Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

The M5 uncertainty competition: Results, findings and conclusions

Spyros Makridakis [a], Evangelos Spiliotis [b,*], Vassilios Assimakopoulos [b], Zhi Chen [c], Anil Gaba [d], Ilia Tsetlin [d], Robert L. Winkler [e]

[a] *Institute For the Future, University of Nicosia, Ringgold ID 121343, Cyprus*
[b] *Forecasting and Strategy Unit, School of Electrical and Computer Engineering, National Technical University of Athens, Greece*
[c] *Department of Analytics and Operations, NUS Business School, National University of Singapore, 119245, Ringgold ID 37580, Singapore*
[d] *INSEAD, 138676, Ringgol ID 52160, Singapore*
[e] *Fuqua School of Business, Duke University, Durham, NC 27708, Ringgold ID 33853, USA*

ARTICLE INFO

Keywords:
Forecasting competitions
M competitions
Uncertainty
Probabilistic forecasts
Time series
Machine learning
Retail sales forecasting

ABSTRACT

This paper describes the M5 “Uncertainty” competition, the second of two parallel challenges of the latest M competition, aiming to advance the theory and practice of forecasting. The particular objective of the M5 “Uncertainty” competition was to accurately forecast the uncertainty distributions of the realized values of 42,840 time series that represent the hierarchical unit sales of the largest retail company in the world by revenue, Walmart. To do so, the competition required the prediction of nine different quantiles (0.005, 0.025, 0.165, 0.250, 0.500, 0.750, 0.835, 0.975, and 0.995), that can sufficiently describe the complete distributions of future sales. The paper provides details on the implementation and execution of the M5 “Uncertainty” competition, presents its results and the top-performing methods, and summarizes its major findings and conclusions. Finally, it discusses the implications of its findings and suggests directions for future research.

1. Introduction

Why is the M5 “Uncertainty” competition important? In recent years, the interest in and use of probabilistic forecasts of different types has increased considerably (Gaba et al., 2017; Winkler et al., 2019). Advances in computing have played a significant role in this, as have the resulting developments in analytics and data science and the general public’s exposure to more and more probabilistic forecasts in the media. Companies have invested in hiring young computer scientists and statisticians who are conversant in the latest techniques. The number of papers on probabilistic forecasting in the forecasting, statistical, and computer science literature has increased dramatically (Gneiting, 2008).

There is a general realization that the world is changing quickly and that understanding uncertainties and the resulting risks is becoming more critical. Point estimates are useful, but probabilistic forecasts are also needed for decision-makers concerned about the uncertainty and the risks of their decisions. Fisher and Raman (1996) and Gaba et al. (2019), for example, discuss the inevitable need to convert point estimates of sales into probabilistic forecasts for making stocking decisions. However, this goes beyond just sales or demand forecasting. Probabilistic forecasts are essential for decision-makers in any situation where robust decisions must be made against future uncertain outcomes such as, for example, in budgeting

[*] Corresponding author.
E-mail address: spiliotis@fsu.gr (E. Spiliotis).

https://doi.org/10.1016/j.ijforecast.2021.10.009

decisions or hedging and trading strategies. This is further illustrated by the current environment, with profound uncertainties involving the COVID-19 pandemic, climate change, the economic situation, and global conflicts. The M5 "Uncertainty" competition is important because it provides a laboratory of sorts to discover new methodologies with the potential to improve probabilistic forecasts and learn how to deal with future uncertainties.

Over the past 40 years, the M competitions have influenced the theory and practice of forecasting by providing insight into what can lead to better forecasts. In the M4 competition, forecasts to deal with uncertainty were included in a limited way. We learned, among other things, that mixtures of forecasts from different types of methods were beneficial and that machine learning (ML) methods and hybrid approaches were promising in helping us to better face the future (Makridakis et al., 2020c).

The M5 "Uncertainty" competition moved forward into a new, major direction compared to the M4. For the first time, participants forecast a real problem many decision-makers face and made a larger number of forecasts dealing with uncertainty. The number of teams providing the required uncertainty forecasts increased from 49 in the M4 competition to almost 900 in the M5. The results of the M5 competition reinforced the key takeaways from the M4 and provided some important new findings that are discussed in this paper.

2. Implementation and execution

The M5 "Uncertainty" competition was organized following the general principles described by Makridakis et al. (2021). As such, the competition began on March 3rd, 2020, when the initial train data set became available to download on the Kaggle platform,[1] and ended on June 30th, 2020, when the final leaderboard was announced. Also, the competition was chronologically divided into two phases used for evaluating the teams, namely the "validation" and "test" phases. The former phase was used to support the participating teams in assessing their forecasting methods by receiving feedback, while the latter was for the final evaluation of their performance. Similarly, the data set used involved the unit sales of 3,049 products sold in the USA. This was organized in the form of grouped time series aggregated based on their type (three categories divided into seven departments) or selling location (ten stores located in three states), thus consisting of 42,840 series and 12 cross-sectional aggregation levels.

The implementation and execution of the M5 "Uncertainty" competition differed from the "Accuracy" one (Makridakis et al., 2020a) only in the following four aspects: (i) submission template, (ii) performance measure, (iii) prizes, and (iv) benchmarks. These aspects are described in the following subsections.

2.1. Submission

All forecasts were submitted through the Kaggle platform using the template provided by the organizers. The template for the M5 "Uncertainty" competition referred to all 42,840 series of the data set, also covering the nine different quantiles that had to be predicted for each series, thus requiring a total of 385,560 forecasts. Note that, based on the submission format, the forecasts submitted by the participating teams could be independent and, therefore, not hierarchically related (Hyndman et al., 2011). This is because there is still not a unique, well-accepted way to define coherency in probabilistic forecasting settings (Panagiotelis et al., 2020), which is in contrast to point forecasts where coherence is clearly defined (forecasts at the lower aggregation levels have to sum up to the ones at the higher levels). In this regard, the submission format of the "Uncertainty" competition differed from that of the "Accuracy" one, with the latter consisting of the forecasts at the lowest aggregation level of the data set (level 12, product-store) that could then be appropriately aggregated (summed up) to derive coherent forecasts for the rest of the levels.

Note that, similar to the "Accuracy" competition, teams were allowed to submit a maximum of five entries per day. However, for their final evaluation, each team had to select a single set of quantile forecasts (one submission). In real life, forecasters face the same problem of choosing a single set of forecasts that they believe will represent the future as precisely as possible. If no particular submission was selected, the system automatically selected the one with the highest performance during the "validation" phase.

2.2. Performance measure

The M5 "Uncertainty" competition considered nine different quantiles, $q(u)$, of nominal values (probability levels) $u \in \{0.005, 0.025, 0.165, 0.250, 0.500, 0.750, 0.835, 0.975, 0.995\}$, for predicting the uncertainty distributions of the realized values of the series of the data set, namely the median and the 50%, 67%, 95%, and 99% central prediction intervals (PIs). The smaller quantiles (of u_1=0.005 and u_2=0.025) correspond to the left side of the distributions, the higher ones (of u_8=0.975 and u_9=0.995) to the right side of the distributions, while the rest (of u_3=0.165, u_4=0.250, u_5=0.500, u_6=0.750, and u_7=0.835) to the middle of the distributions. Therefore, the median and the 50% and 67% central PIs provide a good sense of the middle of the distributions, while the 95% and 99% central PIs provide information about their tails, which are important in terms of the risk of extremely high or extremely low outcomes. Consequently, the nine quantiles considered provide sufficient information about the uncertainty and allow the effective description of the complete distributions.

Note that, typically, in retail sales forecasting applications, forecasters focus on high quantiles (usually of values between 0.925 and 0.995) as they are more helpful in determining safety stocks (Barrow & Kourentzes, 2016). However, since M5 does not focus on a particular

[1] https://www.kaggle.com/c/m5-forecasting-uncertainty.

decision-making problem or define the exact parameters of such a problem (which could also vary for different aggregation levels and series), it becomes evident that all quantiles could be potentially useful for decision making. Moreover, since the objective of the M5 is to predict the uncertainty distributions of realized values of the series as precisely as possible, both sides and ends of the distributions are considered equally relevant.

For the probabilistic forecasts to be precise and informative (Gneiting & Raftery, 2007), they must display a relative frequency (RF, also known as calibration or coverage) of value close to the nominal probability level and a reasonable, relatively small deviation from the realized values. In other words, we expect the probabilistic forecasts to capture uncertainty without significantly deviating from the realized values of the series (Makridakis et al., 2020c). RF is computed as follows

$$\mathrm{RF(u)} = \frac{1}{h}\sum_{t=n+1}^{n+h} \mathbf{1}\left\{y_t \leq q_t(u)\right\}, \tag{1}$$

where y_t is the actual future value of the examined time series at point t, $q_t(u)$ the generated forecast for quantile u, n the length of the training sample (number of historical observations), h the forecasting horizon (28 days), and **1** the indicator function (being 1 if y is within the postulated interval and 0 otherwise). For example, the forecasts that refer to quantile u_6, which has a nominal probability level of 0.75, should be larger than the realized values in 75% of the cases, thus having an RF of 0.75. Moreover, the forecasts should also have the smallest possible deviation from the realized values. These properties are critical in retail sales forecasting applications where the width of the respective PIs (distance between the upper, U, and lower, L, bounds/endpoints) serves as a proxy for holding costs (Svetunkov & Petropoulos, 2018) and the RF as a proxy of the target service level (Trapero et al., 2019). Accordingly, for a measure to properly evaluate the performance of probabilistic forecasts, it has to calculate a penalty that becomes (i) more significant when the future values are outside the specified bound and (ii) larger as the distance between the realized values and the forecasts increases (Makridakis et al., 2020c).

The M5 "Uncertainty" competition evaluated the submitted forecasts using a measure that incorporates the criteria mentioned above, namely the Scaled Pinball Loss (SPL) function. This is a variation of the pinball loss (Gneiting, 2011), modified for scaling the derived errors and allowing comparisons for multiple series of different scales (Hyndman & Koehler, 2006). The measure is calculated for each series i and quantile u as follows:

$$\begin{aligned}&\mathrm{SPL}_i(u)\\&\quad= \frac{1}{h}\frac{\sum_{t=n+1}^{n+h}(y_t - q_t(u))u\mathbf{1}\left\{q_t(u) \leq y_t\right\} + (q_t - y_t)(1-u)\mathbf{1}\left\{q_t(u) > y_t\right\}}{\frac{1}{n-1}\sum_{t=2}^{n}|y_t - y_{t-1}|}.\end{aligned} \tag{2}$$

Note that the denominator of SPL (in-sample, one-step-ahead mean absolute error of the Naive method) is computed only for the periods during which the examined product(s) are actively sold, i.e., the periods following the first non-zero demand observed for the series under evaluation. This is done because many of the products included in the data set started being sold later than the first available date (Makridakis et al., 2021).

After estimating SPL for all 42,840 time series of the competition separately for each quantile (average performance reported for each series across the complete forecasting horizon per quantile), the overall performance of the forecasting method is computed by averaging the SPL scores across all series of the data set and quantiles using appropriate weights. The measure, to be called Weighted SPL (WSPL), is defined as follows:

$$WSPL = \sum_{i=1}^{42,840} w_i \times \frac{1}{9}\sum_{j=1}^{9} SPL_i(u_j), \tag{3}$$

where w_i and $SPL_i(u_j)$ is the weight and the SPL score of the i$^{\text{th}}$ series of the competition for quantile j, respectively. Lower WSPL scores indicate more precise forecasts. WSPL, a weighted variant of proper scoring rules, is the measure officially used by the competition organizers to rank the participating methods and the primary performance measure of the present study.

