on the natural and the logarithmic scale. F, G, H, I, J, K: Analogously for deaths.

4 Discussion

In this paper, we proposed the use of transformations, with a particular focus on the natural logarithmic transformation, when evaluating forecasts in an epidemiological setting. These transformations can address issues that arise when evaluating epidemiological forecasts based on measures of absolute error and their probabilistic generalisations (i.e CRPS and WIS). We showed that scores obtained after log-transforming both forecasts and observations can be interpreted as a) a measure of relative prediction errors, as well as b) a score for a forecast of the exponential growth rate of the target quantity and c) as variance stabilising transformation in some settings. When applying this approach to forecasts from the European COVID-19 Forecast Hub, we found overall scores on the log scale to be more equal across, time, location and target type (cases, deaths) than scores on the natural scale. Scores on the log scale were much less influenced by the overall incidence level in a country and showed a slight tendency to be higher in locations with very low incidences. We found that model rankings changed noticeably.

On the natural scale, missing the peak and overshooting was more severely penalised than missing the nadir and the following upswing in numbers. Both failure modes tended to be more equally penalised on the log scale (with undershooting receiving slightly higher penalties in our example).

Applying a log-transformation prior to the WIS means that forecasts are evaluated in terms of relative errors and errors on the exponential growth rate, rather than absolute errors. The most important strength of this approach is that the evaluation better accommodates the exponential nature of the epidemiological

431 process and the types of errors forecasters who accurately model those processes are expected to make. The
432 log-transformation also helps avoid issues with scores being strongly influenced by the order of magnitude
433 of the forecast quantity, which can be an issue when evaluating forecasts on the natural scale. A potential
434 downside is that forecast evaluation is unreliable in situations where observed values are zero or very small.
435 One could argue that this correctly reflect inherent uncertainty about the future course of an epidemic when
436 numbers are small. Users nevertheless need to be aware that this can pose issues in practice. Including very
437 small values in prediction intervals (see e.g. Figure SI.7) can lead to excessive dispersion values on the log
438 scale. Similarly, locations with lower incidences may get disproportionate weight (i.e. high scores) when
439 evaluating forecasts on the log scale. Bracher et al. (2021a) argue that it is desirable to give large weight to
440 forecasts for locations with high incidences, as this reflects performance on the targets we should care about
441 most. On the other hand, scoring forecasts on the log scale may be less influenced by outliers and better
442 reflect consistent performance across time, space, and forecast targets. Furthermore, decision makers may
443 specifically care about situations in which numbers start to rise from a previously low level.

444 The log-transformation is only one of many transformations that may be useful and appropriate in an
445 epidemiological context. One obvious option is to apply a population standardization to obtain incidence
446 forecasts e.g., per 100,000 population (Abbott et al., 2022). We suggested using the natural logarithm as a
447 variance-stabilising transformation (VST). This is appropriate for variables that are approximately normally
448 distributed and have a quadratic mean-variance relationship with $\sigma^2 = c \times \mu^2$ (this is e.g. approximately
449 true for the negative binomial distribution and large μ). Alternatively, the square-root transformation
450 can be appropriate in the case of a Poisson distributed variable (Dunn and Smyth, 2018). Other VST like
451 the Box-Cox (Box and Cox, 1964) are conceivable as well. If one is interested in multiplicative, rather than
452 exponential growth rates, one could, instead of applying a log transformation, convert forecasts into forecasts
453 for the multiplicative growth rate by dividing numbers by the last value that was observed at the time the
454 forecast was made. Forecasters would then implicitly predict a separate multiplicative growth rate from
455 today to horizon 1, 2, etc. Instead of dividing by the last observed value, another promising transformation
456 would be to divide each forecast by the forecast of the previous week (and analogously for observations), in
457 order to obtain forecasts for week-to-week growth rates. Alternatively, one could also take first differences
458 of values on the log scale. This approach would be akin to evaluating the shape of the predicted trajectory
459 against the shape of the observed trajectory (for a different approach to evaluating the shape of a forecast,
460 see Srivastava et al., 2022). Dividing values by the previous value, unfortunately, is not feasible under the
461 current quantile-based format of the Forecast Hubs, as the growth rate of the α -quantile may be different
462 from the α -quantile of the growth-rate. However, it may be an interesting approach if predictive samples are
463 available or if quantiles for weekwise growth rates have been collected. Potentially, the variance stabilising
464 time-series forecasting literature may be a useful source of other transformations for various forecast settings.

465 It is possible to go beyond choosing a single transformation by constructing composite scores as a weighted
466 sum of scores based on different transformations. This would make it possible to create custom scores and
467 allow forecast consumers to choose and assign explicit weights to different qualities of the forecasts they
468 might care about.

469 Exploring transformations is a promising avenue for future work that could help bridge the gap between
470 modellers and policymakers by providing scoring rules that better reflect what forecast consumers care about.
471 In this paper, we did not make any particular assumptions about policy makers' priorities and preferences.
472 Rather, we aimed to enable users to make an informed choice by showing how different transformations lead
473 to different relative weights for the kinds of prediction errors forecast consumers may care about, such as
474 absolute vs. relative errors or the size of penalties for over- vs. underprediction. In practice, engagement
475 with decision makers is important to determine what their priorities are and how different ways to measure
476 predictive importance should be weighed.

477 We have shown that the natural logarithm transformation can lead to significant changes in the relative
478 rankings of models against each other, with potentially important implications for decision-makers who rely
479 on the knowledge of past performance to make a judgement about which forecasts should inform future
480 decisions. While it is commonly accepted that multiple proper scoring rules should usually be considered

481 when comparing forecasts, we think this should be supplemented by considering different transformations of
482 the data to obtain a richer picture of model performance. More work needs to be done to better understand
483 the effects of applying transformations in different contexts, and how they may impact decision-making.

484 A Supplementary information

485 A.1 Alternative Formulation of the WIS

486 Instead of defining the WIS as an average of scores for individual quantiles, we can define it using an average of
 487 scores for symmetric predictive intervals. For a single prediction interval, the interval score (IS) is computed
 488 as the sum of three penalty components, dispersion (width of the prediction interval), underprediction and
 489 overprediction,

$$490 \quad IS_{\alpha}(F, y) = (u - l) + \frac{2}{\alpha} \cdot (l - y) \cdot \mathbf{1}(y \leq l) + \frac{2}{\alpha} \cdot (y - u) \cdot \mathbf{1}(y \geq u) \quad (22)$$

$$491 \quad = \text{dispersion} + \text{underprediction} + \text{overprediction}, \quad (23)$$

493 where $\mathbf{1}(\cdot)$ is the indicator function, y is the observed value, and l and u are the $\frac{\alpha}{2}$ and $1 - \frac{\alpha}{2}$ quantiles of
 494 the predictive distribution, i.e. the lower and upper bound of a single central prediction interval. For a set
 495 of K^* prediction intervals and the median m , the WIS is computed as a weighted sum,

$$496 \quad \text{WIS} = \frac{1}{K^* + 0.5} \cdot \left(w_0 \cdot |y - m| + \sum_{k=1}^{K^*} w_k \cdot IS_{\alpha_k}(F, y) \right), \quad (24)$$

497 where w_k is a weight for every interval. Usually, $w_k = \frac{\alpha_k}{2}$ and $w_0 = 0.5$.

