INSEAD The Business School for the World®

Working Paper

2022/21/DSC

Evaluating Quantile Forecasts in the M5 Uncertainty Competition

Zhi Chen
National University of Singapore, bizczhi@nus.edu.sg

Anil Gaba
INSEAD, anil.gaba@insead.edu

Ilia Tsetlin
INSEAD, ilia.tsetlin@insead.edu

Robert L. Winkler
Duke University, rwinkler@duke.edu

Version: March 22, 2022
International Journal of Forecasting, forthcoming

Probabilistic forecasts are necessary for robust decisions in the face of uncertainty. The M5 Uncertainty competition required participating teams to forecast nine quantiles for unit sales of various products at various aggregation levels and for different time horizons. This paper evaluates the forecasting performance of the quantile forecasts at different aggregation levels and at different quantile levels. We contrast this with some theoretical predictions, and discuss potential implications and promising future research directions for the practice of probabilistic forecasting

Keywords: Probabilistic Forecasting; Quantiles; Prediction Intervals; Scoring Rules

1. Introduction

In this paper, we investigate the performance of quantile forecasts (a type of probabilistic forecast) based on a rich data set from the M5 Uncertainty competition involving a complex real-world setting (Makridakis et al., 2020). This competition is a welcome step in the direction of exploring the practice of probabilistic forecasting. The data set involved sales of certain products at selected USA stores of the giant retailer Walmart. Each team was required to submit nine quantile forecasts (the 0.005, 0.025, 0.165, 0.25. 0.5, 0.75, 0.835, 0.975, and 0.995 quantiles) for the unit sales of these products at various aggregation levels and different time horizons, leading to nine quantile forecasts for each of the 42,840 quantities. Almost 900 teams from many countries participated in this competition, providing a very rich data set.

Probabilistic forecasts are crucial inputs for decisions in the face of uncertainty because they enable decision makers to better understand the risks and opportunities associated with important decisions. Point forecasts provide useful information but are not helpful in managing the risks of different courses of action. For example, an investor might want to explore a hedging strategy for an investment portfolio. Understanding the chances of the hedging strategy causing a serious drawdown in the investment fund (the risk) or causing higher returns in the future (the opportunity) enables the investor to balance the risk of a serious loss against the opportunity for a big payoff. Probabilities of different levels of demand for a product can help a retailer determine an order quantity, weighing the risk of under-ordering against that of over-ordering. In the COVID-19 pandemic, risks and opportunities associated with different policy decisions in terms of public health and economic activity should be considered before a final decision.

Complete probability distributions are the ideal type of probabilistic forecast for decision makers. However, forecasting full distributions is at best an arduous task, not just cognitively but also with sophisticated models and methods, often necessitating uncomfortable assumptions. Probabilistic forecasts such as quantiles or probability intervals are easier to obtain and also well understood by forecasters and decision makers. Increased exposure to probabilities in the media means that even the general public has a better understanding of probabilistic forecasts (Joslyn & Savelli, 2010; Joslyn & LeClerc, 2012). At the same time, recent developments in analytics, data science, and computer science have made it easier for decision makers to obtain and use probabilistic forecasts (Winkler, et al., 2019). In the M5 Uncertainty competition, the top-performing teams all used machine learning methods (Makridakis et al., 2020).

Providing forecasts for quantities at different hierarchical (aggregation) levels is common in a wide range of different industries such as retail (Kremer et al., 2015, Fildes, et al., 2019), energy (Ben Taieb et al., 2017, Adhikari & Karki, 2019), and tourism (Bertani et al., 2021). An extensive literature on forecast reconciliation in hierarchical forecasting has been created in the past 15 years. Much of this literature focuses on point forecasts, considering bottom-up, top-down, and middle-out approaches (Athanasopoulos, et al., 2009) with a variety of different reconciliation methods such as GLS (Hyndman et al., 2011), Optimal Weighting (Hyndman & Athanasopoulos, 2014), and Trace Minimization (Wickramasuriya et al., 2019). For reviews of forecast reconciliation methods, see Athanasopoulos et al. (2020, Chapter 21), Hollyman et al. (2021), and Hyndman & Athanasopoulos (2021, Chapter 11).

Reflecting the increased interest in probabilistic forecasts, some work on reconciliation has considered such forecasts. For example, see Ben Taieb et al., (2017, 2021), Gilbert et al. (2020), and Panagiotelis et al. (2020). This recent work uses a variety of approaches, with much emphasis on Bayesian models

because the Bayesian approach is fully probabilistic and sophisticated methods can be used to generate full predictive distributions.

Most of the work on forecast reconciliation concerns point forecasts, and the work on probabilistic forecasts focuses on generating forecasts, often full probability distributions. In the context of the M5 Uncertainty competition, the primary focus of this paper is on evaluating the quantile forecasts that have already been provided by the participating teams, not on aggregating those forecasts or generating new forecasts. In a similar vein, Athanasopoulos & Kourentzes (2020) discuss challenges in evaluating hierarchical quantile forecasts. They note the importance of quantile forecasts for a large number of decisions and reference the Pinball Loss of Gneiting (2011), which is used in the M5 Uncertainty competition and in our evaluation of quantile forecasts, but limit themselves to useful general advice without details or examples.

Just as hierarchical point forecasts are common, the same is also true for uncertainty forecasts such as quantile forecasts for different quantile levels (Jain et al., 2013; Gaba et al., 2017). How good is the practice of assessing the tails of the distribution (such as the 0.005 and 0.025 or the 0.975 and 0.995 quantiles) versus the middle of the distribution (such as the 0.5 quantile)? Does this vary at different aggregation levels? These are the types of questions that are worth exploring.

Little is known about how quantile forecasts perform at different hierarchical levels. Exploring this is useful because eliciting forecasts from experts on quantities at different hierarchical levels could easily become a daunting task for the decision makers as well as for the forecasters as the number of products and aggregation levels increase. We work within the constraints posed by the design of the M5 competition, in which the observations at the bottom (most granular) level are added to generate the observations at the next level in a bottom-up fashion. The quantile forecasts submitted by participating teams, on the other hand, are made separately for each level and do not need to add up in this fashion.

Knowing the relative performance of forecasts at different aggregation levels can help the decision maker prioritize the elicitation. It can be helpful to compare the performance of a forecast at the less granular level with a measure of performance of the forecasts below it at the more granular level. If the latter turns out to be better than or at least as good as the former, the decision maker can avoid the burden of eliciting forecasts at the aggregate level and just focus on eliciting the forecasts at a more granular level.

Using the M5 data set, Makridakis et al. (2020) discuss the performance of quantiles and prediction intervals obtained from combining the quantile forecasts. In this paper, we retain our focus on the obtained quantile forecasts rather than prediction intervals. We study the performance of the quantile forecasts with respect to (i) aggregation levels and (ii) quantile levels. In §2, we describe our main measures for evaluating quantile forecasts. To keep this paper self-contained, we briefly outline the characteristics of the data in §3. In §4, we explore the types of questions noted in the previous paragraph. First, in §4.1, we describe our empirical approach with rescaling and present empirical results. Then, in §4.2, we describe our empirical approach without rescaling and present empirical results for that approach. In §4.3, we provide a theoretical proposition for the performance of the quantile forecasts at different aggregation levels and contrast that with our empirical results for the M5 competition data. In §5, we put forward a theoretical proposition for the performance of quantile forecasts at different quantile levels and follow that up with empirical results. A summary and discussion is provided in §6.

2. Measures for Evaluating Quantile Forecasts

In this section, we introduce various performance measures to evaluate quantile forecasts. Suppose that an α ($0 < \alpha < 1$) quantile forecast q_α is provided for an unknown quantity $\tilde{x}$. To measure its performance, we evaluate it in view of the actual realized value x of $\tilde{x}$. Due to the randomness involved, computing the performance measures for just one or a few forecasts is not very meaningful. Instead, we perform such evaluations over a series of quantile forecasts. We begin with the S-score.

S-score: The S-score for a quantile is a strictly proper scoring rule, which is a rule that incentivizes forecasters to report their best estimate of the quantile (Jose & Winkler, 2009; Gneiting, 2011). It measures the overall performance of a quantile forecast, and is defined as

$$S(q_\alpha, x) = \alpha \max(x - q_\alpha, 0) + (1 - \alpha) \max(q_\alpha - x, 0). \qquad (1)$$

The S-score $S(q_\alpha, x)$ is a function of the quantile forecast q_α and the actual realization x. A lower score implies better quantile performance. The best possible score is zero, which is achieved when q_α coincides with the realization x. But in general, x either falls to the left or to the right of q_α. Any deviation of q_α from x increases the S-score, thus worsening the quantile performance. To see the impact such deviations have on the S-score, we next introduce deviation measures: deviations to the left and right, represented by $\Delta_L(q_\alpha, x)$ and $\Delta_R(q_\alpha, x)$, respectfully.

Deviations Δ_L to the left and Δ_R to the right: Deviations to the left and right are defined as $\Delta_L(q_\alpha, x) = \max(q_\alpha - x, 0)$ and $\Delta_R(q_\alpha, x) = \max(x - q_\alpha, 0)$, respectively. From (1),

$$S(q_\alpha, x) = \alpha \Delta_R(q_\alpha, x) + (1 - \alpha)\Delta_L(q_\alpha, x). \qquad (2)$$

Both $\Delta_L(q_\alpha, x)$ and $\Delta_R(q_\alpha, x)$ measure the distance between the quantile forecast q_α and its realization x. For any given q_α and x, only one of $\Delta_L(q_\alpha, x)$ and $\Delta_R(q_\alpha, x)$ can be positive, with the other one being zero (i.e., x can either fall above or below q_α).