Similar to the M5 "Accuracy" competition, the weights are computed based on the last 28 observations of the final training sample of the data set. Specifically, this is based on the cumulative actual dollar sales that each series displayed in that particular period (sum of units sold multiplied by their respective price). This means some slow-moving series may be assigned zero weights, i.e., not contributing to the WSPL score's estimation. Although this was true in 883 cases, i.e., about 2% of the 42,840 series of the competition, the realized dollar sales of these series in the testing period were minor, accounting for less than 1.3% of the total revenue. Note also that, as reported in Table 3 of Makridakis et al. (2021), the dollar sales computed at state, store, product category, and product department level do not change significantly between the "validation" and "test" phase of the competition, meaning that the weights of the WSPL measure remain relatively constant for short periods of time and are therefore indicative of the value that each series represents.

Note that the estimation of WSPL differs from the approaches adopted in the previous M competitions. All errors were computed per series and forecasting horizon in the first three competitions and then equally averaged together. In the M4 competition, the errors were first averaged per series, exactly as done in M5, but then averaged again using equal weights. We believe that the weighting scheme adopted in the M5 competition is more appropriate for successfully identifying forecasting methods that add significant value to retail companies interested in accurately forecasting the series that provide relatively higher revenues. Moreover, since all aggregation levels and quantiles are equally weighted (considered equally important for determining the winning methods), decision-making is effectively supported at all cross-sectional levels by requiring equally precise forecasts for all parts of the uncertainty distributions.

An indicative example for computing WSPL can be found in the Competitors' Guide, available on the M5 website.[2] The code for estimating WSPL, and the exact weight of each series, can be found in the GitHub repository of the competition.[3]

2.3. Prizes

For a team to be eligible for a prize, probabilistic forecasts had to be provided for all 42,840 series of the competition and all nine quantiles considered. Moreover, teams had to give a code for reproducing the forecasts initially submitted to the competition and some documentation describing the forecasting method used.

Just like in M4, objectivity and reproducibility was a prerequisite for collecting any prize (Makridakis et al., 2018a) and therefore, the winning teams, with the exception of companies providing forecasting services and those claiming proprietary software, had to upload their code onto the Kaggle platform no later than 14 days after the end of the competition (i.e., the 14th of July 2020). This material was later uploaded onto the M5 public GitHub repository for individuals and companies interested in using the winning methods while crediting the team that had developed them. Companies providing forecasting services and those claiming proprietary software had to provide the organizers with a detailed description of how their forecasts were made and a source or execution file for reproducing them.

After receiving the code and documentation from all winning teams, the organizers evaluated the reproducibility of their results. Since ML algorithms typically involve random initializations, the organizers considered as fully reproducible any method that displayed a reproducibility rate, i.e. absolute percentage difference of WSPL between the original and reproduced forecasts, lower than 2%.[4] Although all winning methods were found to be fully reproducible, if this were not true, the prizes would have been given to the next best-performing and fully reproducible submission.

The prizes of the M5 "Uncertainty" competition are listed in Table 1. Note that there were no restrictions preventing a team from collecting both a regular and a student prize.[5] Moreover, there were no restrictions preventing a team from collecting a prize in both the M5 "Accuracy" and the M5 "Uncertainty" competitions. The awards were given during the virtual, online M5 conference on October 29th, 2020.

An amount of $40,000 was generously provided by Kaggle, who also waived the fees for hosting the M5 competition. In addition, MOFC and Google generously provided $20,000 each, while Walmart, apart from the M5 data set, also generously provided an amount of $10,000. Finally, the global transportation technology company Uber generously provided $5,000, while IIF generously provided another $5,000. The total amount of $100,000 was equally distributed between the "Accuracy" and "Uncertainty" challenges of the M5 competition.

2.4. Benchmarks

As with the M5 "Accuracy" competition, the organizers considered a set of simple forecasting methods to serve as benchmarks to compare the performance of the forecasting approaches submitted by the participating teams and evaluate potential improvements. In total, there were six benchmarks, including two naive approaches (Naive and seasonal Naive, sNaive), three conventional time series methods, namely exponential smoothing (ETS), simple exponential smoothing (SES), and Autoregressive Integrated Moving Average (ARIMA), and a non-parametric, empirical quantile estimation (Kernel). These benchmarks are described in Appendix C of the supplementary material; their forecasting performance is also evaluated in terms of WSPL and RF.

We will not claim that the selected benchmarks were the most appropriate to use, and we would definitely expect different opinions or perhaps even strong objections to their selection. For instance, with the exception of the Kernel method and the few ETS and ARIMA models that empirically simulate quantiles, the benchmarks considered assume that forecast errors are normally distributed, an assumption that is rarely met in practice when forecasting intermittent demand data (Syntetos & Boylan, 2005). In such cases, the time series data are more likely to follow a Poisson or a negative binomial distribution. This means that relevant theoretical estimations or even empirical simulations would be more appropriate to use, allowing for more meaningful and representative comparisons (Kolassa, 2016).

Nevertheless, we still believe that the examined benchmarks can add value to the analysis of the competition results and facilitate discussion. The reasons are threefold:

- As reported in Makridakis et al. (2021), intermittent demand series are only present at the most granular cross-sectional levels of the M5 data set, i.e., levels 12, 11, and, to a limited extent, level 10. At the rest of the levels, where trend and seasonal patterns become dominant, the series are smooth and display patterns that can be adequately forecast by conventional time series methods. Given that WSPL assigns equal weights to the individual levels, the scores reported for the benchmarks can still allow for valid comparisons, at least when they refer to the overall performance of the methods or their performance at high aggregation levels.
- At levels 12 and 11, one would expect methods that build on empirical estimations or simulations to outperform conventional time series methods that assume normality. However, Table C.1 suggests that the Kernel method produces less accurate forecasts than ETS and ARIMA in terms of WSPL. Therefore, in practice, it could be the case that conventional time

[2] https://mofc.unic.ac.cy/m5-competition/.

[3] https://github.com/Mcompetitions/M5-methods.

[4] Provided that the winning teams had set the seeds used for generating random numbers in their code, it would have been possible in principle to reproduce their results with a reproducibility rate of 0%. Unfortunately, this was not the case.

[5] A student team is one for which at least half of the team members are currently full-time students. Teams that were eligible for the student prize have a name followed by "_STU".

Table 1
The six prizes of the M5 "Uncertainty" competition.

Prize name	Description	Amount
1st prize	Best performing method according to WSPL	$25,000
2nd prize	Second-best performing method according to WSPL	$10,000
3rd prize	Third-best performing method according to WSPL	$5,000
4th prize	Fourth-best performing method according to WSPL	$3,000
5th prize	Fifth-best performing method according to WSPL	$2,000
Student prize	Best performing method among student teams according to WSPL	$5,000
Total		$50,000

series methods do not differ significantly in terms of forecasting performance compared to more theoretically appropriate benchmarks. This finding is in line with the results reported in Spiliotis et al. (2021). In their study, Spiliotis et al. (2021) produced quantile forecasts for the product-store series of the M5 data set (level 12) using both conventional time series methods and empirical computations and simulations. They conclude that the latter, although more promising overall, did not consistently outperform the former.

- Many studies suggest that retail firms continue to use conventional time series methods, such as SES, for forecasting demand (Rostami-Tabar et al., 2013). Although this practice is not recommended, it suggests that evaluating the potential improvements of the winning submissions over such benchmarks could add value to the practice of forecasting, motivating retail firms and suppliers to reconsider their alternatives, if needed.

Drawing from the above, although we believe that the selected benchmarks can provide evidence regarding the potential improvements of the winning submissions of the competition over standard methods frequently used in the industry, pointing towards interesting avenues of future research, we clarify that these findings should be treated with care. Undoubtedly, further analysis of the competition results would be required to reach concrete conclusions on this topic.

3. Participating teams and submissions

The M5 "Uncertainty" competition involved 1,137 participants on 892 teams from 94 countries. Of these teams, 516 entered the competition during the "validation" phase and 376 during the "test" phase. Moreover, 225 teams made submissions during both the "validation" and "test" phase of the competition, while 291 only during the "validation" phase. The participating teams made 9,936 submissions, most of which (about 56%) were submitted during the "validation" phase. Note that most of the teams made a single submission, while the majority of the rest made between 3 and 14 submissions. It is worth mentioning that for 74 participants, including two in the top 10, this was their first time participating in a Kaggle competition.

Similar to the M5 "Accuracy" competition, and due to privacy regulations, no information was made available about the academic background of the participating teams, their experience and skills, and the type of methods utilized (e.g., statistical, ML, combination or hybrid), with the exception of the winning teams and a few more that were willing to share this information with the organizers. However, given that the members of the Kaggle community are typically more experienced in employing ML methods, we assume that most of the teams had (at least) an adequate background in statistics and computer science, and were also familiar with ML forecasting methods.

From the participating teams, 769 (86.2%) managed to outperform the Naive benchmark, 553 (62%) outperformed the sNaive benchmark, and 202 (22.6%) beat the top-performing benchmark (ARIMA). Thus, we find that, proportionally, a larger number of teams outperformed the benchmarks in the M5 "Uncertainty" competition than in the "Accuracy" one. This indicated that some of the methods frequently used in the industry for producing probabilistic forecasts might considerably underestimate uncertainty. However, as discussed earlier, this finding may be due to the selected benchmarks' limitations. Also, similar to the "Accuracy" competition, we find that many teams failed to select the "best" submission made while the competition was active, probably due to misleading validation scores. If all teams had successfully identified their "best" set of forecasts, 87.1%, 71.4%, and 35.1% of the participants would have outperformed the Naive, sNaive, and ARIMA benchmarks respectively, versus 63.7%, 48.8%, and 12.2% in the "Accuracy" challenge.

Fig. 1 summarizes this information, presenting the daily number of submissions made and the cumulative number of participating teams, the number of participants per country, the distribution of the forecasting performance score (WSPL) of the teams that did better than the Naive benchmark, and the forecasting performance of the teams that did better than the top-performing statistical benchmark, along with their respective ranks.

By observing Fig. 1 we notice the following.