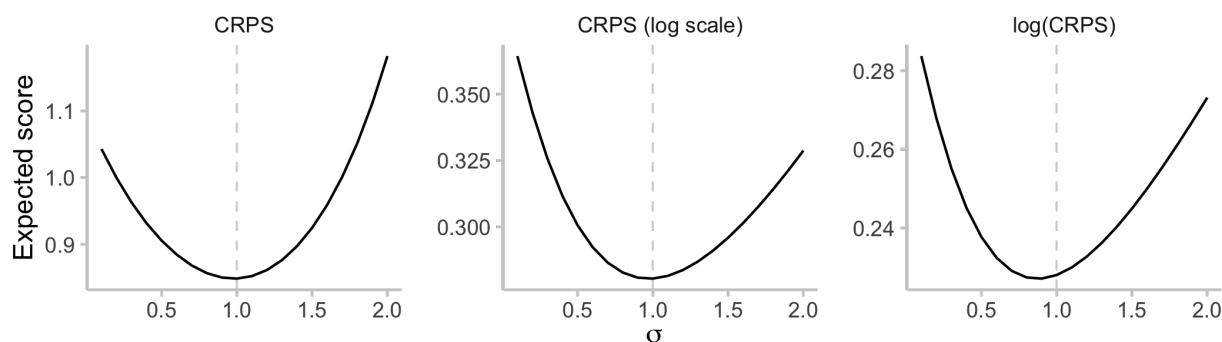


Figure SI.1: Illustration of the effect of applying a transformation after scoring. We assume $Y \sim \text{LogNormal}(0, 1)$ and evaluate the expected CRPS for predictive distributions $\text{LogNormal}(0, \sigma)$ with varying values of $\sigma \in [0.1, 2]$. For the regular CRPS (left) and CRPS applied to log-transformed outcomes (middle), the lowest expectation is achieved for the true value $\sigma = 1$. For the log-transformed CRPS, the optimal value is 0.9, i.e. there is an incentive to report a forecast that is too sharp. The score is therefore no longer proper.

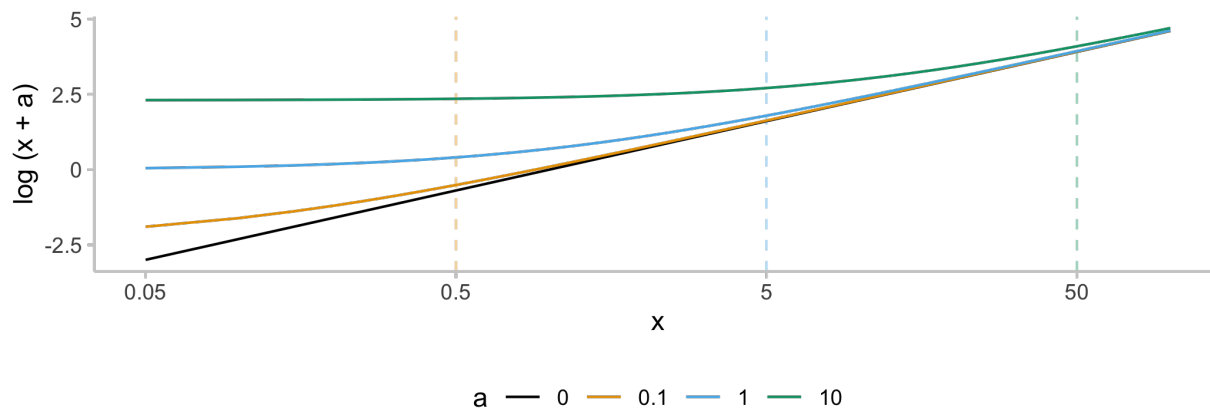


Figure SI.2: Illustration of the effect of adding a small quantity to a value before taking the natural logarithm. For increasing x , all lines eventually approach the black line (representing a transformation with no offset applied). For a given solid line, the dashed line of the same colour marks the x -value that is equal to 5 times the corresponding offset. It can be seen that for a values smaller than one fifth of the transformed quantity, the effect of adding an offset is generally small. When choosing a suitable a , the trade-off is between staying close to the interpretation of a pure log-transformation (choosing a small a) and not giving excessive weights to small observations (by choosing a larger a , see Figure 6).

target type	quantity	measure	natural	log
Cases	Observations	mean	61979	9.19
Cases	Observations	sd	171916	2.10
Cases	Observations	var	29555122130	4.42
Deaths	Observations	mean	220	3.89
Deaths	Observations	sd	435	1.96
Deaths	Observations	var	189051	3.83
Cases	WIS	mean	15840	0.27
Cases	WIS	sd	53117	0.28
Deaths	WIS	mean	31	0.23
Deaths	WIS	sd	65	0.28

Table SI.1: Summary statistics for observations and scores for forecasts from the ECDC data set.

True value	&	Median prediction
> 0		> 100× true value
> 10		> 20× true value
> 50		< 1/50× true value
= 0		> 100

Table SI.2: Criteria for removing forecasts. Any forecast that met one of the listed criteria (represented by a row in the table), was removed. Those forecasts were removed in order to be better able to illustrate the effects of the log-transformation on scores and eliminating distortions caused by outlier forecasters. When evaluating models against each other (rather than illustrating the effect of a transformation), one would prefer not to condition on the outcome when deciding whether a forecast should be taken into account.

