International Journal of Scientific Research and Management (IJSRM)
||Volume||09||Issue||12||Pages||MP-2021-471-489||2021||
Website: www.ijsrm.in ISSN (e): 2321-3418
DOI: 10.18535/ijsrm/v9i12.mp01

Composite Quantile Probability Predictions: Performance and Coherence Analysis of US COVID-19 Confirmed Infection Cases

[1]Mary E. Thomson, [2]Andrew C. Pollock, [3]Jennifer Murray

[1]Northumbria University, UK
[2]Statistical Analyst, UK
[3]Edinburgh Napier University, UK

Abstract
An analytical framework is presented for the evaluation of composite probability forecasts using *empiricalquantiles*. The framework is demonstrated via the examination of forecasts of the changes in the number of US COVID-19 confirmed infection cases, applying 18 two-week ahead quantile forecasts from four forecasting organisations. The forecasts are analysed individually for each organisation and in combinations of organisational forecasts to ascertain the highest level of performance. It is shown that the relative error education achieved by combining forecasts depends on the extent to which the component forecasts contain in dependent information. The implications of the study 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

Over the course of the COVID-19 pandemic, the virus has mutated with new variants having increased the rate and ease of transmission and reproduction (Davies *et al.*, 2021). New variants and the roll out of vaccinations have also altered the symptomatology experienced by patients since the pandemic began (e.g., BMJ, 2021a; SeyedAlingaghi *et al.*, 2021), with more patients who have had vaccination experiencing lower infection rates and milder symptoms (BMJ, 2021b), and increased transmission and infection of new variants being seen in younger people than at the start of the pandemic (e.g., BMJ, 2021c). Further, the roll-out and roll-back of national and local lockdown initiatives influence transmission rates (Kochanczyk & Lipniacki, 2021). This makes forecasting based on symptoms, time and/or location challenging. A concrete outcome measure which is focused on earliest detection within the disease progression cycle within a forecasting model is therefore essential in informing clinical practice and public health policy and planning. One such measure is confirmed infection cases.

Confirmed infection cases are defined as people who have laboratory confirmation of their COVID-19 infection (WHO, 2020), regardless of whether they have shown symptoms or not. Recorded confirmed cases are a leading indicator of future hospitalisations and deaths. Evidence has shown that that the number of days between a sustained rise in cases and an increase in hospitalisations in the US of about 12 days (Kissane *et al.*, 2020). Other evidence has shown that there is a lagged relationship between cases and deaths in the US, with deaths occurring 2 to 8 weeks after the onset of symptoms (Testa *et al.*, 2020). Confirmed cases, therefore, provide an important indicator of future pressures arising from hospitalisations, intensive care demand and deaths. In addition, confirmed cases have a major impact on the economic environment as governments need to respond actively to changing infection case statistics. This has resulted in the implementation of extensive government restrictions on economic activity, particularly on sectors such as hospitality, retail and, travel and tourism. Reliable forecasts of the changing number of COVID-19 infection cases are, therefore, necessary for managing activity in a wide range of domains.

This necessity has resulted in the emergence of forecasting models focussing on factors such as the changing number of confirmed infection cases for future periods. In this regard, forecasts that are as accurate as possible are essential for the implementation of measures to control the virus and provide a leading indication of the future need for medical and associated services, as well as implications for the wider economic environment. Various available forecasting models use cumulative quantiles to provide prediction

intervals for a range of probability values. For example, the Git COVID-19 Forecasting Hub (2021a) provide quantile forecasts on changes in infection cases from a range of organisations that provide forecasts for the US and individual US States. To examine the accuracy of such forecasts we propose an analytical framework for COVID-19 predictions using *empirical quantiles*. The *empirical probability* technique (Pollock *et al*., 2005; 2008; 2010), which was extended to the empirical quantile technique (Thomson *et al*., 2021), is further broadened to incorporate first order correlation values in the present paper. Specifically, this study obtains estimated quantiles for daily changes in the number of COVID-19 confirmed infection cases over short periods of 14-days, with daily changes that exhibit a mean with a degree of first order autocorrelation. The procedure used in this study considers that the series has residual standard errors, generated by a simple *Autoregressive Order One*, *AR(1)* process with a constant, that follow an approximate identically and identically normal distribution. The empirical quantile technique presented here uses the Student t distribution at each time-period (14-day period) to estimate the standard error. This estimated sample standard error, with an adjustment, for each of the two-week periods that can then be used, with the mean daily change, to give an estimated probability distribution that allows the empirical quantiles to be derived for each cumulative quantile probability. Two-week ahead quantile forecasts for changes in confirmed cases for the US from four organisations are examined in the present study using the *Mean Squared Quantile Performance Score*, *MSQPS* (Thomson *et al* (2021). The demonstration of this analytical framework to ascertain the quality of these COVID-19 forecasts was, therefore, our first objective.

A second objective was to provide a mechanism within our framework to determine the most accurate combination of forecasts. **C**ombining forecasts as a method of increasing the forecast accuracy of individual predictions has, in fact, received considerable attention in point forecasting literature. In this context, composite forecasts have been shown to improve forecast accuracy and reduce the variance of errors (e.g., Armstrong, 2001; Taylor & Taylor, 2021; 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**;** Wallis, 2011). Combining forecasts are particularly beneficial when individual forecasts are based on dissimilar information sets (Wallis, 2011). This is because there is likely to be relative inconsistencies between the individual predictions that make up the composite forecasts, which, in turn, augment individual levels of accuracy and cancel out error. It is illustrated that the *Mean Squared Quantile Coherence Score* (*MSQCS*) between pairs of forecasts, introduced in Thomson *et al* (2021) can be applied to illustrate how relative error reduction in combined forecasts depends on the extent to which the component forecasts contain independent information. Accordingly, the second objective aimed to incorporate a measure of coherence (consistency), into the proposed quantile-based analytical framework so that the most dissimilar forecasts could be identified and combined to achieve the most accurate merged result.

The role of coherence can have important implications when considering collaborative efforts between forecasting institutions. In relation to the forecasting of deaths, Reich *et al*. (2019) showed that collaborative efforts in infective disease forecasting between research teams to develop composite or ensemble forecasting approaches brought measurable improvements in forecast accuracy and important reductions in the variability of performance. Cramer *et al.* (2021) and Ray *et al.* (2020) found that, in relation to COVID-19 deaths, ensemble forecasting models showed the best overall accuracy of any model. Taylor & Taylor (2021) found that aggregating COVID-19 quantile forecasts improved accuracy compared with single models. Ray *et al.* (2020) used point forecasts, but Cramer *et al.* (2021) and Taylor & Taylor (2021) analysed quantile probability forecasts using the *Weighted Interval Score* (*WIS*). Research using data from the *Git COVID-19 Forecasting Hub* (2021a) has generally used the *WIS*, as described in Bracher *et al.* (2020). These results emphasise the role that collaboration and active coordination between forecasting organisations can play a vital role in developing the modelling capabilities to support the responses of decision makers to COVID-19 outbreaks. These studies did not, however, directly examine the role of coherence and its ability to identify the most effective combinations.

To proceed along these lines, we next briefly describe how our previous research is extended to achieve these aims. In a point forecasting context, Thomson *et al.* (2019) demonstrated that forecast performance using the Mean Squared Error (*MSE*) could be improved by adding a measure of coherence to the analytic framework and combining the most dissimilar forecasts. Thomson *et al* (2021) then showed that the MSE could be extended to apply to a set of cumulative quantile predictions with respect to each quantile probability using the *MSQPS*. This work is extended by linking the *MSQPS* and the *MSQCS*.

To demonstrate the application of the framework, the current study employed two-week quantile forecasts on the changes in the number of confirmed infection cases from four forecasting models provided from the Git COVID-19 Forecasting Hub (2021a) for the US. These four models made quantile probability predictions with the median and point predictions, and six quantiles. The current study particularly examined 18 two-week ahead forecasts made each Saturday for forecast dates from 15/08/2020 to 10/04/2021 which were evaluated using daily data from 02/08/2020 to 10/04/2021. Empirical quantiles were derived and then used to examine the overall and specific aspects of the four forecasting agencies performance and coherence in this period and to undertake comparisons used to obtain composite forecasts.

The remainder of the paper is set out as follows. Section 2 provides a description of the statistical forecast analysis used to derive empirical quantiles. Section 3 sets out the statistical measures of performance and coherence. Section 4 describes the actual COVID-19 case data used, their background and statistical characteristics. This is followed by Section 5 which describes the forecast data used and presents the results of the study. Finally, discussion and concluding observations are provided in Section 6.

