Eurasian Journal of Business and Management, 9(2), 2021, 139-163
DOI: 10.15604/ejbm.2021.09.02.004

EURASIAN JOURNAL OF BUSINESS AND MANAGEMENT

www.eurasianpublications.com

QUANTILE PROBABILITY PREDICTIONS: A DEMONSTRATIVE PERFORMANCE ANALYSIS OF FORECASTS OF US COVID-19 DEATHS

Mary E. Thomson
Corresponding Author: Northumbria University, UK
Email: mary.thomson@northumbria.ac.uk

Andrew C. Pollock
Statistical Analyst, UK
Email: acpollock@virginmedia.com

Jennifer Murray
Edinburgh Napier University, UK
Email: j.murray2@napier.ac.uk

Received: April 10, 2021 Accepted: May 25, 2021

Abstract

An analytical framework is presented for the evaluation of quantile probability forecasts. It is demonstrated using weekly quantile forecasts of changes in the number of US COVID-19 deaths. Empirical quantiles are derived using the assumption that daily changes in a variable follow a normal distribution with time varying means and standard deviations, which can be assumed constant over short horizons such as one week. These empirical quantiles are used to evaluate quantile forecasts using the Mean Squared Quantile Score (MSQS), which, in turn, is decomposed into sub-components involving bias, resolution and error variation to identify specific aspects of performance, which highlight the strengths and weaknesses of forecasts. The framework is then extended to test if performance enhancement can be achieved by combining diverse forecasts from different sources. The demonstration illustrates that the technique can effectively evaluate quantile forecasting performance based on a limited number of data points, which is crucial in emergency situations such as forecasting pandemic behavior. It also shows that combining the predictions with quantile probability forecasts generated from an Autoregressive Order One, AR(1) model provided substantially improved performance. The implications of these findings are discussed, suggestions are offered for future research and potential limitations are considered.

Keywords: Forecasts, Accuracy, Probability Forecasting, Composite Forecasts, Coherence, Covid-19

1. Introduction

COVID-19 has been described as “a pandemic viral disease with catastrophic global impact” (Larsen *et al.* 2020, pp. 473). It was first declared a global pandemic in March 2020 (WHO, 2020a), with the initial reported death due to COVID-19 announced via Chinese State Media in early January 2020 (Qin and Hernandez, 2020). America’s first case was reported in late January 2020 and by April 2020 the disease was present in all fifty states. By December 2020, there were over 68 million confirmed cases of COVID-19 and around 1.5 million estimated deaths worldwide (WHO, 2020b), despite lockdowns and related containment efforts for over one fifth of the world’s population (Gilbert, 2020). It is estimated to be 2-3 times more contagious than influenza (Anderson *et al.* 2020), and has been described as atypical of adult respiratory distress syndrome (ARDS), with patients often not responding to the standard treatment for ARDS (Cascella *et al.* 2020). Current treatment of COVID-19 relies on supportive therapies and reduction in transmission and critical cases include respiratory failure, septic shock, and/or multiple organ dysfunction or failure (Cascella *et al.* 2020). Predictably, the virus has put colossal strain on health services worldwide in terms of the provision of appropriate numbers of medical staff to treat patients, hospital beds, medicines, and necessary breathing apparatus along with essential personal protection equipment (PPE).

Against this background has been the emergence of forecasting models focusing on factors such as the number of deaths, hospitalizations and infections for future periods, models which are essential for the implementation of measures to control the virus and for the provision of medical and associated services. Many of these models use cumulative quantiles to provide prediction intervals for a range of possible probability values. For example, the COVID-19 Forecasting Hub (2020) provide quantile forecasts from a range of forecasting organizations in relation to the US and US states. Similarly, the *Los Alamos National Laboratory*, *LANL*, forecasting model provides quantile predictions on the number of confirmed cases and deaths for the US as a whole and US states as well as many other countries worldwide (LANL, 2020).

A major difficulty for the evaluation of probability-based quantile forecasts is that there are no directly available actual probabilities at the end of the forecast period with which the forecasts can be directly compared. While it is relatively easy to compare point predictions with the resulting real values, this option is not available for quantile probability forecasts. In this situation, it is desirable to consider the form of probability distribution relevant to the values of the variable of interest and apply this distribution to obtain *empirical quantiles* that can be used in the evaluation of the predictive performance of a set of quantile forecasts.

Evaluating the accuracy of quantile probability predictions typically involves examining the nominal coverage over a set of forecasts. This encompasses using the proportion of times the ex-post value of the variable falls within the specified lower and upper quantile probability interval, which is termed the *empirical coverage*. These probability intervals can be equal or unequal: for instance, a 0.25 probability interval using quartiles, or quantiles with probabilities 0.025, 0.5 and 0.975. There is, however, a trade-off between the interval’s forecast length and its coverage (Granger *et al.* 1989). It is also important that the quantile forecasts meet the Christoffersen (1998) condition that they have good reliability or *calibration*. To evaluate calibration, it is best to have intervals that have the shortest length. Good calibration requires that the nominal and empirical coverage measures be closely correlated for all the quantile probability intervals. There have been some studies that have used theoretical and applied approaches in the evaluation of probability quantile forecasts. These have been mainly based on research by Gneiting *et al.* (2007) and Gneiting and Rafferty (2007). Such studies stress that the aim of quantile forecasting should be to obtain the best *sharpness* (which is related to the concentration of forecasts) with the best *calibration*. Bracher *et al.* (2020) demonstrated how the procedures, set out in Gneiting *et al*. (2007) and Gneiting and Rafferty (2007), could be used in the context of examining epidemic (COVID-19) quantile forecasts. The COVID-19 Forecasting Hub (2020) uses a measure termed the *Weighted Interval Score* (*WIS*), as described in Bracher *et al.* (2020). The *WIS* is a relatively straightforward measure that does not require assumptions about the statistical distribution. However, an important problem with this measure is that its use tends to require a

large number of forecasts, whereas when evaluating quantile forecasts of COVID-19 deaths, there is usually only a limited number of observations.

The issue of the insufficiency of observations in this context was recently emphasized by Petropoulos and Makridakis (2020), who pointed out that forecast accuracy is essentially contingent upon the availability of data to base predictions and assessments of uncertainty; but in the case of pandemics, there tends to be limited data points. Procedures, such as the *WIS,* can also be adversely affected by the underlying statistical distribution of such series, which exhibit gradually changing parameters over short periods of time. A feature of the distribution of changes in the number of COVID-19 deaths is that it is influenced by a range of factors that cause the mean and standard deviation to fluctuate over short periods of time, reflecting environmental influences such as changing government restrictions, altering attitudes that impact human behavior, advances in medical developments and the availability of medical provisions.

The first objective of this demonstration was to provide an innovative framework that can evaluate quantile forecast performance with a limited number of observations on changes in the number of COVID-19 deaths and that considers its statistical distribution. To achieve this, the *empirical probability* technique (e.g., Pollock *et al.* 2005; Pollock *et al*. 2008, Pollock *et.al.* 2010; Thomson *et al.* 2003; Thomson *et al.* 2004; Thomson *et al.* 2013), is extended in the present paper to obtain estimated quantiles. The procedure requires the series under consideration to have first differences, over short periods, which are approximately independently and identically normally distributed. Under these conditions, empirical ex-post quantile values, at each time-period, can then be used to evaluate the performance of quantile forecasts. The empirical quantile technique uses the Student t distribution at each time-period. Daily changes of the actual series are assumed to be approximately independent and normally distributed, with a time-varying mean and standard deviation, but with approximately stable parameters over short periods of one week. This allows estimated sample standard deviations to be obtained for each weekly period that can be used, with the weekly mean daily change, to give an estimated probability distribution that allows the empirical quantiles to be derived for each cumulative quantile probability.

The second objective was to extend available forecast evaluation methods to quantile probability predictions to examine overall accuracy and specific underlying aspects of accuracy. The performance of point forecasts is usually measured by a numerical value, or scoring rule, using a set of forecasts and numerical outcomes, such as the *Mean Squared Error* (*MSE*). In this study the *Mean Squared Quantile Score* (*MSQS*) is used, which is related to the *MSE,* to provide measure that can be applied to the forecast quantiles, with empirical quantiles used in place of the actual values. In addition, the *MSQS* is decomposed to allow identification of specific aspects of performance that can be used to highlight strengths and weaknesses of the forecasts. This decomposition follows a form originally adapted from Yates' (1982) covariance approach decomposition of the Mean Probability Score, previously utilized to evaluate point and probability forecasts in various financial contexts (e.g., Pollock and Wilkie, 1996; Thomson *et al*. 2003, 2004, 2013; Wilkie and Pollock, 1996; Wilkie-Thomson *et al.* 1997). In the present study, the approach is extended to analyze quantile forecasts. The *MSQS* is the sum of three component measures termed: *bias squared* (a squared measure of calibration or bias); *resolution variation* (the variation in the actual values multiplied the square of unity minus a *resolution* term); and *error variation* (the variation of the actual values that is not explained by variation in the forecasts).

These performance measures use empirical quantiles derived in probability form compatible with probability density forecasts. The empirical quantiles take advantage of the precise values of the data and their approximate statistical distribution, which allows stable performance measures for a given set of quantile forecasts. Empirical quantiles also allow the *MSQS* and its component measures to be analyzed separately at each cumulative quantile probability. As forecast and empirical quantile values can be directly compared, narrow quantile probability intervals do not require an increased number of observations. The width of the quantile probability intervals and the number of quantiles have no direct impact on the calculation of the *MSQS* and its components at each cumulative quantile probability.

In the current demonstration, the framework is applied to weekly quantile cumulative forecasts in relation to changes in the number of confirmed deaths provided by the *Los Alamos National Laboratory* (LANL) for the US. The LANL forecasting model makes these cumulative

quantile probability predictions from one day ahead to over six weeks using the median and 22 quantiles (LANL, 2020). This study specifically examined 17 one-week ahead forecasts made each Wednesday for forecast dates from 08/07/2020 to 28/10/2020. These were evaluated using daily data from 02/06/2020 to 28/10/2020 employing the framework set out above. Empirical quantiles were derived and then used to examine the overall and specific aspects of the model's performance in this period and to undertake a comparison with similar quantile probability forecasts derived from a simple *Autoregressive Order One*, *AR(1)* model. The AR(1) model was selected because it was considered to be the most suitable naïve model to use in the circumstances for comparison purposes.