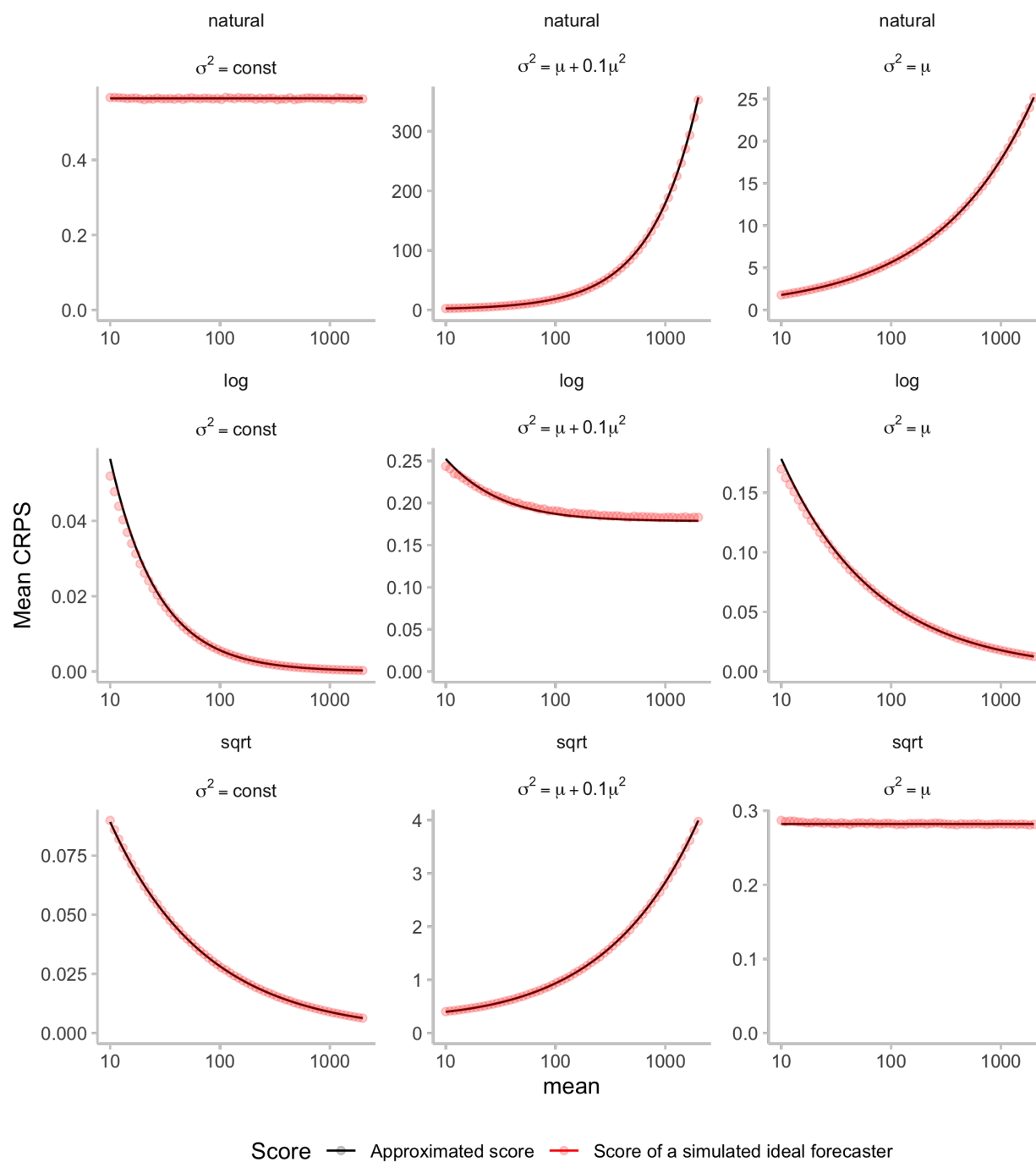


Figure SI.3: Visualisation of expected CRPS values against approximated scores using the approximation detailed in Section 2.4 (see also Figure 2). Expected CRPS scores are shown for three different distributions once on the natural scale (top row) and once scored on the log scale (bottom row).

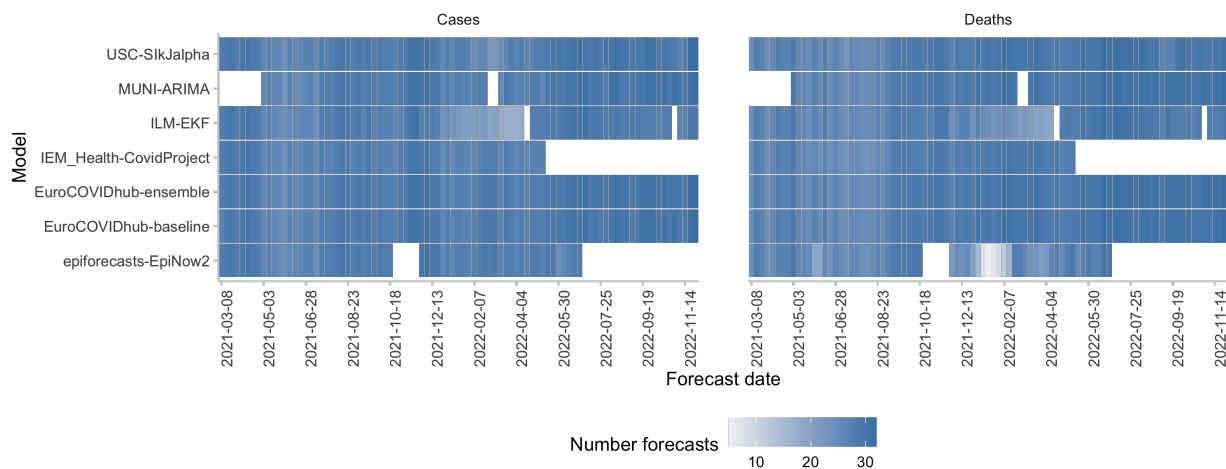


Figure SI.4: Number of forecasts available from different models for each forecast date.