2. Statistical Analysis Using Empirical Quantiles

A major problem with examining quantile case 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 easy to compare point predictions with the resulting actual 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 the predictive performance of a set of quantile forecasts.

To examine quantile forecasts for changes in the number of confirmed infection cases, empirical quartiles were derived from the actual values to allow the quantile forecasts to be evaluated. The framework extends the empirical quantile probability technique set out in Thomson *et al*. (2021), to consider situations where changes in the variable follow an *autoregressive first order*, *AR(1)*, process This section describes the derivation of empirical quantiles and illustrates that a simple statistical test can be applied to the *AR(1)* model coefficient for each 14-day period that can be used to aid interpretation.

2.1 Derivation of Empirical Quantiles

The quantile forecasts available for COVID-19 infection cases available were in the form of changes or first differences rather than the cumulative number of cases over weekly periods. In examining these quantile forecasts, it was, however, necessary to use actual data in the form of cumulative total COVID-19 infection cases which were available daily. Therefore, to obtain two-week estimated quantiles it is necessary to consider the actual changes in the cumulative case variable, $\Delta X_{j,i} = X_{j,i} - X_{j,i-1}$ for day, *i*, *i=1,2,…,n*, of the series over a two-week period, *j, j=1,2,…,k*. For instance, for a two-week period of 14 days, *n=14*, the sum of the actual daily changes, over the two-week 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 demonstrated in this study that two-week data on changes in the number of cases tends to exhibit a high degree of first order autocorrelation which also occurs, albeit, to a lesser extent, in daily data within a two-week period. It is, therefore, necessary to take this into account when deriving empirical quantiles over short horizons, such as two-weeks. In this study, therefore, empirical quantiles derived using the assumption that daily changes in the number of cases follow a first order autoregressive process, with a constant included, with residual errors that follow normal distribution. This distribution can be considered to have time varying means, first autocorrelation values and residual standard errors, which can be assumed constant over short horizons, such as a two-week period, but exhibit variation over longer horizons. That is the daily changes, follow an *autoregressive first order, AR(1)* process with stable mean and first order autocorrelation and residual standard errors that are approximately independently normally distributed over the period, *i, i=1 to n*, for week, *j*. This estimated relationship is defined in equation (1):

$$\Delta X_{j,i} = a_j + b_j \Delta X_{j,i-1} + v_{j,i} \quad (1)$$

where a_j is a constant, b_j is the first order autocorrelation coefficient and v_j error term, with

$$b_j = C(\Delta X_{j,i}, \Delta X_{j,i-1}) / V(\Delta X_{j,i-1})$$

$$a_j = M(\Delta X_{j,i}) - b_j M(\Delta X_{j,i-1})$$

where

$M(\Delta X_{j,i}) = \frac{1}{n}\sum_{i=1}^{n} \Delta X_{j,i}$ and $M(\Delta X_{j.i-1}) = \frac{1}{n}\sum_{i=1}^{n} \Delta X_{j,i-1}$

$V(\Delta X_{j,i-1} = \left(\frac{1}{n}\sum_{i=1}^{n} \Delta X_{j,i-1}^2\right) - M^2(\Delta X_{j,i-1})$

$C(\Delta X_{j,i}, \Delta X_{j,i-1}) = \left(\frac{1}{n}\sum_{i=1}^{n} \Delta X_{j,i}\, \Delta X_{j,i-1}\right) - M(\Delta X_{j,i}, \Delta X_{j,i-1})$

v denotes a normally distributed residuals with zero mean, that is:

$v_{j,i} = \Delta X_{j,i} - a_j - b_j \Delta X_{j,i-1}$

The procedure used to obtain empirical quantiles is set out below.

The mean values, M_j, of the changes, $\Delta X_{j,i}$, for a two-week period, j, $j=1,2,\ldots,k$, is defined as $M(\Delta X_{j,i})$.

The volatility of the changes, $\Delta X_{j,i}$, over a one-week, j, termed the *adjusted root mean squared error* (*ARMSE*), S_j, is calculated using the *root mean squared error* (RMSE) of the residuals $v_{j,i}$, multiplied by an adjustment term that involves the square root of $1/(1- b_j^2)$, S_j. The adjustment considers that the theoretical variance of $\Delta X_{j,i}$, in a stationary AR(1) process is the error variance divided by unity less the square of the first order autoregressive coefficient. S_j, as defined in equation (2):

$$S_j = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n} v_{j,i}^2} * \sqrt{\frac{1}{1-b_j^2}} \qquad (2)$$

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 two-week change in the actual value, ΔX_j, would be below the value $d_{\alpha,j}$, is defined in equation (3):

$$d_{\alpha,j} = n\left\{M_j + F_{n-2}^{-1}(\alpha) * \frac{S_j}{\sqrt{n}}\right\} \qquad (3)$$

where $F_{n-2}^{-1}(\alpha)$ is the inverse cumulative distribution function of the Student t-distribution with *n-2* 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 residual standard error, S_j, is used to give a value $d_{0,k}$, as presented in equation (4):

$$d_{0,j} = F_{n-2}\left\{\sqrt{n} * \left(-\frac{M_j}{S_j}\right)\right\} \qquad (4)$$

where F_{n-2} denotes the cumulative distribution function of the Student t-distribution with *n-2* degrees of freedom.

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

$$e_{\alpha,j} = n\{M_j + F_{n-2}^{-1}[\alpha(1 - d_{0,j}) + d_{0,j}] * (\frac{S_j}{\sqrt{n}}\} \qquad (5)$$

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

2.2 Statistical Tests on the Autoregressive First Order Model Coefficients

The statistical significance of the first order autoregressive coefficient b_j, set out in equation (1), can also be obtained.

The standard error of b_j, $se(b_j)$, is defined in equation (6):

$$se(b_j) = \frac{S_j}{\{s(\Delta X_{j,i-1}) * \sqrt{n-1}\}} \qquad (6)$$

where *s* denotes the sample standard deviation,

$$s(\Delta X_{j,i-1}) = \sqrt{\left\{\frac{1}{n-1}\sum_{i=1}^{n}[\Delta X_{j,i-1} - M(\Delta X_{j,i-1})]^2\right\}}$$

The Student t-value, $t(b_j)$, can be calculated using the ratio of the coefficient b_j, divided by the standard error, $se(b_j)$, to give a quantity $t(b_j)$ defined in Equation (7):

$$t(b_j) = \frac{b_j}{se(b_j)} \quad (7)$$

The one tailed probability value on the right tail for $t_j(b_j)$, $(pv(b_j)$, can be calulated as defined in equation (8):

$$pv(b_j) = 1 - F_{n-2}[t(b_j)] \quad (8)$$

where F_{n-2} is the cumulative distribution function of the Student t-distribution with $k–2$ degrees of freedom.

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

3. Performance, Coherence Measures and Composite Forecasts

To evaluate performance, the forecasts, $q_{\alpha,g,j}$, at each cumulative quantile probability, α, for each forecaster, g, $g=1,2,\ldots,G$, and each observation period, j, can be compared with the empirical quantile values, $e_{\alpha,j}$. This can be carried out for each forecaster g. The median or point forecast, $q_{0.5,g,j}$ or $q_{p,g,j}$, and the actual forecast change, $e_{0.5,j}$, respectively, for two-week period j, provide a direct evaluation of the accuracy of the median/point forecast. The empirical quantile value at quantile probability 0.5, $e_{0.5,j}$, is the same as the actual change. Composite forecasts can be similarly compared using simple averages of a group of forecasters. For instance, composite forecasts for each quantile, α, $q_{\alpha,m,j}$, for period j for all forecasters, denoted m, are obtained by taking the simple average of these individual forecasts for period j. That is: $q_{\alpha,m,j} = \frac{1}{G}\sum_{i=1}^{g} q_{\alpha,g,j}$. The simple arithmetic average has been shown to be the most popular method. Evidence for point estimates suggests that it is difficult to outperform this simple average (Schnaars, 1986; Clemen, 1989; Makridakis & Hibon, 2000; Stock & 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. Thomson *et al.* (2021) effectively used the simple average in the context of cumulative quantile predictions.

To evaluate paired coherence between two forecasters, the forecasts for a forecaster g, $q_{\alpha,g,j}$, can be compared with forecaster h, $q_{\alpha,h,j}$, where $g=1,2,\ldots,G-1$, $h=2,3,\ldots,G$, $h>j$. The performance and coherence measures used in this analysis are discussed below.