The third objective of this study was to consider if forecast performance could be improved by combining the LANL model predictions with another simple model to obtain composite forecasts. For this, predictions from the simple *AR(1)* model, without a constant, were employed, over the period under consideration. The importance of combining point forecasts as a means of increasing the forecast performance of individual predictions has received considerable attention in the literature. Composite forecasts have been shown to improve forecast accuracy and reduce the variance of errors (e.g., Armstrong, 2001; Armstrong *et al.* 2015; Cramer *et al.* 2021; Goodwin, 2015; Graefe *et al.* 2014; Green *et al.* 2015; Harvey, 2001; Lubecke *et al.,* 1995; Ray *et al.* 2020; Reich *et al.* 2019; Thomson *et al.* 2019). This aspect of composite forecasts is especially beneficial when individual forecasts are based on dissimilar information sets (Wallis, 2011), as there is likely to be relative inconsistencies between the individual predictions that make up the composite forecasts. It was shown in Thomson *et al.* (2019) that forecast performance using the *MSE* and component performance measures for point forecasts can be improved by combining forecasts. In the present study, this is extended to a set of cumulative quantile predictions with respect to each quantile probability using the *MSQS*. In the formation of composite forecasts, the simple arithmetic average has been shown to be the most popular method. Evidence for point estimates, in fact, suggests that it is difficult to outperform this simple average (Schnaars, 1986; Clemen, 1989; Makridakis and Hibon, 2000; Stock and Watson, 2004). The simple arithmetic average has the advantage of impartiality and robustness and frequently produces good results (Clemen, 1989; De Menezes *et al*. 2000). In addition, simple averages do not require estimation of covariances across model errors (Timmerman, 2006), and so they are easy to use in practice. Accordingly, the simple arithmetic average is used in the current study, but this is extended to the application of composite quantile forecasts derived at each cumulative quantile probability. It is illustrated that the framework used in this study, based on empirical quantiles and the *MSQS* and component measures, can be extended to evaluate cumulative quantile forecast performance for the composite forecasts that illustrate the importance of incoherence (inconsistency) between the two sets of forecasts, thus extending the analysis set out in Thomson *et al*. (2019). There is evidence that composite models can improve virus forecasting performance. Reich *et al.* (2019), in the context of infectious disease, showed that collaborative efforts between research teams to develop composite or ensemble forecasting approaches can bring measurable improvements in forecast accuracy and important reductions in the variability of performance. Cramer *et al.* (2021) and Ray, et al. (2020) found that COVID-19 ensemble forecasting models showed the best overall accuracy of any model. These results emphasize the role that collaboration and active coordination between forecasting organizations can play a vital role in developing the modeling capabilities to support the responses of decision makers to pandemic outbreaks. However, to date, studies in this context have not accounted for the superior accuracy enhancement that can result from purposefully combining diverse forecasts.

To sum up, the fulfilment of these objectives makes several important contributions to the forecasting literature and to practice. First, as detailed above, the framework fills a research gap in terms of the need to develop a method of evaluating quantile forecasts with a limited number of data points. Second, it extends existing quantile forecasting evaluation methods in a manner that allows the identification of specific strengths and shortcomings in performance, which, in turn, can enable forecasting practitioners to improve their overall prediction accuracy. Third, the technique demonstrates, and can account for, the superior accuracy that can result from

deliberately combining dissimilar forecasts, which, again, is likely to be extremely useful to forecasting practitioners.

The remainder of the paper is set out as follows. Section 2 provides a description of the statistical forecast analysis using empirical quantiles. Section 3 sets out the statistical measures of performance. Section 4 explains the *AR(1)* and composite model specifications and analysis. Section 5 describes the data used, their background and statistical characteristics. This is followed by Section 6, which presents the results of the demonstration. Finally, the discussion is presented in Section 7 and concluding observations are provided in Section 8.

2. Statistical analysis using empirical quantiles

To examine quantile forecasts for changes in the number of deaths, empirical quantiles were derived from the actual values to allow the forecasts to be evaluated. The framework uses the empirical probability technique set out in, for example, Pollock *et al.* (2008), which is extended to quantile forecasts. In this section, the derivation of empirical quantiles is described.

When examining cumulative quantile forecasts, it is usually convenient to examine changes or first differences in the variable under consideration, rather than the actual values. To obtain weekly estimated quantiles, it is necessary to consider the actual changes in a variable, $\Delta X_{j,i} = X_{j,i} - X_{j,i-1}$ for day, i, $i=1,2,\ldots,n$, of the series over a weekly period, j, $j=1,2,\ldots,k$. For a week of 7 days, $n=7$, the sum of the actual daily changes, over the weekly period, is simply the change in the variable for the whole week, that is $\sum_{i=1}^{n} \Delta X_{j,i} = X_{j,n} - X_{j,0}$. It is further considered that the changes are approximately independently and identically normally distributed with stable means and standard deviations over the period, i, $i=1$ *to* n, for week, j. The procedure used to obtain empirical quantiles is set out below.

The mean values, m_j, of the changes, $\Delta X_{j,i}$, for week, j, $j=1,2,\ldots,k$, was calculated, as defined in equation (1):

$$m_j = \frac{1}{n}\sum_{i=1}^{n} \Delta X_{j,i} \tag{1}$$

The volatility of the changes, $\Delta X_{j,i}$, for week, j, was calculated using the standard deviations, s_j, as defined in equation (2):

$$s_j = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} (\Delta X_{j,i} - m_j)^2} \tag{2}$$

To obtain empirical quantiles, $d_{\alpha,j}$, for a specific quantile cumulative probability, α, where $\alpha = 1,2,\ldots,h$, and α represents the cumulative probability that the weekly change in the actual value, $\Delta X_{j,i}$, would be below the value $d_{\alpha,j}$, is defined in equation (3):

$$d_{\alpha,j} = n\left\{m_j + F_{n-1}^{-1}(\alpha) * \frac{S_j}{\sqrt{n}}\right\} \tag{3}$$

where $F_{n-1}^{-1}(\alpha)$ is the inverse cumulative distribution function of the Student t-distribution with $n-1$ degrees of freedom for the quantile probability, α.

There is, however, one further modification required when changes in the variable under consideration cannot take negative values, specifically when the first differences of the actual values of the changes in the variable follow a normal distribution with a left truncation at zero. Therefore, a left truncated Student t distribution can be considered more appropriate, which would tend to have a higher mean and smaller standard deviation than the non-truncated Student t distribution. In this study, to evaluate the impact of truncating on the distribution, the cumulative probabilities that a value for a non-truncated t distribution, for week, j, with mean, m_j, and standard deviation, s_j, is used to give a value $d_{0,j}$, as presented in equation (4):

$$d_{0,j} = F_{n-1}\left\{\sqrt{n} * \left(-\frac{m_j}{s_j}\right)\right\} \quad (4)$$

where F_{n-1} denotes the cumulative distribution function of the Student t-distribution with *n-1* degrees of freedom.

Equation (3) can be modified to consider that the distribution could be truncated at zero to give adjusted empirical quantile values, $e_{\alpha,k}$, as set out in equation (5):

$$e_{\alpha,j} = n\left\{m_j + F_{n-1}^{-1}\left[\alpha\left(1 - d_{0,j}\right) + d_{0,j}\right] * \left(\frac{S_j}{\sqrt{n}}\right)\right\} \quad (5)$$

where F_{n-1}^{-1} is the inverse cumulative distribution function of the Student t distribution with *n-1* degrees of freedom.

These weekly empirical quantile values, $e_{\alpha,j}$, can be directly compared with the weekly forecast quantile values, $q_{\alpha,j}$. For each week *j*, the forecast and empirical quantile distributions, using the $q_{\alpha,j}$ and $e_{\alpha,j}$ values, over the range of values of α, can be used to compare the two distributions. The empirical quantile value at quantile probability 0.5, $e_{0.5,j}$, is the same as the actual change. The forecast change at quantile probability 0.5, $q_{0.5,j}$, represent the median forecast change, which can also be regarded as the point forecast change. Therefore, the forecast and the actual forecast change, $q_{0.5,j}$ and $e_{0.5,j}$ respectively, for week *j*, provide a direct evaluation of the accuracy of the median/point forecast.

The following section describes the derivation of performance measures used in the framework.

3. Performance measures

To evaluate performance, the forecasts, $q_{\alpha,j}$, at each cumulative quantile probability, α are compared with the empirical quantile values, $e_{\alpha,j}$. This is carried out using the LANL model forecasts, the AR(1) model forecasts and the composite forecasts. The simple AR(1) model is used in this study as a simple model of comparison that incorporates weekly positive autocorrelation which is highlighted in the following section. The AR(1) model used did not include a constant and was based on estimation that required using actual weekly changes to provide the quantile probability estimates. The composite model used a simple average of the LANL and AR(1) forecasts at each quantile.

3.1. The mean squared quantile score

The overall performance of a set of forecasts for each quantile cumulative probability, α, is measured by the *mean squared quantile score* ($MSQS_\alpha$), which is the average of the squared forecast errors, where the forecast error is measured as the forecast quantile value minus the empirical quantile value. The $MSQS_\alpha$ is defined in equation (6):

$$MSQS_\alpha = \frac{1}{k}\sum_{j=1}^{k}(q_{\alpha,j} - e_{\alpha,j})^2 \quad (6)$$

A value of zero would imply that forecast quantile values are identical to the empirical values (indicating perfect accuracy); hence, the higher the value of the $MSQS_\alpha$, the poorer the forecast performance.

The $MSQS_\alpha$ is an overall performance measure which can be decomposed to identify specific components that reflect the multidimensional aspects of accuracy. The decomposition used in the present study involves *bias squared* (BS_α), *resolution variation* (RV_α) and *error variation* (EV_α). The $MSQS_\alpha$ decompositions are presented in equation (7):

$$MSQS_\alpha = BS_\alpha + RV_\alpha + EV_\alpha \quad (7)$$

These three components are discussed next.

3.2. Bias squared

Bias (B_α) or calibration is measured by the difference in the mean of forecast quantile values at α, $M(q_\alpha)$ and the mean of the empirical quantiles values, $M(e_\alpha)$. *Bias* occurs when the mean quantile forecast value at α is either too low, reflecting under-estimation ($B_\alpha<0$), or too high, reflecting over-estimation ($B_\alpha>0$), of the empirical quantile values. The *bias squared* value is simply the square of *bias* and is a specific component of the $MSQS_\alpha$ decomposition. A value on this measure of zero implies no bias.

Bias squared (BS_α) at α is defined in equation (8):

$$BS_\alpha = B_\alpha^2 \tag{8}$$

where $B_\alpha = M(q_\alpha) - M(e_\alpha)$, $M(q_\alpha) = \frac{1}{k}\sum_{j=1}^{k} q_{\alpha,j}$ and $M(e_\alpha) = \frac{1}{k}\sum_{j=1}^{k} e_{\alpha,j}$

3.3. Resolution variation

The resolution variation component of performance RV_α is related to *resolution* or *slope* (SL_α), which is a measure of discrimination that reflects the ability to detect and make appropriate adjustments of the correct size, compatible with the empirical quantile change. *Resolution* is measured by the slope coefficient that relates to a linear relationship between the forecast and the actual values. *Resolution* is particularly important in situations when it is essential to correctly identify the size of the change to allow discrimination between large and small movements. This is a critical aspect of performance that reveals the forecaster's level of expertise. A value of unity indicates perfect resolution.

The *resolution variation* component of performance (RV_α) is the square of unity minus the *resolution* or *slope* (SL_α) term, multiplied by the variance of the empirical quantiles, $V(e_\alpha)$. As *resolution* or *slope* approaches unity, RV_α approaches zero.

Resolution variation for performance (RV_α) is defined in equations (9):

$$RV_\alpha = (1 - SL_\alpha)^2 V(e_\alpha) \tag{9}$$

where $SL_\alpha = C(q_\alpha, e_\alpha)/V(e_\alpha)$ and $V(e_\alpha) = (\frac{1}{k}\sum_{j=1}^{k} e_{\alpha,j}^2) - M(e_\alpha)^2$

$C(q_\alpha, e_\alpha) = (\frac{1}{k}\sum_{j=1}^{k} q_{\alpha,j}\, e_{\alpha,j}) - M(q_\alpha)\, M(e_\alpha)$

3.4. Error variation

The error variation for performance (EV_α) is variation in the forecast quantile values that is not explained by variation in the empirical values. Error variation is measured by the scatter (SC_α) term. Error variation can arise when forecasters use diverse strategies in forming their predictions or identify patterns in the series that are not relevant. A value of zero indicates no error variation.