Depending on the value of α, the impacts of $\Delta_L(q_\alpha, x)$ and $\Delta_R(q_\alpha, x)$ on the S-score are different. For the 0.01 quantile, the penalty if $x \le q_\alpha$ is $\Delta_L(q_\alpha, x)$ multiplied by $1 - \alpha = 0.99$, and the penalty if $x \ge q_\alpha$ is $\Delta_R(q_\alpha, x)$ multiplied by $\alpha = 0.01$. For left-tail quantiles such as the 0.01 quantile, the S-score is heavily penalized if $x \le q_\alpha$ but much less heavily punished if $x \ge q_\alpha$ by the same amount. This makes intuitive sense; in this way the scoring rule incentivizes forecasters to push their quantile forecasts to the left in order to avoid the heavier penalties. The reverse is true for right-tail quantiles, with an incentive to push quantile forecasts to the right. For the median ($\alpha = 0.5$), the penalties due to deviations of the same size to the left or right are identical. As a result, forecasters should strive to provide a median forecast that is as close to the realized value as possible. For any value of α, the incentive is to push the forecasts to their best estimate of the quantile, because the rule is strictly proper.

Sum of Deviations Δ: Discussing the sum of deviations for a single quantile forecast is not meaningful. Instead, we define the sum of deviations over a series of *n* forecasts with the same value of α, some of them potentially deviating to the left and others to the right. This measure is useful in showing an aggregate picture in terms of how far the quantile forecasts are from their realizations over a series of quantile forecasts.

We have shown that deviations play an important role in determining the overall quantile performance, as measured by the S-score. However, the size of the deviations alone conveys only a partial picture. We are also interested in whether the quantile performs in a way that is consistent with its quantile label, α. We can do this by measuring the proportion of realized values less than or equal to the forecasts for a given α.

Degree of Miscalibration: Over a series of n quantile forecasts q_α with the same α, we can compute the relative frequency of times the realized values fall below the forecasts (RF_α). For well-calibrated forecasts, $RF_\alpha = \alpha$. In this way, the forecasts optimally balance the penalties due to deviations to the left and to the right. Any deviation of RF_α from α represents miscalibration (Murphy & Winkler, 1984; O'Hagan et al., 2006), defined as $MC_\alpha = RF_\alpha - \alpha$.

All of these measures except for the degree of miscalibration are sensitive to scales. As a result, we cannot make direct comparisons of them at different aggregation levels. We address this issue in §4.

3. Data Description

In this section, we describe the data used for our analysis. Following Makridakis et al. (2020), we use quantile forecasts provided by the top 50 teams participating in the M5 uncertainty competition. Each team provided forecasts for the daily unit sales of 3,049 products sold in the USA over a 28-day duration, aggregated based on their type (category and department) or selling location (store and state). Each team forecasted a total of 42,840 quantities distributed over 12 aggregation levels. For each quantity, each team provided daily forecasts for 28 days. Level 1 (referred to as L1) is the most aggregate level, representing the total aggregate sales. L12 is the most granular level, representing the sales of each individual product in an individual store. As we move from L1 to L12, the quantities generally become more granular, although the nesting of hierarchies is not perfect. For each quantity on a given day, nine quantile forecasts (the 0.005, 0.025, 0.165, 0.25, 0.5, 0.75, 0.835, 0.975, and 0.995 quantiles) were provided. Table 1 summarizes the exact number of forecasting quantities and description for each aggregation level.

Table 1: A summary of data descriptions

Aggregation Level	No. of Quantities	Description
L1	1	Unit sales
L2	3	Unit sales, aggregated by state
L3	10	Unit sales, aggregated by store
L4	3	Unit sales, aggregated by category
L5	7	Unit sales, aggregated by department
L6	9	Unit sales, aggregated by state and category
L7	21	Unit sales, aggregated by state and department
L8	30	Unit sales, aggregated by store and category
L9	70	Unit sales, aggregated by store and department
L10	3049	Unit sales, aggregated by product
L11	9147	Unit sales, aggregated by product and state
L12	30490	Unit sales, aggregated by product and store
Total	42840	

4. How Quantile Performance Changes with Aggregation Levels

We start by defining some terminology that will be used in the presentation of our empirical results. We will use teams as the unit of analysis and express things on a "per team" basis. Empirical results will be presented using (1) performance measures that are rescaled as described in §4.1 to allow a direct comparison of the measures across different aggregation levels and (2) performance measures that are not rescaled.

For the S-score, the key statistics are found by summing the S-scores across the 50 teams and then taking the average per team. The same approach is used for the measures involving Deviations. For the empirical analysis with rescaling in §4.1, the key statistics are defined as follows, where the **T** in **TR** stands for Team and the **R** stands for Rescaled:

ScoreTR = Average Sum of Rescaled S-scores Per Team

DeviationsTR = Average Sum of Rescaled Deviations Per Team

DeviationsTRright = Average Sum of Rescaled Deviations to the Right Per Team

DeviationsTRleft = Average Sum of Rescaled Deviations to the Left Per Team

For the empirical analysis *without* rescaling in §4.2, the key statistics are similar without the **R**:

ScoreT = Average Sum of S-scores Per Team

DeviationsT = Average Sum of Deviations Per Team

DeviationsTright = Average Sum of Deviations to the Right Per Team

DeviationsTleft = Average Sum of Deviations to the Left Per Team

The degree of miscalibration is not sensitive to scaling, so the key statistic is always **Degree of Miscalibration** as defned in §2.

Now we discuss the first research question: How does quantile performance change with aggregation level? Different aggregation levels have different units. For example, L1 is the most aggregate level. Unit sales at L1 represents the sales aggregated across all three states and all products. Its magnitude is much larger than that at L12, the most granular level, which represents the sales of a particular product in a particular store. Hence, a direct comparison of the average quantile forecast performance across different aggregation levels is not meaningful.

To address this issue, we adopt two different approaches. In §4.1, following the literature (see, e.g., Athanasopoulos & Kourentzes, 2020; Makridakis et al., 2020), we normalize the performance measures for different aggregation levels with appropriate rescaling factors. In this way, rescaled performance measures can be directly compared across different aggregation levels. However, the empirical results could be sensitive to the choice of rescaling method (Athanasopoulos & Kourentzes, 2020). In §4.2, we propose an alternative approach without rescaling the forecasts. In §4.3, we further complement our empirical analyses with some theoretical results and discuss the implications for hierarchical forecasts.

4.1. Empirical Analysis With Rescaling

In this section, we briefly introduce the empirical approach with rescaling in §4.1.1 and then present the empirical results in §4.1.2.

4.1.1. Description of Empirical Approach

Following the empirical approach used in the M5 Uncertainty competition, we rescale various performance measures introduced in §2 except for the degree of miscalibration. For any quantity i, the rescaling factor is given by $\frac{1}{n-1}\sum_{t=2}^{n}|x_{i,t}-x_{i,t-1}|$, where n is the length of the training sample (number of historical observations) and $x_{i,t}$ is the actual realization of quantity i at time t in the historical data. In this way, the rescaling factor can be interpreted as the in-sample, one-step-ahead mean absolute error of

the Naïve forecasting method, which treats the actual realization of any period as the point forecast for the next period. Hence, the rescaled S-score, referred to as the Weighted Scaled Pinball Loss or WSPL in Makridakis et al.(2020), is the actual S-score in Equation (2) divided by the rescsaling factor and being weighted appropriately according to the dollar sales. The rescaled versions of the deviation measures introduced in §2 can be computed in similar ways. Such a rescaling approach is also recommended by Athanasopoulos & Kourentzes (2020). For more detailed procedures of rescaling, see §2.2 of Makridakis et al. (2020).

4.1.2. Empirical Results

Figures 1-5 demonstrate how various rescaled performance measures vary with respect to the aggregation levels. As we move from L1 (aggregate) to L12 (granular) in Figure 1, the rescaled S-scores **ScoreTR** tend to be higher (worse) as we move from L1 to L12. This makes intuitive sense, as it is harder to forecast quantities at more granular levels as compared to those at more aggregate levels. Similar trends have been observed in the interval forecasts (Makridakis et al., 2020). The segments in Figure 1 that shift downward to L4, L6, and L8 reflect the fact that the nesting of hierarchies is not perfect, as noted in §3.

The higher **ScoreTR** at more granular levels is primarily explained by the worsening of all rescaled deviation measures: the rescaled total deviations **DeviationsTR** (Figure 2) tend to become larger at more granular levels than those at more aggregate levels, as do the rescaled deviations to the left **DeviationsTRleft** (Figure 3) and to the right **DeviationsTRright** (Figure 4).

Moreover, according to Figure 5, **Degree of Miscalibration** tends to be more severe in the middle part of the distribution (e.g., the 0.165, 0.25, 0.5, 0.75, and 0.835 quantiles) than the tails, especially from L1 to L6. Beyond L6, the degree of miscalibration exhibits less variability – in general, it is between -0.05 and 0.05. Overall, the quantile forecasts are remarkably well calibrated, with the average degree of miscalibration being around 0.04 across all quantiles and aggregation levels. This is consistent with the findings in Makridakis et al. (2020) that prediction intervals derived from these quantile forecasts are well calibrated.

INSERT FIGURES 1, 2, 3, 4, and 5

There is an exception in the general trends discussed above. For the left-tail quantiles 0.005, 0.025, 0.165, and 0.25, Figure 1 shows that the rescaled S-scores improve considerably with granularity from L10 to L12. This can be partly explained by the exceptional patterns observed in the corresponding rescaled deviations measures: the rescaled deviation measures improve (i.e., become smaller) from L10 to L12 for the left-tail quantiles, as is evident from Figures 2-4.

To gain a deeper understanding about this exception, we revisit the rescaled performance measures. They equal the original performance measures in §2 divided by the rescaling factor $\frac{1}{n-1}\sum_{t=2}^{n}|x_{i,t}-x_{i,t-1}|$. The rescaling factor, in the form of absolute deviations, can be perceived as a measure of variability for the underlying quantity (Athanasopoulos & Kourentzes, 2020). Because more granular quantities tend to exhibit higher variability relative to their scale, the rescaling factor seems to favor granular quantities at lower aggregation levels. As a result, the extent to which the trends discussed above (especially the exceptions that the left-tail quantile forecasts seem to improve from L10 to L12) are genuine or driven by the rescaling factor is unclear. In §4.2 we address this question by analyzing the data without applying the rescaling factor.