- The majority of the teams made most of their submissions during the "validation" phase, when the public leaderboard was available and live feedback could be received. During the "test" phase, most teams probably used their private cross-validation (CV) strategies to fine-tune their methods, mainly submitted on the last day of the competition.
- The majority of the participants originated from Japan (16%), the USA (16%), India (11%), Russia (9.5%), and China (7%). Thus, similar to the "Accuracy" competition, we conclude that there is a large, active community interested in forecasting in both

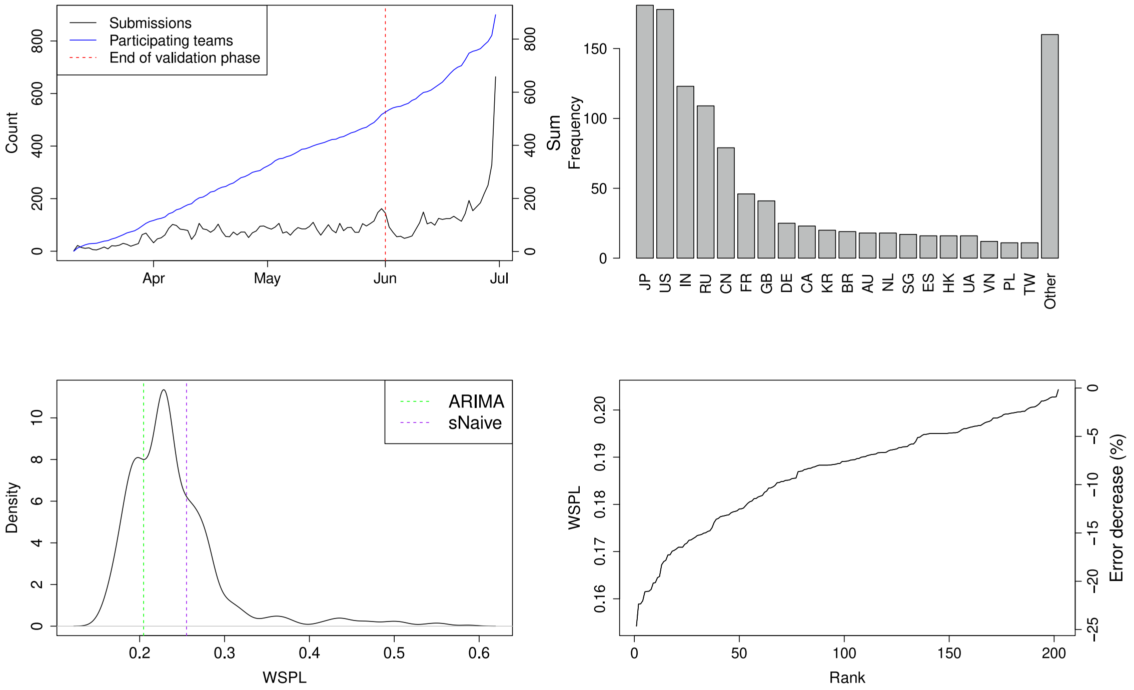

Fig. 1. Summary of the participating teams and submissions made. Top left: The daily number of submissions made (black line, measured based on the left vertical axis) and the cumulative number of participating teams (blue line, measured based on the right vertical axis). The red dotted line indicates the end of the "validation" phase; Top right: Number of participants per country (top 20 in terms of participation), as estimated based on their IP address; Bottom left: The distribution of the forecasting performance (WSPL) achieved by the teams that did better than the Naive benchmark. The green dotted line indicates the forecasting performance of the ARIMA benchmark, while the purple dotted line the forecasting performance of sNaive; Bottom right: The forecasting performance (WSPL) and ranks of the teams that did better than the top-performing benchmark (ARIMA). Percentage improvements over ARIMA are also reported. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

developed and developing countries. However, by comparing the absolute number of participating teams between the two forecasting challenges, we find that fewer practitioners were motivated to estimate the uncertainty than producing the point forecasts, probably due to the higher complexity of the former challenge.

- A significant number of teams managed to outperform the top-performing benchmark of the competition, with the majority of them being more precise than the ARIMA method by about 9%.
- From the 202 teams that managed to outperform all benchmarks of the competition, 10 displayed an improvement greater than 20%, 33 greater than 15%, 67 greater than 10%, and 136 greater than 5%. These improvements suggest that these methods can potentially result in better forecasting performance than relatively simple forecasting approaches frequently used in the industry. Moreover, the five winners of the competition, plus the top-performing student that ranked 7th, were among the ten teams to accomplish a performance improvement greater than 20%, thus achieving a difficult yet clear victory over the rest.

As with the Makridakis et al., 2020a study, the various tables presented in the remainder of this paper focus on the top 50 performing teams of the competition and the benchmarks considered by the organizers, where appropriate. This is done because it is practically impossible to analyze and report the results of all the teams that participated in the competition, considering that very few teams were willing to share detailed information about the methods utilized. This limited what could be learned from analyzing their submissions.

4. Results, winning submissions, and key findings

4.1. Results

4.1.1. Quantile forecasts

Table 2 presents the forecasting performance (WSPL) achieved by the top 50 teams of the "Uncertainty" competition, both overall and across the 12 aggregation levels (for more details about the structure of the M5 data set, please refer to Makridakis et al., 2021). The last column of the table displays the overall (42,840 series) percentage improvement of the teams over the top-performing statistical benchmark (ARIMA), which performance is displayed at the bottom of the table.

Table 2 indicates that all top 50 submissions improve the overall forecasting performance of the top-ranked benchmark by more than 12.5%. At the same time, the improvements are greater than 20% for the top 10 performing methods and an impressive 24.6% for the winning team. This improvement is slightly greater than the

Table 2
The performance of the top 50 teams of the M5 "Uncertainty" competition in terms of WSPL. The results are presented both per aggregation level and overall. Overall percentage improvements are also reported in comparison to the top-performing benchmark (ARIMA). Column-wise minimum values are displayed in bold.

Rank	Team	Aggregation level												Average	
		1 Total	2 State	3 Store	4 Category	5 Department	6 State Category	7 State Department	8 Store Category	9 Store Department	10 Product	11 Product State	12 Product Store		Improvement over ARIMA (%)
1	Everyday Low SPLices	0.074	0.096	0.113	0.091	0.112	0.116	0.136	0.135	0.159	0.284	0.270	0.265	**0.154**	24.6%
2	GoodsForecast - Nick Mamonov	0.067	0.097	0.118	0.085	0.112	0.117	0.141	0.139	0.164	0.302	0.284	0.280	0.159	22.3%
3	Ouranos	0.082	0.100	0.113	0.093	0.112	0.117	0.135	**0.131**	**0.155**	0.303	0.287	0.280	0.159	22.3%
4	Marisaka Mozz	0.074	0.096	0.127	0.097	**0.103**	0.115	0.132	0.144	0.166	0.300	0.283	0.279	0.160	22.0%
5	IHiroaki	0.078	0.101	0.125	0.089	0.114	0.118	0.141	0.139	0.164	0.300	0.282	0.286	0.161	21.1%
6	WalSmart	0.081	0.097	0.119	0.102	0.121	0.120	0.142	0.141	0.175	0.304	0.274	**0.264**	0.162	21.0%
7	Ka Ho_STU	0.080	0.099	0.122	0.098	0.115	0.118	0.139	0.140	0.162	0.299	0.282	0.285	0.162	21.0%
8	Kazu	0.088	0.100	0.131	0.099	0.119	0.117	0.143	0.148	0.177	0.284	0.270	0.270	0.162	20.8%
9	Praveen Adepu	0.066	0.096	0.125	0.083	0.129	0.121	0.152	0.145	0.171	0.310	0.286	0.277	0.163	20.2%
10	golubyatniks	0.080	0.105	0.125	0.094	0.116	0.123	0.145	0.142	0.166	0.302	0.284	0.280	0.163	20.1%
11	Andrij	0.085	0.108	0.129	0.089	0.107	0.124	0.140	0.143	0.164	0.315	0.290	0.280	0.164	19.6%
12	Tobias Tesch_STU	0.070	0.106	0.130	0.083	0.114	0.124	0.143	0.145	0.167	0.311	0.293	0.289	0.165	19.5%
13	Wal Dash Mart	0.109	0.118	0.129	0.114	0.127	0.130	0.144	0.144	0.163	0.292	0.271	0.265	0.167	18.3%
14	A Certain Uncertainty	0.069	0.097	0.129	0.096	0.121	0.125	0.149	0.154	0.177	0.317	0.295	0.287	0.168	18.0%
15	Astral	0.088	0.104	0.127	0.096	0.119	0.124	0.146	0.145	0.166	0.316	0.295	0.292	0.168	17.8%
16	toshi_k	0.108	0.118	0.130	0.111	0.126	0.128	0.144	0.143	0.163	0.298	0.281	0.282	0.169	17.3%
17	Random_prediction	0.073	0.105	0.131	0.097	0.129	0.135	0.155	0.155	0.176	0.303	0.287	0.287	0.169	17.3%
18	EXTASY	0.105	0.109	0.142	0.106	0.124	0.122	0.144	0.151	0.172	0.307	0.283	0.276	0.170	16.9%
19	AkiraIshikawa	0.099	0.110	0.122	0.102	0.113	0.129	0.148	0.147	0.157	0.294	0.307	0.315	0.170	16.8%
20	Peter CXL	0.105	0.109	0.127	0.094	0.111	0.121	0.143	0.144	0.170	0.334	0.302	0.286	0.171	16.7%
21	Takuma Omiya	**0.057**	0.092	0.137	**0.071**	0.109	**0.106**	0.140	0.150	0.178	0.359	0.327	0.324	0.171	16.5%
22	KBU	0.057	0.092	0.137	0.071	0.109	0.106	0.140	0.150	0.178	0.359	0.327	0.324	0.171	16.5%
23	lkdy	0.057	0.092	0.137	0.071	0.109	0.106	0.140	0.150	0.178	0.359	0.327	0.324	0.171	16.5%
24	Jacques Peeters	0.119	0.114	0.139	0.127	0.120	0.140	0.153	0.153	0.168	0.290	**0.270**	0.267	0.172	16.2%
25	Rob Mulla	0.074	0.108	0.135	0.091	0.117	0.127	0.146	0.149	0.173	0.329	0.306	0.307	0.172	16.1%
26	Costas Voglis	0.081	0.098	**0.110**	0.088	0.109	0.114	**0.131**	0.142	0.159	0.326	0.359	0.350	0.172	15.8%
27	RandomObserver	0.081	0.110	0.134	0.101	0.129	0.140	0.158	0.159	0.178	0.305	0.288	0.286	0.172	15.7%
28	Appian	0.081	0.112	0.139	0.097	0.130	0.140	0.162	0.159	0.179	0.306	0.286	0.282	0.173	15.6%
29	rapidsai	0.104	0.112	0.133	0.105	0.125	0.125	0.144	0.149	0.166	0.309	0.298	0.305	0.173	15.4%
30	lz1997	0.104	0.099	0.133	0.098	0.126	0.135	0.146	0.153	0.169	0.327	0.302	0.288	0.173	15.3%
31	Bo Peng	0.094	0.121	0.138	0.105	0.126	0.140	0.153	0.154	0.172	0.306	0.288	0.286	0.174	15.2%
32	shirokane_friends_from_acc	0.064	**0.089**	0.114	0.082	0.105	0.111	0.134	0.136	0.159	0.363	0.358	0.369	0.174	15.2%
33	MPWARE	0.124	0.115	0.131	0.121	0.119	0.132	0.141	0.146	0.161	0.297	0.289	0.310	0.174	15.0%
34	Alberto Benayas	0.088	0.134	0.145	0.100	0.105	0.135	0.147	0.158	0.174	0.325	0.297	0.281	0.174	15.0%
35	lpyczn	0.099	0.113	0.149	0.112	0.136	0.132	0.153	0.165	0.179	0.295	0.280	0.278	0.174	14.8%
36	PreciseMen	0.097	0.118	0.141	0.103	0.129	0.134	0.155	0.156	0.176	0.313	0.290	0.281	0.174	14.8%
37	Volo	0.118	0.122	0.134	0.122	0.130	0.135	0.147	0.149	0.167	0.306	0.283	0.286	0.175	14.5%
38	Football_Winer	0.096	0.117	0.140	0.097	0.135	0.141	0.161	0.161	0.180	0.314	0.291	0.282	0.176	13.9%
39	sibmike_STU	0.076	0.094	0.112	0.085	0.105	0.111	0.133	0.133	0.156	0.373	0.365	0.378	0.177	13.6%
40	Chris X	0.100	0.115	0.148	0.104	0.135	0.133	0.156	0.161	0.181	0.319	0.292	0.281	0.177	13.5%
41	KrisztianSz_STU	0.107	0.122	0.135	0.121	0.138	0.136	0.158	0.153	0.171	0.309	0.290	0.289	0.177	13.3%
42	Onl	0.095	0.123	0.145	0.104	0.131	0.150	0.165	0.164	0.183	0.308	0.286	0.276	0.178	13.3%
43	shuheioka	0.068	0.092	0.112	0.082	0.106	0.111	0.134	0.133	0.157	0.381	0.372	0.382	0.178	13.2%
44	Dennis Igshv	0.085	0.126	0.146	0.108	0.123	0.132	0.146	0.157	0.173	0.343	0.306	0.287	0.178	13.2%
45	slaweks	0.143	0.141	0.141	0.136	0.135	0.141	0.152	0.152	0.169	**0.282**	0.272	0.269	0.178	13.1%
46	valeska	0.114	0.129	0.155	0.105	0.117	0.136	0.149	0.155	0.169	0.326	0.298	0.285	0.178	12.9%
47	M0T0	0.080	0.104	0.122	0.109	0.128	0.126	0.149	0.147	0.167	0.325	0.331	0.353	0.178	12.8%
48	ArshjotKhehra	0.092	0.113	0.136	0.119	0.149	0.143	0.172	0.161	0.188	0.308	0.283	0.277	0.179	12.8%
49	Konrad Banachewicz	0.095	0.114	0.140	0.106	0.146	0.133	0.162	0.159	0.189	0.315	0.294	0.292	0.179	12.7%
50	Lindada	0.100	0.118	0.150	0.121	0.145	0.138	0.171	0.168	0.189	0.297	0.279	0.272	0.179	12.5%
203	ARIMA - Benchmark	0.158	0.148	0.163	0.147	0.167	0.170	0.202	0.178	0.201	0.322	0.298	0.302	0.205	–