Error variation for performance (EV_α) is defined in equation (10):

$$EV_\alpha = SC_\alpha = V(u_\alpha) \tag{10}$$

where $u_{\alpha,j} = q_{\alpha,j} - A_\alpha - SL_\alpha\, e_{\alpha,j}$ and $V(u_\alpha) = V(q_\alpha) - SL_\alpha^2\, V(e_\alpha)$

$A_\alpha = M(q_\alpha) - SL_\alpha\, M(e_\alpha)$

$V(q_\alpha) = (\frac{1}{k}\sum_{j=1}^{k} q_{\alpha,j}^2) - M(q_\alpha)^2$

4. The AR(1) and composite models

When analyzing performance, it is useful to compare the LANL model forecasts with an appropriate simple model. In this study, the AR(1) model was used to provide a comparison benchmark to highlight possible forecasting strengths and weaknesses of the LANL model

forecasts. The AR(1) model incorporates the weekly first order positive autocorrelation in the changes in the number of deaths highlighted in the following section. It is also appropriate to consider if combining this model with the LANL model can produce more accurate composite forecasts and improve on forecast performance. While it is relatively easy to apply the analysis set out above to the LANL model, where the quantile predictions are directly available for each week, the predictions for the AR(1) and composite models need to be generated so that overall accuracy and its components can be examined at each quantile probability.

4.1. The AR(1) model

The AR(1) is used, without a constant, based on the estimation from the current and previous weekly actual changes, which are used to make one-week ahead forecasts. This AR(1) model generates a phi value (φ_j) for each week *j,* which is used to give a one-week ahead point or median forecast at *j+1*, denoted $a_{0.5,j+1}$, where $a_{0.5,j+1} = \varphi_j e_{0.5,j}$, with φ_j defined for a *T* week period in equation (11):

$$\varphi_j = \frac{\sum_{i=0}^{T-2}(e_{0.5,j-i} * e_{0.5,j-1-i})}{\sum_{i=0}^{T-2}(e^2{}_{0.5,j-1-i})} \quad (11)$$

A moving 10-week period (*T=10)* is used in this study via data for weeks *j-9* to *j* to calculate the quantile forecast value at *j+1*. Performance measures for median/point quantile cumulative probability, $\alpha=0.5$, that is, $a_{0.5,j}$, can be directly compared with the actual values, $e_{0.5,j}$. Values of phi, (φ_j), are interpreted as an indicator measure as to whether the weekly change in the number of deaths is increasing ($\varphi_j>1$), or decreasing ($\varphi_j<1$). The AR(1) model is also used in this study to generate quantile predictions for each of the 22 quantile probabilities at α, $a_{\alpha,j+1}$, using the assumption that the error terms are independently and identically normally distributed. The quantile predictions at α, $a_{\alpha,j+1}$, are defined for a *T* week period in equation (12):

$$a_{\alpha,j+1} = \varphi_j a_{0.5,j} + \{F_{T-2}^{-1}(\alpha) * \sqrt{\left[\frac{\sum_{i=0}^{T-2}(e_{0.5,j-1} - \varphi_{j-1} e_{0.5,j-1-i})^2}{T-2}\right]}\} \quad (12)$$

where F_{T-2}^{-1} is the inverse cumulative distribution function of the Student t distribution with *T-2* degrees of freedom.

4.2. The composite model

Composite forecasts for each week, *j,* and each quantile probability α, $c_{\alpha,j}$, were obtained by taking a simple average of the LANL model forecast, $l_{\alpha,j}$, and the AR(1) model forecast, $a_{\alpha,j}$, for each quantile probability, α, and each week, *j*, that is, $c_{\alpha,j}=(l_{\alpha,j}+a_{\alpha,j})/2$.

It is important to consider whether the LANL model quantile forecast performance for each of the quantile probabilities can be improved by integrating its predictions into composite forecasts using the AR(1) model forecasts. To do this, it is also essential to understand the role of coherence between the LANL and AR(1) model forecasts at each quantile probability. It can be shown that the composite *MSQS* and its components between two sets of forecasts is a function of the average of two forecast measures minus a measure of paired coherence between the two divided by four (Thomson *et al*. 2019). This can be easily extended to the *MSQS*. Coherence measures compatible with the *MSQS* and its components can, therefore, easily be obtained from the *MSQS* of the LANL, AR(1) and composite models*,* which can be extended to the components of the *MSQS*. These paired coherence measures between the LANL and AR(1) models can be defined as the *Mean Squared Quantile Score for Coherence*, $MSQSC_{\alpha,L,A}$, *Bias Squared for Coherence*, $BSC_{\alpha,L,A}$, *Resolution Variation for Coherence*, $RVC_{\alpha,L,A}$, and *Error Variation for Coherence*, $EVP_{\alpha,L,A}$. These are defined in equations (13a to 13d):

$$MSQSC_{\alpha,L,A} = 4\left\{\left[\frac{MSQS_{\alpha,L} + MSQS_{\alpha,A}}{2}\right] - MSQS_{\alpha,C}\right\} \quad (13a)$$

$$BSC_{\alpha,L,A} = 4\left\{\left[\frac{BS_{\alpha,L} + BS_{\alpha,A}}{2}\right] - BS_{\alpha,C}\right\} \quad (13b)$$

$$RVC_{\alpha,L,A} = 4\left\{\left[\frac{RV_{\alpha,L} + RV_{\alpha,A}}{2}\right] - RV_{\alpha,C}\right\} \quad (13c)$$

$$EVC_{\alpha,L,A} = 4\left\{\left[\frac{EV_{\alpha,L} + EV_{\alpha,A}}{2}\right] - EV_{\alpha,C}\right\} \quad (13d)$$

The subscripts *L, A* and *C* relate to the LANL, AR(1) and composite models, respectively.

A value of zero indicates perfect coherence. The higher the value, the lower the coherence between the two sets of forecasts at the quantile probability, *α*. When combining forecasts, it is desirable to have forecasts that show low coherence and good performance.

The relative percentage improvement of composite forecasts, compared with the average of the individual forecasts on a particular measure, can be used to obtain a dimensionless measure of coherence. The *Relative Percentage* improvement performance of the composite forecasts ($RPMSQS_{\alpha,C}$, $RPBS_{\alpha,C}$, $RPRV_{\alpha,C}$, and $RPEV_{\alpha,C}$) can be used to compare the composite performance measures ($MSQS_{\alpha,C}$, $BS_{\alpha,C}$, $RV_{\alpha,C}$, and $EV_{\alpha,C}$) with the mean of the LANL ($MSQS_{\alpha,L}$, $BS_{\alpha,L}$, $RV_{\alpha,L}$, and $EV_{\alpha,L}$) and the AR(1) performance measures ($MSQS_{\alpha,A}$, $BS_{\alpha,A}$, $RV_{\alpha,A}$ and $EV_{\alpha,A}$). These are defined in equations (14a to 14d):

$$RPMSQS_{\alpha,C} = \left\{1 - \frac{MSQS_{\alpha,C}}{\frac{MSQS_{\alpha,L} + MSQS_{\alpha,A}}{2}}\right\} * 100 \quad (14a)$$

$$RPBS_{\alpha,C} = \left\{1 - \frac{BS_{\alpha,C}}{\frac{BS_{\alpha,L} + BS_{\alpha,A}}{2}}\right\} * 100 \quad (14b)$$

$$RPRV_{\alpha,C} = \left\{1 - \frac{RV_{\alpha,C}}{\frac{RV + RV_{\alpha,A}}{2}}\right\} * 100 \quad (14c)$$

$$RPEV_{\alpha,C} = \left\{1 - \frac{EV_{\alpha,C}}{\frac{EV_{\alpha,L} + EV_{\alpha,A}}{2}}\right\} * 100 \quad (14d)$$

These measures show the relative percentage improvement of the composite measures for the *MSQS* and its components. The higher the value, the lower the relative coherence between the two sets of forecasts at the quantile probability, *α*. In the discussion below, the results for these coherence measures are integrated into the results of the performance measures.

5. The data and their statistical characteristics

The framework set out in Section 2 is applied to provide empirical quantiles for weekly changes on the number of deaths from COVID-19 for the US. The data and their characteristics are described below.

5.1. The data

In the application of quantitative analysis to data on changes in the cumulative number of deaths from COVID-19 time series, it is important to understand the nature of the data. The limited testing and the challenges in determining the reasons for the causes of death implies that the number of confirmed deaths may not represent the true number of deaths and a time lag exists from

symptoms to death that range from two to eight weeks for COVID-19 (Richie, 2020). There would also be an additional short lag arising from the time between a death occurring and its recording. Therefore, the number of reported deaths would be, to a degree, less than the actual number of deaths at given point in time. This needs to be considered when evaluating past and current figures and in making future predictions. It should also be pointed out that the number of recorded deaths involving COVID-19 reflect deaths where the virus was a contributing factor, not necessarily the exclusive cause (Banerjee *et al.* 2020). For instance, in their reporting of deaths due to COVID-19, the *Centres for Disease Prevention and Control*, CDC (2020a) include estimates of excess deaths involving cases where COVID-19 was not mentioned on the death certificate as the specific cause of death but where it was a co-morbid and likely contributing factor, in addition to deaths where COVID-19 is noted as the cause. There is, therefore, not only uncertainty in the path of the natural number of deaths, but also uncertainty in its measurement.

The framework, set out above, is applied to weekly quantile cumulative forecasts on the number of confirmed deaths, in the US, provided by the *Los Alamos National Laboratory*, LANL (2020), the AR(1) model and composite forecasts during a period from 02/07/2020 to 28/10/2020 with weeks numbered 1 to 17. Specifically, the forecasts relate to the cumulative number of confirmed US deaths reported by *Johns Hopkins University* (*JHU*) *Coronavirus Research Centre* which provides the data measurements that the forecasts generated by the LANL COVID-19 Forecasting Model, the AR(1) model and the composite forecasts can be validated against. The LANL model can be described as a statistical dynamic growth model that considers susceptibility of the population to COVID-19. It is a probability-based model designed to allow for uncertainty in the future path of the cumulative number of deaths from COVID-19. The LANL model assumes that interventions currently in operation will continue. The outputs from LANL include ex-post actual values as well as its cumulative quantile forecasts. LANL also provides quantile forecasts on the cumulative number of confirmed deaths for most countries, and all US states, from one day ahead to over six weeks ahead using their forecasting model. Their results present forecasts using the median and 22 quantiles. The cumulative quantile probabilities, α, are as follows: 0.01, 0.025, 0.05, 0.1, 0.15, and in units of 0.5 to 0.9, 0.95, 0.975 and 0.99. The data are updated twice a week and presented in a form compatible with Microsoft Excel. This study uses weekly forecast periods ending on Wednesday for consecutive weeks.