4.2. Empirical Analysis Without Rescaling

Without rescaling, the question remains as to how to make sure that comparisons across different aggregation levels is fair, given they have different units. We first describe our empirical approach without rescaling in §4.2.1, followed by the empirical results in §4.2.2.

4.2.1. Description of Empirical Approach

To ensure that the units are normalized to the same magnitude, the hierarchical data structure is helpful. In particular, we note that L1 is the most aggregate level and levels L2-L12 are breakdowns of L1 in different granularities. With respect to each of the quantile measures Δ_L, Δ_R, Δ, and the S-score for any level other than L1, we sum the quantile measure for a given quantile across different quantities within that aggregation level. In this way, all aggregation levels other than L1 are converted to the units of L1, establishing a fair basis for comparison across different aggregation levels.

To illustrate the approach described above, we use L1 and L2 in a simplified example. The L1 data consist of quantile forecasts q_α for total sales x, where x equals the sum of the total L2 sales in each of the three individual states, $x = \sum_i x_i, i \in \{CA, TX, WI\}$). The S-score for q_α in L1 is then $S(q_\alpha, x)$. The L2 data consist of quantile forecasts for total sales in each of the states, denoted by $q_{\alpha,i}$, where $i \in \{CA, TX, WI\}$. For L2 forecasts, the sum of the S-scores is given by $\sum_i S(q_{\alpha,i}, x_i)$, $i \in \{CA, TX, WI\}$, which can be perceived as a proxy for L2 forecast performance. In this way, the comparison of S-scores for L1 and L2 boils down to the comparison between the S-score of the L1 forecast and the sum of S-scores of the L2 forecasts, i.e., between $S(q_\alpha, x)$ and $\sum_i S(q_{\alpha,i}, x_i)$, $i \in \{CA, TX, WI\}$). Comparisons of deviation measures Δ_L, Δ_R, and Δ across different aggregation levels can be made in a similar fashion. Table 2 provides an example with 0.005 quantile forecasts. The first row shows the L1 forecast, the realized quantity, and the various measures corresponding to the forecast. The next three rows show the same for the three L2 forecasts. And, the fifth row, labeled summed L2 scores, simply sums up the performance measures of the three L2 forecasts. In this example, the L1 forecast in Row 1 outperforms the summed L2 scores in Row 5 in terms of smaller deviation measures as well as a smaller S-score.

Table 2: A sample of 0.005 quantile forecasts

Aggregation Level	Description	q	x	Δ_L	Δ_R	Δ	S-score
L1	Total Sales	41064	45704	0	4640	4640	23
L2	CA Sales	17324	20730	0	3406	3406	17
L2	TX Sales	11158	12673	0	1515	1515	8
L2	WI Sales	12411	12301	110	0	110	110
summed L2 scores				110	4921	5031	135

4.2.2. Empirical Results

For the empirical approach without rescaling, Figures 6-9 show how the S-score **ScoreT** and the deviation measures **DeviationsT**, **DeviationsTright**, and **DeviationsTleft** change with respect to the aggregation levels for any given quantile level $\in \{0.005, 0.025, 0.165, 0.25, 0.5, 0.75, 0.835, 0.975, 0.995\}$. The general trends of the empirical results are consistent with those under rescaling in the respective Figures 1-4: the S-score and the deviation measures worsen from L1 (aggregate) to L12 (granular).

INSERT FIGURES 6a-c, 7, 8, and 9

To estimate how these measures vary, we compute their standard deviations and divide these standard deviations by the square root of $n = 50$, the number of teams, to find their standard errors. Then we use a common rule of thumb, defining the margins of error as 2 standard errors to obtain error bounds for the

measures. For example, Figures 6b and 6c present **ScoreT** for the 0.005 and 0.165 quantiles at aggregation levels L10-L12, which are the same as the values for L10-L12 in Figure 6a, adding the corresponding error bounds in Figures 6b and 6c. Figure 6c for the 0.165 quantile is quite representative of quantiles where the error bounds are very tight and do not overlap for different aggregation levels. In contrast, Figure 6b for the 0.005 quantile is representative of quantiles where the error bounds for granular levels do overlap for different aggregation levels.

However, there are some notable differences. Comparing the rescaled S-score (Figure 1) and the S-scores without rescaling (Figure 6a), Figure 1 shows that from L10 to L12, the rescaled S-scores for all the examined left-tail quantiles (the 0.005, 0.025, 0.165 and 0.25 quantiles) improve. In Figure 6a without rescaling, this is only true for the two leftmost quantile forecasts (the 0.005 and 0.025 quantiles). Moreover, similar patterns seem to show up in the deviation measures. For instance, comparing the rescaled total deviations (Figure 2) with those without rescaling (Figure 7), Figure 2 shows that from L10 to L12, the rescaled total deviations for all the examined left-tail quantile forecasts (the 0.005, 0.025, 0.165, and 0.25 quantiles) improve; but this is actually the opposite for the left-tail quantile forecasts without rescaling (total deviations tend to get worse from L1 to L12 as shown in Figure 7). In a way, such observed discrepancies confirm our earlier conjecture that the rescaling factor tends to favor quantities at more granular levels. But overall, the main takeaway by comparing the figures with and without rescaling is that the choice of a rescaling factor does have considerable effects on the results but does not seem to impact the general trends observed. In other words, the empirical approaches with and without rescaling seem to produce consistent results in general.

One puzzle still remains: Figure 6a shows that **ScoreT** for 0.005 and 0.025 quantile forecasts improves with granularity from L10 to L12 but the respective **DeviationsT** goes in the opposite direction, as seen in Figure 7. In general, we would expect the improved ScoreT from L10 to L12 to be associated with improved DeviationsT. But we observe the opposite. Even though the error bounds for **ScoreT** in this case are overlapping (as observed in part in Figure 6b), it is interesting to reconcile this seemingly inconsistent pattern between the S-score and total deviations in this dataset. We do not claim that in another identical run of this competition, we will observe this anomaly again. In this paper, our focus is simply on the descriptive statistics of the observed M5 data.

We further plot the deviations to the left and to the right for the left-tail quantiles in Figure 10 and 11. This clearly shows that **DeviationsTleft** decreases from L10 to L12 for the 0.005 and 0.025 quantiles (Figure 10) but **DeviationsTright** increases (Figure 11). This happens for the left-tail quantiles due to truncation at zero because sales cannot be negative. Moreover, this phenomenon is particularly salient at the extreme granular levels beause the sparsity of data at such levels causes spikes at zero sales, capping the deviations to the left. Also, a deviation to the left is a more serious mistake for these quantiles with very low values of α, as it will be multiplied by a heavy penalty $(1-\alpha)$ to get the S-score..

INSERT FIGURES 10 and 11

4.3. Further Explanation of the Empirical Results

In §4.1 and §4.2, we have established empirically that quantile forecasts tend to perform better at a more aggregate level. Due to truncation and sparsity, there may be some exceptions for the left-tail quantile forecasts with respect to more granular levels such as the 0.005 and 0.025 quantiles for L10, L11, and L12. In this section, we seek to rationalize the abovementioned patterns, building on the empirical approach without the rescaling introduced in §4.2.

In §4.2, the comparison is between (a) the performance measures for L1 forecasts and (b) the sum of the same performance measures for forecasts at more granular levels. For example, in Table 2, we compare the performance measures for a given quantile of (a) Row 1, involving the L1 forecasts and (b) Row 5, involving the sums of the performance measures for the L2 forecasts from Rows 2-4. In Row 6, we now we add another measure (c) by summing not the performance measures but the quantiles from Rows 2-4 and comparing the performance of these summed quantiles with the performance of the L1 quantile forecast of total sales from Row 1. Summing the quantiles in this way mimics the way the actual sales are summed in the M5 competition. This can be thought of as the most naïve bottom-up foreast reconciliation approach for uncertainty forecasts.

To gain intuition, we first look at some illustrative examples. The example in Table 3 builds on the earlier example in Table 2 by inserting a new row (Row 6), by summing the 3 L2 0.005 quantile forecasts for the 3 individual states in Rows 2-4. We then evaluate these summed L2 quantiles using the total sales from L1 to find the S-score, and the deviation measures for summed L2 quantiles can be found accordingly. They are being treated as summed quantiles for L1 (from L2). If we compare the summed L2 quantiles (Row 6) with the L1 quantile forecast (Row 1), we find that the L1 forecast is better than the summed L2 quantiles in terms of both weakly smaller deviations and a lower S-score. Table 4 provides yet another example in the same spirit. Of course, in other examples the summed quantiles might perform better.

Table 3: A sample of 0.005 quantile forecasts (Example 1)

Aggregation Level	**Description**	**q**	**x**	$\boldsymbol{\Delta_L}$	$\boldsymbol{\Delta_R}$	**Δ**	**S-score**
L1	Total Sales	41064	45704	0	4640	4640	23
L2	CA Sales	17324	20730	0	3406	3406	17
L2	TX Sales	11158	12673	0	1515	1515	8
L2	WI Sales	12411	12301	110	0	110	110
summed L2 scores				110	4921	5031	135
summed L2 quantiles	Total Sales	40893	45704	0	4811	4811	24

Table 4: A sample of 0.005 quantile forecasts (Example 2)

Aggregation Level	**Description**	**q**	**x**	$\boldsymbol{\Delta_L}$	$\boldsymbol{\Delta_R}$	**Δ**	**S-score**
L1	Total Sales	34141	40385	0	6244	6244	31
L2	CA Sales	14226	17292	0	3066	3066	15
L2	TX Sales	8653	11372	0	2719	2719	14
L2	WI Sales	9720	11721	0	2001	2001	10
summed L2 scores				0	7786	7786	39
summed L2 quantiles	Total Sales	32599	40385	0	7786	7786	39

To formalize the patterns observed in the above examples in Tables 3 and 4, we show in Proposition 1 that the performance measures ($\boldsymbol{\Delta_L}$, $\boldsymbol{\Delta_R}$ and **S-Score**) of the summed quantiles weakly outperform the sum of the respective measures for the more granular forecasts.