one achieved in the M4 competition, where the winning method (Smyl, 2020) performed 22.4% better than the top-performing benchmark (ETS; Hyndman et al., 2002) for the case of the 95% central PI. These results confirm the findings of the M4 competition, suggesting that ML and hybrid methods can potentially improve the precision of probabilistic forecasts (Makridakis et al., 2020c). They also indicate that retail and logistic companies could utilize such innovative forecasting approaches in practice, thus meeting target service levels without necessarily increasing holding costs.

Another interesting finding, which is in accordance with that of the M5 "Accuracy" competition, is that the winning team (*Everyday Low SPLices*), although the best-performing across all aggregation levels, did not display the most precise forecasts for any of the 12 aggregation levels considered. This is also true for the runner-up (*GoodsForecast - Nick Mamonov*), while the third winning team (*Ouranos*) only managed to provide the best forecasts for levels 8 (store-category) and 9 (store-department). Interestingly, the most precise forecasts for each aggregation level are primarily identified across the teams that ranked between the 21st and 32nd position, with the M4 winner, Slawek Smyl, ranked 45th in this challenge, providing the "best" forecasts at level 10 (product). Thus, our results suggest that there may be "horses for courses" (Petropoulos et al., 2014) in demand forecasting, both for the case of point and probabilistic forecasts, and that depending on the aggregation level examined and the particularities of the predicted series, (Spiliotis et al., 2020a; Theodorou et al., 2021), different forecasting methods may be more appropriate for supporting decisions and optimizing performance. Although this finding points towards an interesting avenue for future research, further analysis would be required to ascertain whether the effect mentioned above can be attributed to chance and to what extent (e.g., by testing the significance of the differences of the submissions for each quantile and aggregation level separately using statistical tests).

This finding is better visualized in Fig. 2, which presents the distribution of WSPL for the top 50 performing teams per aggregation level, along with the performance of the ARIMA benchmark and the winning team. As seen, although the winning method is consistently among the top-performing ones across all aggregation levels, it does not provide the "best" results in any of them, doing

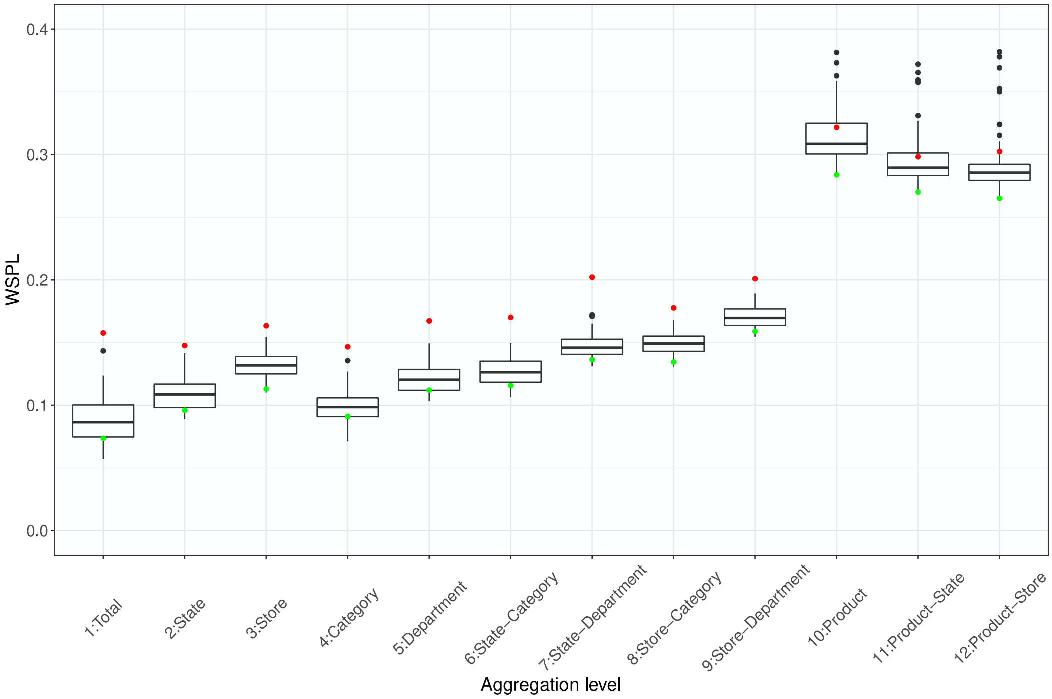

Fig. 2. Forecasting performance (WSPL) of the top 50 performing teams of the M5 "Uncertainty" competition. The results are reported per aggregation level and box-plots are used to display the distribution of the average errors recorded for the examined methods (minimum value, 1st quantile, median, 3rd quantile, maximum value, and outliers, noted with black dots). The red dots indicate the performance of the top-performing benchmark of the competition (ARIMA), while the green dots the performance of the winning team (*Everyday Low SPLices*). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

however impressively well at the lowest levels. Also, it becomes evident that the performance of the methods deteriorates at lower aggregation levels where randomness increases and the time series patterns are more difficult to capture due to intermittency and erraticness (Makridakis et al., 2021). Moreover, although all teams outperform the top-performing benchmark at levels 1 to 9, the improvements reported for the other levels are less significant, with many teams performing even worse than the benchmark at levels 10, 11, and 12. For instance, the average improvement of the methods over the ARIMA benchmark is about 44% at level 1, 26% at levels 2–7, 16% at levels 8 and 9, and drops to 2% at levels 10, 11, and 12. Therefore, we find that the gains of the top-performing methods mainly refer to the top and middle parts of the hierarchy, and are somewhat limited in terms of WSPL at the product, product-state, and product-store levels.

To investigate the performance of the top 50 performing methods for different parts of the distributions, we aggregate the WSPL results by quantile and compare them with those of the top-performing benchmark. The results are summarized in Fig. 3, visualizing the distribution of WSPL for the complete data set but separately for each quantile. We observe that, similar to Fig. 2, the winning submission does not always provide the most precise forecasts for all quantiles. However, this time the winner provides the "best" results for the left side of the distributions. Moreover, we observe that although the top-performing benchmark is outperformed by all teams at quantiles 0.165, 0.250, 0.500, and 0.750, the improvements reported for the other levels are less significant, with many teams performing even worse than the benchmark. This is particularly true for quantiles 0.005 and 0.995, where 46% and 32% of the methods fail to provide more precise forecasts than the ARIMA method. Accordingly, the percentage of teams outperformed by the benchmark at quantiles 0.025 and 0.975 is 22% and 26%, respectively. Therefore, the average improvements of the 50 methods over the ARIMA benchmark are negative at u_1 (−8%), become positive at u_2 (11%), increase at $u_3 - u_5$ (21%), then drop at $u_6 - u_8$ (10%), and become negative again at u_9 (−6%). Thus, we conclude that although the submitted methods provided better forecasts for the middle of the distributions, they offered limited or even negative improvements for their tails. Detailed results on the WSPL scores achieved by the top 50 methods for each quantile are provided in Appendix A of the supplementary material.

Fig. 4 presents the results of a similar analysis that investigates the performance of the top 50 methods per quantile but focuses on the RF of the respective forecasts, i.e., their calibration, instead of the WSPL scores. In this regard, teams that perform well will display RF values close to the nominal probability level and vice versa. Note that the same weighting scheme used for estimating WSPL is employed to compute the overall calibration of each method. By observing Fig. 4, we find that, on average, the calibration of the top-performing methods

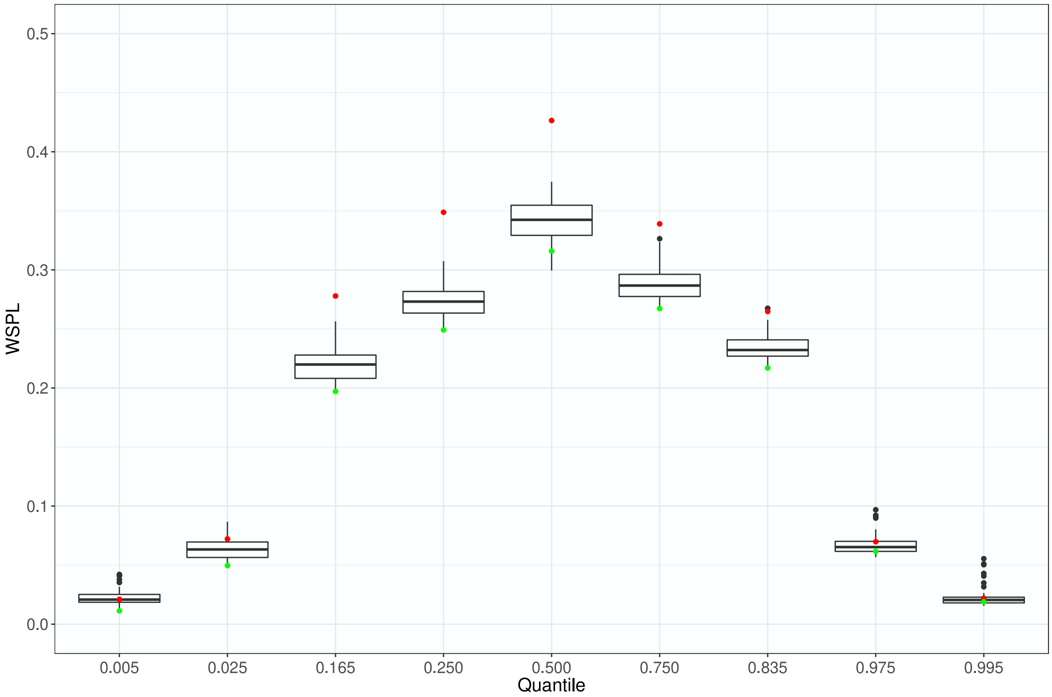

Fig. 3. Forecasting performance (WSPL) of the top 50 performing teams of the M5 "Uncertainty" competition. The results are reported per quantile and box-plots are used to display the distribution of the average errors recorded for the examined methods (minimum value, 1st quantile, median, 3rd quantile, maximum value, and outliers, noted with black dots). The red dots indicate the performance of the top-performing benchmark of the competition (ARIMA), while the green dots the performance of the winning team (*Everyday Low SPLices*). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

is impressive, with the majority of the teams achieving RF values close to the nominal probability level. This is particularly true for the tails and mid-right of the distributions. For instance, *Football_Winner*, ranked 38*th*, provides an impressive calibration of 0.004 and 0.024 for the left tail of the distributions, while the M4 winner, *slaweks*, an impressive calibration of 0.496 and 0.996 for quantiles u_5 and u_9, respectively. However, there are notable variations among the teams, mainly at quantiles 0.165, 0.250, and 0.500. Thus, we conclude that most of the teams managed to calibrate their forecasts effectively. However, this proved to be a challenging task for the middle of the distributions, probably because, in contrast to the tails, quantiles u_3 to u_7 are not naturally bounded. Detailed results on the calibration achieved by the top 50 methods are provided in Appendix B of the supplementary material.