3.1 The Mean Squared Quantile Performance Score

The overall performance of a set of forecasts for each quantile cumulative probability, α, for forecaster g, can be measured by the *mean squared quantile performance score* ($MSQPS_{\alpha,g}$), 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 $MSQPS_{\alpha,m}$ for the composite forecaster can be similarly defined. The $MSQPS_{\alpha,g}$ and $MSQPS_{\alpha,m}$ are defined in equations (9a and 9b) respectively:

$$MSQPS_{\alpha,g} = \frac{1}{k}\sum_{j=1}^{k}(q_{\alpha,g,j} - e_{\alpha,j})^2 \quad (9a)$$

$$MSQPS_{\alpha,m} = \frac{1}{k}\sum_{j=1}^{k}(q_{\alpha,m,j} - e_{\alpha,j})^2 \quad (9b)$$

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 *MSQPS* at quantile α the poorer the forecast performance.

3.2 Mean Squared Quantile Coherence Score between Two Individual Forecasters

The forecasts for *forecaster g* ($g=1,2,\ldots,G-1$), at quantile α, denoted $q_{\alpha,g,j}$, can be compared with that of *forecaster h* ($h=2,3,\ldots,G$, $h>g$), denoted $q_{\alpha,h,j}$, to examine coherence (consistency) between the predictions. When predictions are totally coherent for a specific period, j, at quantile α, these values should be equal, i.e., $q_{\alpha,g,j} = q_{\alpha,h,j}$, so that situations where the values are not equal reflect a degree of lack of coherence (inconsistency). The forecasts for all j periods ($j=1,2,\ldots,k$), can be compared, for each pair of forecasters g and h, at each quantile α, to provide a measure for a set of paired forecasts. There will be $G(G-1)/2$ sets of paired values for a total for G forecasters.

The *Mean Squared Quantile Coherence Score* (*MSQCS*), is an overall measure of coherence obtained from forecasts, $q_{\alpha,g,j}$ and $q_{\alpha,h,j}$,. The *MSQCS*, at quantile α, between each pair of forecasters, g and h, over the j periods, is defined in equation (10):

$$MSQCS_{\alpha,g,h} = \frac{1}{k}\sum_{j=1}^{k}(q_{\alpha,g,j} - q_{\alpha,h,j})^2 \quad (10)$$

A value of zero would imply that the two individuals, *g* and *h*, have made perfectly coherent predictions.

3.3 Linking Coherency and Performance Measures

A statistical link exists between the performance measures for individual forecasters, composite forecasts, and the coherence between the individual pairs of forecasts, set out in equations (9a, 9b and 10). The performance measures for composite forecasts measures can be directly obtained from the performance measures of the individual forecasters and the paired coherency measures. Composite forecasts tend to give measurably better predictions than individual forecasts because of lack of coherence, at quantile α, between the paired sets of forecasts between the individuals. $MSQPS_{\alpha,m}$ is the sum of the $MSQPS_{\alpha,g}$ for all g forecasters divided by G less the sum of the $MSQCS_{\alpha,g,h}$ for all pairs of forecasters divided by the square of G. This is illustrated in equation (11):

$$MSQPS_{,\alpha,m} = \frac{1}{G}\sum_{g=1}^{G} MSQPS_{\alpha,g} - \frac{1}{G^2}\sum_{g=1}^{G-1}\sum_{h=2,h>g}^{G} MSQCS_{\alpha,g,h} \quad (11)$$

Equation (11) illustrates that the *MSQPS* for composite forecasts, at quantile α, $MSQPS_{\alpha,m}$, is lower than the mean *MSQPS* of individual forecasters, $MSQPS_{\alpha,g}$, due to non-zero paired coherence values reflected in the sum of the paired *MSQCS* values that have a negative effect on the composite *MSQPS*.

3.4 The Relative Percentage Improvement of Composite Forecasts

The relative improvement in performance achieved by the composite forecaster compared with the relevant average of the individual forecasts can be obtained using the paired coherence measures. This is demonstrated using equation (11) by dividing the second term in this equation by the first term and then multiplying the result by 100 to give a value in percentage terms. This gives the following *Relative Percentage Mean Squared Quantile Score*, $RPMSQPS_{\alpha,m}$ measure for composite forecasts, for each quantile α, set out below in equation (12).

$$RPMSQPS_{\alpha,m} = 100 * \left\{ \frac{\frac{1}{G^2}\sum_{g=1}^{G-1}\sum_{h=2,h>g}^{G} MSCE_{\alpha,g,h}}{\frac{1}{G}\sum_{g=1}^{G} MSPEI_{\alpha,g}} \right\} \quad (12)$$

Equation (12) shows the relative percentage improvement of the *MSQPS* made by using composite forecasts compared with the mean of the *MSQPS* of the relevant individual forecasters at each quantile α.

4. The COVID-19 Case Data and Its Statistical Characteristics

The framework set out in Section 2 was applied to provide empirical quantiles for the two-week changes on the number of COVID-19 cases for the US. The data and their characteristics are described below.

4.1 The COVID-19 Case Data

To understand the scale of the COVID-19 pandemic at any given time it is important to be able to obtain a measure of the total number of people affected by the virus. This can be considered in many ways, including numbers of confirmed infection cases, numbers of deaths, people affected at an economic level or using other socio-economic metrics. For the purpose of this study, we determined that the total number of cases of infection at a given point in time (i.e., confirmed cases) was the best metric to inform forecasting models as it provides a balance between a concrete and clinically meaningful outcome measure to inform the statistical modelling effectively while also allowing relatively early intervention measures to be applied in practice to support prevention of death. We recognise that the number of confirmed cases will differ from the number of actual cases present in a population: in practice, the total number of cases will not be known precisely, therefore, the number of confirmed cases is usually used. Richie and Roser (2021) points out that for a case to be confirmed, the infected person needs to have a positive result from a laboratory test (or test of similar accuracy). The test result would be irrespective of whether the person shows symptoms of COVID-19 or not. The number of confirmed cases is, therefore, likely to be lower than the actual number of cases as not everyone with the virus is tested.

In the application of quantitative analysis to examine COVID-19 cases it is usually desirable to consider changes in the cumulative number of COVID-19 confirmed cases, rather than the cumulative total. The data provided by forecaster available through the Git COVID-19 Forecasting Hub (2021a) are in this form. Richie and Roser (2021) also point out that the number of confirmed cases reported by an institution, for

example, Johns Hopkins University, for a given day, would not necessarily represent the actual number of cases confirmed on that day, due to reporting lags between the identification of a new case and its inclusion in the reported statistics. There are sometimes also lags in reporting of confirmed cases at a national level (Sutherland, Headicar, & Delong, 2021), with lags in reporting occurring, for example, over weekend periods, with increases in confirmed cases seen at the start of the week following weekend testing/reporting lags, among other variables. It is therefore important to consider daily data across a time period to take these potential lags into consideration, and to explicitly test for these as part of the modelling process (as described in Section 4.2).

The framework, set out in Sections 2 and 3, was applied to changes in the two-week quantile case forecasts on the number of confirmed number of cases, in the US, provided by four institutions. The two-week ahead incident forecasts and composite forecasts were analysed using actual daily data during a period from 02/08/2020 to 10/04/2021 with forecasts for two-week periods ending from 15/08/2020 to 10/04/2021 with the two-week periods numbered 1 to 18. Specifically, the daily cumulative and incident number of confirmed US cases were obtained from the data reported by *Johns Hopkins University* (*JHU*) *Coronavirus Research Centre*. The data used came from the *Los Alamos National Laboratory* (LANL, 2021) website so that the data available at the time of the forecast and the end of the forecast period were consistent with values available at that time rather than being affected by subsequent updates. This institution provides daily data on both cumulative and incident confirmed US COVID-19 cases available in Microsoft Excel format. As this data is subject to regular updates that are back dated it was necessary to splice the data between the revisions using the LANL data when this occurred. The adjustment involved using the last day overlap of the cumulative case data denoted day t, for the original data, $X_{O,t}$, and the revised data, $X_{R,t}$. The resulting daily change for day t, was obtained as $\Delta X_t = X_{O,t} - X_{O,t-1}$, and day t+1, was obtained as $\Delta X_{t+1} = X_{R,t+1} - X_{R,t}$. One other adjustment was made to the daily change on 19/02/2021 with the value of 106 thousand replaced by a value of 79 thousand, as there appears to be an administrative data revision that can be identified from subsequent data updates.