5.2. Statistical characteristics of the data

The time series of the changes in the number of deaths for weekly periods ending from 18/03/2020 to 28/10/2020 are presented in Figure 1, with weeks numbered from -15 to 17. The LANL model forecasts are also presented on quantiles 0.025, 0.15, 0.5, 0.85 and 0.975, with the weeks when the LANL model provided forecasts being numbered from -8 to 17. The changes the US number of deaths showed typical characteristics of data on epidemics and pandemics. There was a rapidly increasing number of deaths in the initial stage with the cumulative deaths showing exponential growth with the size of the changes related to the actual levels of the number of deaths. This first phase was reflected in the weekly changes that occurred which showed a rapid increase from 82 in week -15 (18/03/2020) to 18,284 in week -10 (22/04/2020). The actual cumulative number of deaths did not remain exponential after the initial phase had past, and the rate of growth slowed down with the changes reaching a turning point at week -10. This was accompanied by measures to control the disease that gradually took effect. The rate of growth eventually slowed and the actual changes in the number of deaths showed a marked fall after week -10. This second phase was reflected in the weekly changes that occurred until week 0 (01/07/2020), which had a value of 3,657. In the third phase, the actual changes stabilize and showed relatively smaller rises and falls. This phase occurred from week 1 onwards to week 17 (28/10/2020), when the value was 5,509. The third phase is characterized by restrictions being subject to adjustments and medical advances influencing the number of deaths. The performance analysis of the LANL model forecasts undertaken in this study was restricted to this phase. The LANL model was updated from version one to version two after this period, so this provides a convenient end point. This third phase was probably the most difficult of the three phases to forecast as the series showed a clear turning point, but it was considered that this was appropriate

period to demonstrate the framework set out in this study. It also needs to be pointed out that, given aggregate US data was used in this study, the pattern of deaths varied considerably across the different US states. For instance, for week -7 (13/05/2020) and below, 40% to 55% of the deaths were explained by two states, New Jersey and New York and in week 3 (22/07/2020) to week 16 (21/10/2020), 25% to 50% of deaths were explained by three states, California, Florida and Texas. LANL provide forecasts for all US states, although these were not directly used in this study.

The daily adjusted changes in the number of confirmed deaths published with the LANL forecasts (LANL, 2020) were used to obtain the weekly empirical quantiles. This involved using these changes in the number of deaths for each day for 17 consecutive weeks (weeks 1 to 17) ending on Wednesdays. There was, however, an issue regarding the recorded deaths for the different days of the week. While it would be expected that the natural occurrence of changes in the number of deaths would be similar for each day, over a one-week period, the recorded changes in deaths reflect administrative factors associated with the death notifications and recording.

The data were particularly affected by weekend effects. The recorded values for Sunday and Monday, which largely reflect the occurrence of deaths that occurred on Saturday and Sunday, respectively, due to a lag between recorded deaths and announcements, were approximately a half of the values of the other five days of the week. A simple adjustment was made to consider this by using the combined values for Sunday and Monday as one day, resulting effectively in a six-day week. These six values were used in calculating the weekly empirical quantiles. The mean values of the six revised daily values for the 16-week period prior to the forecast period were relatively similar and ranged from 1019 (Saturday) to a value of (1435) Thursday. The mean values for the whole 33-week period were also relatively similar for the six revised daily values which ranged from 910 (Saturday) to a value of 1234 (Wednesday). A one-way ANOVA supported the assumption of the equality of the daily means with the F statistic clearly non-significant in both cases. In the case of the 16-week period, all the paired t-test values were non-significant and in the 33-week period, all except one of the paired Student t test values showed non-significant for the pairs of values. This was the Wednesday / Saturday pair. There was a tendency for values on Saturday to be lower than other days and the results partly reflect the low value of 242 that occurred on 04/07/2020 due to the Friday 03/07/2020, Independence Day holiday. Levene's (1960) test for inequality of variance also clearly showed non-significance in both cases. Lilliefors (1967) test for normality was applied to the six daily changes for each of the 33-weekly periods with all the statistics showing non-significance, except week 13 (30/09/2020) which was significant at 5%. One from 33 is slightly less than would be expected by chance. The results did not show evidence of overall autocorrelation, for the six daily changes for each of the 33-weekly periods. The results also show that the impact of adjusting the empirical distribution, to take account of the changes in the cumulative number of deaths being restricted to non-negative values, had little effect on the results. The probability of the unadjusted distribution having a value below zero was below 0.0005 for the 16 values used to calculate the empirical quantile values, the only exception being for week 1 (08/07/2020), which had a value of 0.0013, which reflected the Independence Day holiday effect.

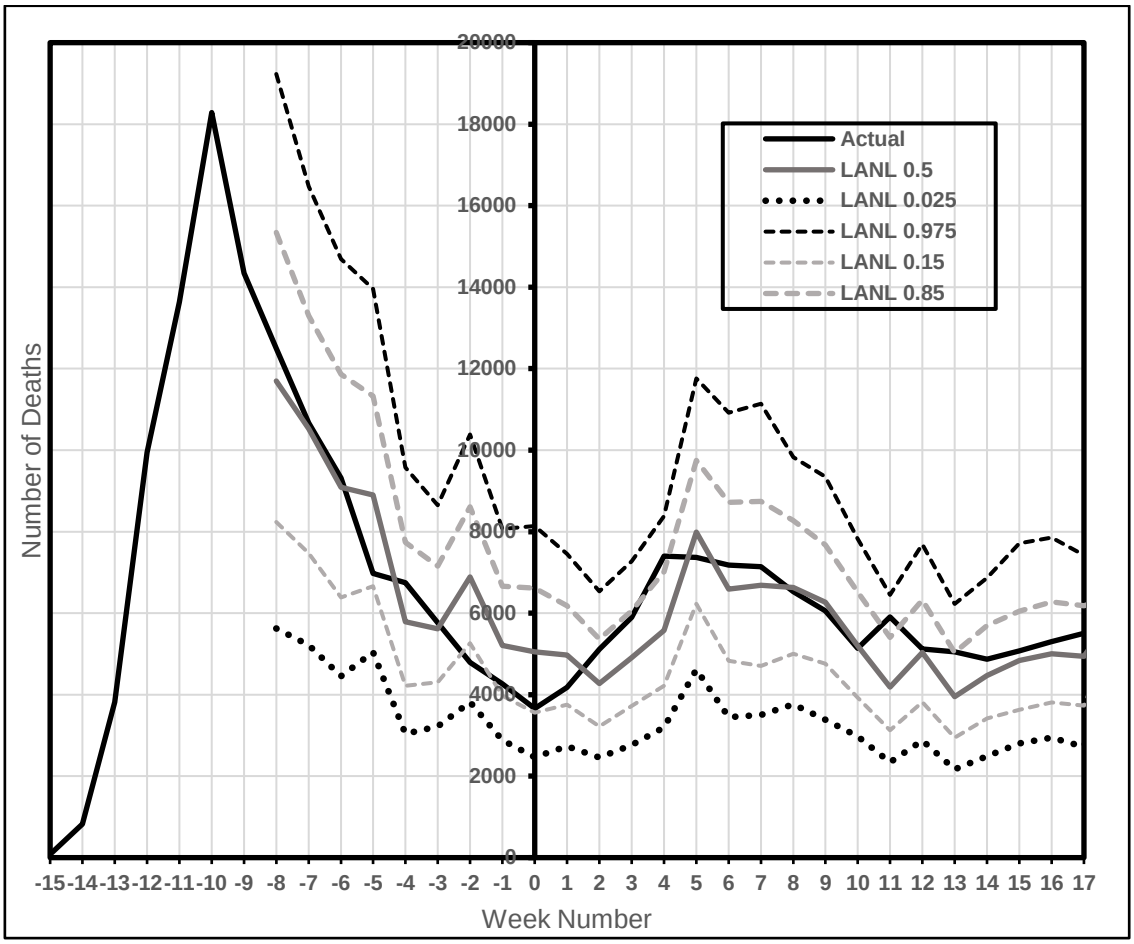

Figure 1. Weekly actual values and LANL forecasts for quantiles 0.025, 0.15, 0.5, 0.85 and 0.975

These statistics support the use of the six daily changes and the assumption that these changes approximately follow a normal distribution over the week, which are used to obtain the mean and standard deviation estimates to derive the empirical quantiles.

The analysis using the accuracy measures was restricted to the daily period from 02/07/2020 to 28/10/2020, for forecast weeks ending 08/07/2020 to 28/10/2020, numbered 1 to 17, relevant to the LANL forecasts in the third phase of the pandemic. Table 1 presents the dates, week numbers, means and standard deviations and the weekly changes in the number of deaths, as well other statistics that are discussed in the next section. The means presented in Table 1 are essentially the weekly change divided by 6, therefore they show essentially an identical pattern to the weekly changes. The mean values increased from week 1 to week 4 and then showed a general decline to week 14, after which the mean values showed a small rise. The standard deviation was also variable. The relatively large standard deviation of 311 in week 1 reflects the relatively small value of 242 on Saturday 04/07/2020 and the relatively large value of 1195 on Tuesday 07/07/2020, a few days later that was likely to be affected by the Independence Day holiday. The results tend to support the view that the parameters of the underlying distribution were subject to changes from one week to the next.

Table 1. Weekly statistics and median / point forecasts

Date	Week	Daily	Daily	Weekly	AR(1)	AR(1)	LANL	Comp	AR(1)	LANL	Comp
	No.	Mean	SD	Change	Phi	F'cast	F'cast	F'cast	F'cast	F'cast	F'cast
									% Error	% Error	% Error
08/07/2020	1	697	311	4182	0.859	3134	4975	4054	-25.1	19.0	-3.0
15/07/2020	2	853	116	5115	0.877	3592	4270	3931	-29.8	-16.5	-23.1
22/07/2020	3	984	132	5901	0.901	4485	4911	4698	-24.0	-16.8	-20.4
29/07/2020	4	1233	230	7397	0.999	5316	5577	5447	-28.1	-24.6	-26.4
05/08/2020	5	1229	166	7371	1.004	7389	7986	7687	0.2	8.3	4.3
12/08/2020	6	1197	178	7180	1.023	7403	6590	6996	3.1	-8.2	-2.6
19/08/2020	7	1191	166	7143	1.040	7346	6682	7014	2.8	-6.5	-1.8
26/08/2020	8	1089	132	6531	1.031	7427	6633	7030	13.7	1.6	7.6
02/09/2020	9	1009	87	6054	1.027	6731	6266	6498	11.2	3.5	7.3
09/09/2020	10	857	279	5139	1.005	6217	5197	5707	21.0	1.1	11.1
16/09/2020	11	984	228	5904	1.005	5165	4186	4675	-12.5	-29.1	-20.8
23/09/2020	12	854	181	5121	0.983	5933	5038	5486	15.9	-1.6	7.1
30/09/2020	13	841	149	5048	0.958	5032	3950	4491	-0.3	-21.7	-11.0
07/10/2020	14	812	99	4869	0.952	4835	4467	4651	-0.7	-8.3	-4.5
14/10/2020	15	845	158	5071	0.955	4637	4834	4735	-8.6	-4.7	-6.6
21/10/2020	16	884	139	5304	0.956	4844	5005	4924	-8.7	-5.6	-7.2
28/10/2020	17	918	71	5509	0.973	5072	4943	5007	-7.9	-10.3	-9.1
Mean		969	166	5814	0.973	5562	5383	5473	-4.6	-7.1	-5.8

The weekly series were also examined for evidence of autocorrelation over the whole 33-week period and the 17-week period relating to the third phase, when the LANL forecasts were used in the performance analysis. The ACF values were 0.810 and 0.701 respectively which were significant at the 1% level based on a Bartlett (1946) test, indicating strong positive autocorrelation. This has important implications for the interpretation and evaluation of the performance analysis and supports the use of the AR(1) model as a benchmark of comparison for the LANL model. It also supports the use of a composite forecasts that integrate the forecasts from the LANL and AR(1) models. The performance analysis of forecasts from the LANL, AR(1) and composite models are discussed in the next section.