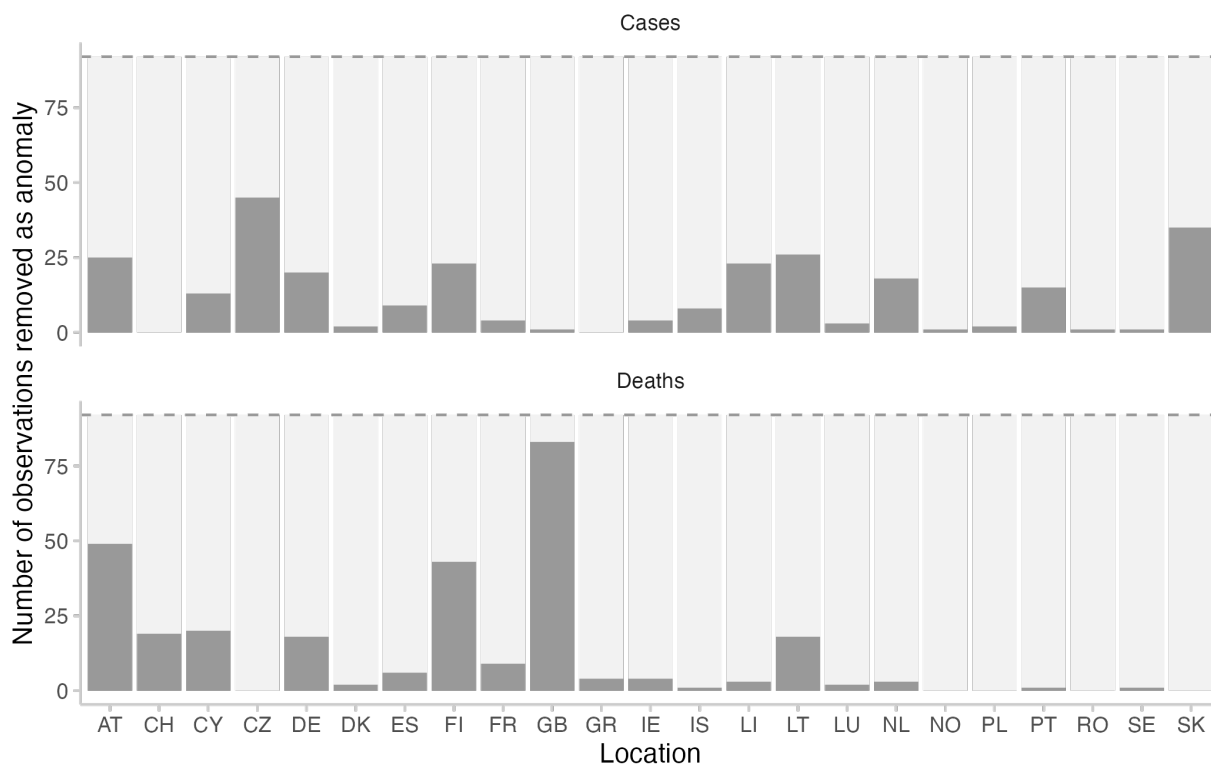


Figure SI.5: Number of observed values that were removed as anomalous. The values were marked as anomalous by the European Forecast Hub team.

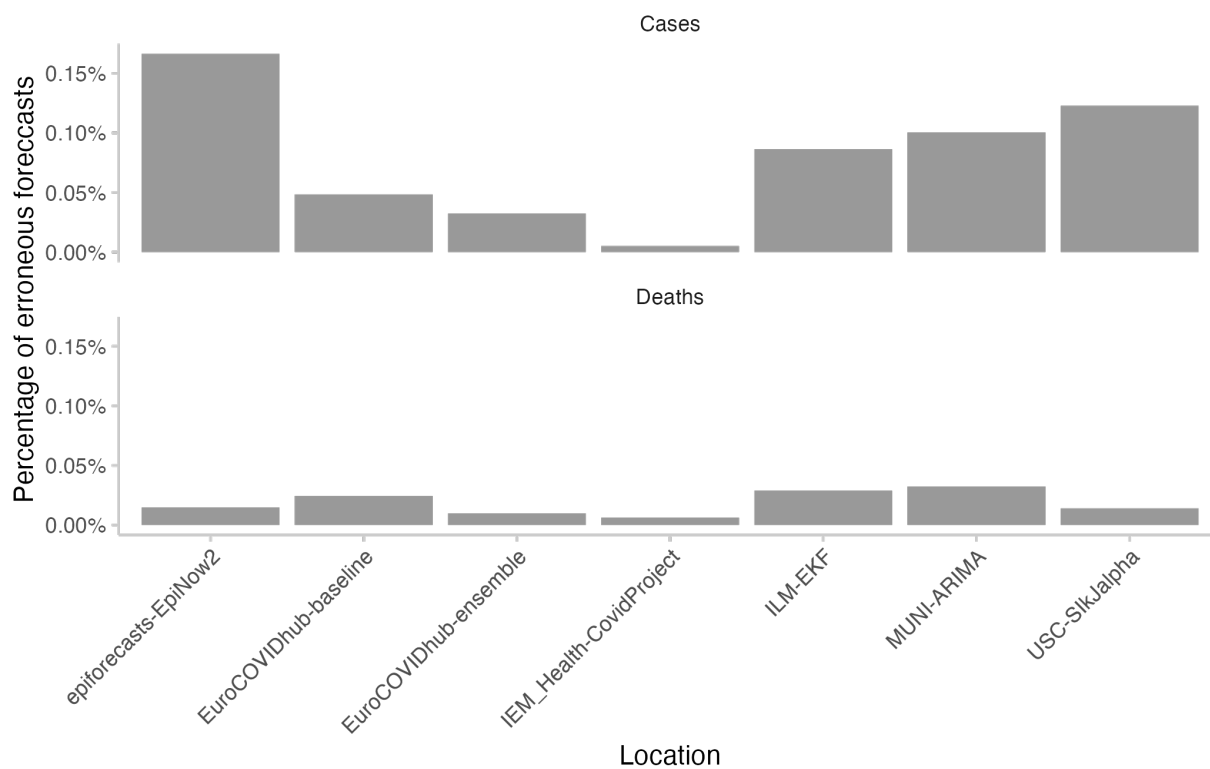


Figure SI.6: Number of forecasts marked as erroneous and removed. Forecasts that were in extremely poor agreement with the observed values were removed from the analysis according to the criteria shown in Table SI.2.