4.2 Statistical Characteristics of the US COVID-19 Case Data

The time series of the changes in the number of confirmed infection cases were analysed over an extended period from 01/03/2020 to 10/04/2020 which consisted of 58 weekly periods ending on Saturday or 29 two-week periods also ending on Saturday. The two-week periods with weeks ending on 14/03/2020 to 10/04/2021 were numbered from

-10 to 18. The forecast performance and coherence analysis were restricted to the 18 two-week periods ending on 15/08/2020 to 10/04/2021, numbered 1 to 18. The changes in the number of cases for the 29 two-week periods is presented in Figure 1, showing the partition of the data used for the performance and coherence analysis section. Figure 1 also shows the point forecasts from four forecasting models which are discussed in the next section. The data showed relatively low changes in case values that were below 1 million for two-week periods numbered 6 and below, although the values in the early periods could be particularly affected by testing limitations. For the two-week periods numbered 7 to 10 the data showed a rapid increase in the cases to around 3 million, before falling back to around 2.75 million in the two-week period 11. There was then a subsequent rise in two-week period 12 to around 3.25 million after which there was a very sharp fall to two-week period 16 to under 1 million although there was a small rise after this period to two-week period 18. The two-week periods numbered 1 to 18 gave a series that showed sharp upward movement reflecting the rapid rise in infections in November and December 2020 followed by a sharp downward movement in January and February 2021 that reflected the impact of vaccinations and government restrictions. The series showed a clear turning point at the two-week period 12, hence it was considered that this was appropriate period to demonstrate the framework set out in this study. The above characteristics of the changes in the infection cases over this period illustrate that this was a difficult series to forecast for this period.

The daily changes in the number of confirmed cases were used to obtain the two-week empirical quantiles. This involved taking first differences of the actual values of the cumulative number of cases on each of the 14 days, for each of the 18 consecutive, two-week periods (numbered 1 to 18). To consider if the data were affected by any weekday or weekend effects, a simple one-way ANOVA was used. The one-way ANOVA was applied to the daily case changes with respect to for the extended period from 01/03/2020 to 10/04/2021. The results supported the assumption of the equality of the daily means with the F statistic

clearly non-significant. In addition, the paired Student t-test values also showed non-significant for all the pairs of values. Levene's test (Levene, 1960) for inequality of variance also clearly showed non-significance. Therefore, there does not appear to be any significant day-of-the-week effects on the recording of changes in US COVID-19 cases over the period.

In Section 2, it was set out that the framework considers that the daily changes in cases can follow an AR(1) process with stable mean and first order autocorrelation and with the residual standard errors considered to be approximately independently normally distributed over each of the two-week periods. Lilliefors test (Lilliefors, 1967) and the Shapiro-Wilk test (Shapiro and Wilk, 1985) for normality were applied to the 14 daily residuals for each of the two-week periods from 1 to 18. The statistics showed non-significance, except for two of the two-week periods numbered 7 and 17, which were significant at 5%. Week 7 was significant due to two outlier large negative values that occurred on Sunday, 25/10/2020 and Saturday, 31/10/2020, and week 17 was significant due to two outlier values, one negative on Sunday, 21/03/2021 and one positive on Wednesday, 24/03/2021. On balance, it can, therefore, be concluded that the assumption that the *root mean squared error* (*RMSE*) of the 14 daily values of the residuals followed an approximate normal distribution over the 18 two-week periods, was reasonable, which was used in the derivation of the empirical quantiles. The results also showed that the impact of adjusting the empirical distribution, to take account of the changes in the cumulative number of cases being restricted to non-negative values, had virtually no effect on the results.

Table 1 presents values for the date, two-week period numbers, the two-week changes in case numbers in thousands, the daily means and the residual *RMSEs* of cases, the AR(1) coefficient and its probability value, and the constant coefficient, for the 18 two-week periods for two two-week periods ending from 15/08/2020 to 10/04/2021.

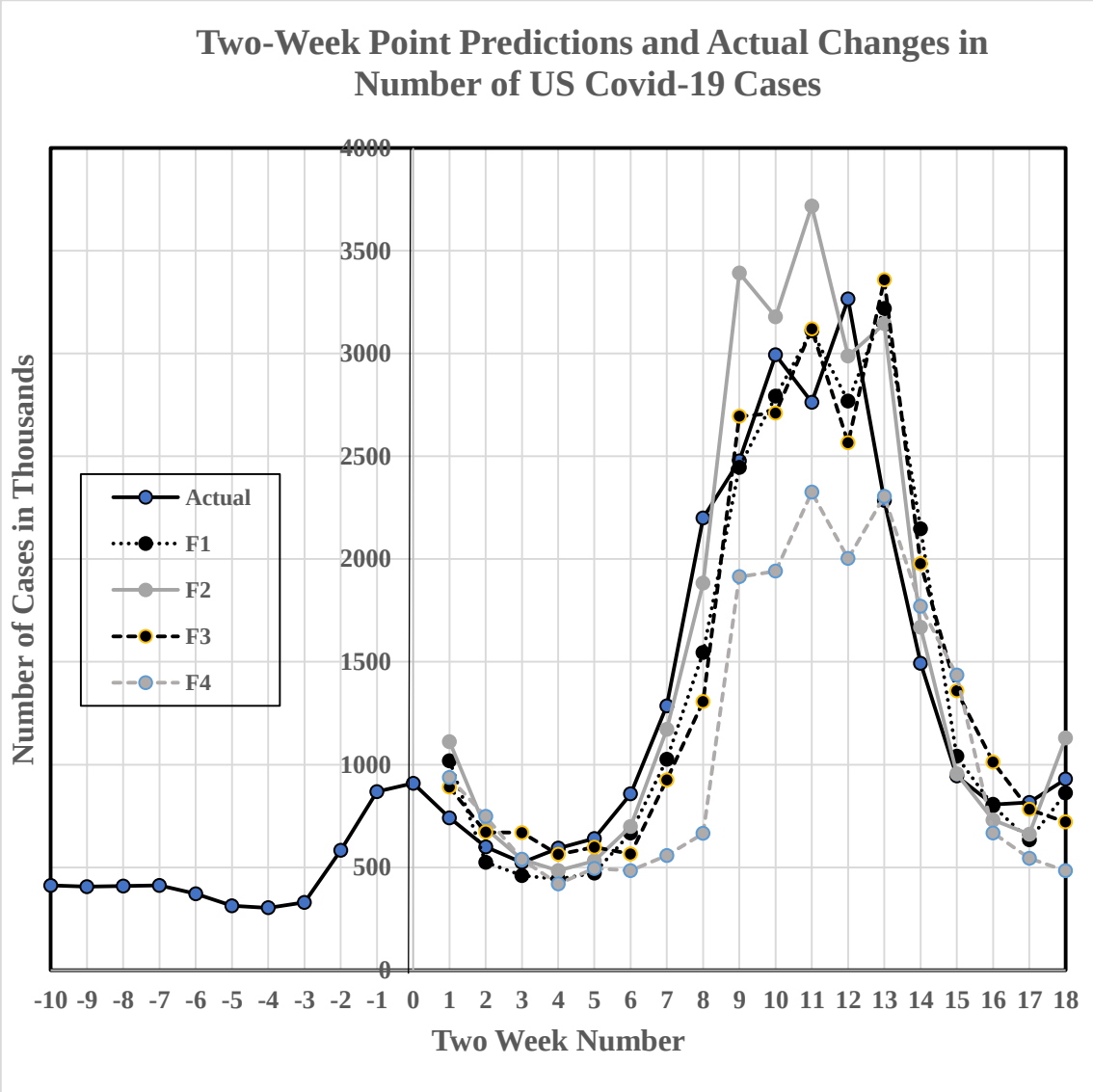
Table 1 : Statistics on the Actual Data for the 18 Two-Week Periods

Date	Two Week No	Actual Change Thous	Daily Mean Thous	Daily ARMSE Thous	AR(1) Coeff b	AR(1) PV b	Const Coeff a
15/08/2020	1	741	53	6	-0.110	0.645	59
29/08/2020	2	600	43	5	0.322	0.127	29
12/09/2020	3	524	37	7	0.388	0.074	23
26/09/2020	4	593	42	7	0.180	0.271	35
10/10/2020	5	640	46	7	0.648	0.010	17
24/10/2020	6	857	61	8	0.906	0.000	8
07/11/2020	7	1286	92	12	0.979	0.000	5
21/11/2020	8	2200	157	17	0.752	0.001	42
05/12/2020	9	2477	177	34	0.283	0.176	128
19/12/2020	10	2993	214	23	0.343	0.121	140
02/01/2021	11	2762	197	46	-0.392	0.857	272
16/01/2021	12	3265	233	30	0.351	0.086	149
30/01/2021	13	2285	163	18	0.352	0.089	104
13/02/2021	14	1493	107	15	0.508	0.023	50
27/02/2021	15	944	67	8	0.282	0.137	48
13/03/2021	16	806	58	7	0.396	0.077	34
27/03/2021	17	816	58	13	0.330	0.124	39
10/04/2021	18	929	66	14	0.023	0.468	65
Mean	0	1529	109	16	0.387	0.168	71