6. Performance analysis results

This section presents the performance analysis results using the framework set out in Sections 2, 3 and 4 comparing the forecast quantiles, $q_{\alpha,j}$, for the LANL, AR(1) and composite models with the empirical quantiles, $e_{\alpha,j}$, over the 17-week period, with forecasts for weeks ending 08/07/2020 to 28/10/2020, with weeks numbered 1 to 17.

6.1. The weekly median forecast values

The LANL, AR(1) and composite model median forecasts and the actual changes in the number of deaths for each of the 17 weeks with 5% and 30% confidence bands are presented in Figure 1. The forecasts were reasonably close to the actual values in this period with two values outside the 30% confidence bands and no values outside the 5% confidence bands. Table 1 shows that for the median/point forecasts of the LANL model, the forecast error, as a percentage of the actual value, was less than 10% for ten of the weeks (weeks 5, 6, 7, 8, 9, 10, 12, 14, 15 and 16) and less than 20% for a further four of the weeks (weeks 1, 2, 3 and 17) with 12 out of these 17 under-estimating the actual value. The forecast error was greater than 20% for three weeks, with weeks 4, 11 and 13, all showing under-estimates of the actual value (25%, 29% and 22% respectively).

The mean difference in the percentage errors showed an average under-estimate of 7.1% in the LANL median forecasts. The AR(1) model showed a similar distribution with a lower mean under-estimate of 4.6%. The composite forecast mean showed an under-estimate that was between the two of 5.8%. The correlation between the LANL and AR(1) models' forecasts errors (0.366) was non-significant. This reflects some degree of lack of coherence between the two sets of median forecasts.

The phi values for the AR(1) model for each week are also presented in Table 1 which shows values below and above one, with the lowest values occurring from weeks 1 to 3, which reflect the falling values in number of deaths in the previous second phase of the pandemic. Values above one occurred for weeks 5 to 11. The values for weeks between 4 and 17 were relatively close to unity being between 0.952 and 1.040.

6.2. The means and variances of the forecast distributions and empirical quantiles

The means and variances for the forecast quantile probabilities over the 17-week period are presented for the LANL, AR(1) and composite models in Tables 2, 3 and 4, respectively, as well as the values for the MSQS and its components. The median predictions (the 0.5 quantile values) for the LANL, AR(1) and composite models provide a direct evaluation of the median or point accuracy of the LANL, AR(1) and composite model forecasts as the empirical quantile value at the 0.5 probability is the same as the actual change. As the quantile cumulative probabilities, α, move towards both tails of the distribution, the importance of the empirical quantile assumptions become more and more relevant. The LANL model showed that the forecast quantile values had, generally, higher errors at the tails, reflecting a greater spread of forecast distributions compared with the empirical distributions. The forecast distributions were longer tailed than the empirical distributions, with longer upper tails than lower tails. The AR(1) model forecast distributions also show longer tails than the empirical distributions, but these were symmetric. The composite forecast distributions were an average of the two.

The mean values reflect the characteristics of the forecasting distributions over the quantile probabilities. As would be expected, empirical quantiles show higher means at the low probabilities and lower means at the higher probabilities with the forecasts for the three models showing a similar pattern. The AR(1) model had much closer values to the empirical quantiles than the LANL model, with the composite model having a value between the two. The values at the 0.7 quantile probability were similar for the empirical quantiles and the three models.

The variance values also reflect the characteristics of the forecasting distributions over the quantile probabilities. The empirical quantiles had similar variances over the quantile probability range with slightly higher variance values at the extreme tails. The LANL model showed a much greater range of values for the variances with lower values than the empirical quantiles for quantile probabilities of 0.3 and below and higher values for 0.35 and above. The AR(1) model showed a smaller range of values for the variances, which were generally much higher than the empirical quantiles at all probabilities than the LANL model. The AR(1) model, therefore, showed much more variation in the forecasts compared with the LANL model. The composite model variances had values that were generally between the two, although closer to the LANL than the AR(1) values.

6.3. Mean squared quantile score

The Mean Squared Quantile Score (*MSQS*) measures overall performance and is the sum of *bias squared*, *resolution variation* and *error variation*. The best value is zero on this measure. Table 2 shows that the *MSQS* shows U-shaped distribution over the quantile probabilities for the LANL model with the lowest value of 526 thousand at the 0.65 quantile probability. Table 3 shows that the AR(1) model has a relatively flat distribution. The LANL model had lower values for probabilities between 0.45 and 0.8. At the 0.5 quantile probability the LANL model had a *MSQS* value of 672 thousand with 69% of this explained by the *error variation component* and the AR(1) model had a value of 828 thousand with 92% of this explained by *error variation*. The LANL model's better performance in the 0.45 to 0.8 quantile probability range reflects the better

performance on *error variation*, with similar performance on *bias squared*. The LANL model's poorer performance outside this range essentially reflects the poorer performance on *bias squared* and, to a lesser extent, *resolution variation* below the 0.45 quantile probability.

Table 2. Performance statistics, LANL model

Measure	Mean	Mean	Var	Var	Bias	Bias Sq	Res'n	RV	EV	MSQS
Quantile	LANL	Emp	LANL	Emp	LANL	LANL	LANL	LANL	LANL	LANL
			Thou's	Thou's		Thou's		Thou's	Thou's	Thou's
0.010	2,642	4,453	278	1,242	-1,811	3,279	0.297	614	168	4,062
0.025	3,010	4,771	340	1,131	-1,762	3,103	0.366	455	189	3,747
0.050	3,356	4,996	455	1,062	-1,640	2,691	0.457	313	233	3,236
0.100	3,757	5,215	563	1,006	-1,457	2,124	0.531	221	279	2,625
0.150	4,049	5,345	647	979	-1,296	1,680	0.584	169	313	2,162
0.200	4,295	5,440	718	961	-1,146	1,313	0.630	131	336	1,780
0.250	4,505	5,519	799	949	-1,014	1,028	0.676	100	366	1,493
0.300	4,702	5,587	870	940	-885	783	0.711	78	395	1,257
0.350	4,883	5,648	937	933	-766	586	0.749	59	414	1,059
0.400	5,052	5,706	985	927	-653	427	0.775	47	428	902
0.450	5,220	5,760	1,054	922	-540	292	0.812	32	445	769
0.500	**5,383**	**5,814**	**1,120**	**919**	**-431**	**186**	**0.845**	**22**	**464**	**672**
0.550	5,545	5,868	1,194	916	-322	104	0.884	12	479	595
0.600	5,717	5,923	1,257	914	-206	42	0.917	6	489	538
0.650	5,890	5,980	1,335	913	-90	8	0.948	2	515	526
0.700	6,078	6,042	1,414	912	36	1	0.971	1	554	557
0.750	6,280	6,110	1,489	913	170	29	0.998	0	579	608
0.800	6,518	6,188	1,624	916	330	109	1.051	2	613	724
0.850	6,779	6,284	1,761	922	495	245	1.091	8	664	917
0.900	7,138	6,414	2,001	934	724	523	1.152	22	762	1,307
0.950	7,690	6,633	2,275	966	1,057	1,117	1.178	31	934	2,082
0.975	8,276	6,859	2,757	1,013	1,416	2,006	1.214	46	1,265	3,317
0.990	8,991	7,182	3,359	1,106	1,809	3,271	1.225	56	1,698	5,025

Table 4 shows that the composite model has better performance on the *MSQS* than the LANL model at all quantile probabilities, although the values between quantile probabilities 0.6 to 0.7 were close to the LANL model. Table 3 shows that the *MSQS for consistency*, *MSQSC*, has relatively high values at the tails which largely reflects values for *bias squared for consistency*, *BSC*, and *error variation for consistency*, *EVC*. There are lower values for the *MSQSC* for quantile probabilities between 0.35 and 0.8, with the *EVC* being the most important component in this range. The *relative percentage improvement in MSQS for composite forecasts*, *RPMSQS*, is between 17% and 24% for quantile probabilities of 0.15 and above. The *MSQS* for the composite forecasts shows a considerable general improvement on the LANL model forecasts across all quantile probabilities. By definition, the composite forecast performance, as measured by the *MSQS*, has values better than the average of *MSQS* values of the LANL and AR(1) models at all quantile probabilities, the magnitude of the difference depending on the degree of the lack of coherence between the LANL and AR(1) forecasts on the *MSQS* at the quantile probabilities.

6.4. Bias and bias squared

Bias (*B*) or *calibration* is a performance measure of under- or over-estimation in forecasting. *Bias squared* (*BS*) and *bias* are important where quantile forecasts are required that do not show general over/under-estimation of the changes to the number of deaths over a given time, as this could influence the use of measures to control the spread of the virus and the provision of medical

services. If *bias* is small, any under-estimation that occurred in some weeks would be offset by over-estimation in other weeks, and vice versa. A value of zero indicates no bias.

Table 3. Performance statistics, AR(1) model

Measure	Mean	Mean	Var	Var	Bias	Bias Sq	Res'n	RV	EV	MSQS
Quantile	AR(1)	Emp	AR(1)	Emp	AR(1)	AR(1)	AR(1)	AR(1)	AR(1)	AR(1)
			Thou's	Thou's		Thou's		Thou's	Thou's	Thou's
0.010	3,444	4,453	1,112	1,242	-1,009	1,017	0.562	238	720	1,976
0.025	3,892	4,771	1,201	1,131	-880	774	0.654	135	717	1,626
0.050	4,224	4,996	1,279	1,062	-772	596	0.727	79	717	1,392
0.100	4,563	5,215	1,369	1,006	-652	425	0.803	39	720	1,184
0.150	4,772	5,345	1,430	979	-573	328	0.850	22	724	1,074
0.200	4,929	5,440	1,479	961	-511	261	0.884	13	728	1,003
0.250	5,060	5,519	1,522	949	-459	211	0.911	7	733	951
0.300	5,174	5,587	1,560	940	-413	170	0.935	4	738	913
0.350	5,278	5,648	1,596	933	-370	137	0.956	2	744	883
0.400	5,376	5,706	1,631	927	-329	108	0.975	1	750	859
0.450	5,470	5,760	1,665	922	-290	84	0.993	0	757	841
0.500	**5,562**	**5,814**	**1,700**	**919**	**-252**	**64**	**1.009**	**0**	**764**	**828**
0.550	5,654	5,868	1,735	916	-214	46	1.025	1	772	818
0.600	5,748	5,923	1,772	914	-175	31	1.041	2	782	814
0.650	5,846	5,980	1,811	913	-134	18	1.057	3	792	813
0.700	5,950	6,042	1,854	912	-92	8	1.072	5	805	819
0.750	6,064	6,110	1,902	913	-45	2	1.088	7	821	831
0.800	6,195	6,188	1,959	916	7	0	1.104	10	842	852
0.850	6,353	6,284	2,029	922	69	5	1.121	13	872	890
0.900	6,562	6,414	2,126	934	147	22	1.137	18	918	957
0.950	6,900	6,633	2,292	966	267	71	1.151	22	1,013	1,107
0.975	7,233	6,859	2,466	1,013	373	139	1.146	22	1,135	1,296
0.990	7,680	7,182	2,716	1,106	498	248	1.112	14	1,348	1,609

The LANL model forecasts presented in Table 2 show a negative *bias* at the 0.5 quantile probability of 431 that reflects under-estimation of changes. In relation to the other quantile probabilities, the *bias* values indicate a general under-estimation of the empirical values for probabilities of 0.65 and below with over-estimation above this value. *Bias* values for the LANL model ranged from an under-estimate of 1,811 deaths per week at probability 0.01 to an over-estimate of 1.809 at 0.99 with *bias* having values between -500 and 500 for probabilities between 0.5 and 0.85. The AR(1) model forecasts presented in Table 3 show a *bias* value at the 0.5 quantile probability that illustrates a small under-estimate of 252. The AR(1) model had slightly worse bias in absolute terms, albeit relatively small, at the 0.65 and 0.7 quantile probabilities, but for all the other probabilities the AR(1) model *bias* is much closer to zero than the LANL model. The AR(1) model shows a range with values increasing from an under-estimate of 1009 at probability 0.01 to an over-estimate of 498 at probability 0.99. The composite forecasts give values between the LANL and AR(1) models for all quantile probabilities, with values closer to zero for all quantile probabilities compared with the LANL model. Table 4 shows that the composite forecasts generally improve on *bias* relative to the LANL model.