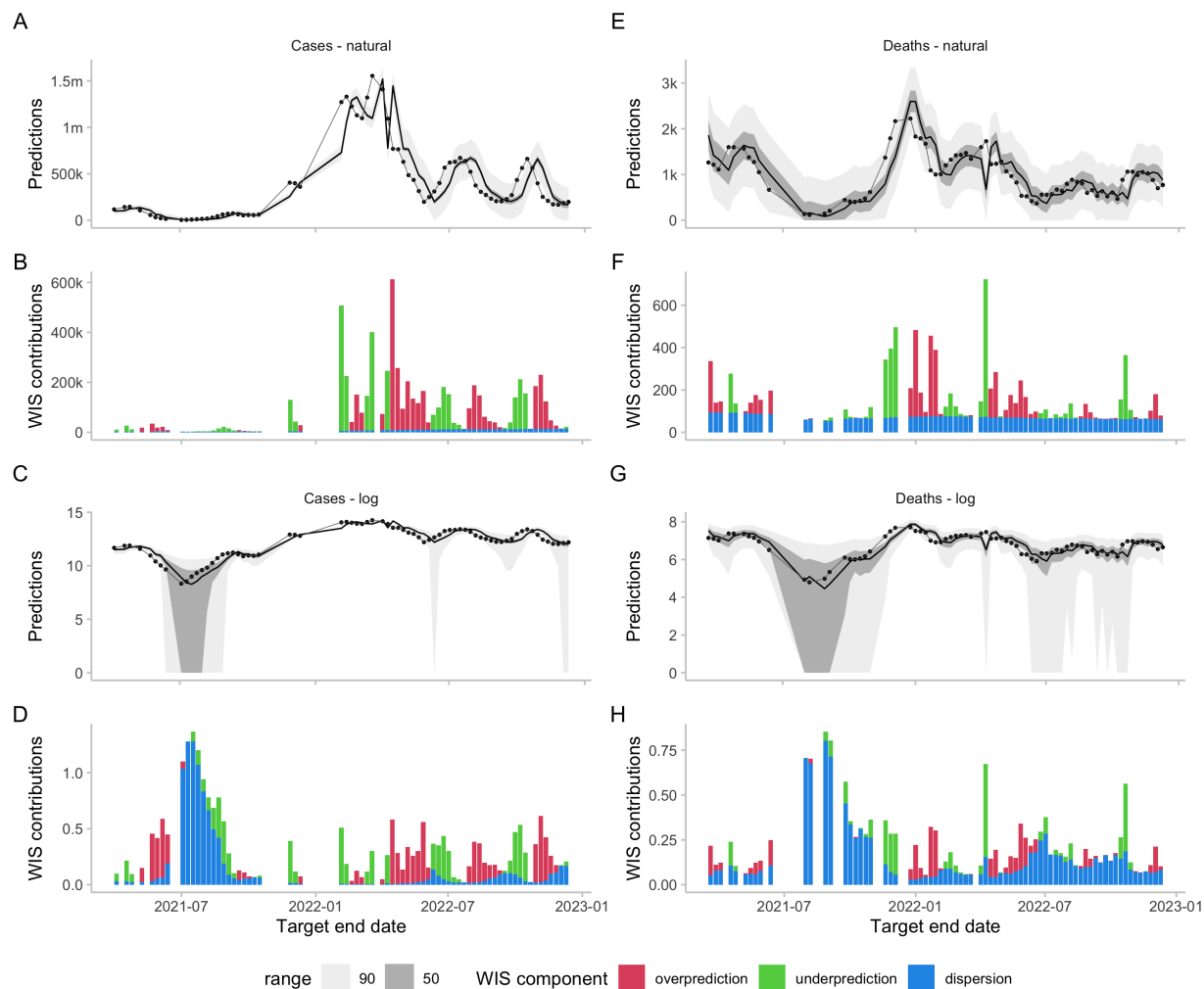


Figure SI.7: Forecasts and scores for two-week-ahead predictions from the EuroCOVIDhub-baseline made in Germany. The model had zero included in some of its 50 percent intervals (e.g. for case forecasts in July 2021), leading to excessive dispersion values on the log scale. One could argue that including zero in the prediction intervals constituted an unreasonable forecast that was rightly penalised, but in general care has to be taken with small numbers. One potential way to do deal with this could be to use a higher a value when applying a transformation $\log(x + a)$, for example $a = 10$ instead of $a = 1$. A, E: 50% and 90% prediction intervals and observed values for cases and deaths on the natural scale. B, F: Corresponding scores. C, G: Forecasts and observations on the log scale. D, H: Corresponding scores.

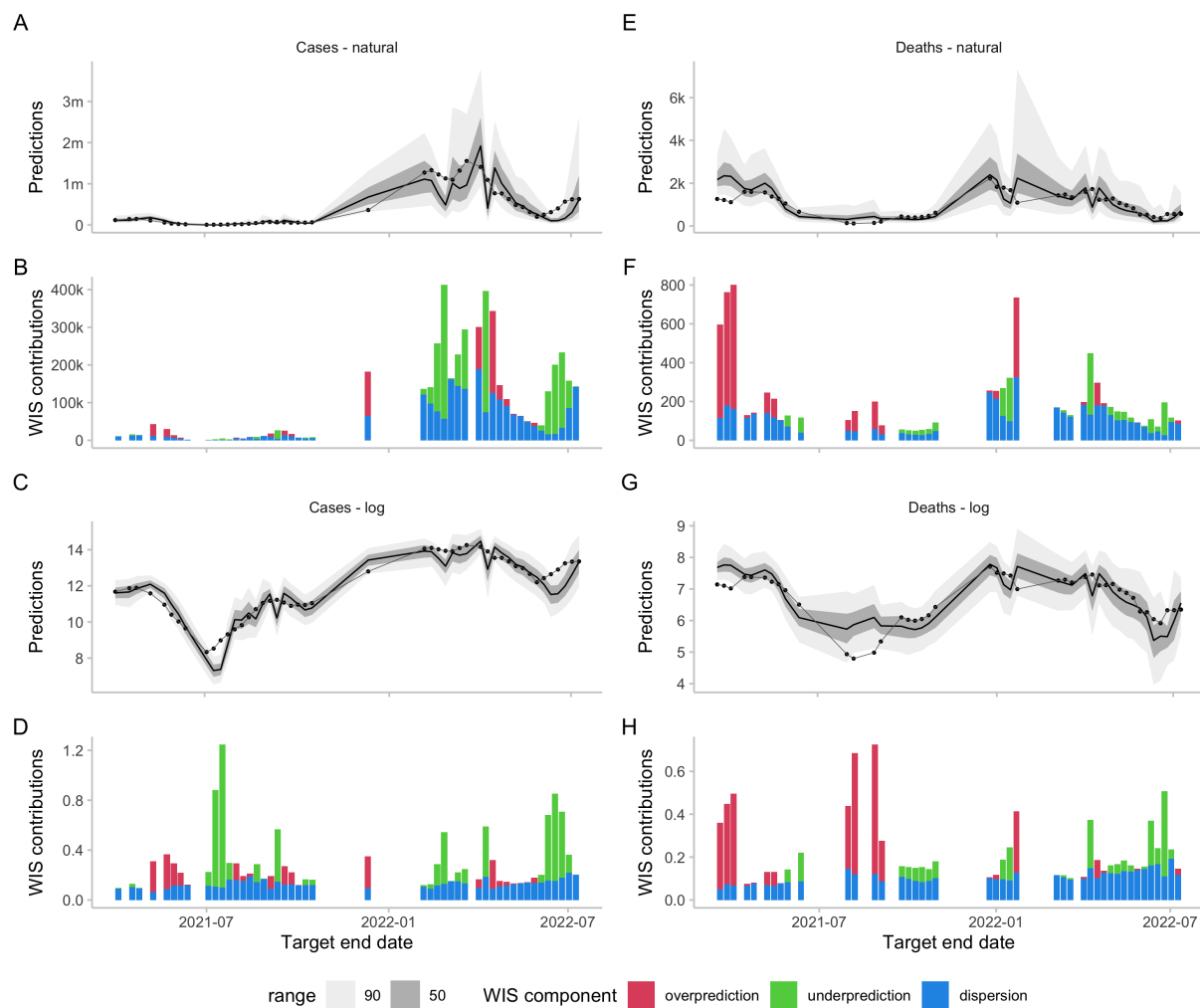


Figure SI.8: Forecasts and scores for two-week-ahead predictions from the epiforecasts-EpiNow2 model (Abbott et al., 2020) made in Germany. A, E: 50% and 90% prediction intervals and observed values for cases and deaths on the natural scale. B, F: Corresponding scores. C, G: Forecasts and observations on the log scale. D, H: Corresponding scores.