The means presented in Table 1 are the two-week change divided by 14, therefore they show an identical pattern to the two-week changes which were discussed in relation to Figure 1. The Adjusted RMSE was variable reflecting the differing variation of the daily changes for each of the two-week periods, although the

values tended to show overall that the magnitude of values were reasonably closely related to that of the mean. The AR(1) coefficients were quite variable with a minimum (negative) value in the two-week periods numbered 11 of -0.392 and a maximum (positive) value of 0.979 in two-week period 7. The values were positive for 16 out of the 18 values. The values were significant at 1% for the two-week periods numbered 5, 6, 7 and 8, and significant at 5% for two-week period numbered 14. The high positive values between the periods numbered 5 and 8 were associated with the periods of rapid rise in the number of cases. The constant coefficient values are also variable and appear inversely related to the AR(1) coefficient values. These results tend to support the view that the parameters of the underlying distribution were subject to considerable changes from each two-week period to the next. These results underly the assumptions used in the derivation of the empirical quantiles that were used in the performance analysis presented in the next section.

Figure 1



Another characteristic of the data on two-week changes in the number of infection cases is that the data exhibits strong first order autocorrelation. The ACF value for the 18 two-week periods gives a value of 0.851 which was significant at the 1% level based on a Bartlett test (Bartlett, 1946). To test the null hypothesis that a unit root was present in an AR model the Dickey - Fuller (DF) test (Dickey and Fuller, 1981) and Augmented DF (ADF) test was used. The alternative hypothesis is that the series is stationary. The result illustrated that an ADF test, with a constant and a lag of 2 was significant at 5%, but not at 10%. The ADF test, however, can be week for small samples (Afriyie, Twumasi-Ankrah, Gyamfi, Arthur, & Pels, 2020) so the results are inconclusive. The high first order autocorrelation, together with the changing

parameters, implies that these factors need to be incorporated into quantile probability forecasts on two-week changes in COVID-19 infection cases.

5. The Forecast Data and the Performance and Coherence Analyses Results

This section presents the characteristics of the forecast data and the performance and coherence analysis results using the framework set out in Sections 2 and 3. The forecast quantiles, $q_{\alpha,g,j}$, from four models, *g=1 to 4*, and composites, $q_{\alpha,m,j}$, were compared with the empirical quantiles, $e_{\alpha,j}$. The analysis was applied to forecasts for over 18 two-week periods from 15/08/2020 to 10/04/2021 with weeks numbered 1 to 18. The characteristics of the forecast data and the application of the performance and coherence analysis is discussed below.

5.1 The Forecast Data

The forecasts data used in this study were obtained from data available in Microsoft Excel format from Git COVID-19 Forecasting Hub (2021a) using case incident quantile forecasts for quantile probabilities 0.025, 0.1,0.25, 0.5, 0.75, 0.9 and 0.975 and point forecasts. The two-week ahead forecasts were obtained using the sum of the data on the '1 week ahead incident case' and '2 week ahead incident case' for four forecasters which used the same forecasts dates. The Forecasting Hub Excel spreadsheets did not give the cumulative cases data or the input actual values when the forecasts were made, hence this could have had an influence on the results. That is, the forecasters could have used actual case values when the forecasts were made that could have been different from the actual case values used in this analysis. The forecast data used came from four models for 18 two-week periods ending on Saturdays from 15/08/2012 to 10/04/2021 for two-week periods numbered 1 to 18. These models were the Covid19Sim -Simulator denoted F1, the Karen – pypm denoted F2, the IowaStateLW-STEM denoted F3 and the JHU-IDD-CovidSP denoted F4 with the data obtained from Git COVID-19 Forecasting Hub (2021b,c,d,e respectively). There was, however, no forecasts available for F4 for two weeks before week 8, (21/11/2020), hence the relevant forecasts available from three weeks before were used. This study only used the model forecasts at face value and there was no direct attempt to consider the different models' specifications.

5.2 The Two-Week Point Forecast Values

The results shown in Figure 1 and Table 2 show the two-week ahead point predictions for periods numbered 1 to 18. Table 2 shows numerical two-week changes, the forecasts from the four forecast organisations and the percentage forecast errors. The results showed that the four model forecasts, F1 to F4 had predictions that were reasonably close to the actual changes. There were, however, some exceptions in particular periods. The four models showed that 28 out of 72 forecasts had moderate percentage forecast errors that were outside the +/–25% range and 4 out of 72 showed forecast errors that were outside the +/–50% range. Specifically, for the 18 two-week period, there were 6 moderate percentage forecast errors for F1, 4 for F2, 8 for F3 and 11 for F4 and 1 large forecast error for F2 and 3 for F4. Figure 1 and Table 2 showed that F4 tended to underestimate the major peak.

Table 2 : Actual and Point Forecasts and Percent Forecast Errors for the 18 Two-Week Periods

Date	Two Week No	Actual Change Thous	F1 F'cast Thous	F2 F'cast Thous	F3 F'cast Thous	F4 F'cast Thous	F1 % Error	F2 % Error	F3 % Error	F4 % Error
15/08/2020	1	741	1019	1111	890	937	38	50	20	26
29/08/2020	2	600	525	690	671	748	-13	15	12	25
12/09/2020	3	524	460	538	669	539	-12	3	28	3
26/09/2020	4	593	443	484	563	420	-25	-18	-5	-29
10/10/2020	5	640	472	531	598	494	-26	-17	-6	-23
24/10/2020	6	857	667	699	565	484	-22	-19	-34	-44
07/11/2020	7	1286	1025	1171	925	558	-20	-9	-28	-57
21/11/2020	8	2200	1546	1882	1307	665	-30	-14	-41	-70
05/12/2020	9	2477	2446	3391	2694	1913	-1	37	9	-23
19/12/2020	10	2993	2792	3178	2710	1941	-7	6	-9	-35
02/01/2021	11	2762	3111	3717	3119	2325	13	35	13	-16
16/01/2021	12	3265	2768	2988	2565	2002	-15	-8	-21	-39
30/01/2021	13	2285	3218	3144	3358	2305	41	38	47	1
13/02/2021	14	1493	2148	1668	1977	1770	44	12	32	19
27/02/2021	15	944	1040	955	1358	1435	10	1	44	52
13/03/2021	16	806	799	730	1012	667	-1	-9	26	-17
27/03/2021	17	816	633	661	782	544	-22	-19	-4	-33
10/04/2021	18	929	861	1129	721	484	-7	22	-22	-48
Mean		1529	1530	1680	1561	1200	-2	6	5	-14

5.3 The Two-Week Empirical and Forecast Quantile Values

There were also marked differences between the cumulative quantile forecast values between the four forecasters. While the mean difference between the 0.975 and 0.025 quantile values for the 18 values for these empirical quantiles was 324 thousand, the mean values for F1 and F3 were much less, 165 and 166 respectively, but the mean values for F2 and F4 were much higher, 594 and 526 respectively. This reflects the general characteristic that for this and the other quantile ranges F1 and F3 gave much narrower bands than F2 and F4. Table 3 presents the actual change and the values of the 0.025 and 0.975 quantiles for the empirical and the four forecast models for the 18 two-week periods. The results show that only in 3 two-week periods did the actual value fall in this 95% quantile range for F1, 11 for F2, 3 for F3 and 5 for F4. The results suggest that the performance on this aspect of F1, F3 and F4 was very poor in setting their forecast quantiles and F2 performance could only be described as poor to moderate. This characteristic is also a feature of the other, 80% and 50% quantile ranges. This general poor performance partly reflects the time series characteristics of the series which showed a rapid rise followed by rapid fall as well as difficult to predict movements at the beginning and the end of the series.