Proposition 1. Suppose that there are n quantities and for each quantity i, the α ($0 < \alpha < 1$) quantile forecast is $q_{\alpha,i}$ and the corresponding realized value is x_i, where $i = 1, 2, \ldots, n$. Let $q_{\alpha,c} = \sum_{i=1}^{n} q_{\alpha,i}$ and $x = \sum_{i=1}^{n} x_i$. Then the following relationships hold:

(i) $\sum_{i=1}^{n} \Delta_L(q_{\alpha,i}, x_i) \geq \Delta_L(q_{\alpha,c}, x)$.

(ii) $\sum_{i=1}^{n} \Delta_R(q_{\alpha,i}, x_i) \geq \Delta_R(q_{\alpha,c}, x)$.

(iii) $\sum_{i=1}^{n} S(q_{\alpha,i}, x_i) \geq S(q_{\alpha,c}, x)$.

Proof: (i) Let $y_i = q_{\alpha,i} - x_i$, so that $q_{\alpha,c} - x = \sum_{i=1}^{n} y_i$. It remains to show that $\sum_{i=1}^{n} \max(y_i, 0) \geq \max\left(\sum_{i=1}^{n} y_i, 0\right)$. Note that $\max(y_i, 0)$ is convex in y_i and $\max(ty_i, 0) = t\max(y_i, 0)$ for any $t > 0$. From Jensen's inequality, it then follows that $\sum_{i=1}^{n} \frac{1}{n} \max(y_i, 0) \geq \max\left(\frac{1}{n}\sum_{i=1}^{n} y_i, 0\right) = \frac{1}{n}\max\left(\sum_{i=1}^{n} y_i, 0\right)$. Canceling $\frac{1}{n}$ on both sides of the inequalities, we conclude the proof. For (ii), the proof follows the steps used to prove (i). Then (iii) directly follows from (i) and (ii). □

To further understand the intuition in Proposition 1, we revisit the examples in Tables 3 and 4, using use L1 and L2 forecasts as illustrative examples. We do so for ease of exposition. The arguments hold in general when it comes to comparing performance between forecasts at a more aggregate level with those at a more granular level.

First, consider the example in Table 3. Among the three L2 forecasts (Rows 2 to 4), the directions of deviations are different, with the realizations falling above the forecasts in CA and TX but below the forecast in WI. In summing up the three forecasts and evaluating the summed quantile forecasts against the total sales, the summed quantiles benefit from the effect of pooling, where granular level deviations occurring in different directions cancel out to a certain extent. As a result, the summed quantiles yield smaller deviations and a lower S-score relative to the sums of those measures for the granular quantities; compare Rows 5 and 6 in Table 3.

Table 4 paints a slightly different picture. In contrast with Table 3, the deviations all happen in the same direction (the realizations are above the forecasts). Hence, the pooling benefit is muted. The performance measure for the summed quantiles are simply as good as, not better than, the sum of the respective measures for granular quantities. This insight is in the same spirit as the wisdom of the crowd: there is a greater value in aggregating forecasts if the crowd is diverse, with their forecasts falling on both sides of the realized value (Grushka-Cockayne et al., 2017).

The key takeaway from the examples in Tables 3 and 4 is that the performance of summed quantiles is closely related to the dependence among different quantities at a more granular level. When some quantile forecasts of different quantities fall above and some fall below below the realized value at a more granular level, summed quantiles have smaller deviations and overall S-score relative to the sum of these performance measures at the more granular level. However, when forecasts across different quantities show a strong dependence, in the sense that most of them fall on the same side of the realized value, summed quantiles offer little improvement.

Proposition 1 shows that the performance measures for the summed quantiles weakly outperform the sums of the performance measures of the more granular forecasts. If we compare (a) the performance of the actual forecast for the less granular level, (b) the sum of the performance measures of the forecasts at the more granular level, and (c) the performance of the summed quantiles, then (c) outperforms (b) by Proposition 1. Unfortunately, it is impossible to derive a formal theoretical relationship between (a) and (c). Intuitively, we would expect that (a) would outperform (c) because this is the actual forecast directly provided by the forecasters, potentially derived from sophisticated state-of-the art forecasting methods, taking into account such considerations as dependence at the more granular level. As a result, (a) should also be expected to outperform (b). Indeed, this is what typically happens empirically, but not always.

The practical implication of a reversal of the rank order between (a) and (b) is far-reaching. When (b) actually outperforms (a), that implies that (c) should outperform (a), since by Proposition 1, (c) outperforms (b). An empirical example from our analysis involves the most granular levels (L10-L12). In

the context of the M5 Uncertainty competition, this implies that for the 0.005 and 0.025 quantile forecasts among the most granular quantities (e.g., L10-L12), the summed quantiles for L10 will outperform the actual L10 forecasts provided separately by the participants. This is demonstrated in Figure 12 in terms of the S-scores. Moreover, Figure 13 shows that it is demonstrated even more convincingly in terms of the degree of miscalibration. As the quantile levels increase beyond around 0.08, this result gets reversed.

This example suggests that in some situations, there is no value in making separate quantile forecasts for L10 since even the simple approach of the summed quantiles for L10 from L12 performs better. Based on the M5 Uncertainty competition dataset, this seems to happen in situations with (i) extremely left-tail quantiles and (ii) for the very granular quantities. More empirical studies may be required to validate the robustness of this result.

INSERT FIGURES 12 and 13

5. How Quantile Performance Changes with Quantile Levels

In this section, we explore how quantile forecast performance changes with respect to the quantile levels, as measured by $\alpha \in (0,1)$. Proposition 2 presents the theoretical prediction, followed by the empirical results displayed on Figure 15.

For any given $\alpha \in (0,1)$, suppose that q_α is the theoretical α quantile that minimizes the expected S-score $E(S(q_\alpha, x))$. Then the following result holds.

Proposition 2. $E(S(q_\alpha, x))$ is concave in $\alpha \in (0,1)$. It first increases and then decreases in α. The maximum is reached at $\alpha = P(\tilde{x} < \mu)$, where $E(\tilde{x}) = \mu$.

Proof: (i) By the Envelope Theorem (Mas-Colell et al., 1995), $\frac{\partial q_\alpha}{\partial \alpha} = 0$. Then $\frac{\partial E(S(q_\alpha, x))}{\partial \alpha} = \max(x - q_\alpha, 0) - \max(q_\alpha - x, 0) = \mu - q_\alpha$, which is decreasing in α. Then the expected S-score is concave in α, and the maximum is reached at $\mu = q_\alpha$, or $\alpha = P(\tilde{x} < \mu)$. □

Proposition 2 shows that the expected S-score exhibits an inverted-U shape with respect to the quantile levels. It implies that the middle part of the distribution tends to have higher scores than the tails, with the maximum score occurring at $\alpha = P(\tilde{x} < \mu)$. When the unknown quantity $\tilde{x}$ is symmetrically distributed, the mean of $\tilde{x}$ is equivalent to the median; as a result, the median (i.e., $\alpha = P(\tilde{x} < \mu) = 0.5$) also leads to the highest score.

Figure 14 shows **ScoreT** for a given aggregation level as a function of quantile levels. The results are consistent with the theoretical prediction in Proposition 2 that the S-scores exhibit an inverted-U shape with respect to the quantile levels.

INSERT FIGURE 14

6. Conclusion

In this paper, we investigate the relative performance of quantile forecasts at various aggregation levels and at different quantile levels based on a rich data set from the M5 Uncertainty competition. The participating teams in this competition used quantitative methods such as statistical modeling, machine learning algorithms, and combinations of forecasts from different methods to generate quantile forecasts for unit sales of consumer goods. The teams were required to submit nine quantile forecasts for each of the forecasting quantities at various aggregation levels and different forecasting horizons.

In our empirical analysis, we first attempt to make the performance at different hierarchical levels comparable by rescaling the performance measures. Working within the framework of the M5 Uncertainty competition, we use the rescaling factor used for the interval forecasts in Makridakis et al. (2020). As we move from the most aggregate level L1 to the most granular level L12, the rescaled S-scores and all of the rescaled deviation measures become larger (worse) for the quantile forecasts as we move from the most aggregate level L1 to the most granular level L12, with the exception of the left-tail quantile forecasts at the granular levels L10 to L12.

The degree of miscalibration does not require rescaling. Overall, we find the quantile forecasts to be remarkably well calibrated. Poorer calibration is observed only for the quantiles in the middle part of the distribution at the most aggregate levels. This is consistent with the findings in Makridakis et al. (2020) that prediction intervals derived from these quantile forecasts are well calibrated, since the more extreme quantiles that form the prediction intervals are well calibrated.

Next, we use a different approach to enable a fair comparison across the twelve different aggregation levels in the competition without rescaling the forecasts. We first sum up performance measures across different quantities within an aggregation level and compare the summed measures with the performance measures of the direct forecast at a more aggregate level. The empirical results are generally consistent with those for the rescaling: the S-scores and the deviation measures worsen at more granular levels, except for the left-tail quantile foreasts at the granular levels L10 to L12.

We also explore a naïve bottom-up forecast reconciliation approach, where we sum the quantiles from one level and evaluate the summed quantiles from the more granular level (e.g., the summed L2 quantiles) using the total sales from the more aggregate level (e.g., the total sales for L1). We show theoretically (Propostion 1) and empirically that the summed quantiles perform at least as well as the sum of the performance measures from the more granular level (e.g., the summed L2 scores). However, we would expect the actual forecasts provided by the forecasters at a given aggregate level to perform better than the summed quantiles approach, since the forecasters presumably use more sophisticated methods than the summed quantiles approach. Indeed, this is what usually happens empirically.