498 References

- 499 Abbott, S., Hellewell, J., Sherratt, K., Gostic, K., Hickson, J., Badr, H. S., DeWitt, M., Thompson, R.,
500 EpiForecasts, and Funk, S. (2020). *EpiNow2: Estimate Real-Time Case Counts and Time-Varying Epi-*
501 *demiological Parameters*. R package, <https://doi.org/10.5281/zenodo.3957490>.
- 502 Abbott, S., Sherratt, K., Bosse, N., Gruson, H., Bracher, J., and Funk, S. (2022). Evaluating an epidemio-
503 logically motivated surrogate model of a multi-model ensemble.
- 504 Bartlett, M. S. (1936). The Square Root Transformation in Analysis of Variance. *Supplement to the Journal*
505 *of the Royal Statistical Society*, 3(1):68–78.
- 506 Bellégo, C., Benatia, D., and Pape, L. (2022). Dealing with Logs and Zeros in Regression Models.
- 507 Bolin, D. and Wallin, J. (2023). Local scale invariance and robustness of proper scoring rules. *Statistical*
508 *Science*, 38(1):140–159. DOI: 10.1214/22-STS864.
- 509 Bosse, N. I., Gruson, H., Cori, A., van Leeuwen, E., Funk, S., and Abbott, S. (2022). Evaluating Forecasts
510 with scoringutils in R.
- 511 Box, G. E. P. and Cox, D. R. (1964). An Analysis of Transformations. *Journal of the Royal Statistical*
512 *Society. Series B (Methodological)*, 26(2):211–252.
- 513 Bracher, J., Ray, E. L., Gneiting, T., and Reich, N. G. (2021a). Evaluating epidemic forecasts in an interval
514 format. *PLoS computational biology*, 17(2):e1008618.
- 515 Bracher, J., Wolfram, D., Deuschel, J., Goergen, K., Ketterer, J. L., Ullrich, A., Abbott, S., Barbarossa,
516 M. V., Bertsimas, D., Bhatia, S., Bodych, M., Bosse, N. I., Burgard, J. P., Castro, L., Fairchild, G., Fiedler,
517 J., Fuhrmann, J., Funk, S., Gambin, A., Gogolewski, K., Heyder, S., Hotz, T., Kheifetz, Y., Kirsten, H.,
518 Krueger, T., Krymova, E., Leithäuser, N., Li, M. L., Meinke, J. H., Miasojedow, B., Michaud, I. J.,
519 Mohring, J., Nouvellet, P., Nowosielski, J. M., Ozanski, T., Radwan, M., Rakowski, F., Scholz, M., Soni,
520 S., Srivastava, A., Gneiting, T., and Schienle, M. (2022). National and subnational short-term forecasting
521 of COVID-19 in Germany and Poland, early 2021. *Communications Medicine*. DOI: 10.1038/s43856-022-
522 00191-8.
- 523 Bracher, J., Wolfram, D., Deuschel, J., Görgen, K., Ketterer, J. L., Ullrich, A., Abbott, S., Barbarossa,
524 M. V., Bertsimas, D., Bhatia, S., Bodych, M., Bosse, N. I., Burgard, J. P., Castro, L., Fairchild, G.,
525 Fuhrmann, J., Funk, S., Gogolewski, K., Gu, Q., Heyder, S., Hotz, T., Kheifetz, Y., Kirsten, H., Krueger,
526 T., Krymova, E., Li, M. L., Meinke, J. H., Michaud, I. J., Niedzielewski, K., Ożański, T., Rakowski,
527 F., Scholz, M., Soni, S., Srivastava, A., Zieliński, J., Zou, D., Gneiting, T., and Schienle, M. (2021b).
528 Short-term forecasting of COVID-19 in Germany and Poland during the second wave – a preregistered
529 study. *medRxiv*, page 2020.12.24.20248826.
- 530 CDC (2022). Cdcepi/Flusight-forecast-data. CDC Epidemic Prediction Initiative. Data repository, <https://github.com/cdcepi/Flusight-forecast-data>.
531
- 532 Cramer, E., Ray, E. L., Lopez, V. K., Bracher, J., Brennen, A., Rivadeneira, A. J. C., Gerding, A., Gneiting,
533 T., House, K. H., Huang, Y., Jayawardena, D., Kanji, A. H., Khandelwal, A., Le, K., Mühlemann, A.,
534 Niemi, J., Shah, A., Stark, A., Wang, Y., Wattanachit, N., Zorn, M. W., Gu, Y., Jain, S., Bannur, N.,
535 Deva, A., Kulkarni, M., Merugu, S., Raval, A., Shingi, S., Tiwari, A., White, J., Woody, S., Dahan, M.,
536 Fox, S., Gaither, K., Lachmann, M., Meyers, L. A., Scott, J. G., Tec, M., Srivastava, A., George, G. E.,
537 Cegan, J. C., Dettwiller, I. D., England, W. P., Farthing, M. W., Hunter, R. H., Lafferty, B., Linkov,
538 I., Mayo, M. L., Parno, M. D., Rowland, M. A., Trump, B. D., Corsetti, S. M., Baer, T. M., Eisenberg,
539 M. C., Falb, K., Huang, Y., Martin, E. T., McCauley, E., Myers, R. L., Schwarz, T., Sheldon, D., Gibson,
540 G. C., Yu, R., Gao, L., Ma, Y., Wu, D., Yan, X., Jin, X., Wang, Y.-X., Chen, Y., Guo, L., Zhao, Y., Gu,
541 Q., Chen, J., Wang, L., Xu, P., Zhang, W., Zou, D., Biegel, H., Lega, J., Snyder, T. L., Wilson, D. D.,
542 McConnell, S., Walraven, R., Shi, Y., Ban, X., Hong, Q.-J., Kong, S., Turtle, J. A., Ben-Nun, M., Riley,
543 P., Riley, S., Koyluoglu, U., DesRoches, D., Hamory, B., Kyriakides, C., Leis, H., Milliken, J., Moloney,