Bias Squared shows a U-shaped distribution for the LANL model with higher values the closer the cumulative probabilities to the tails, with the lowest value at the 0.7 probability. *Bias squared* for the AR(1) model shows a flatter distribution and lower values than for the LANL model, except at the 0.65 and 0.7 quantile probabilities. The composite model has values between the two. Table 5 shows high values for *BSC* at the tails, with *relative percentage improvement in bias squared for composite forecasts*, *RPBS*, having values between 7% and 13% for quantile

probabilities of 0.3 and below, and values between 24% and 31% for quantile probabilities of 0.9 and above. On *bias squared* the composite forecasts certainly improved on the LANL model forecasts especially towards the tails, with values better than the average of *bias squared* values of the LANL and AR(1) models, reflecting a degree of lack of coherence.

Table 4. Performance statistics, composite model

Measure	Mean	Mean	Var	Var	Bias	Bias Sq	Res'n	RV	EV	MSQS
Quantile	Comp	Emp	Comp	Emp	Comp	Comp	Comp	Comp	Comp	Comp
			Thou's	Thou's		Thou's		Thou's	Thou's	Thou's
0.010	3,043	4,453	546	1,242	-1,410	1,987	0.429	405	317	2,709
0.025	3,451	4,771	623	1,131	-1,321	1,744	0.510	272	329	2,345
0.050	3,790	4,996	727	1,062	-1,206	1,455	0.592	176	354	1,986
0.100	4,160	5,215	822	1,006	-1,055	1,112	0.667	111	374	1,598
0.150	4,410	5,345	892	979	-934	873	0.717	78	389	1,340
0.200	4,612	5,440	955	961	-829	687	0.757	57	404	1,147
0.250	4,782	5,519	1,017	949	-736	542	0.794	40	420	1,002
0.300	4,938	5,587	1,070	940	-649	421	0.823	29	434	884
0.350	5,081	5,648	1,120	933	-568	322	0.852	20	442	785
0.400	5,214	5,706	1,162	927	-491	241	0.875	14	452	708
0.450	5,345	5,760	1,212	922	-415	172	0.902	9	461	642
0.500	**5,473**	**5,814**	**1,262**	**919**	**-342**	**117**	**0.927**	**5**	**472**	**594**
0.550	5,600	5,868	1,316	916	-268	72	0.955	2	481	555
0.600	5,732	5,923	1,365	914	-190	36	0.979	0	490	526
0.650	5,868	5,980	1,422	913	-112	13	1.002	0	505	518
0.700	6,014	6,042	1,478	912	-28	1	1.021	0	526	528
0.750	6,172	6,110	1,537	913	63	4	1.043	2	543	549
0.800	6,357	6,188	1,633	916	169	28	1.077	5	570	604
0.850	6,566	6,284	1,731	922	282	80	1.106	10	604	694
0.900	6,850	6,414	1,893	934	435	190	1.145	20	670	879
0.950	7,295	6,633	2,079	966	662	438	1.164	26	769	1,234
0.975	7,754	6,859	2,361	1,013	895	801	1.180	33	951	1,784
0.990	8,335	7,182	2,746	1,106	1,153	1,330	1.169	32	1,235	2,596

6.5. Resolution and resolution variation

Resolution variation (RV) and resolution or slope (SL) are important where forecasts are required that identify and distinguish between weeks when high or low changes in the number of deaths is likely to occur. *Resolution* is relevant to situations when there is a need to discriminate between large and small changes so that measures or medical resources could be appropriately adjusted. With *resolution,* the best possible value is unity.

Table 2 shows that at the 0.5 quantile probability, the LANL model has a *resolution* value of 0.845 compared with 1.009 for the AR(1) model. The ACF value at the 0.5 quantile probability was 0.701, which was significant at the 1% level based on a Bartlett (1946) test, indicating there was a high degree of autocorrelation in the weekly actual changes in US deaths over the 17-week period. The LANL model shows a range of values for *resolution* increasing from a relatively low value of 0.297 at probability 0.01 to a value of 1.225 at probability 0.99. The results indicate that the LANL model response to a one-unit actual change was a forecast change that was less than unity for quantile probabilities of 0.75 or less, but greater than unity above this probability. Table 3 shows that the AR(1) model has values generally much closer to unity than the LANL model at most quantile probabilities, except between 0.65 and 0.85, with values between 0.80 and 1.16 for

all probabilities of 0.1 and above. Table 4 shows that the composite model has values which are essentially the average of the two. The resolution results suggest that the LANL model did not appear to take into consideration the full effect of the positive autocorrelation in the actual changes in the number of deaths, particularly for quantile probabilities below 0.3.

Table 5. Coherence measures and relative percentage improvement of the composite forecasts

Measure	BSC	RVC	EVC	MSQSC	RPBS	RPRV	RPEV	RPMSQS
Quantile					%	%	%	%
0.010	644	87	507	1,239	7.5	5.1	28.6	10.3
0.025	777	94	494	1,366	10.0	8.0	27.3	12.7
0.050	754	77	481	1,312	11.5	9.9	25.3	14.2
0.100	649	75	500	1,224	12.7	14.3	25.0	16.1
0.150	523	69	519	1,110	13.0	18.0	25.0	17.2
0.200	403	62	512	977	12.8	21.4	24.1	17.5
0.250	308	53	520	880	12.4	24.6	23.7	18.0
0.300	223	47	532	803	11.7	28.6	23.5	18.5
0.350	157	40	548	744	10.8	33.0	23.6	19.2
0.400	105	37	549	691	9.8	39.0	23.3	19.6
0.450	62	30	561	653	8.3	46.0	23.3	20.3
0.500	**32**	**25**	**566**	**623**	**6.4**	**56.0**	**23.1**	**20.8**
0.550	12	18	576	606	4.0	70.9	23.0	21.4
0.600	1	14	582	597	0.7	89.7	22.9	22.1
0.650	2	11	593	606	3.7	99.8	22.7	22.6
0.700	16	9	614	640	84.1	84.9	22.6	23.3
0.750	46	7	628	681	74.8	52.1	22.4	23.7
0.800	105	3	631	738	47.9	10.6	21.7	23.4
0.850	182	1	655	838	36.4	1.9	21.3	23.2
0.900	332	0	681	1,013	30.4	0.3	20.3	22.4
0.950	624	1	819	1,443	26.2	0.7	21.0	22.6
0.975	1,088	5	997	2,090	25.4	3.4	20.8	22.7
0.990	1,718	14	1,153	2,886	24.4	10.1	18.9	21.7

Resolution variation is a relatively small component of the *MSQS* for all three models. *Resolution variation* for the LANL model showed a general decline from a value of 614 thousand at quantile probability 0.01 to a value of almost zero at quantile 0.75 and then a relatively small rise. The AR(1) model has much smaller values which are below 100 thousand for probabilities of 0.05 and above. The much better *resolution variation* values for the AR(1) model reflect its ability to identify the first order positive autocorrelation. The composite model has values between the two. Table 5 shows relatively high values for *resolution variation for consistency*, *RVC*, for quantile probabilities of 0.2 or below, but the *RVC* was only a small component of the *MESC.* Given these relatively low values the *relative percentage improvement in resolution variation for composite forecasts*, *RPRV* are not important. On *resolution variation* the composite forecasts certainly improved on the LANL model forecasts, especially for the quantile probabilities below 0.5, but this only had a limited impact on the *MSQS*.

6.6. Error variation

Error variation (EV) is important where it is required to have forecasts with low unexplained variation in the changes in the number of deaths, which is variation in the quantile forecasts that is not explained by variation in the empirical quantile values. The poorer the performance on *error*

variation, the greater the size of medical provisions that would be necessary to compensate for unpredicted movements in the changes in the number of deaths.

Table 2 for the LANL model shows an increase in values from 168 thousand at quantile probability 0.01 to 1,698 thousand at probability 0.99. Tables 2 and 3 show that the LANL model values are better than the AR(1) for all quantile probabilities except at 0.975 and 0.99. At the 0.5 quantile probability the value for the LANL model is 464 thousand and the AR(1) model is 764 thousand. The LANL model values move closer to the AR(1) model values as the quantile probabilities values increase. Table 4 shows that the composite model has values that are close to the LANL model with higher values at quantile probabilities at 0.6 and below, but lower values at 0.65 and above. Table 3 shows that the *EVC*, has values between 480 and 1,160 at all quantile probabilities, with values for the *relative percentage improvement in resolution variation for composite forecasts*, *RPEV*, between 18% and 29%. This is an important component of the *MSQS* with the composite forecaster clearly having values better than the average of *error variation* values of the LANL and AR(1) models due to a degree of lack of coherence. This accounts for the composite forecasts being much closer to the better LANL forecasts than the AR(1) forecasts. The results can partly be explained by the differences between the variances of the LANL and AR(1) models at most quantile probabilities.

7. Discussion

An analytical framework is demonstrated for the evaluation of forecasts presented in the form of quantiles in relation to weekly changes in the number of *US*, *COVID-19 Deaths* for forecast dates from 08/07/2020 to 28/10/2020, with weeks numbered 1 to 17. The framework compares weekly changes in the quantile forecasts from the LANL model (LANL, 2020) with *empirical quantiles*, obtained using the Student t distribution of daily changes (first differences) in the number of deaths over the 17 weekly periods. These empirical quantiles are then used to evaluate the forecasts using the *Mean Squared Quantile Score* (*MSQS*) which is further decomposed into sub-components involving *bias*, *resolution,* and *error variation*. The LANL model forecasts are also compared with quantile forecasts from a simple AR(1) model, and then used to generate composite quantile predictions. The importance of coherence in the relationship between the LANL and AR(1) forecasts are examined in the context of the composite forecasts. In relation to these analyses, the three objectives outlined in the introduction are now considered.