However, if we make the same comparisons for L10 from L12, this can be reversed for the extreme left-tail quantile forecasts (0.005, 0.025), as illustrated in Figures 12 and 13. This finding is intriguing. For example, when we evaluate the summed L12 quantiles using the total sales from L10, they perform better than the actual L10 forecasts. This implies that there is no value in making direct quantile forecasts for L10 instead of just using the summed quantiles from the more granular level L12. Empirically, however, this holds only in situations with (i) extremely left-tail quantiles and (ii) for the very granular quantities.

On the performance of quantile forecasts at different quantile levels, we first show in Proposition 2 that quantile performance exhibits an inverted-U shape with respect to the quantile levels. In other words, the scores are higher in the middle part of the distribution versus the tails. This theoretical result is consistent with the M5 data.

In this paper, our focus is on evaluating the performances of uncertainty forecasts provided in the M5 competition, not on generating forecasts. However, the results point to a number of promising avenues for future research on generating forecasts in a hierarchical setting. There clearly is potential for improving forecasting performance by using different approaches suggested in the recent literature on forecast reconciliation for uncertainty forecasts.. We briefly explore only one way of doing this, by simply taking a bottom-up approach of summing up the relevant quantile forecasts at a given granular level and using that to generate quantile forecasts for a more aggregate level. Our results for even this simple approach suggest that in some instances (e.g., for the left-tail quantiles at the most granular levels), construction of

forecasts at a given aggregate level from the more granular levels could outperform the corresponding direct aggregate forecasts. Such efforts might not only considerably impove forecasting performance but could also substantially reduce the enormous burden of generating hierarchical forecasts at all levels, for example, by using granular forecasts to construct forecasts at a more aggregate level in some situations.

Finally, another useful direction for future research would be to consider in more detail the impact of dependence among the time series at the same level (e.g., spatial dependence) on actual forecasts, summed quantiles, or other measures. For a quantity like sales, such dependence is likely. This dependence can have an impact on the resulting performance and can be modeled in different ways. For example, Gilbert et al. (2020) use copulas to model the dependence. Copulas can provide more flexibility and can separate the modeling of dependence from the modeling of location and spread. All that might lead to developing new reconciliation approaches, involving both bottom-up and top-down accounting of dependence.

References

Adhikari, S., & Karki, R. (2019). Integrated disturbance modeling of wind-integrated power systems to quantify the operational reliability benefits of flywheel energy storage, *IEEE Transactions on Sustainable Energy*,*10*(3), 1152-1160.

Athanasopoulos, G., Ahmed, R.A., & Hyndman, R.J. (2009). Hierarchical forecasts for Australian domestic tourism. *International Journal of Forecasting*, *25*(1), 146-166.

Athanasopoulos, G., Gamakurmara, P., Panagiotelis, A., Hyndman, R.J., & Affan, M. (2020). Hierarchical forecasting. In Fuleky, P. (Ed.), *Macroeconomic Forecasting in the Era of Big Data: Theory and Practice*, Springer, Cham, Switzerland: Springer.

Athanasopoulos, G., & Kourentzes, N. (2020). On the evaluation of hierarchical forecasts, Working Paper 02/20, Department of Econometrics & Business Statistics, Monash University, Monash, Australia.

Ben Taieb, S., Taylor, J.W., & Hyndman, R.J. (2017). Coherent probabilistic forecasts for hierarchical time series. *Proceedings of the 34th International Conference on Machine Learning*, Sydney, Australia, PMLR 70, 3348–3357.

Ben Taieb, S., Taylor, J.W., & Hyndman, R.J. (2021). Hierarchical probabilistic forecasting of electiricty demand with smart meter data. *Journal of the American Statistical Association*, *116*(533), 27-43.

Bertani, N., Satopää, V., & Jensen, S. (2021). Joint bottom-up method for hierarchical time-series: Application to Australian Tourism. Working Paper.

Fildes, R., Ma, S., & Kolassa, S. (2019). Retail forecasting: Research and practice. *International Journal of Forecasting*, https:doi.org/10.1016/j.ijforecast2019.06.004.

Gaba, A., Tsetlin, I., & Winkler, R.L. (2017). Combining interval forecasts. *Decision Analysis, 14*(1), 1–20.

Gilbert, C., Browell, J., & McMillan, D. (2020). Leveraging turbine-level data for improved probabilistic wind power forecasting. *IEEE Transactions on Sustainable Energy*, *11*(3), 1152-1160.

Gneting, T. (2011). Quantiles as optimal point forecasts. *International Journal of Forecasting, 27*(2), 197-207.

Grushka-Cockayne, Y., Lichtendahl, K.C. Jr., Jose, V.R.R., & Winkler, R.L. (2017). Quantile evaluation,

sensitivity to bracketing, and sharing business payoffs. *Operations Research*, *65*(3), 712-728.

Hollyman, R., Petropoulos, F., & Tipping, M.E. (2021). Understanding forecast reconciliation. *Euroopean Journal of Operational Research*, *294*(1), 149-160.

Hyndman, R.J., Ahmed, R.A., Athanasopoulos, G., & Shang, H.L. (2011). Optimal combination forecasts for hierarchical time series. *Computational Statistics and Data Analysis*, *55*(9), 2579-2589.

Hyndman, R.J., & Athanasopoulos, G. (2014). Optimally reconciling forecasts in a hierarchy. *Foresight: The International Journal of Applied Forecasting*, *35* 42-48.

Hyndman, R.J., & Athanasopoulos, G. (2021). *Forecasting: Principles and Practice*, 3rd Ed., OTexts: Melbourne, Australia. Otexts.com/fpp3.

Jain, K., Mukherjee, K., Bearden, J.N., & Gaba, A. (2013). Unpacking the future: A nudge toward wider subjective confidence intervals. *Management Science, 59*(9), 1970–1987.

Jose, V.R.R., & Winkler, R.L. (2009). Evaluating quantile assessments. *Operations Research, 57*(5), 1287-1297.

Joslyn, S.L., & LeClerc, J.E. (2012). Uncertainty forecasts improve weather-related decisions and attenuate the effects of forecast error. *Journal of Experimental Psychology: Applied*, *18*(1), 126-140.

Joslyn, S, & Savelli, S. (2010). Communicating forecast uncertainty: Public perception of weather forecast uncertainty. *Metoeorological Applications*, *17*, 180-195.

Kremer, M., Siemsen, E., & Thomas, D.J. (2015). The sum and its parts: Judgmental hierarchical forecasting. *Management Science*, *62*(9), 2745–2764.

Makridakis, S., Spiliotis, E., Assimakopoulos, V., Chen, Z., Gaba, A., Tsetlin, I., & Winkler, R.L. (2020). The M5 Uncertainty competition: Results, findings and conclusions. *International Journal of Forecasting*, forthcoming. 10.1016/j.ijforecast.2021.10.009.

Mas-Colell, A., Whinston, M.D., & Green, J.R. (1995). *Microeconomic Theory*, Oxford University Press, Oxford, U.K

Murphy, A.H., & Winkler, R.L. (1984). Probability forecasting in meteorology. *Journal of the American Statistical Association, 79*(387), 489-500.

O'Hagan, A., Buck, C.E., Daneshkhah, A., Elser, J.R., Garthwaite, P.H., Jenkinson, D.J., Oakley, J.E., & Rakow, T. (2006). *Uncertain Judgements: Eliciting Experts' Probabilities*, John Wiley & Sons, New York.

Panagiotelis, A., Gamakumara, P., Athanasopoulos, G., & Hyndman, R.J. (2020). Probabilistic forecast reconciliation: Properties, evaluation and score optimisation. Working Paper 26/20, Department of Econometrics & Business Statistics, Monash University, Monash, Australia.

Wickramasuriya, S.L., Athanasopoulos, G, & Hyndman, R.J. (2019). Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization. *Journal of the American Statistical Association*, *114*(526), 804-819.

Winkler, R.L., Grushka-Cockayne, Y., Lichtendahl, K.C. Jr., & Jose, V.R.R. (2019). Probability forecasts and their combination: A research perspective. *Decision Analysis*, *16*(4), 239-260.

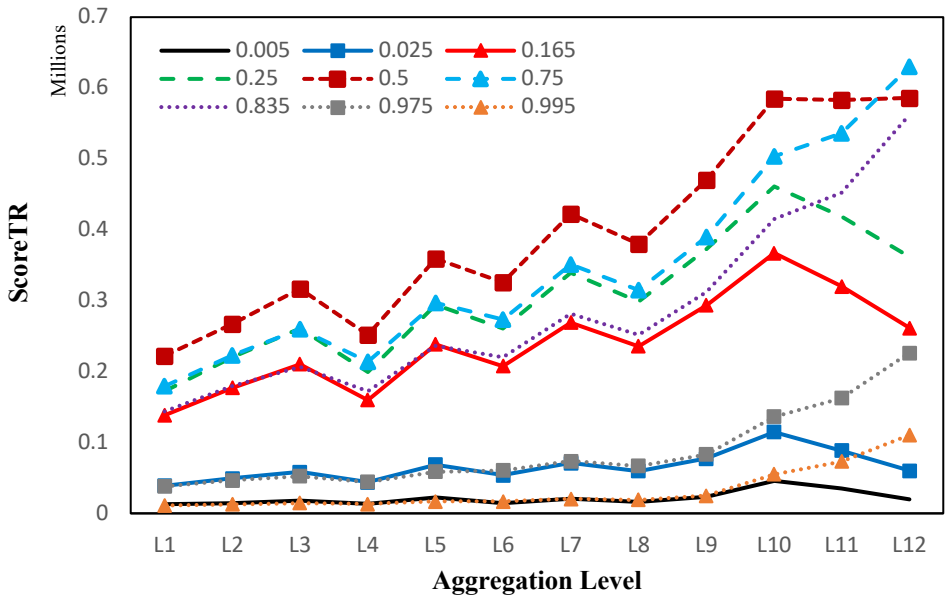

Figure 1: **ScoreTR** Within a Given Aggregation Level as a Function of Quantile Levels