Table 3 : Actual, and 0.025 and 0.975 Empirical and Forecast Quantile Values for the 18 Two-Week Periods

Date	Two Week No	Actual Change Thous	Actual 0.025 Quant Thous	Actual 0.095 Quant Thous	F1 0.025 Quant Thous	F1 0.975 Quant Thous	F2 0.025 Quant Thous	F2 0.975 Quant Thous	F3 0.025 Quant Thous	F3 0.975 Quant Thous	F4 0.025 Quant Thous	F4 0.975 Quant Thous
15/08/2020	1	741	691	790	887	1138	1013	1252	835	951	755	1193
29/08/2020	2	600	561	639	504	562	638	798	606	743	625	897
12/09/2020	3	524	460	588	376	493	487	677	628	750	452	656
26/09/2020	4	593	537	649	435	446	460	624	535	597	345	503
10/10/2020	5	640	568	711	458	481	525	674	554	648	383	625
24/10/2020	6	857	699	1016	643	680	665	838	516	624	373	628
07/11/2020	7	1286	826	1746	969	1040	1051	1336	875	1008	416	739
21/11/2020	8	2200	1984	2415	1458	1662	1762	2102	1243	1373	490	894
05/12/2020	9	2477	2189	2766	1236	2632	3131	3599	2536	2915	1512	2467
19/12/2020	10	2993	2797	3188	2702	2853	2780	3725	2569	2871	1505	2456
02/01/2021	11	2762	2356	3169	3095	3116	3028	4763	2979	3297	1757	2950
16/01/2021	12	3265	3006	3524	2682	2781	2490	3732	2466	2667	1554	2583
30/01/2021	13	2285	2124	2446	3192	3259	2691	3829	3216	3500	1890	2780
13/02/2021	14	1493	1355	1630	2113	2182	1448	2196	1884	2073	1444	2193
27/02/2021	15	944	875	1013	908	1133	804	1431	1323	1394	1158	1831
13/03/2021	16	806	747	865	763	825	545	1012	742	1064	546	812
27/03/2021	17	816	701	930	615	654	460	1066	776	788	453	690
10/04/2021	18	929	813	1046	830	901	703	1726	715	725	384	618

5.4 MSQPS for Individual Forecasters

The *MSQPS* for the four forecast models, F1, F2, F3 and F4, are presented in Table 4. The best value on this measure is zero, hence the best performing model is identified as that having the minimum *MSQPS*. F2 and F4 also gave separate point forecasts as well as the 0.5 quantile probability. For all four forecast models the results showed that F1 had the best performance for quantile probabilities of 0.25 and above and point forecasts, and F2 had the best performance on quantile probabilities of 0.1 and below. F4 had the poorest performance of the four forecast models at all quantile probabilities.

Table 4 : The MSQS Individual Forecasters and Previous Empirical Forecaster

MSQPS					
Forecaster	**F1**	**F2**	**F3**	**F4**	**Pre Emp**
Quantile					
0.025	203,414	126,155	176,319	422,875	227,601
0.1	142,243	129,926	174,557	424,409	215,842
0.25	133,262	141,955	180,127	413,961	209,344
0.5	134,707	167,156	191,711	395,989	205,345
0.75	144,983	226,356	205,577	378,267	204,261
0.9	159,149	294,112	219,834	365,636	205,932
0.975	176,194	390,317	241,099	368,929	211,690
Point	134,707	166,974	191,711	386,466	205,345

As a standard of comparison, a simple naïve hypothetical forecaster was also presented in Table 4. The model used the empirical quantiles generated at the previous two-week period, as a predictor for the next two-week period. The results showed that this simple model produced better *MSQSP* values for the point estimates and all quantiles for F4, the 0.75 quantile for F3, and the 0.9 and 0.975 for F2 and F3. The reasonable performance of this naïve forecast model partly reflects the strong first order autocorrelation over the 18 two-week actual values.

5.5 MSQPS and RPMSQS for Composite Paired Forecasters

The *MSQPS* for paired composite forecasts and their relative improvement *RPMSQPS* for the six pairs of the forecast models, F1, F2, F3 and F4, are presented in Table 5, denoted C followed by 1, 2, 3, 4. Of the six paired forecaster composite models the C12 composite performed the best at all quantile probabilities and the point forecasts, although the performance at the 0.9 and 0.975 quantile probabilities were poorer than F1. The relative improvement, as measured by the *RPMSQPS*, for the C12 pair at all quantile probabilities was between 14 and 39, reflecting some degree of lack of coherence. Some of the composite forecasts performed worse than the naïve forecaster. This occurred for C14 at the 0.75 to 0.975 quantiles, C23 and C24 at the 0.975 quantile, and C34 at the 0.25 to 0.975 quantile and the point forecast.

Table 5 : MSQPS and RPMSQPS for Paired Forecasts

Forecaster	**C12**	**C13**	**C14**	**C23**	**C24**	**C34**
Quantile						
MSQPS						
0.025	101,164	159,295	196,241	128,245	138,164	193,520
0.1	117,707	150,730	178,268	128,368	139,196	212,357
0.25	117,948	149,135	186,068	134,704	135,174	223,649
0.5	123,953	155,766	199,223	147,790	135,173	233,854
0.75	141,614	167,577	214,273	173,248	148,197	244,620
0.9	161,702	181,857	227,905	196,732	169,609	257,481
0.975	184,776	199,738	246,193	225,822	212,081	281,242
Point	123,591	155,766	196,650	147,506	138,548	231,481
RPMESPM						
0.025	39	16	37	15	50	35
0.1	14	5	37	16	50	29
0.25	14	5	32	16	51	25
0.5	18	5	25	18	52	20
0.75	24	4	18	20	51	16
0.9	29	4	13	23	49	12
0.975	35	4	10	28	44	8
Point	18	5	25	18	50	20

The relative improvement shows the highest level of coherence occurred for the C13 pair with values of 4 and 5 for all quantile probabilities except 0.025. On the other hand, the C24 pair showed low levels of coherence, with relative improvement values between 44 and 52 at all quantile probabilities. These results imply that there was considerable variation in the paired performance and coherence which has important implications in the formation of composite forecasts.

5.6 MSQPS and RPMSQS for Composite Triple or More Forecasters

The *MSQPS* for triple or more composite forecasts and their relative improvement *RPMSQPS* for four groups of the forecast models, F1, F2, F3 and F4, are presented in Tables 6 denoted C followed by 1, 2, 3, 4. Regarding the five groups of composite models, presented in Table 6, the C124 composite performed the best out of the four for the point forecast and for all quantile probabilities, except at the 0.975 quantitative probability where the C123 composite performed marginally better. The C124 composite performed better at the 0.1, 0.75 and 0.9 quantile probabilities than all individual and composites models. The relative improvement of C124 composite gave values between 50, 44 and 42 at the 0.1, 0.75 and 0.9 quantile probabilities respectively illustrating a lack of coherence overall between these three forecast models. It is interesting to note that the overall composite C1234, did not produce the best forecast in comparison with the all the other fourteen possible individual and composite forecasters using the four forecast models. The overall composite did, however, perform better, on average, than 74% of these other forecaster

combinations, with a range from 64% for the 0.5 quantile probability and point estimates to 86% for the 0.9 quantile probability. The C134 composite forecaster performed worse than the naïve forecaster at the 0.9 to 0.975 quantiles.

Table 6 : MSQPS and RPMSQPS for Triple or More Composite Forecast Models

Forecaster	C123	C124	C134	C234	C1234
Quantile					
MSQPS					
0.025	116,547	109,981	154,846	123,819	113,062
0.1	126,721	116,012	158,246	132,310	122,764
0.25	127,978	118,620	167,562	137,562	128,006
0.5	135,163	126,172	181,441	145,824	137,744
0.75	150,316	140,748	197,451	161,562	152,984
0.9	165,341	157,552	213,817	179,523	168,980
0.975	181,526	181,862	235,830	208,471	190,396
Point	134,896	127,446	180,301	147,221	138,403
RPMESPM					
0.025	31	56	42	49	51
0.1	15	50	36	46	44
0.25	16	48	31	44	41
0.5	18	46	25	42	38
0.75	22	44	19	40	36
0.9	26	42	14	39	35
0.975	33	42	10	37	35
Point	18	44	24	41	37

Combining the information from Tables 4 to 6, the results imply that the use of the overall composite may not be the best method to obtain the best possible forecasts. For the four forecast models and composites the best forecast came from the combination C12 for quantile probabilities 0.025, 0.25 and 0.5 and point forecasts, combination C124 for quantile probabilities 0.1, 0.75 and 0.9 and individual forecaster F1 for quantile probabilities 0.975. These results emphasized the important role of coherence in deciding composite forecast combinations. For instance, although individual forecaster F3 performed much better that F4, F3's strong coherence with F1 resulted in this forecaster not figuring in any of the best forecast combinations. On the other hand, F4's lower coherence with F1 resulted in F4 being included in 3 of the best forecast combinations.