The first objective of this study was to provide a framework to evaluate quantile forecast performance that could be applied with a limited number of observations. The method described above was applied to daily first differenced COVID-19 data on the number of deaths in the US, over 17 weekly periods, reported by JHU. The framework involved using the assumption that the distribution of the daily values, over a one-week period, followed a normal distribution, with the values used to obtain standard deviation estimates via the Student t distribution to acquire empirical quantiles. A simple adjustment procedure was used to consider the fact that the recorded values on Sundays and Mondays were approximately one half of the recorded values on the other weekdays. The statistical characteristics of the data were examined, and the results implied that the series satisfied the assumptions of approximate normality. Empirical quantiles were then derived for the median and 22 quantiles. It was then illustrated that these empirical quantiles could be used to evaluate the quantile forecasts.

The second objective was to extend available forecast evaluation methods to analyze quantile forecasts to examine overall accuracy, as well as specific aspects of accuracy. The *Mean Squared Quantile Score* (*MSQS*) measure was used to compare the forecast quantiles with the empirical quantiles. The *MSQS* was decomposed to allow identification of specific aspects of performance that can be used to highlight strengths and weaknesses of the forecasts using the procedure set out in Thomson, et al. (2019), but extended to quantile forecast analysis. The three components of the *MSQS, squared bias, resolution variation* and *error variation*, were examined over the 17-week period by using the quantile predictions from the LANL model and the empirical quantiles. The performance of the LANL model was compared with performance analysis of an AR(1) model, applied without a constant, and based on a moving period of 10 weeks using the values from the current and 9 previous weeks. The AR(1) model was used as the weekly changes

in the US number of deaths showed strong first order autocorrelation. The results reveal that the LANL model showed adequate overall performance, as measured by the *MSQS*, that was better than the AR(1) model for quantile probabilities between 0.45 and 0.8, but poorer outside this range. The LANL model's forecast performance, therefore, could be fairly described as moderate. The AR(1) model had *bias* values closer to zero than the LANL model for all quantile probabilities, except the 0.65 and 0.7 probabilities. The LANL model showed negative *bias* for quantile probabilities of 0.65 and below, and positive bias for 0.7 and above. This gave a *bias squared* U-Shaped distribution. *Resolution* was much better for the AR(1) model compared with the LANL model for most quantile probabilities, except between 0.65 and 0.85, which was also, of course, the case for *resolution variation*. This could reflect that the LANL model did not appear to fully consider the first order autocorrelation in the series. On the other hand, *error variation* was better for the LANL model as compared with the AR(1) model for all quantile probabilities except 0.975 and 0.99. This can be partly explained by the higher variation in the AR(1) forecasts.

The results indicate poorer values at the lower and higher probabilities for the LANL model compared with the empirical quantiles that partly reflect the fact that the LANL model quantiles showed more variation and a forecasting distribution with longer tails, particularly on the upper tail, than the empirical quantiles, which had symmetric distributions. The increased variation or spread of the forecast quantiles compared with the empirical quantiles would reflect the level of uncertainty incorporated into the LANL model quantile forecasts. The longer upper tails of the forecast distribution are an issue that relates particularly to the construction of the LANL model, which is described in LANL (2020). The longer upper tails could reflect a higher level of uncertainty surrounding large changes compared with small changes in the number of deaths. Considering the practical consequences and/or implications for these findings, the greater variation of the *MSQS* values across the quantiles for the LANL model in comparison to the AR(1) model and the subsequent moderate forecasting performance of the LANL model may make the planning for medical provisions difficult. With the lack of consistent forecasting performance across all the quantiles, decision makers may be led to over- or under-compensate depending on which quantile is considered, and this would have a direct impact on service provision, patient experience, and potentially patient death. This reflects the levels of uncertainty as incorporated into the LANL model, as discussed previously.

The third objective of this study was to consider if forecast performance could be improved by combining the LANL model predictions with the AR(1) model predictions over the period under consideration. The results overall show that the use of a composite forecasts involving the LANL model and a simple AR(1) model result in substantial improvements in forecast performance. A major reason for this is that the AR(1) model forecasts show a considerable lack of coherence with the LANL model forecasts, arising from the *error variation* over all quantile probabilities and *bias squared* at the tails. The composite forecasts incorporate the information input of the LANL and AR(1) model forecasts. This better performance of the composite model was achieved despite the AR(1) forecasts having considerably more *error variation* than the LANL forecasts. These results support the view that even models that perform very well overall or on certain components, can often be improved by combining the forecasts with other models, provided there exists a lack of coherence between the two sets of forecasts. These improvements in forecasting performance could potentially result in real improvements in supporting policy and clinical decision making, and, in this instance, would better inform service provision related to COVID-19, reducing pressure on services and medical staff, and subsequently reducing the risk of death related to COVID-19.

8. Conclusion

This study has provided an approach using empirical quantiles to evaluate quantile forecasts on the number of recorded deaths from COVID-19. The LANL model was used in this study as it provided regular and comprehensive data on COVID-19 number of death predictions for the US, using the median and 22 quantiles that were available twice weekly in a format very suitable for statistical analysis. This has allowed the framework set out in this paper to be easily applied in a relevant and practical manner.

LANL also provides forecasts on the number of deaths for many countries and US states. In addition, LANL provides forecasts for horizons of up to six weeks and the analysis set out in this study could easily be extended to evaluate these forecasts. In this situation, empirical quantiles could be obtained for longer periods than one week by using a combination of the weekly periods to reduce the problem of changing means and standard deviations from one week to the next (Pollock, et al*.,* 2010). This analysis could be used to evaluate these forecasts and similar forecasts for other COVID-19 forecasting centers.

The framework set out in this study can be certainly applied to the more common forecasting situations when point or median predictions with confidence intervals are used. There are a considerable number of institutions that provide median estimates and 95% confidence intervals for COVID-19 deaths. CDC (2020b) publish these forms of forecasts for many institutions that involve 95% confidence intervals together with quartiles for the US and individual US states. The COVID-19 Forecasting Hub (2020) provides more comprehensive data from these sources, including the median and 22 quantiles for a large range of institutions on the number of deaths for the US and US states.

Several modifications and extensions to the current framework can be suggested. One modification and extension would involve application of the framework to grouped probability forecasts. That would involve the derivation of grouped empirical probabilities on a similar basis to that used to obtain the empirical quantiles. This framework could be used, for instance, to evaluate the performance of forecast changes in the weekly number of deaths from COVID-19 expressed as probabilities for a range of mutually exclusive bands representing changes, for example, bands of 0 to <500, 500 to <1000, and so on. This would allow performance to be evaluated using a much smaller number of observations than the common method of analyzing probability forecasts using dichotomous, ex-post, values of zero (when the actual value does not fall into the forecast group) or unity (when the actual value falls into the forecast group) assigned, to each group, at each observation forecast period. Empirical probabilities would, in this instance, be an extension to the first objective discussed above.

Another modification could involve the extension of the integration of performance and coherence measures to composite forecasts obtained from more than two models. This would allow the examination of agreement or disagreement between forecasts from different institutions and permit the benefits of using composite forecasts obtained from multiple sources to be examined. For example, using the COVID-19 forecast data on the number of deaths available for the US and US states from CDC (2020b) and the COVID-19 Forecasting Hub (2020) to obtain the best composite forecasts and identify reasons for changes in performance and coherence as the forecast horizon increases. This could be used to extend the work of Ray, et al. (2020), who found in their analysis, using over two and a half thousand composite / ensemble forecasts with data from the COVID-19 Forecasting Hub (2020), which forecast performance deteriorated as the prediction horizon increased from one to four weeks.

The procedure set out in this study could be extended and applied to a wide range of forecasting situations, where probability predictions are involved, and the forecast observation period can be partitioned into an adequate number of smaller periods. The statistical distribution can be approximated from sets of smaller period data points and used to obtain values for the longer forecast period data points. For instance, using weekly data to obtain quarterly empirical quantiles that could be used to evaluate quarterly quantile forecasts. In this study, the distribution was assumed to be normal, but other identifiable distributions could be used.

The empirical probability approach set out in this paper does have potential limitations, however. It assumes that the values over a short period, such as a week, are approximately independently and identically normally distributed. Before applying the technique, it is necessary to examine the data to verify that this assumption is appropriate. A further possible limitation is that the performance and coherence analysis used in this study uses measures associated with quadratic loss functions, specifically the *MSQS* and *MSQSC* and related measures. In some situations, other forms of loss function may be more appropriate. If the conditions are not satisfied, it may be more appropriate to use an alternative approach to analyze quantile forecasts such as the *WIS*, although these measures can be limited particularly in relation to coherence.

In addition, the framework set out in the paper has been applied in a within-sample situation. It could be useful to examine the stability of these measures in out-of-sample or rolling sample situations. It could be useful, possibly as a future extension, to consider how the performance and coherence results from different models change over time, for instance, in relation to the different phases in the path of changes in the number of COVID-19 deaths.

Regardless of the effects of these potential limitations, the framework provides a powerful diagnostic tool that can be used in a multitude of practical situations when predictions are presented in the form of quantiles. This can be a valuable aid to decision making and to policy makers where there is a need to be able to evaluate quantile probability predictions with a relatively small number of past values and examine specific aspects of forecast performance as well as overall accuracy so that the quality of forecasts can be determined. In relation to forecasts that use COVID-19 data, the evaluation of performance is crucial to the accurate prediction of changes in the number of recorded deaths, hospitalizations, and infections for planning purposes. For instance, such forecasts can provide hospitals, policy makers, and governments with crucial information about how current resources should be used, such as medical staff, protective equipment, intensive-care hospital beds and ventilators. The accessibility of such resources often needs to be quickly adjusted so that they are readily available when required.

Acknowledgement: The authors would like to thank the US LANL COVID-19 Forecasting Team for their carefully considered and relevant comments on an earlier version of this research paper. We would like to particularly thank them for highlighting a range of issues that proved to be invaluable to us in terms of moving our research work forward. The comments and advice from the LANL COVID-19 Forecasting Team are very much appreciated.

References

Anderson, R. M., Heesterbeek, H., Klinkenberg, D. and Déirdre Hollingsworth, T., 2020. How will country-based mitigation measures influence the course of the COVID-19 epidemic? *The Lancet,* 395(10228), pp. 931-934. https://doi.org/10.1016/S0140-6736(20)30567-5

Armstrong, J. S., 2001. Combining forecasts. In: J. S. Armstrong, ed. 2001. *Principles of forecasting: A handbook for researchers and practitioners*. Norwell, MA: Kluwer Academic Publishers. pp. 417–439. https://doi.org/10.1007/978-0-306-47630-3

Armstrong, J. S., Green, K. C. and Graefe, A., 2015. Golden rule of forecasting: be conservative. *Journal of Business Research*, 68, pp. 1717–1731. https://doi.org/10.1016/j.jbusres.2015.03.031

Banerjee, A., Pasea, L., Harris, S., Gonzalez-Izquierdo, A., Torralbo, A., Shallcross, L., Noursadeghi, M., Pillay, D., Sebire, N., Holmes, C., Pagel, C., Wong, W. K., Langenberg, C., Williams, B., Denaxas, S., Hemingway, H., 2020. Estimating excess 1-year mortality associated with the COVID-19 pandemic according to underlying conditions and age: A population-based cohort study. *The Lancet,* 395(10238), pp. 1715–1725. https://doi.org/10.1016/S0140-6736(20)30854-0

Bartlett, M. S., 1946. On the theoretical specification of sampling properties of autocorrelated time series. *Journal of the Royal Statistical Society, Series B,* 8, pp. 27–41. https://doi.org/10.2307/2983611

Bracher, J., Ray, E. L., Gneiting, T. and Reich, N. G., 2020. Evaluating epidemic forecasts in an interval format. *arXiv*, 05/02/2021. [online] Available at: <https://arxiv.org/abs/2005.12881> [Accessed on 17 March 2021].