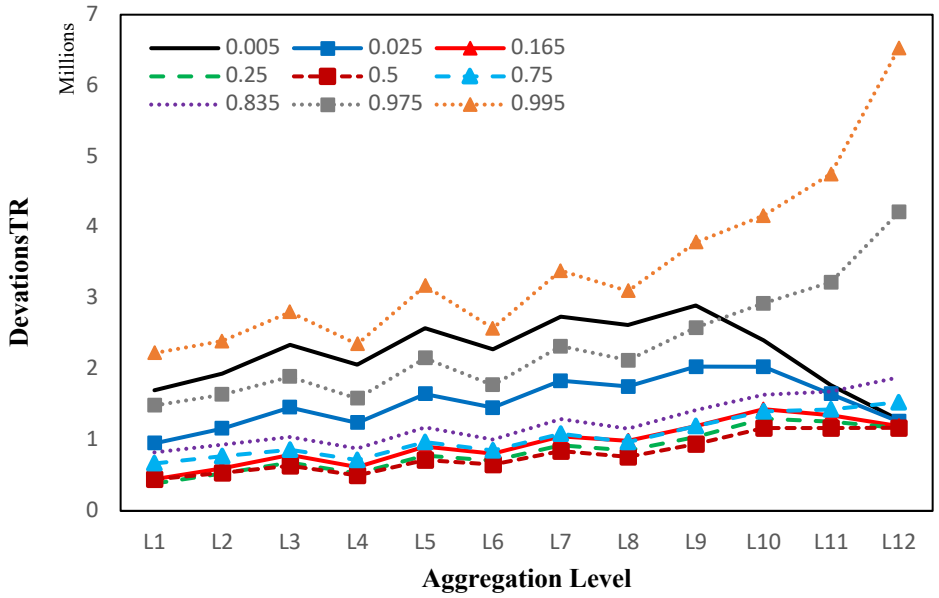

Figure 2: **DeviationsTR** Within a Given Aggregation Level as a Function of Quantile Levels

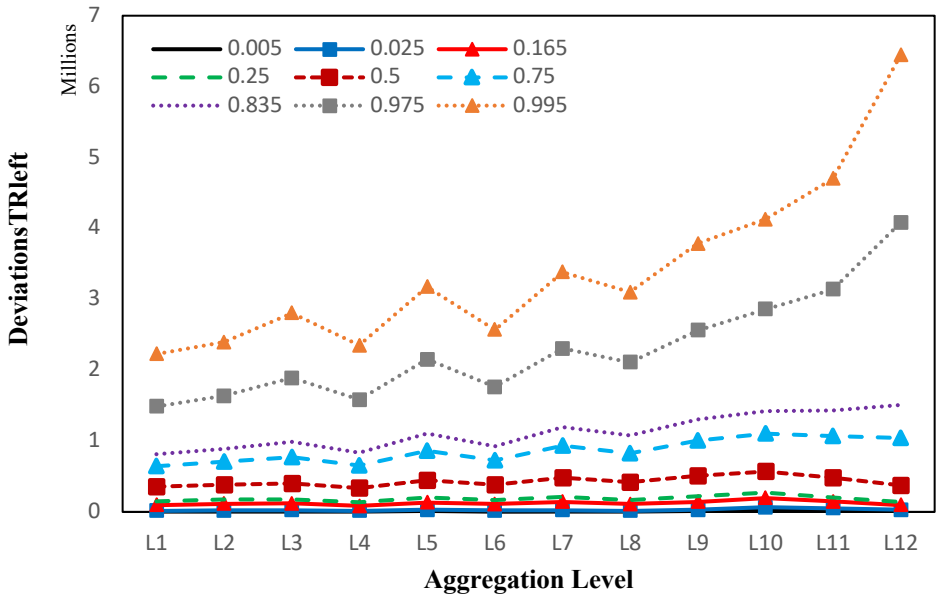

Figure 3: **DeviationsTRleft** Within a Given Aggregation Level as a Function of Quantile Levels

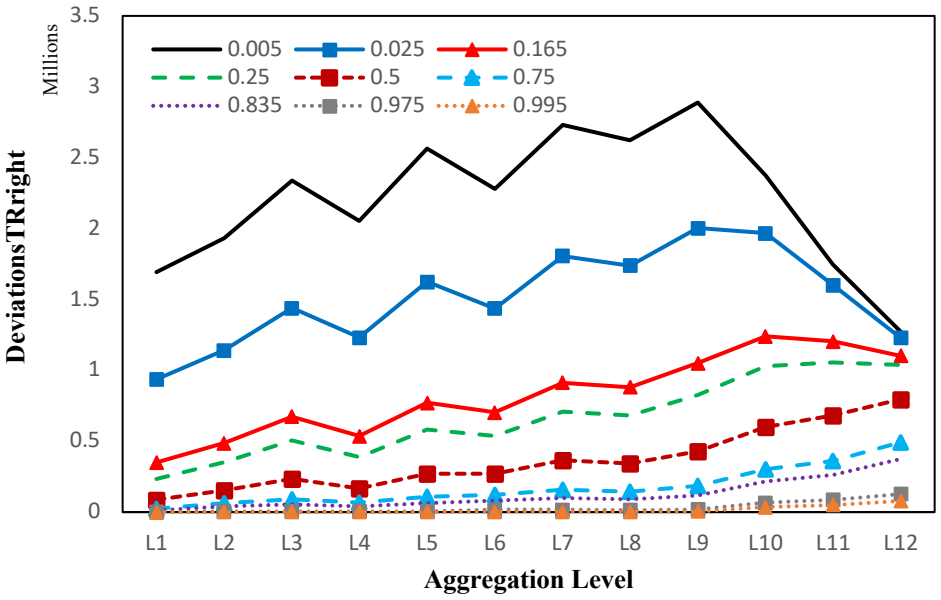

Figure 4: **DeviationsTRright** Within a Given Aggregation Level as a Function of Quantile Levels

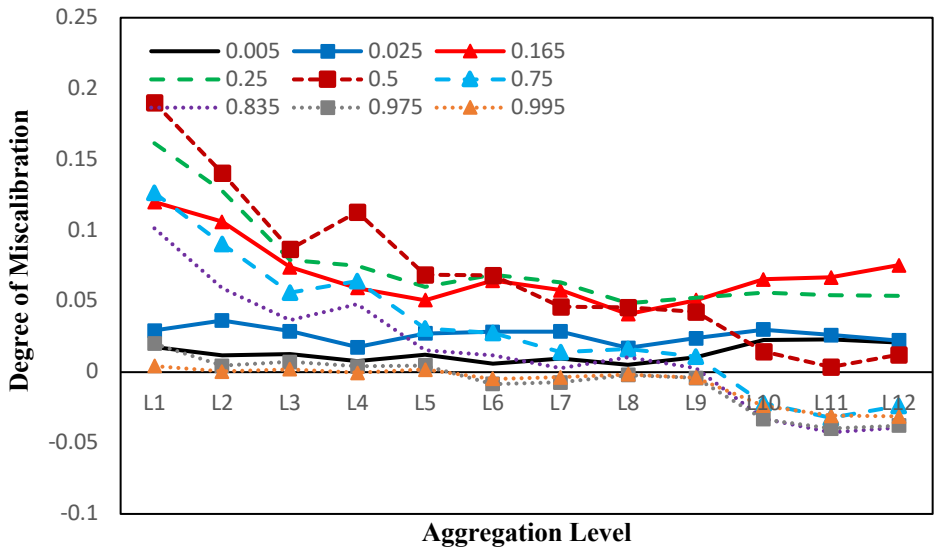

Figure 5: **Degree of Miscalibration** Within a Given Aggregation Level as a Function of Quantile Levels

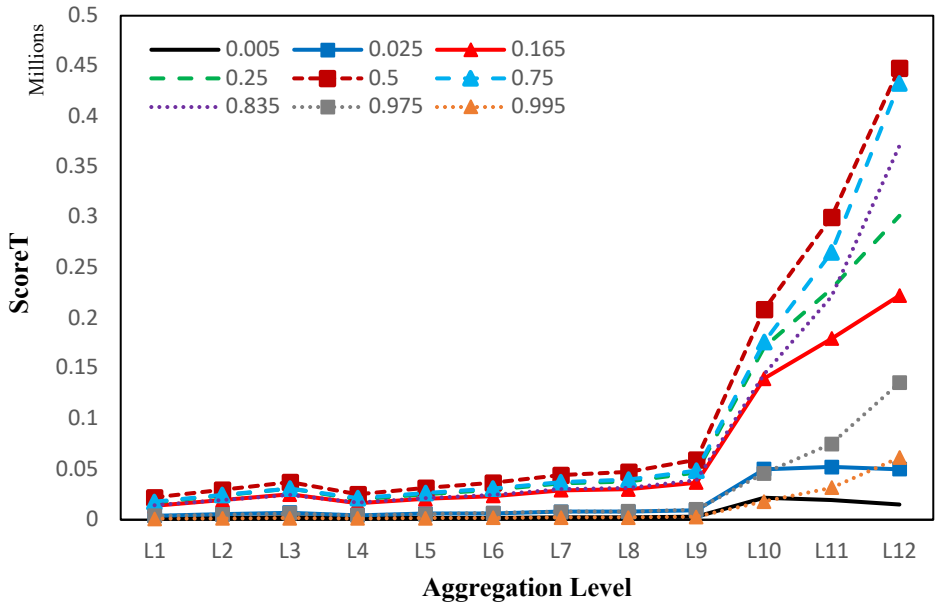

Figure 6a: **ScoreT** Within a Given Aggregation Level as a Function of Quantile Levels