- 544 M., Morgan, J., Ozcan, G., Schrader, C., Shakhnovich, E., Siegel, D., Spatz, R., Stiefeling, C., Wilkinson,
545 B., Wong, A., Gao, Z., Bian, J., Cao, W., Ferres, J. L., Li, C., Liu, T.-Y., Xie, X., Zhang, S., Zheng,
546 S., Vespignani, A., Chinazzi, M., Davis, J. T., Mu, K., y Piontti, A. P., Xiong, X., Zheng, A., Baek, J.,
547 Farias, V., Georgescu, A., Levi, R., Sinha, D., Wilde, J., Penna, N. D., Celi, L. A., Sundar, S., Cavany,
548 S., España, G., Moore, S., Oidtman, R., Perkins, A., Osthus, D., Castro, L., Fairchild, G., Michaud, I.,
549 Karlen, D., Lee, E. C., Dent, J., Grantz, K. H., Kaminsky, J., Kaminsky, K., Keegan, L. T., Lauer, S. A.,
550 Lemaitre, J. C., Lessler, J., Meredith, H. R., Perez-Saez, J., Shah, S., Smith, C. P., Truelove, S. A., Wills,
551 J., Kinsey, M., Obrecht, R. F., Tallaksen, K., Burant, J. C., Wang, L., Gao, L., Gu, Z., Kim, M., Li,
552 X., Wang, G., Wang, Y., Yu, S., Reiner, R. C., Barber, R., Gaikedu, E., Hay, S., Lim, S., Murray, C.,
553 Pigott, D., Prakash, B. A., Adhikari, B., Cui, J., Rodriguez, A., Tabassum, A., Xie, J., Keskinocak, P.,
554 Asplund, J., Baxter, A., Oruc, B. E., Serban, N., Arik, S. O., Dusenberry, M., Epshteyn, A., Kanal, E.,
555 Le, L. T., Li, C.-L., Pfister, T., Sava, D., Sinha, R., Tsai, T., Yoder, N., Yoon, J., Zhang, L., Abbott,
556 S., Bosse, N. I., Funk, S., Hellewel, J., Meakin, S. R., Munday, J. D., Sherratt, K., Zhou, M., Kalantari,
557 R., Yamana, T. K., Pei, S., Shaman, J., Ayer, T., Adee, M., Chhatwal, J., Dalgic, O. O., Ladd, M. A.,
558 Linas, B. P., Mueller, P., Xiao, J., Li, M. L., Bertsimas, D., Lami, O. S., Soni, S., Bouardi, H. T., Wang,
559 Y., Wang, Q., Xie, S., Zeng, D., Green, A., Bien, J., Hu, A. J., Jahja, M., Narasimhan, B., Rajanala, S.,
560 Rumack, A., Simon, N., Tibshirani, R., Tibshirani, R., Ventura, V., Wasserman, L., O’Dea, E. B., Drake,
561 J. M., Pagano, R., Walker, J. W., Slayton, R. B., Johansson, M., Biggerstaff, M., and Reich, N. G. (2021).
562 Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the US. *medRxiv*,
563 page 2021.02.03.21250974.
- 564 Cramer, E., Reich, N. G., Wang, S. Y., Niemi, J., Hannan, A., House, K., Gu, Y., Xie, S., Horstman,
565 S., aniruddhadiga, Walraven, R., starkari, Li, M. L., Gibson, G., Castro, L., Karlen, D., Wattanachit,
566 N., jinghuichen, zyt9lsb, aagarwal1996, Woody, S., Ray, E., Xu, F. T., Biegel, H., GuidoEspaña, X, X.,
567 Bracher, J., Lee, E., har96, and leyouz (2020). COVID-19 Forecast Hub: 4 December 2020 snapshot.
- 568 Diks, C., Panchenko, V., and van Dijk, D. (2011). Likelihood-based scoring rules for comparing density
569 forecasts in tails. *Journal of Econometrics*, 163(2):215–230.
- 570 Dunn, P. K. and Smyth, G. K. (2018). *Generalized Linear Models With Examples in R*. Springer.
- 571 Dushoff, J. and Park, S. W. (2021). Speed and strength of an epidemic intervention. *Proceedings of the*
572 *Royal Society B: Biological Sciences*, 288(1947):20201556.
- 573 European Covid-19 Forecast Hub (2021). European Covid-19 Forecast Hub. <https://covid19forecasthub.eu/>.
- 574 Flores, B. E. (1986). A pragmatic view of accuracy measurement in forecasting. *Omega*, 14(2):93–98.
- 575 Fuglstad, G.-A., Simpson, D., Lindgren, F., and Rue, H. (2015). Does non-stationary spatial data always
576 require non-stationary random fields? *Spatial Statistics*, 14:505–531.
- 577 Gneiting, T. (2011). Making and Evaluating Point Forecasts. *Journal of the American Statistical Association*,
578 106(494):746–762.
- 579 Gneiting, T. and Raftery, A. E. (2005). Weather Forecasting with Ensemble Methods. *Science*,
580 310(5746):248–249.
- 581 Gneiting, T. and Raftery, A. E. (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal*
582 *of the American Statistical Association*, 102(477):359–378.
- 583 Good, I. J. (1952). Rational Decisions. *Journal of the Royal Statistical Society. Series B (Methodological)*,
584 14(1):107–114.
- 585 Gostic, K. M., McGough, L., Baskerville, E., Abbott, S., Joshi, K., Tedijanto, C., Kahn, R., Niehus, R.,
586 Hay, J., de Salazar, P., Hellewell, J., Meakin, S., Munday, J., Bosse, N. I., Sherrat, K., Thompson, R. N.,
587 White, L. F., Huisman, J. S., Scire, J., Bonhoeffer, S., Stadler, T., Wallinga, J., Funk, S., Lipsitch, M., and
588 Cobey, S. (2020). Practical considerations for measuring the effective reproductive number, Rt. *medRxiv*.