To further investigate the differences reported between the top 50 submissions, as well as the top-performing benchmark, we employ the multiple comparisons with the best (MCB) test (Koning et al., 2005). Specifically, the test computes the average ranks of the forecasting methods according to SPL across the complete data set of the competition. It concludes whether or not these are statistically different. Fig. 5 presents the results of the analysis. If the intervals of the two methods do not overlap, this indicates a statistically different performance. Thus, methods that do not overlap with the gray interval of the figure are considered significantly worse than the best, and vice versa. Although we believe that the results of the MCB test are indicative and useful for investigating the relative performance of the submitted methods, we should clarify that the test presupposes that the forecasts being compared are independent. Since the forecasts used in our case for conducting the MCB test refer to grouped time series and because multiple quantile forecasts are evaluated for each series, strictly speaking, this assumption does not hold, and the conclusions drawn may not be entirely valid.

As seen, team *Wal Dash Mart*, ranked 13th in terms of WSPL, displays the lowest average rank according to MCB and provides significantly better forecasts than the rest of the examined methods. This indicates that it managed to produce the most precise forecasts in the vast majority of the series. As such, similar to the "Accuracy" competition, we find that none of the five winning teams managed to perform significantly better than the rest of the participating teams, with the winner and the runner-up ranking 8th and 26th, respectively, according to MCB. Thus, it is confirmed that the winning teams developed methods that mostly predicted more accurately the relatively more expensive and fast-moving products, which account for larger weights in the WSPL measure, thus

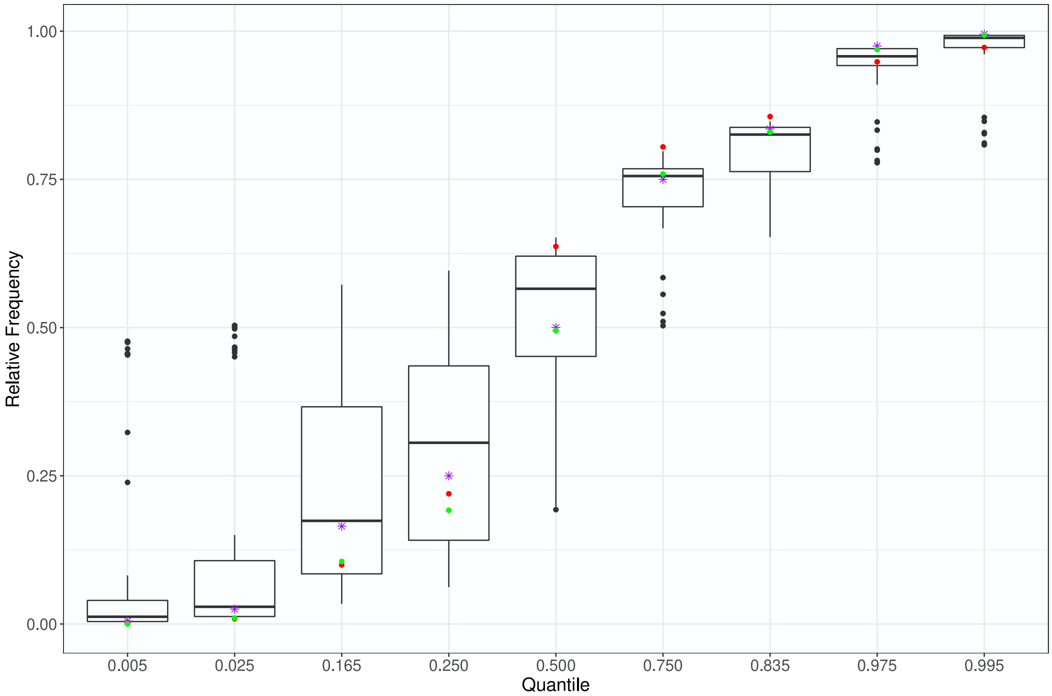

Fig. 4. Forecasting performance (calibration) of the top 50 performing teams of the M5 "Uncertainty" competition. The results are reported per quantile, and box-plots are used to display the distribution of the average relative frequency values recorded for the examined methods (minimum value, 1st quartile, median, 3rd quartile, maximum value, and outliers, noted with black dots). The red dots indicate the performance of the top-performing benchmark of the competition (ARIMA), the green dots the performance of the winning team (*Everyday Low SPLices*), while the purple dots the nominal probability level of the quantile. The same weighting scheme used for estimating WSPL is employed for computing the overall relative frequency of each method. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

providing sub-optimal results for the rest of the series. This highlights the fact that the objective of the competition (producing accurate forecasts across all aggregation levels and especially for high-valued series) was critical for determining the winning submissions and that, depending on the evaluation measure employed, different methods could be identified as the "best". On the other hand, given that the winning methods were flexible and highly parameterized with the objective to minimize WSPL, it could be the case that different optimizations of the same methods could result in "optimal" forecasts according to different criteria. Note also that the benchmark provides significantly better forecasts than 12 of the top 50 methods, mostly the ones that are ranked below the 20th position.

Finally, we investigate the impact of the forecasting horizon's length on the performance achieved by the top 50 methods of the competition. To do so, we first compute the WSPL score of these methods for each forecasting horizon and series separately and then aggregate the results per aggregation level and horizon. A summary of the results is presented in Fig. 6. We observe that, as with the "Accuracy" challenge, although in most of the aggregation levels, the performance of the methods remains constant and is even slightly reduced in some cases. This is not true for the lowest aggregation levels (10, 11, and 12) where the accuracy significantly deteriorates as the forecasting horizon increases. This can be attributed to the intermittency and erraticness of the series, which is limited at the top levels of the data set but significant at the lowest ones (Makridakis et al., 2021). It is also notable that, in many aggregation levels (especially at the lowest ones), the errors display some sort of periodicity (larger errors are observed during the weekends). This indicates that part of the seasonality present in the data was not appropriately captured by the forecasting methods of the competition, even the top-ranked ones.

4.1.2. Prediction intervals

In this section, we analyze the performance of the central PIs of the top 50 submissions of the M5 "Uncertainty" competition. Interval forecasts are closely related to quantile forecasts. A central $(1-\alpha)\times 100\%$ PI, where $\alpha \in (0, 1)$, consists of two endpoints, namely the lower and upper bounds $[L(\alpha), U(\alpha)]$. The former endpoint corresponds to the $\alpha/2$ quantile and the latter to the $(1-\alpha/2)$ quantile. For instance, the endpoints of a 95% ($\alpha = 0.05$) PI are the 0.025 and 0.975 quantiles. Building on the measures used in Section 2.2 for evaluating quantile forecasts, we briefly introduce performance measures for the case of the interval forecasts and then proceed with our analysis. For a further discussion on these measures, please see Gaba et al. (2017).

- **S-score:** The S-score is a strictly proper scoring rule that measures the overall performance of an interval

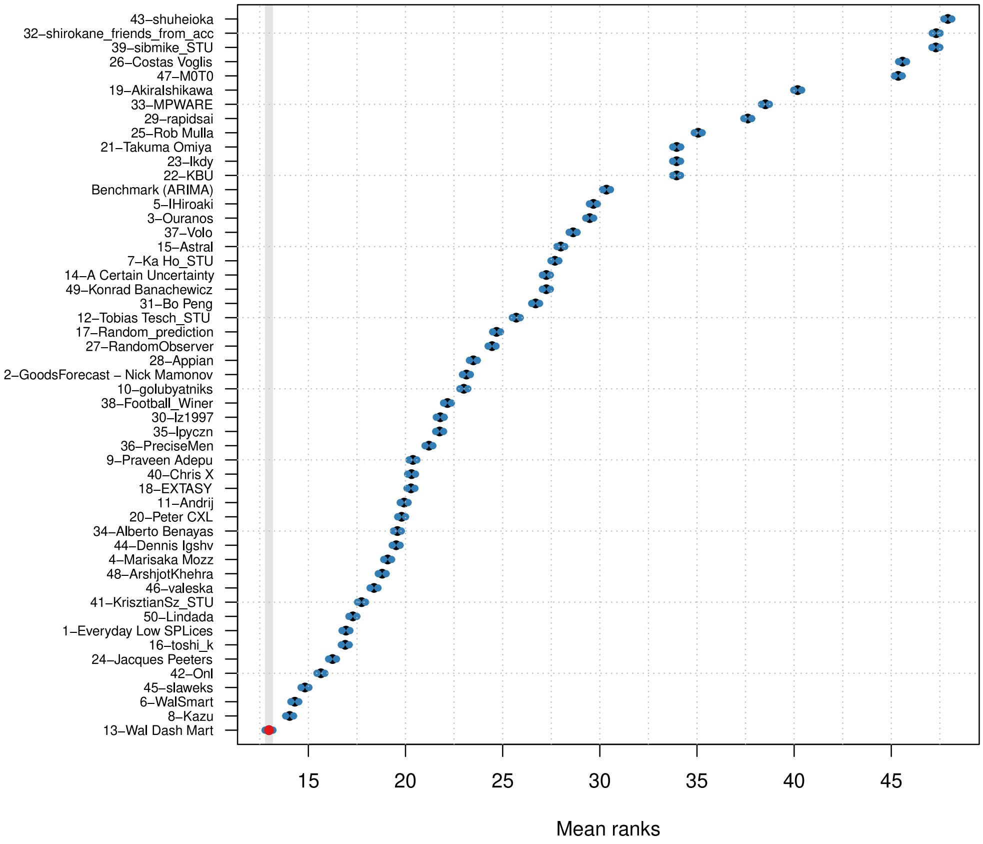

Fig. 5. Average ranks and 95% confidence intervals of the top 50 performing teams of the M5 "Uncertainty" competition, plus the top-performing benchmark (ARIMA) over all series: multiple comparisons with the best (SPL used for ranking the methods) as proposed by Koning et al. (2005). The overall rank of the teams in terms of WSPL is displayed to the left of their names.

forecast (Gaba et al., 2017; Jose & Winkler, 2009). For a $(1-\alpha)\times 100\%$ PI, the S-score is defined as:

$$\text{S-score}(\alpha) = \frac{1}{h}\sum_{t=n+1}^{n+h} \frac{\alpha}{2}(U_t(\alpha) - L_t(\alpha)) + (L_t(\alpha) - y_t)\mathbf{1}\{y_t < L_t(\alpha)\} + (y_t - U_t(\alpha))\mathbf{1}\{y_t > U_t(\alpha)\}\,. \tag{4}$$

The S-score is greater or equal to zero, zero indicating perfect foresight and higher values indicating poorer interval performance. Note that strictly proper scoring rules for quantiles differ from the commonly used strictly proper scoring rules for probabilities (Winkler, 1996), which are not appropriate to evaluate quantile forecasts (Jose & Winkler, 2009). Note also that S-score(α) is equal to the sum of the pinball loss functions of quantiles $\alpha/2$ and $(1-\alpha/2)$, which is essentially the unscaled version (numerator) of the SPL score.