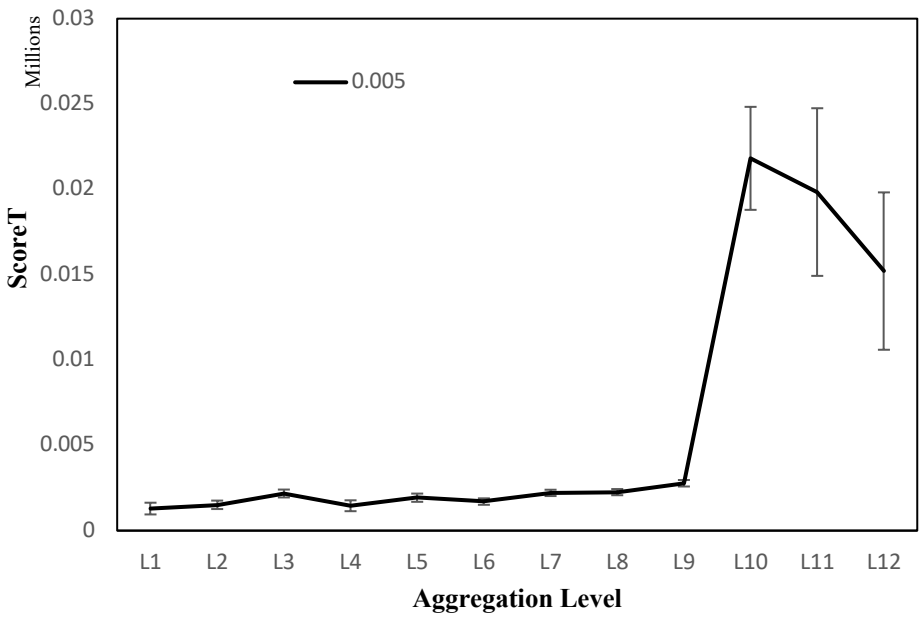

Figure 6b: **ScoreT** Within a Given Aggregation Level for the 0.005 Quantile with Error Bounds

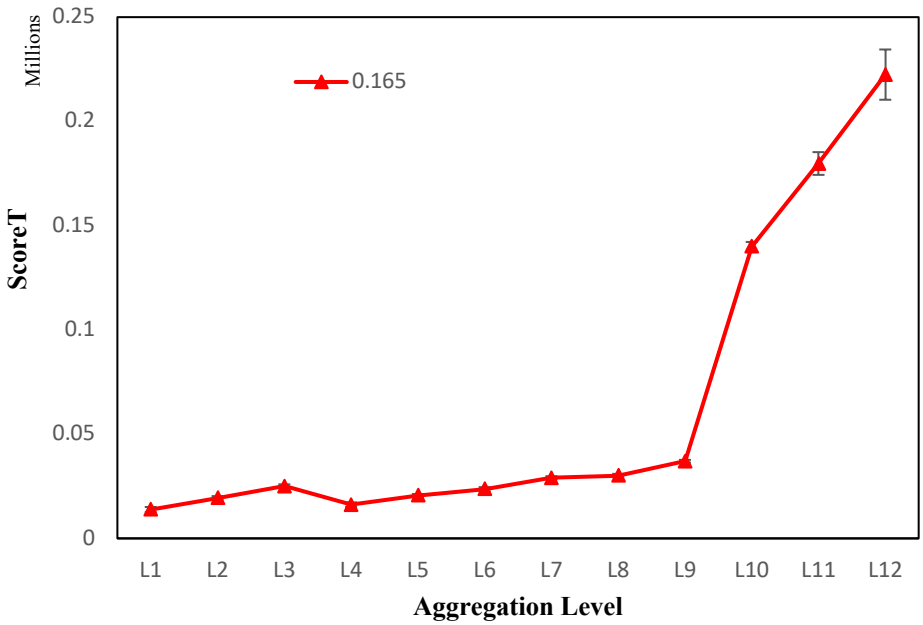

Figure 6c: **ScoreT** Within a Given Aggregation Level for the 0.165 Quantile with Error Bounds

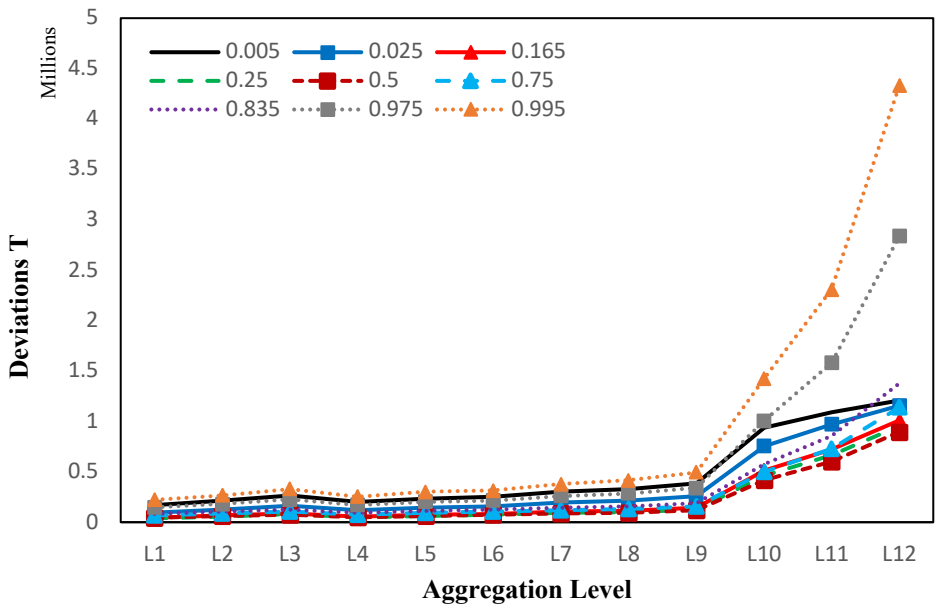

Figure 7: **DeviationsT** Within a Given Aggregation Level as a Function of Quantile Levels

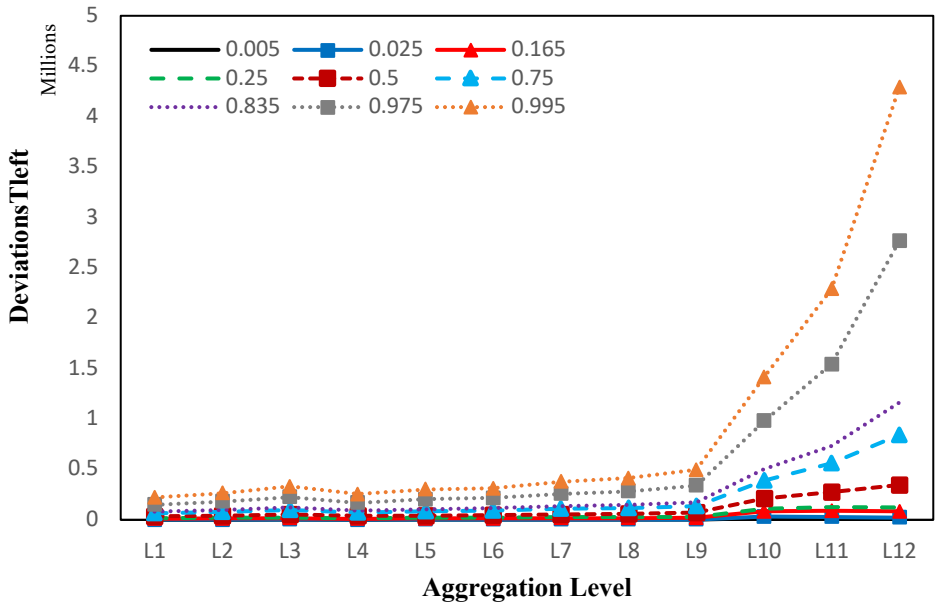

Figure 8: **DeviationsTleft** Within a Given Aggregation Level as a Function of Quantile Levels

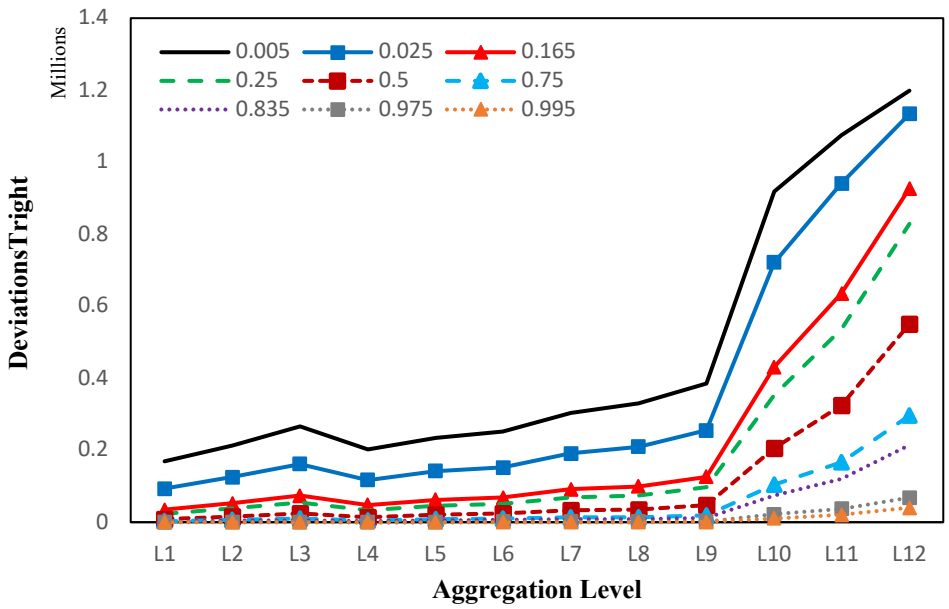

Figure 9: **DeviationsTright** Within a Given Aggregation Level as a Function of Quantile Levels