- 589 Held, L., Meyer, S., and Bracher, J. (2017). Probabilistic forecasting in infectious disease epidemiology: The
590 13th Armitage lecture. *Statistics in Medicine*, 36(22):3443–3460.
- 591 Johansson, M. A., Apfeldorf, K. M., Dobson, S., Devita, J., Buczak, A. L., Baugher, B., Moniz, L. J.,
592 Bagley, T., Babin, S. M., Guven, E., Yamana, T. K., Shaman, J., Moschou, T., Lothian, N., Lane, A.,
593 Osborne, G., Jiang, G., Brooks, L. C., Farrow, D. C., Hyun, S., Tibshirani, R. J., Rosenfeld, R., Lessler,
594 J., Reich, N. G., Cummings, D. A. T., Lauer, S. A., Moore, S. M., Clapham, H. E., Lowe, R., Bailey,
595 T. C., García-Díez, M., Carvalho, M. S., Rodó, X., Sardar, T., Paul, R., Ray, E. L., Sakrejda, K., Brown,
596 A. C., Meng, X., Osoba, O., Vardavas, R., Manheim, D., Moore, M., Rao, D. M., Porco, T. C., Ackley,
597 S., Liu, F., Worden, L., Convertino, M., Liu, Y., Reddy, A., Ortiz, E., Rivero, J., Brito, H., Juarrero, A.,
598 Johnson, L. R., Gramacy, R. B., Cohen, J. M., Mordecai, E. A., Murdock, C. C., Rohr, J. R., Ryan, S. J.,
599 Stewart-Ibarra, A. M., Weikel, D. P., Jutla, A., Khan, R., Poultney, M., Colwell, R. R., Rivera-García,
600 B., Barker, C. M., Bell, J. E., Biggerstaff, M., Swerdlow, D., Mier-y-Teran-Romero, L., Forshey, B. M.,
601 Trtanj, J., Asher, J., Clay, M., Margolis, H. S., Hebbeler, A. M., George, D., and Jean-Paul Chretien
602 (2019). An open challenge to advance probabilistic forecasting for dengue epidemics. *Proceedings of the
603 National Academy of Sciences*, 116(48):24268–24274.
- 604 Lehmann, E. L. (1950). Some Principles of the Theory of Testing Hypotheses. *The Annals of Mathematical
605 Statistics*, 21(1):1 – 26.
- 606 Lerch, S., Thorarinsdottir, T. L., Ravazzolo, F., and Gneiting, T. (2015). Forecaster’s Dilemma: Extreme
607 Events and Forecast Evaluation.
- 608 Löwe, R., Mikkelsen, P. S., and Madsen, H. (2014). Stochastic rainfall-runoff forecasting: Parameter estima-
609 tion, multi-step prediction, and evaluation of overflow risk. *Stochastic Environmental Research and Risk
610 Assessment*, 28(3):505–516.
- 611 Mayr, J. and Ulbricht, D. (2015). Log versus level in VAR forecasting: 42 million empirical answers—Expect
612 the unexpected. *Economics Letters*, 126:40–42.
- 613 R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical
614 Computing, Vienna, Austria.
- 615 Reich, N. G., Lessler, J., Funk, S., Viboud, C., Vespignani, A., Tibshirani, R. J., Shea, K., Schienle, M.,
616 Runge, M. C., Rosenfeld, R., Ray, E. L., Niehus, R., Johnson, H. C., Johansson, M. A., Hochheiser, H.,
617 Gardner, L., Bracher, J., Borcherding, R. K., and Biggerstaff, M. (2022). Collaborative hubs: Making the
618 most of predictive epidemic modeling. *American Journal of Public Health*, 112(6):839–842.
- 619 Sherratt, K., Gruson, H., Grah, R., Johnson, H., Niehus, R., Prasse, B., Sandman, F., Deuschel, J., Wolfram,
620 D., Abbott, S., Ullrich, A., Gibson, G., Ray, EL., Reich, NG., Sheldon, D., Wang, Y., Wattanachit, N.,
621 Wang, L., Trnka, J., Obozinski, G., Sun, T., Thanou, D., Pottier, L., Krymova, E., Barbarossa, MV.,
622 Leithäuser, N., Mohring, J., Schneider, J., Wlazlo, J., Fuhrmann, J., Lange, B., Rodiah, I., Baccam,
623 P., Gurung, H., Stage, S., Suchoski, B., Budzinski, J., Walraven, R., Villanueva, I., Tucek, V., Šmíd, M.,
624 Zajíček, M., Pérez, Á. C., Reina, B., Bosse, NI., Meakin, S., Di Loro, A., Maruotti, A., Eclerová, V., Kraus,
625 A., Kraus, D., Pribylova, L., Dimitris, B., Li, ML., Saksham, S., Dehning, J., Mohr, S., Priesemann, V.,
626 Redlarski, G., Bejar, B., Ardenghi, G., Parolini, N., Ziarelli, G., Bock, W., Heyder, S., Hotz, T., E., S. D.,
627 Guzman-Merino, M., Aznarte, JL., Moriña, D., Alonso, S., Álvarez, E., López, D., Prats, C., Burgard, JP.,
628 Rodloff, A., Zimmermann, T., Kuhlmann, A., Zibert, J., Pennoni, F., Divino, F., Català, M., Lovison,
629 G., Giudici, P., Tarantino, B., Bartolucci, F., Jona, L. G., Mingione, M., Farcomeni, A., Srivastava,
630 A., Montero-Manso, P., Adiga, A., Hurt, B., Lewis, B., Marathe, M., Porebski, P., Venkatramanan, S.,
631 Bartzuk, R., Dreger, F., Gambin, A., Gogolewski, K., Gruzziel-Slomka, M., Krupa, B., Moszynski, A.,
632 Niedzielewski, K., Nowosielski, J., Radwan, M., Rakowski, F., Semeniuk, M., Szczurek, E., Zielinski, J.,
633 Kisielewski, J., Pabjan, B., Holger, K., Kheifetz, Y., Scholz, M., Bodych, M., Filinski, M., Idzikowski,
634 R., Krueger, T., Ozanski, T., Bracher, J., and Funk, S. (2022). Predictive performance of multi-model
635 ensemble forecasts of COVID-19 across European nation.
- 636 Srivastava, A., Singh, S., and Lee, F. (2022). Shape-based Evaluation of Epidemic Forecasts.

- 637 Taylor, J. W. (1999). Evaluating volatility and interval forecasts. *Journal of Forecasting*, 18(2):111–128.
- 638 Timmermann, A. (2018). Forecasting Methods in Finance. *Annual Review of Financial Economics*,
639 10(1):449–479.
- 640 Wallinga, J. and Lipsitch, M. (2007). How generation intervals shape the relationship between growth rates
641 and reproductive numbers. *Proceedings of the Royal Society B: Biological Sciences*, 274(1609):599–604.
- 642 Winkler, R. (1996). Scoring rules and the evaluation of probabilities. *TEST*, 5(1):1–60.