The first term of the S-score represents a penalty for interval width, while the second and third terms are a penalty for the times that the realized values fall outside the specified interval. A wider interval is penalized more on the width. Still, due to missing the realized values, the penalty is simultaneously reduced as the realized values are less likely to fall outside the specified bounds. Therefore, the choice of an interval can be viewed as a trade-off between these two penalties. Proper interval forecasts are sharp in the sense of having narrow intervals and well-calibrated by having an RF close to $(1-\alpha)$.

- **Degree of Miscalibration:** The degree of miscalibration is defined as the difference between the RF reported for the examined forecasting method and the nominal probability level, i.e., $(1-\alpha)$. Any deviation of RF from $(1-\alpha)$ represents miscalibration. More specifically, we say that intervals are overconfident or underconfident if the degree of miscalibration is negative or positive, respectively. Miscalibration typically results from inappropriate interval widths or the location of intervals.
- **Interval Width:** Interval width, $U-L$, indicates the degree of uncertainty about the forecast series. Wider intervals reflect higher uncertainty and vice versa.

Note that, except for the degree of miscalibration, the measures introduced above are sensitive to scales. Thus, to enable fair comparisons across different series and

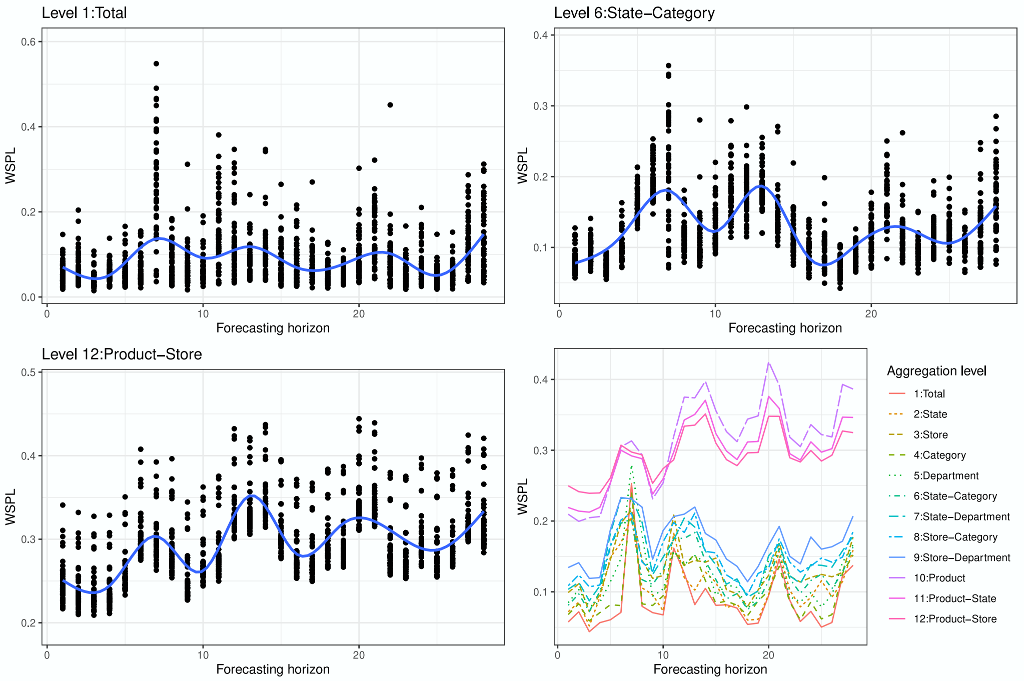

Fig. 6. Forecasting horizon length's impact on the precision of uncertainty estimation. Top left: Forecasting performance (WSPL) of the top 50 performing methods of the competition per forecasting horizon for the top level of the data set. The blue line represents LOESS (locally estimated scatterplot smoothing). Top right: Similar to the top-left figure, but this time the results are reported for the middle aggregation level of the data set (state-category). Bottom left: Similar to the top-left figure, but this time the results are reported for the lowest aggregation level of the data set (product-store); Bottom right: The average performance of all top 50 performing methods across all 12 aggregation levels and forecasting horizons. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

aggregation levels, we consider the scaled version of the measures, exactly as done for the case of the SPL score. Moreover, since we are interested in investigating and understanding the pure, overall performance of the PIs, we treat all series equally. This is equivalent to assigning the same weight w_i of Eq. (3) to all 42,840 series. Therefore, we compute the average of the introduced measures across all 50 submissions over different series and aggregation levels.

In the M5 "Uncertainty" competition, participants submitted nine quantile forecasts that allow the construction of four PIs, namely the 50%, 67%, 95%, and 99% central PIs. Fig. 7 demonstrates how the examined interval performance measures change with respect to different aggregation levels for the four different probability levels.

By observing Fig. 7 we find that the overall performance of the interval forecasts (scaled S-score) tends to get worse for each of the four probability levels as the granularity increases, i.e., from level 1 to 12. This deterioration can be attributed to the degree of miscalibration being worse for more granular levels while at the same time the intervals become more misplaced. These findings make intuitive sense, as granular quantities tend to exhibit more randomness (and thus are harder to forecast), whereas such irregularities are likely to be smoothed out at more aggregated levels. These results confirm and extend our earlier findings on quantiles: regardless of using weighted scores for quantiles (WSPL) or simple averaged scores for intervals (scaled S-score), probabilistic forecasts tend to perform worse at more granular levels. Such decline in performance is also reflected in the degree of miscalibration.

Observe that regardless of the probability and aggregation levels, all interval forecasts exhibit some degree of overconfidence, with the RF of the methods being lower than the nominal one. However, the degree of overconfidence is mostly mild across different aggregation and probability levels, with the worst case (95% PIs at level 12) showing a degree of miscalibration of −15%. This is in line with the findings shown in Fig. 4, where quantiles were, on average, well-calibrated. Furthermore, for a given probability level, if we average the degree of miscalibration across different aggregation levels for each of the four probability levels, there is no noticeable difference across different probability levels, as shown in Fig. 8. Overall, the grand average is about −4.6%, implying that 90% prediction intervals capture realized values around 85.4% of the time. This finding is remarkable and in stark contrast with well-documented findings that subjective forecasts by human experts tend to be severely overconfident. For example, Gaba et al. (2017) report that in making 90% interval forecasts for a wide range of financial products, analysts' interval forecasts have an RF of just 37% on average. Grushka-Cockayne et al. (2017) mention overfitting and overconfidence in models too instead of just subjective forecasts by human experts. These results support the earlier suggestion that ML algorithms can provide more accurate and objective interval forecasts.

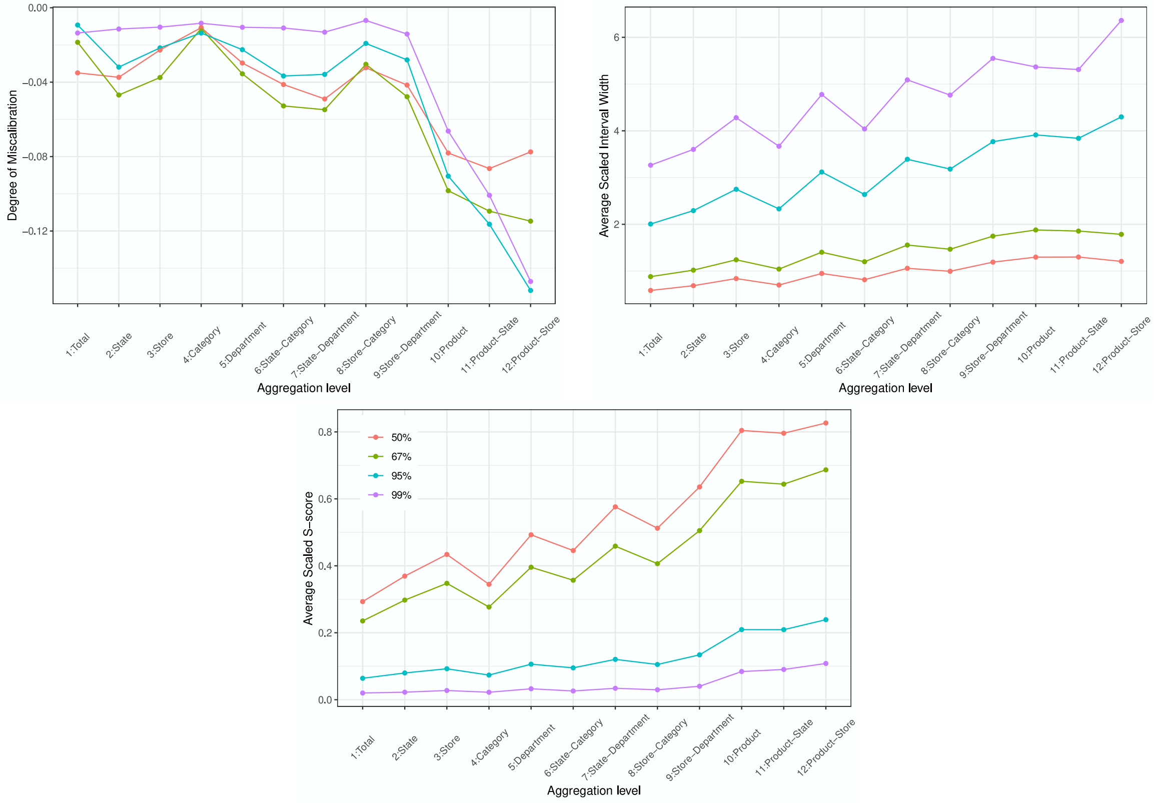

Fig. 7. Average forecasting performance of the top 50 performing teams of the M5 "Uncertainty" competition. The results are reported per prediction interval, while the performance is measured in terms of degree of miscalibration (top left), interval width (top right), and S-score (bottom).

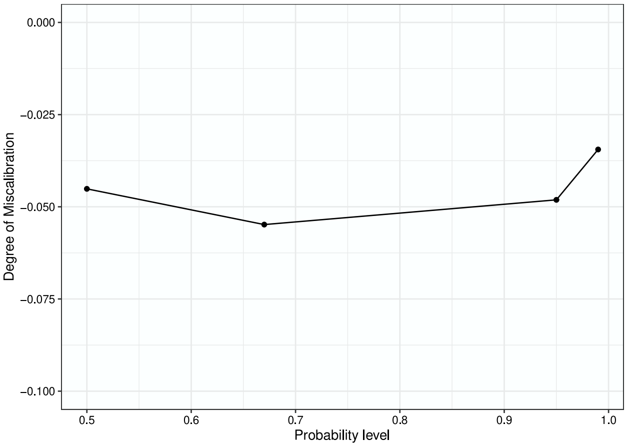

Fig. 8. Average degree of miscalibration of the top 50 performing teams of the M5 "Uncertainty" competition for a given probability level.

4.2. Winning submissions

Unfortunately, similar to the "Accuracy" challenge, a very limited number of teams that participated in the M5 "Uncertainty" competition were willing to share with the organizers and the Kaggle community the description of their methods and the code used for their implementation. Therefore, the organizers tried to reach at least the top 50 performing teams of the competition through e-mails to collect this material. From these teams, only 15 provided the requested information, either by replying directly to the organizers' e-mails or by posting their solutions onto the public discussions and notebooks available on Kaggle. Nevertheless, given that these methods achieved significantly better performance than the benchmarks considered and hundreds of other participating teams, we still believe that they can provide many valuable lessons to advance the theory and practice of forecasting.