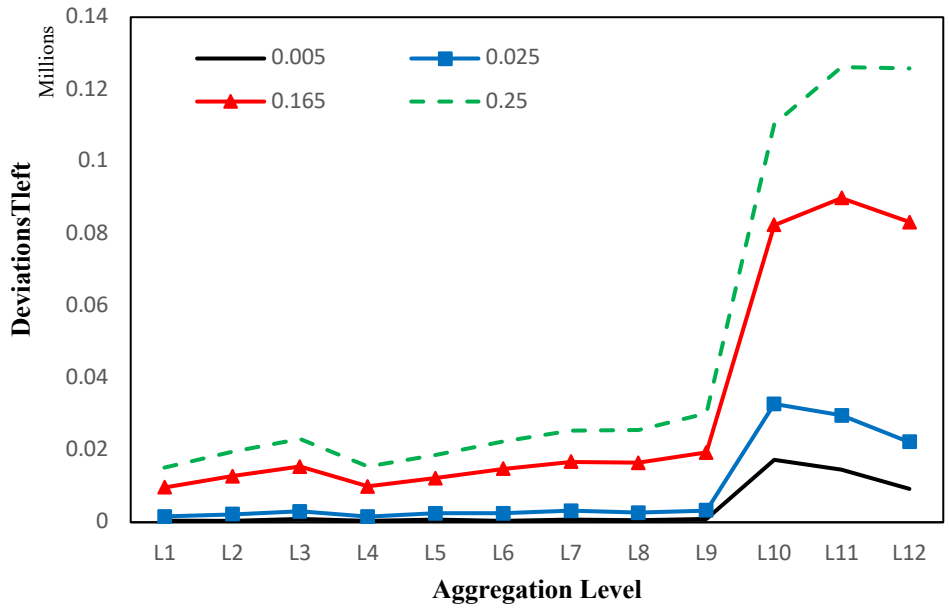

Figure 10: **DeviationsTleft** Within a Given Aggregation Level as a Function of Quantile Levels for the Four Left-tail Quantiles

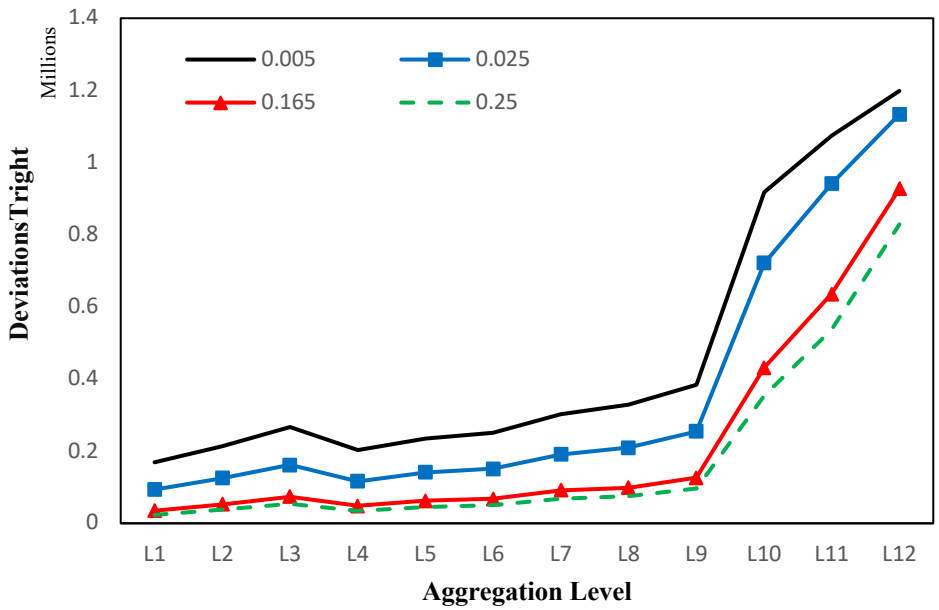

Figure 11: **DeviationsTright** Within a Given Aggregation Level as a Function of Quantile Levels for the Four Left-tail Quantiles

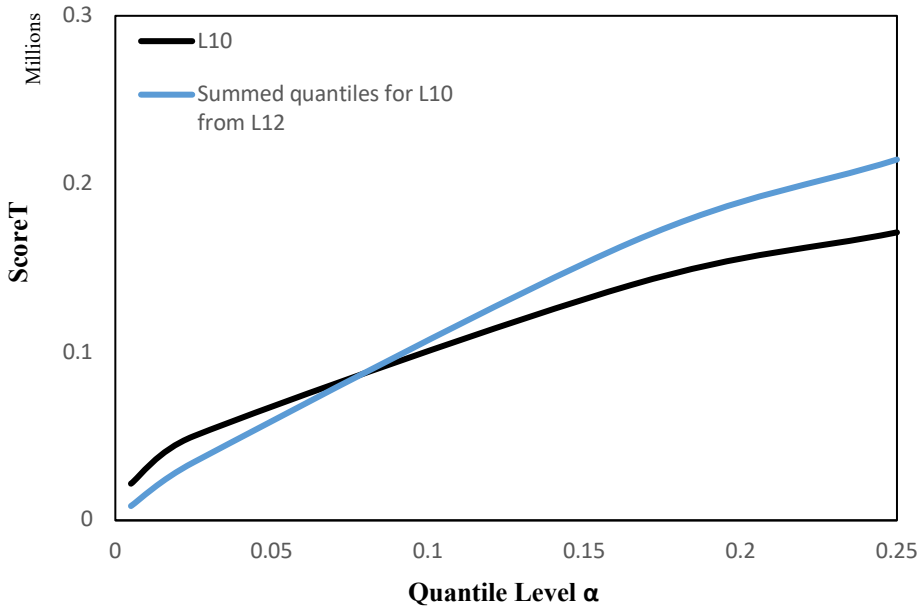

Figure 12: **ScoreT** for L10 Forecasts and Summed Quantiles Per Team for L10 (from L12) as a Function of Quantile Levels for Left-tail Quantiles

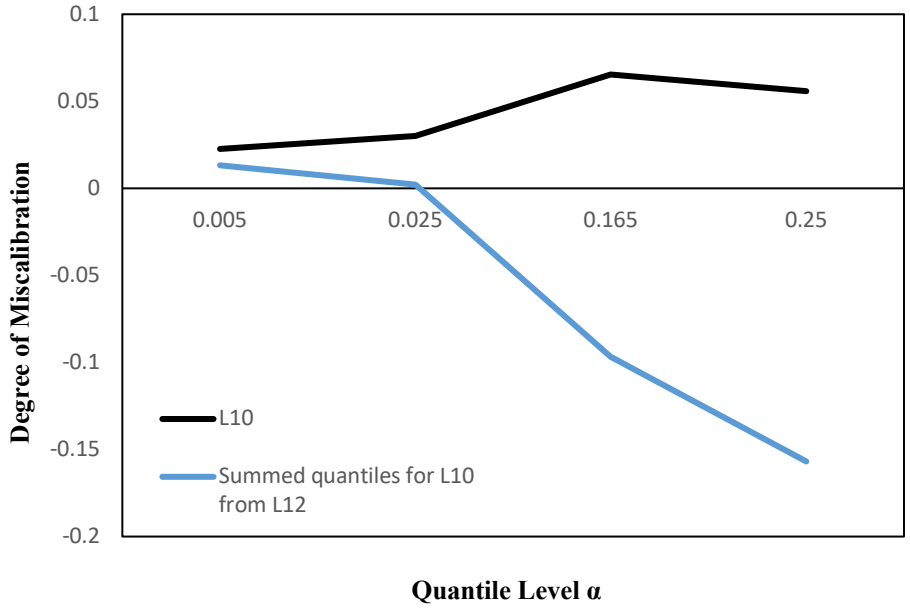

Figure 13: **Degree of Miscalibration** for L10 Forecasts and Summed Quantiles for L10 (from L12) as a Function of Quantile Levels for Left-tail Quantiles

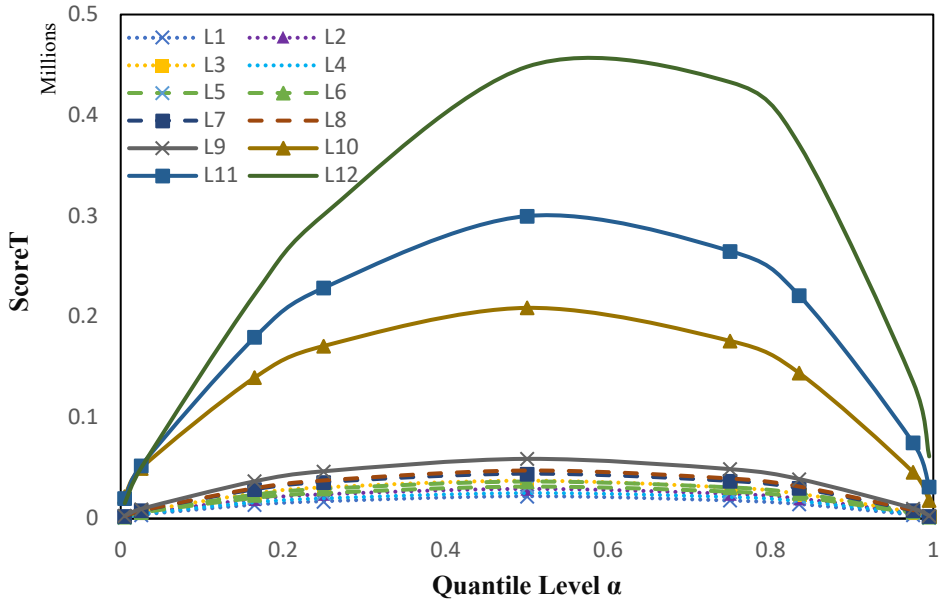

Figure 14: **ScoreT** Within a Given Quantile Level as a Function of Quantile Levels for Twelve Different Aggregation Levels