6. Discussion and Conclusion

An analytical framework was demonstrated for the evaluation of forecasts presented in the form of quantiles was applied to quantile forecasts of the two-week changes in the number of US, COVID-19 confirmed infection cases for forecast dates from 15/08/2020 to 10/04/2020, with two-week periods numbered 1 to 18. The framework compared two-week changes in the quantile forecasts from four forecasting organisations that make data available through the Git COVID-19 Forecasting Hub (2021a), and composite forecasts, using empirical quantiles. The empirical quantiles were obtained using daily changes (first differences) in the number of cumulative infection cases that incorporated first order autocorrelation using an AR(1) model with a constant. The procedure used the daily residual standard errors from this first order autoregressive process for each of the two-week periods, which were assumed to follow an approximate identically and identically normal distribution. In practice, the empirical quantile technique used the Student t distribution at each time-period (14-day period) to estimate the standard error. This allowed the estimated sample standard error to be obtained for each two-week period that was used after an adjustment, with the mean daily change, to give an estimated probability distribution that allowed the empirical quantiles to be derived for each cumulative quantile probability. This distribution of daily changes of the actual series was considered

to follow an approximate independent normal distribution, with a time-varying mean and first order autocorrelation and standard errors, but with approximately stable parameters over short periods of two-weeks. These empirical quantiles were then used to the undertake performance evaluation of the individual organisations and composite forecasts using the Mean Squared Quantile Performance Score (MSQPS). The framework was then extended to undertake coherence (consistency) evaluation between pairs of forecasters using the Mean Squared Quantile Coherence Score (MSQCS) which was integrated with the MSQPS to aid decision making on the most appropriate composition for composite forecasts.

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 with data that exhibits first order autocorrelation as well as a mean. The method described above was applied to daily first differenced COVID-19 data on the number of confirmed infection cases in the US, over 18 two-week periods, reported by JHU and available from LANL (2021). The framework used daily first differences over each two-week period to obtain estimates of the distribution mean and first order autocorrelation with the residuals used to obtain an estimate of the adjusted standard error from which empirical quantile cumulative probabilities were derived for each of the two-week periods. The statistical characteristics of the data were examined, and the results implied that the residual series satisfied the assumptions of approximate normality of the residuals over the 18 two-week periods. The results showed that the means and AR(1) model and constant coefficients were variable over this period. Empirical quantiles were derived for the median/point estimate and the six quantiles, and it was illustrated that these empirical quantiles could be used to undertake comparisons of the quantile forecasts with empirical quantiles.

The second objective was to extend available forecast evaluation methods used by Thomson et al. (2019; 2021) to cumulative quantile probability predictions to examine performance and coherence (consistency), within an integrated framework, which can be used in forming appropriate composite forecasts. The individual forecasters' performance was evaluated using the point estimate forecast errors at each two-week period and the MSQPS for the whole 18 two-week period for each of the four forecasting institutions. The results indicated that model forecast error were greater when the series exhibited accelerating or decelerating changes in the number of cases.

The results also showed that for the individual forecasters, F1 performed the best and F4 performed the poorest. The composite forecasts using F1 and F2 improved on the performance of individual forecasts, except at the 0.9 and 0.975 quantile probabilities, which can be explained by the good performance of these two forecasters and a degree of lack of coherence between them. The best composite forecasts using F1 and F2, and F4 occurred for the 0.1, 0.75 and 0.9 quantiles. Of particular interest was the distinct lack of coherence between F4 and F1 and F4 and F2, which allowed F4 to be included in the best composite forecasts for three quantiles, although F4 was the worse of the four individual forecasters except at the 0.975 quantile. This highlights that understanding coherence is a key factor in deciding the composition of composite forecasts.

Overall, the results show that the composite forecast using all forecasts may gave better forecasts than the individual forecasts in general, although the best forecaster could still be an individual forecaster in some situations. In addition, composite forecasts using all forecasters are unlikely to provide the best composite forecast due to the coherence, which can show substantial differences between forecasters. This has implications in the use composite or ensemble forecasts that use all individual forecasters as this is not likely to be optimal. Our findings, to some extent, supports the results of Reich et al. (2019) for infectious diseases and Cramer et al. (2021), Ray et al. (2020) and Taylor & Taylor (2021) for COVID-19 deaths who found that ensemble forecasting models showed the best overall accuracy of compared with specific models. 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 using coherence information.

The framework set out in this study can be applied to a wide range of forecasting situations when point or median predictions with cumulative quantiles or confidence intervals are used. The Git COVID-19 Forecasting Hub (GitHub, 2021a) provides comprehensive data from a wide range of forecasting institutions, including the median and various quantiles for a large range of institutions on the number of cases, hospitalisation and deaths for the US and US states. The Los Alamos National Laboratory (LANL,

2021) have provided data and forecasts for US states and most major countries of the world, which includes the median and 22 quantiles, for COVID-19 cases and deaths. These rich datasets in combination with the forecasting approach outlined in the current paper, afford future researchers and policy makers the ability to examine the usefulness of the framework that this paper has proposed on various aspects of COVID-19 cases, hospitalisations, and deaths to support and inform their decision making for their specific purposes.

Several modifications and extensions to the framework set out in this study can be suggested. Empirical quantiles are compatible with grouped empirical probabilities for analysing forecasts given in the form of grouped probabilities. One straightforward modification and extension could be the application of this modified framework to grouped probability individual and composite 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 and coherence of forecast changes in the number of confirmed infection cases from COVID-19 expressed as probabilities for a range of mutually exclusive bands representing changes. This would allow performance to be evaluated using a much smaller number of observations than the common method of analysing 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 observation period. Empirical probabilities would, in this instance, be an extension to the first objective discussed above.

The procedure set out in this study could be extended and applied to a wide range of forecasting situations where point and quantile 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 it was considered that the series exhibited an AR(1) process with a constant and the distribution of residuals assumed to be normal. The framework could be extended to other AR processes.

Another modification could involve the extension of the integration of performance and coherence measures to composite forecasts obtained from more than four 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 cases available for the US and US States from the Git COVID-19 Forecasting Hub (2021a) 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 Git COVID-19 Forecasting Hub (2021a), which forecast performance deteriorated as the prediction horizon increased from one to four weeks.

The empirical probability approach set out in this paper does have some limitations. It assumes that the values over a short period, such as a two-week period, that the residuals are approximately independently and identically normally distributed. Before applying the technique, it is appropriate to examine the data to verify that this assumption is appropriate. A further limitation is that the performance and coherence analysis used in this study uses measures associated with quadratic loss functions, specifically the MSQSP 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 analyse quantile forecasts such as the WIS, although these measures can be limited particularly in relation to coherence.

A further limitation is that the framework set out in the paper is that it has only 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, as a future extension, to consider how the performance and coherence results from different models change over time, for instance, in relation to the distinct phases in the path of changes in the number of COVID-19 confirmed infection cases and/or deaths. Extending the framework to consider changes in confirmed cases related to events such as protests and large gatherings or considering location-based changes (State-level differences in restrictions in the US, for example) may also add value in understanding more local level forecasts, should data be available at this more granular level. These extensions could further help support policy and clinical-level decision making when forecasting resource-need within emergency care, intensive care planning, and community level health interventions.

To summarize, 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 with data that exhibits characteristics such as autocorrelation. This can be a valuable aid in decision making and to policy makers where there is a need to be able to evaluate individual and composite quantile probability predictions, considering coherence, 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 essential as the ability to accurately predict changes in the number of recorded deaths, hospitalisation and cases is essential 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. These forecasts can also have major implications on the economic environment as governments need to respond actively to the changing infection statistics, which has resulted in the implementation of extensive government restrictions have resulted in major impacts on economic activity.