Cascella, M., Rajnik, M., Cuomo, A., Dulebohn, S. C. and Di Napoli, R., 2020. *Features, evaluation, and treatment of coronavirus (COVID-19)*. Florida: StatPearls Publishing.

CDC, 2020a. *Excess deaths associated with COVID-19*. [online] Available at: <https://www.cdc.gov/nchs/nvss/vsrr/covid19/excess_deaths.htm> [Accessed on 12 January 2021].

CDC, 2020b. *COVID-19 forecasts deaths*. [online] Available at: <https://www.cdc.gov/coronavirus/2019-ncov/covid-data/forecasting-us.html> [Accessed on 12 January 2021].

Christoffersen, P., 1998. Evaluating interval forecasts. *International Economic Review,* 39, pp. 841–462. https://doi.org/10.2307/2527341

Clemen, R. T., 1989. Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, 5*,* pp. 559–583. https://doi.org/10.1016/0169-2070(89)90012-5

Cramer, E. Y., Ray, E. L., Lopez, V. K., Bracher, J., Brennen, A., Rivadeneira, A. J. C., *et al*. 2021. Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the US. *MedRxiv*, 05/02/2021. Available at: <https://www.medrxiv.org/content/10.1101/2021.02.03.21250974v1> [Accessed on 24 April 2021].

De Menezes, L. M., Bunn D. W. and Taylor, J. W., 2000. Review of guidelines for the use of combined forecasts, *European Journal of Operational Research*, 120, pp. 190–204. https://doi.org/10.1016/S0377-2217(98)00380-4

Gneiting, T., Balabdaoui, F. and Rafferty, A. E., 2007. Probability forecasts, calibration and sharpness*. Journal of the Royal Statistical Society, Series B*, 69, pp. 243–268. https://doi.org/10.1111/j.1467-9868.2007.00587.x

Gneiting, T. and Rafferty, E., 2007. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association,* 102, pp. 359–378. https://doi.org/10.1198/016214506000001437

Gilbert, D., 2020. Which countries are under lockdown – and is it working? *The Telegraph*, [online] 19 April. Available at: <https://www.telegraph.co.uk/news/2020/04/16/countries-in-lockdown-denmark-germany/> [Accessed on 24 April 2021].

Goodwin, P., 2015. Is a more liberal approach to conservatism needed in forecasting? *Journal of Business Research*, 68, pp. 1753–1754. https://doi.org/10.1016/j.jbusres.2015.01.060

Graefe, A., Armstrong, J. S., Jr., Jones, R. J. and Cuzan, A. G., 2014. Combining forecasts: an application to elections. *International Journal of Forecasting*, 30, pp. 43–54. https://doi.org/10.1016/j.ijforecast.2013.02.005

Granger, C. W. J., Kamstra, M. and White, H., 1989. Interval forecasting: An analysis based upon ARCH-quantile estimators. *Journal of Econometrics,* 40, pp. 87–96. https://doi.org/10.1016/0304-4076(89)90031-6

Green, K. C., Armstrong, J. S. and Graefe, A., 2015. Golden rule of forecasting rearticulated: Forecast unto others as you would have them forecast unto you. *Journal of Business Research*, 68, pp. 1768–1771. https://doi.org/10.1016/j.jbusres.2015.03.036

Harvey N., 2001. Improving judgment in forecasting. In: J. S. Armstrong, ed. 2001. *Principles of Forecasting*. International Series in Operations Research and Management Science, vol 30. Boston, M.A: Springer. pp. 59-80. https://doi.org/10.1007/978-0-306-47630-3_4

Larsen, J. R., Martin, M. R., Martin, J. D., Kuhn, P. and Hicks, J. B., 2020. Modeling the onset of COVID-19. *Frontiers in Public Health, 8*, 473. https://doi.org/10.3389/fpubh.2020.00473

LANL, 2020. LANL COVID-19 cases and deaths forecasts. *Los Alamos National Laboratory*, [online] Available at: <https://covid-19.bsvgateway.org/> [Accessed on 12 January 2021].

Levene, H., 1960. Robust tests for equality of variances. In: I., Olkin, G., S. G., Ghurye, W. Hoeffding, W. G., Madow, and H. B., Mann, eds. 1960. *Contributions to probability and statistics: Essays in honor of Harold Hotelling*. Stanford University Press. pp. 278–292.

Lilliefors, H. W., 1967. On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association,* 62*,* pp. 399–402. https://doi.org/10.1080/01621459.1967.10482916

Lubecke, T. H., Markland, R. E., Kwok, C. C. Y. and Donohue, J. M., 1995. Forecasting foreign exchange rates using objective composite models. *Management International Review,* 35, pp. 135–152.

Makridakis, S. and Hibon, M., 2000. The M3 – competition: Results, conclusions and implications. *International Journal of Forecasting*, 16, pp. 451–476. https://doi.org/10.1016/S0169-2070(00)00057-1

Petropoulos, F. and Makridakis, S., 2020. Forecasting the novel coronavirus COVID-19. PLoS ONE 15(3), p. e0231236. https://doi.org/10.1371/journal.pone.0231236

Pollock, A. C., Macaulay, A., Thomson, M. E. and Onkal, D., 2005. Performance evaluation of judgemental directional exchange rate predictions. *International Journal of Forecasting*, 21, pp. 473–489.

Pollock, A. C., Macaulay, A., Thomson, M. E. and Onkal, D., 2008. Using weekly empirical probabilities in currency analysis and forecasting. *Frontiers in Finance and Economics*, 5, pp. 26–55. https://doi.org/10.1016/j.ijforecast.2004.12.006

Pollock, A. C., Macaulay, A., Thomson, M. E., Gonul, M. S. and Onkal, D., 2010. Evaluating strategic directional probability predictions of exchange rates. *International Journal of Applied Management Science*, 2, pp. 282–304. https://doi.org/10.1504/IJAMS.2010.033569

Pollock, A. C. and Wilkie, M. E., 1996. The quality of bank forecasts: The dollar-pound exchange rate, 1990-1993. *European Journal of Operational Research*, 91, pp. 306–313. https://doi.org/10.1016/0377-2217(95)00287-1

Qin, A. and Hernandez, J. C., 2020. China reports first death from new virus. *The New York Times*, [online] 10 January. Available at: <https://www.nytimes.com/2020/01/10/world/asia/china-virus-wuhan-death.html> [Accessed on 21 January 2021].

Ray, E. L., Wattanachit, N., Niemu, J., Kanji, A. H., House, K., Cramer, E. Y., *et al*., 2020. Ensemble forecasts of coronavirus disease 2019 (COVID-19) in the U.S. *MedRxiv*, 22/08/2020. Available at: <https://www.medrxiv.org/content/10.1101/2020.08.19.20177493v1>

Reich, N. G., McGowan, C. J., Yamana, T. K., Tushar, A., Ray, E. L., Osthus, D., *et al.*, 2019. Accuracy of real-time multi-model ensemble forecasts for seasonal influenza in the U.S. *PLOS Computational Biology*, 15(11), e1007486. https://doi.org/10.1371/journal.pcbi.1007486

Richie, H., 2020. Coronavirus source data. *Our World in Data.* Available at: <https://ourworldindata.org/coronavirus-source-data> [Accessed on 12 January 2021].

Schnaars, S. P., 1986. An evaluation of rules for selecting an extrapolative model of yearly sales forecasts. *Interfaces*, 16, pp. 100–107. https://doi.org/10.1287/inte.16.6.100

Stock, J. H. and Watson, M. W., 2004. Combining forecasts of output growth in a seven-country data set. *Journal of Forecasting*, 23, pp. 405–430. https://doi.org/10.1002/for.928

The COVID-19 Forecast Hub, 2020. COVID19-Forecast-Hub. *The Reich Lab at UMass-Amherst, Github*, [online] Available at: <https://github.com/reichlab/covid19-forecast-hub/> [Accessed on 12 January 2021].

Thomson, M. E., Onkal, D., Pollock, A. C. and Macaulay, A., 2003. The influence of trend strength on directional probabilistic currency predictions. *International Journal of Forecasting*, I19, pp. 241- 276. https://doi.org/10.1016/S0169-2070(01)00132-7

Thomson, M. E., Pollock, A. C., Henriksen, K. B. and Macaulay, A., 2004. The influence of forecast horizon on the currency predictions of experts, novices and statistical models. *European Journal of Finance*, 10, pp. 290-307. https://doi.org/10.1080/13518470110047620

Thomson, M. E, Pollock, A. C., Gonul, M. S. and Onkal, D., 2013. Effects of trend strength and direction on performance and consistency in judgmental exchange rate forecasting. *International Journal of Forecasting*, 29, pp. 337–353. https://doi.org/10.1016/j.ijforecast.2012.03.004

Thomson, M. E., Pollock, A. C., Onkal, D. and Gonul, M. S., 2019. Combining forecasts: Performance and coherence. *International Journal of Forecasting*, 21, pp. 473–489.

Timmerman, A., 2006. Forecast combinations. In: G., Elliot, C. W. J., Granger, and A., Timmerman, eds. 2006. *Handbook of economic forecasting, Volume 1*. Amsterdam: Elsevier. pp. 135-196. https://doi.org/10.1016/S1574-0706(05)01004-9

Wallis, K. F., 2011. Combining forecasts – forty years on. *Applied Financial Economics,* 21, pp. 33–41. https://doi.org/10.1080/09603107.2011.523179

Wilkie, M. E. and Pollock, A. C., 1996. An application of probability judgement accuracy measures to currency forecasting. *International Journal of Forecasting*, 12, pp. 91–118. https://doi.org/10.1016/0169-2070(94)02001-9

Wilkie-Thomson, M. E., Onkal-Atay, D. and Pollock, A. C., 1997. Currency forecasting: An investigation of extrapolative judgment. *International Journal of Forecasting*, 13, pp. 509–526. https://doi.org/10.1016/S0169-2070(97)00036-8

WHO, 2020a. WHO Director-General's opening remarks at the media briefing on COVID-19, 11 March 2020. *WHO*, [online] Available at: <https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020>

WHO, 2020b. WHO Coronavirus Disease (COVID-19) Dashboard. *WHO*, [online] Available at: <https://covid19.who.int/?gclid=CjwKCAjw74b7BRA_EiwAF8yHFDMZBxZWsyF2bctUo9fr-rFqk12S3HmBqwqxHpKoLgQv21KtV3U6WxoCCgAQAvD_BwE> [Accessed on 16 September 2020].

Yates, J. F., 1982. External correspondence: decompositions of the mean probability score. *Organisational Behavior and Human Performance,* 30, pp. 132–156. https://doi.org/10.1016/0030-5073(82)90237-9