The forecasting methods of the five winning teams, plus the top-performing student submission, are summarized below. Note that as with the M5 "Accuracy" challenge, most of the winning methods utilized LightGBM, an ML algorithm for performing non-linear regression using gradient-boosted trees and the "standard" method of choice in Kaggle's recent forecasting competitions (for more information about LightGBM, please refer to Bojer & Meldgaard, 2021; Ke et al., 2017; Makridakis et al., 2020a).[6] The rest of the teams based their solutions mostly on Long Short-Term Memory Neural Networks (LSTM NNs; Hochreiter & Schmidhuber, 1997). In contrast to feed-forward NNs, these contain feedback connections to take into account previous states along with the current input before producing the final output, while also avoiding problems related to long-term dependencies (Makridakis et al., 2018b).

- **First place (*Everyday Low SPLices*; A. David Lander & Russ Wolfinger):** The winning team trained gradient-boosted models for each quantile and aggregation level, involving a total of 126 models. The features of the models were crafted and selected in a forward stepwise fashion based on CV, using one fold for each calendar year. Features included the days of the week and month, the number of days forward being forecast, Supplemental Nutrition Assistance Program (SNAP) activities and holidays, rolling means, medians, and quantiles of the past sales, and statistics about the skewness and kurtosis of the series, along with the percentage of zero observations. Neither price nor any information about product identity was employed. All model hyper-parameters were independently selected using a randomized search across a wide range of LightGBM settings. Series were normalized by dividing their values by the average first differences of the series to facilitate training. Novel sampling and augmentation techniques were used to ensure model generalization. Specifically, series were oversampled, and both their features and targets were jointly shifted using Gaussian noise to produce models that generalized across both time and categories. The final forecasts incorporated blends of multiple folds, test-time augmentation, and reconciliation of the forecast levels.
- **Second place (*GoodsForecast - Nick Mamonov*; Nikolay Mamonov):** The runner-up utilized a hybrid approach of gradient boosting models, simple forecasting time series models, and statistical methods that predicted the probability distributions of the realized values of the series and applied corrections through external adjustments. Specifically, recursive LightGBM models were first used to produce point forecasts at the lowest aggregation level. Given that the point forecasts were biased, they were then adjusted using a factor that was determined so that the aggregated base forecasts had the same value as the point forecasts produced for the top level directly, specified using various singular spectrum analysis models. Then, probability distributions were forecast using a histogram technique which puts more weight on the observations of the same seasonal period with the one being forecast, especially the most recent ones. Finally, the estimated distributions were calibrated so that their median fitted the point forecasts produced by the LightGBM models. Additional adjustments were considered based on CV results.
- **Third place (*Ouranos*; Ioannis Nassios):** This method produced probabilistic forecasts by generating point forecasts and tuning them for each quantile using appropriate coefficients determined based on the last 28 days of the train set. The point forecasts were generated using the contribution of *Ouranos* in the final submission by the *Nodalpoints* team, which ranked 21st in the "Accuracy" challenge. This contribution consists of one LightGBM model, trained per store, and an average of three Keras NNs, trained across all series, combined using a weighted geometric mean approach. The point forecast models were trained on three different date-based train-validation splits. What proved to be of great importance for improving the overall precision of this method was the utilization of the weighted average for the product-store level of the tuned quantiles with pure statistics over the last year.
- **Fourth place (*Marisaka Mozz*; Mori Masakazu):** This method was based on two NNs that directly forecast the sales at each aggregation level of the data set. The NNs had the same architecture and consisted of an embedding layer that encoded categorical features, two LSTM layers that process time series data, and seven fully-connected layers, one for each day of the week, that modeled seasonality and generated the final forecasts. The first NN, trained using the negative log-likelihood loss function of Student's t-distribution, was specialized for predicting levels 1–9, while the other, using the weighted negative log-likelihood loss function of the negative binomial distribution, for forecasting levels 10–12. The series were scaled and processed before training using power transformation to facilitate training. Oversampling techniques were also considered for mitigating parameter uncertainty and overfitting. CV was performed by exploiting the last four or eight weeks of data.
- **Fifth place (*IHiroaki*; Hiroaki Ikeshita):** This method involved the construction of 280 LightGBM models, generating separate point forecasts for each store and day. The models were trained using a loss function that approximated the WRMSSE measure of the "Accuracy" challenge and a set of features that included past sales, averages of past sales, categorical variables, and calendar information. After producing the point forecasts, the residual errors were computed for three CV folds of 28 days each.

[6] https://lightgbm.readthedocs.io/en/latest/index.html.

These errors were then used to empirically estimate uncertainty and generate the final probabilistic forecasts.

- **Seventh place - Best student submission (*Ka Ho_STU*; Ka Ho Tsang):** This method involved the simple, equally-weighted combination of two sequence-to-sequence LSTM models; the first was trained to produce point forecasts at a product-store level, while the other at a department-store level. The probabilistic forecasts were then generated by collecting the residual errors of the two models for the past 112 days and constructing the respective i.i.d. normal distributions. The input of the models included a window of 28 days of time series data and a rolling average of past sales, categorical variables, and information about the day of the week, SNAP activities, and holidays. The first model was trained using the Adam algorithm and a custom loss function that approximated WSPL, while the second used a simple mean squared error one with no weighting. No pre-processing occurred for the product-store model, while the series were first scaled and then detrended for the department-store one.

Regarding the rest of the top 50 performing methods for which a method description was available, we should mention that almost all of them adopted similar approaches to the winning submission, involving LightGBM and XGBoost models (trained either across all series or for particular subsets/pools of them and forecasting horizons), NNs (exploiting either standard LSTM architectures or deeper, more advanced ones), and combinations of both. The only exception was probably the *WalSmart* team, ranked 6th, who employed a multi-stage, local-level state-space model, trained using the negative binomial distribution, to generate probabilistic forecasts via Monte Carlo simulations. Finally, we should note that, based on the public discussions made on Kaggle's forum, many teams tried to apply other forecasting methods, mainly statistical ones, or mix them with the above-mentioned ML methods. However, since these approaches failed to provide competitive results, the community largely ignored them, especially during the "test" phase of the competition.

4.3. Key findings

Below is a summary of the findings related to the performance of the winning methods:

Finding 1: The superiority of ML methods. This finding is aligned with that of the M5 "Accuracy" challenge, indicating that ML methods can provide better forecasts than standard, statistical approaches, both point and probabilistic ones. All winning teams in the "Uncertainty" competition employed LightGBM models, NNs, or simple combinations of those, that allowed them to effectively process numerous, correlated series and incorporate useful exogenous/explanatory variables. Moreover, although the base models used by the winning teams to produce their forecasts were similar, there were some significant variations among their final approaches in the way the uncertainty was computed, the training and optimization processes were utilized, the CV strategies were employed, and the techniques used for calibrating the final forecasts. This highlights numerous ways to produce precise probabilistic forecasts, even when a limited pool of ML methods is available, thus encouraging further research in the field. We should clarify that this finding is not drawn based on the improvements that the winning methods reported over the benchmarks. As discussed earlier, the latter may display several limitations when used for predicting intermittent demand data. Instead, it derives from the fact that none of the winning submissions and almost none of the top-performing ones employed a statistical method.

Finding 2: The value of "cross-learning": Similar to the "Accuracy" challenge, all winning methods in the "Uncertainty" competition employed "cross-learning" approaches that produced probabilistic forecasts by learning from the series of the complete data set (or subsets of it) and their inter-dependencies. Thus, it is confirmed that "cross-learning" is the dominant way of applying ML forecasting methods and that, if implemented effectively (Semenoglou et al., 2021), can result in significantly better forecasts than traditional methods that are trained in a series-by-series fashion (Spiliotis et al., 2020b). This is particularly true in applications like those in M5, where the data set consisted of many aligned and highly-correlated series, structured hierarchically.

Finding 3: The significant differences between the winning methods and benchmarks used. Unfortunately, the research done in the field of retail sales probabilistic forecasting is relatively limited compared to its point forecast counterpart (Fildes et al., 2019). Therefore, only a few well-established benchmarks could be used to compare the performance of the winning methods. Nevertheless, the M5 "Uncertainty" competition involved six simple benchmarks frequently used in sales forecasting applications. The results indicate that the winning submissions provided significantly more precise forecasts in terms of ranks when compared to these benchmarks and were also, on average, more than 20% better in terms of WSPL, also providing a better calibration. Although the differences were slight at lower aggregation levels and the tails of the distributions, the results demonstrate their superiority and motivate additional research in the area of ML forecasting methods that can be used to effectively estimate uncertainty. Equally importantly, it can be confirmed that available forecasting methods do not necessarily underestimate uncertainty, as has been the case in the past (Makridakis et al., 1987). This reaffirmed the M4 competition finding when, for the first time, the top two winning methods achieved phenomenal precision in predicting the 95% central PIs. Yet, given the limitations of the examined benchmarks, making them potentially inappropriate for predicting intermittent demand data, we should treat this finding with care and conduct further analysis using other, more suitable benchmarks to validate its extent and make concrete conclusions.

Finding 4: The value added by effective cross-validation strategies and augmentation. The main issue with ML methods is that, due to their high flexibility, they can easily overfit the data, thus producing inaccurate forecasts. This is especially true when dealing with noisy series or complex forecasting tasks, where small changes in the method's settings may result in major differences in post-sample performance. Thus, adopting effective CV strategies is critical for objectively simulating post-sample accuracy, avoiding overfitting, mitigating uncertainty, and optimizing the hyper-parameters, architecture, and input of the ML methods. As with the "Accuracy" challenge, all winning methods in the "Uncertainty" competition adopted such CV strategies and based their estimates on these strategies. Moreover, for the first time in the history of the M competitions, some teams considered augmentation techniques to increase the original size of the data set and deal with overfitting more effectively, thus enhancing the robustness of the methods utilized. Undoubtedly, identifying ways to efficiently exploit oversampling techniques in forecasting applications becomes a fruitful area for future research.

Finding 5: The importance of exogenous/explanatory variables. In retail sales forecasting applications, where demand is highly influenced by external factors like promotions and calendar effects, incorporating additional, explanatory variables in the forecasting models along with historical time series data is critical for improving overall forecasting performance (Fildes et al., 2019). This becomes evident in the results of the M5 "Uncertainty" competition where, similar to the "Accuracy" challenge, all winning teams used exogenous/explanatory variables, such as information about prices, holidays, SNAP activities, and seasonal, calendar effects, to improve the performance of their methods. Although some teams decided not to use all these variables as input to their models, all of them exploited part of the external information provided, determined mostly based on the results of the CV strategies.

Finding 6: The importance of identifying "horses for courses". Given that different series may display different features, and each forecasting method is typically designed to capture some of these (Spiliotis et al., 2020a; Theodorou et al., 2021), accurate forecasting frequently depends on the right selection of an appropriate forecasting method (Fildes & Petropoulos, 2015; Kourentzes et al., 2019). Nowadays, the "horses for courses" concept (Petropoulos et al., 2014) is well established in the literature, and several approaches have been proposed for effectively selecting or combining methods based on the particular characteristics of the predicted series. For instance, such a meta-learning approach was used by Montero-Manso et al. (2020), the runner-up of the M4 competition, who trained an XGBoost model in a "cross-learning" fashion to appropriately combine nine different forecasting methods. The M5 competition, and especially the "Uncertainty" one, reconfirms the potential value of identifying "horses for courses" which, in addition, should be extended for the case of different quantiles and cross-sectional levels. The present study results suggest that even a top-performing method is unlikely to provide the best forecasts in all cases, particularly when forecasts are required for different parts of the uncertainty distribution and various aggregation levels. Nevertheless, as explained in Section 4.1.1, further analysis would be required to determine whether the differences reported for each aggregation level and quantile are significant or not.