References

1. Afriyie, J. W., Twumasi-Ankrah, S., Gyamfi, K. B., Arthur, D. & Pels, W. A. (2020). Evaluating the performance of unit root tests in single time series processes. *Mathematics and Statistics, 8,* 656-664.
2. Armstrong, J. S. (2001). Combining forecasts. In J. S. Armstrong (Ed.), *Principles of forecasting: a handbook for researchers and practitioners* (pp. 417-439). Norwell, MA: Kluwer Academic Publishers.
3. Armstrong, J. S., Green, K. C., & Graefe, A. (2015). Golden rule of forecasting: Be conservative. *Journal of Business Research, 68,* 1717-1731.
4. Bartlett, M. S. (1946). On the theoretical specification of sampling properties of autocorrelated time series. *Journal of the Royal Statistical Society, Series B, 8,* 27-41.
5. BMJ. (2021a). Covid-19: Sore throat, fatigue, and myalgia are more common with new UK variant. *BMJ, 372,* n288.
6. BMJ. (2021b). Covid-19: Moderna and Pfizer vaccines prevent infections as well as symptoms, CDC study finds. *BMJ, 373,* n888.
7. BMJ. (2021c). Covid-19: Delta variant is now UK's most dominant strain and spreading through schools. *BMJ, 373,* n1445.
8. Bracher, J., Ray, E. L., Gneiting, T. & Reich, N. G. (2020). Evaluating epidemic forecasts in an interval format. *arXiv*. Doi: 10.1371/journal.pcbi.1008618
9. Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting, 5,* 559-583.
10. Cramer, E. Y. *et al.* (2021). Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the US. *MedRxiv*. https://doi.org/10.1101/2021.02.03.21250974
11. Davies, N. G. *et al.* (2021) Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. *Science, 372(6538),* eabg3055.
12. De Menezes, L. M., Bunn D. W., & Taylor, J. W. (2000). Review of guidelines for the use of combined forecasts. *European Journal of Operational Research, 120,* 190-204.
13. Dickey, A. D. & Fuller, W. A. (1981). Likelihood ratio statistics for autoregressive time series with a unit root. *Econometrica, 49,* 1057-1072.
14. Git COVID-19 Forecasting Hub. (2021a). Data on COVID-19 forecasts for the US. Retrieved from https://github.com/reichlab/covid19-forecast-hub/
15. Git COVID-19 Forecasting Hub. (2021b). Forecast data for the Covid19Sim – Simulator. Retrieved from https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/Covid19Sim-Simulator
16. Git COVID-19 Forecasting Hub. (2021c). Forecast data for the Karen – pypm model. Retrieved from https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/Karlen-pypm
17. Git COVID-19 Forecasting Hub. (2021d). Forecast data for the IowaStateLW-STEM model. Retrieved from https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/IowaStateLW-STEM

18. Git COVID-19 Forecasting Hub. (2021e). Forecast data for the JHU-IDD-CovidSP model. Retrieved from https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/JHU_IDD-CovidSP
19. Goodwin, P. (2015). Is a more liberal approach to conservatism needed in forecasting? *Journal of Business Research, 68,* 1753-1754.
20. Graefe, A., Armstrong, J. S., Jr., Jones, R. J., & Cuzan, A. G. (2014). Combining
21. forecasts: An application to elections. *International Journal of Forecasting,*
22. *30*, 43-54.
23. Green, K. C., Armstrong, J. S., & Graefe, A. (2015). Golden rule of forecasting
24. rearticulated: Forecast unto others as you would have them forecast unto you. *Journal*
25. *of Business Research, 68,* 1768-1771.
26. Harvey N. (2001). Improving judgment in forecasting. In: J. S. Armstrong (Ed.), *Principles of Forecasting. International Series in Operations Research & Management Science, vol 30* (pp. 59-80). Boston, MA: Springer.
27. Kissane, E., Rivera, J. M., Pearlstein, J, Walker, P. & Zonis, N. (2020). Cases matter. The COVID tracking project. Retrieved from https://covidtracking.com/analysis-updates/cases-matter
28. Kochanczyk, M., Lipniacki, T. (2021). Pareto-based evaluation of national responses to COVID-19 pandemic shows that saving lives and protecting economy are non-trade-off objectives. *Scientific Reports, 11,* 2425.
29. Levene, H. (1960). Robust tests for equality of variances. In I. Olkin, H. Hotelling, *et al.* (Eds.), *Contributions to probability and statistics: Essays in honor of Harold Hotelling* (pp. 278-292). California: Stanford University Press.
30. Lilliefors, H. W. (1967). On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association, 62,* 399-402.
31. LANL. (2021). Los Alamos National Laboratory. Retrieved from https://covid-19.bsvgateway.org/
32. Lubecke, T. H., Markland, R. E., Kwok, C. C. Y., & Donohue, J. M. (1995). Forecasting foreign exchange rates using objective composite models. *Management International Review, 35,* 135-152.
33. Makridakis, S. & Hibon, M. (2000). The M3-competition: Results, conclusions and implications. *International Journal of Forecasting, 16,* 451-476.
34. Pollock, A. C., Macaulay, A., Thomson, M. E., & Önkal, D. (2005). Performance evaluation of judgemental directional exchange rate predictions. *International Journal of Forecasting, 21,* 473-489.
35. Pollock, A. C., Macaulay, A., Thomson, M. E., & Önkal, D. (2008). Using weekly empirical probabilities in currency analysis and forecasting. *Frontiers in Finance and Economics, 5,* 26-55.
36. Pollock, A. C., Macaulay, A., Thomson, M. E., Gönül, M.S., & Önkal, D. (2010). Evaluating strategic directional probability predictions of exchange rates. *International Journal of Applied Management Science, 2,* 282-304.
37. Ray, E. L. *et al.* (2020). Ensemble forecasts of coronavirus disease 2019 (COVID-19) in the U.S. *MedRxiv*. https://doi.org/10.1101/2020.08.19.20177493
38. Reich, N. G. *et al.* (2019). Accuracy of real-time multi-model ensemble forecasts for seasonal influenza in the U.S. *PLOS Computational Biology, 15(11),* e1007486.
39. Richie, H. & Roser, M. (2021). Coronavirus source data. Our World in Data. Retrieved from https://ourworldindata.org/coronavirus-source-data
40. Schnaars, S. P. (1986). An evaluation of rules for selecting an extrapolative model of yearly sales forecasts. *Interfaces, 16,* 100-107.
41. SeyedAlinaghi, S., *et al.* (2021). Characterization of SARS-CoV-2 different variants and related morbidity and mortality: A systematic review. *European Journal of Medical Research, 26,* 51.
42. Shapiro, S. S. & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika, 52,* 591-611.
43. Stock, J. H. & Watson, M. W. (2004). Combining forecasts of output growth in a seven-country data set. *Journal of Forecasting, 23,* 405-430.
44. Sutherland, E., Headicar, J., & Delong, P. (2021). Coronavirus (COVID-19) Infection Survey
45. technical article: waves and lags of COVID-19 in England, June 2021. Office for National Statistics. Retrieved from

https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/articles/coronaviruscovid19infectionsurveytechnicalarticle/wavesandlagsofcovid19inenglandjune2021
46. Taylor, K. S. & Taylor, J. W. (2021). Combining probabilistic forecasts of COVID-19 mortality in the Unites States. *European Journal of Operational Research.* Retrieved from https://doi.org/10.1016/j.ejor.2021.06.044.
47. Testa, C. C., Krieger, N. Chen, J. T., & Hanage, W. P. (2020). Visualizing the lagged connection between COVID-19 cases and deaths in the United States: An animation using per capita state-level data (January 22, 2020 – July 8, 2020). *HCPS Working Paper, 19(4).* Retrieved from https://cdn1.sph.harvard.edu/wp-content/uploads/sites/1266/2020/07/HCPDS-WP_19_4_testa-et-al_Visualizing-Lagged-Connection-Between-COVID-19-Cases-and-Deaths-in-US_final_07_10_with-cover.pdf
48. Thomson, M. E., Pollock, A. C., Önkal, D., & Gönül, M. S. (2019). Combining forecasts: performance and coherence. *International Journal of Forecasting*, *21*, 473-489.
49. Thomson, M. E, Pollock, A. C., & Murray, J. (2021). Quantile probability predictions: A demonstrative performance analysis of forecasts of US COVID-19 deaths. *Eurasian Journal of Business and Management, 9(2),* 139-163.
50. Timmerman, A. (2006). Forecast combinations. In G. Elliot, C. W. J. Granger, & A Timmerman (Eds.), *Handbook of Economic Forecasting, Volume 1* (pp. 135-194).Amsterdam: North-Holland Publishing Co.
51. Wallis, K. F. (2011). Combining forecasts – forty years on. *Applied Financial Economics, 21,* 33-41.
52. WHO. (2020). WHO COVID-19 case definitions. Retrieved from https://www.who.int/publications/i/item/WHO-2019-nCoV-Surveillance_Case_Definition-2020.2