5. Discussion, limitations, advantages, and directions for future research

The "Uncertainty" competition was less popular than the "Accuracy" one, attracting 892 teams that entered 9,936 submissions, i.e., 1/5th of the teams and 1/10th of the submissions of the "Accuracy" challenge. This is consistent with the number of major forecasting competitions that have taken place in the past if we consider that many competitions have focused on point forecasts (Hyndman, 2020), while only three on probabilistic ones (Hong et al., 2016, 2019; Makridakis et al., 2020c). There are two possible reasons for that. The first is the enthusiasm for making predictions versus the unwillingness to consider the uncertain, threatening future, which creates anxiety and complicates decision-making. The other factor could have been the much greater effort required to compete in the "Uncertainty" challenge, requiring a total of 28 days × 9 quantiles × 42,840 series = 10,795,680 forecasts, versus 28 days × 30,490 series = 853,720 forecasts for the "Accuracy" one. In practice, both challenges are equally important. Therefore, future forecasting competitions and research should focus on producing accurate point forecasts and the precise prediction of probability distributions, stressing the latter's implications. Pretending that uncertainty does not exist is like putting one's head in the sand. On the positive side, we are pleased that the M5 "Uncertainty" competition attracted many participants while also contributing to this particular area of forecasting and highlighting the importance of uncertainty estimation, especially among the members of the ML community where probabilistic forecasting remains largely unexplored.

5.1. Discussion

What has become clear from the "Uncertainty" challenge is the superior performance of LightGBM. This tree-based method was widely used by the majority of the top-performing methods of the competition. Interestingly, the same method excelled in the "Accuracy" competition, leaving little doubt that retail firms, at least, should consider adopting ML methods to support decisions related to their inventory and supply chain management. Equally important, statistical methods, traditionally used for many years by numerous firms to estimate the uncertainty in demand, were outperformed by ML methods. Yet, the competition results indicate that these improvements may be closely related to the cross-sectional level examined and the quantile for which probabilistic forecasts are produced.

Table 3 provides a simple comparison between the performance of the Kernel method, widely used for estimating the uncertainty of intermittent demand series (Trapero et al., 2019), and that of the Naive, sNaive, SES, ETS, and ARIMA statistical benchmarks (for more information about these benchmarks, please see Appendix C

Table 3
Percentage improvements (according to WSPL) reported between the Kernel method and the rest of the statistical benchmarks of the competition. The results are reported both overall and per aggregation level. Positive numbers indicate that the examined benchmark performs better than Kernel and vice versa. Column-wise minimum values are displayed in bold.

Methods compared	Aggregation level												Average
	1 Total	2 State	3 Store	4 Category	5 Department	6 State Category	7 State Department	8 Store Category	9 Store Department	10 Product	11 Product State	12 Product Store	
Kernel vs. Naive	−29.8%	−35.5%	−27.8%	−31.6%	−25.4%	−37.1%	−33.7%	−28.8%	−28.7%	−41.4%	−47.0%	−58.0%	−34.2%
Kernel vs. sNaive	63.1%	56.5%	55.2%	58.5%	55.5%	50.5%	47.0%	49.1%	44.0%	6.6%	−2.2%	−15.6%	42.4%
Kernel vs. SES	−29.8%	5.2%	37.4%	6.2%	−7.1%	30.5%	19.7%	32.3%	32.9%	12.0%	9.2%	2.0%	12.5%
Kernel vs. ETS	**72.0%**	65.6%	65.1%	68.7%	65.1%	60.4%	**56.2%**	60.7%	55.6%	**19.8%**	**13.3%**	**3.8%**	53.5%
Kernel vs. ARIMA	68.6%	**68.3%**	**66.3%**	**69.9%**	**66.7%**	**62.3%**	55.9%	**61.8%**	**56.8%**	16.9%	12.2%	3.1%	**53.8%**

Table 4
Percentage improvements (according to WSPL) reported between the Kernel method and the rest of the statistical benchmarks of the competition. The results are reported both overall and per quantile. Positive numbers indicate that the examined benchmark performs better than Kernel and vice versa. Column-wise minimum values are displayed in bold.

Methods	Quantile									Average
compared	0.005	0.025	0.165	0.250	0.500	0.750	0.835	0.975	0.995	
Kernel vs. Naive	−44.8%	−47.6%	−8.9%	8.3%	−10.0%	−64.9%	−81.8%	−99.1%	−149.8%	−34.2%
Kernel vs. sNaive	−7.3%	11.6%	32.1%	37.9%	52.1%	49.4%	43.7%	32.0%	8.9%	42.4%
Kernel vs. SES	−19.3%	−8.7%	21.3%	30.2%	21.8%	2.2%	−3.8%	−11.2%	−48.7%	12.5%
Kernel vs. ETS	**13.1%**	30.5%	48.2%	**52.3%**	**59.7%**	57.4%	53.9%	45.8%	20.1%	53.5%
Kernel vs. ARIMA	12.5%	**32.9%**	**49.1%**	51.8%	56.9%	**57.7%**	**57.1%**	**52.8%**	**26.5%**	**53.8%**

of the supplementary material). As seen, although Kernel is on average 34.2% more accurate than the Naive benchmark, providing better forecasts across all aggregation levels, it is 42.4% worse than sNaive, a naive method accounting for seasonality. Specifically, Kernel is 63.1% less accurate than sNaive at the highest aggregation level, about 52% less accurate at levels 2 to 9, 6.6% less accurate at level 10, and 2.2% and 15.6% more accurate at levels 11 and 12, respectively. These results highlight the limitations of simple non-parametric approaches like Kernel and their inability to correctly deal with seasonality, especially at higher aggregation levels where the seasonal component becomes dominant. This becomes evident when Kernel is compared with SES, ETS, and ARIMA (ETS and ARIMA account for seasonality and trend while SES accounts for none of these features). Our results indicate that, on average, SES, ETS, and ARIMA are 12.5%, 53.5%, and 53.7% more accurate than Kernel, with SES being approximately 14% and 8% better at levels 1–9 and 10–12, respectively, while ETS and ARIMA about 64% and 12% better at levels 1–9 and 10–12, respectively. Note also that ETS and ARIMA outperform Kernel across all 12 aggregation levels.

Table 4 provides similar results to Table 3, but this time for the case of the nine quantiles considered in the competition instead of the 12 aggregation levels. As seen, Kernel provides superior forecasts than the Naive method for all quantiles except that of probability level $u_4 = 0.250$. However, its performance is worse than SES and particularly than sNaive, ETS, and ARIMA, which account for seasonality in the data. Interestingly, ETS and ARIMA outperform Kernel in all nine quantiles, with the highest improvement being observed for the median. However, these improvements significantly decrease when moving towards the distribution's tails, with ETS and ARIMA being on average 22% and 36% better than the Kernel for the left and right tail, respectively. Again, these results highlight the limitations of simple approaches but, more importantly, the difficulties present, in general, to precisely estimate uncertainty equally well for all parts of the distribution of the realized values.

Table 5 presents the WSPL scores of the winning method, *Everyday Low SPLices*, and the top-performing statistical benchmark, ARIMA, as well as the corresponding percentage improvements of the former over the latter, both overall and at each aggregation level separately. Moreover, it reports the WSPL scores achieved when (i) the "best" performing method at each aggregation level is selected (BEST), (ii) the forecasts of these eight "best" performing methods are combined using equal weights (COMB), and (iii) the forecasts of the top five performing methods of the competition are combined using equal weights (COMB-Top5). We find that the average improvement of the winning method over ARIMA is 24.6%, starting at 53.3% and 35% at levels 1 and 2, and then gradually decreasing in an almost linear way to 11.8%, 9.4%, and 12.3% at levels 10, 11, and 12, respectively. Similar conclusions are drawn for the case of the BEST approach, where overall performance is 4.4% higher than that of the winning method. However, these improvements mainly refer to the higher aggregation levels, i.e., levels 1–6, and are minor at the three lowest ones. Interestingly, on average, the COMB approach leads to an improvement of 5.9% over the winning method but, similarly to BEST, these improvements refer to the top and the middle parts of the hierarchy (levels 1–9). At the same time, they are negative at levels 10, 11, and 12. The same is true for the COMB-Top5 approach, which displays a slightly better overall performance than COMB (6.2% improvement over the winner), but worse results for levels 10, 11, and 12.

Finally, Table 6 shows the same information as Table 5, but this time for the case of the nine quantiles considered in the competition instead of the 12 aggregation levels. Observe that the winning method achieves significant improvements over the benchmark at the lower quantiles,

Table 5

Percentage improvements (according to WSPL) reported between (i) the winning team (*Everyday Low SPLices*), (ii) the best-performing team of each aggregation level, i.e. *Ouranos* for levels 8 and 9, *Marisaka Mozz* for level 5, *WalSmart* for level 12, *Takuma Omiya* for levels 1, 4, and 6, *Jacques Peeters* for level 11, *Costas Voglis* for levels 3 and 7, *shirokane_friends_from_acc* for level 2, and *slaweks* for level 10, (iii) the simple, equal weighted combination of these eight methods, and (iv) the simple, equal weighted combination of the top five performing methods of the competition over the ARIMA benchmark. Column-wise minimum values are displayed in bold.

Methods compared	Aggregation level												Average
	1 Total	2 State	3 Store	4 Category	5 Department	6 State Category	7 State Department	8 Store Category	9 Store Department	10 Product	11 Product State	12 Product Store	
ARIMA	0.158	0.148	0.163	0.147	0.167	0.170	0.202	0.178	0.201	0.322	0.298	0.302	0.205
Winning team	0.074	0.096	0.113	0.091	0.112	0.116	0.136	0.135	0.159	0.284	0.270	0.265	0.154
Best team per level (BEST)	**0.057**	0.089	0.110	0.071	0.103	0.106	0.131	0.131	0.155	**0.282**	**0.270**	**0.264**	0.147
Combination of best teams (COMB)	0.058	**0.080**	**0.102**	**0.071**	**0.090**	**0.099**	**0.119**	**0.122**	**0.146**	0.291	0.280	0.282	0.145
Combination of top five teams (COMB-Top5)	0.058	0.082	0.104	0.073	0.096	0.101	0.123	0.123	0.148	0.285	0.271	0.270	**0.145**
Improvement of winner over ARIMA	53.3%	35.0%	30.8%	37.9%	33.0%	31.9%	32.5%	24.3%	20.9%	11.8%	9.4%	12.4%	24.6%
Improvement of BEST over ARIMA	**63.8**%	39.9%	32.6%	51.4%	38.2%	37.4%	35.2%	26.3%	23.1%	**12.3%**	**9.6%**	**12.8%**	28.0%
Improvement of COMB over ARIMA	62.9%	**45.6%**	**37.4%**	**51.4%**	**46.0%**	**41.7%**	**41.2%**	**31.5%**	**27.5%**	9.4%	6.0%	6.8%	29.1%
Improvement of COMB-Top5 over ARIMA	63.1%	44.5%	36.3%	50.0%	42.4%	40.7%	39.2%	30.6%	26.3%	11.3%	9.1%	10.6%	**29.3%**

Table 6

Percentage improvements (according to WSPL) reported between (i) the winning team (*Everyday Low SPLices*), (ii) the best-performing submission of each quantile, i.e. *Everyday Low SPLices* for quantiles 0.005, 0.165, and 0.250, *GoodsForecast - Nick Mamonov* for quantile 0.500, *IHiroaki* for quantile 0.835, *Kazu* for quantile 0.025, *Tobias Tesch_STU* for quantile 0.750, and *Jacques Peeters* for quantiles 0.975 and 0.995, (iii) the simple, equal weighted combination of these six methods, and (iv) the simple, equal weighted combination of the top five performing methods of the competition over the ARIMA benchmark. Column-wise minimum values are displayed in bold.

Methods	Quantile									Average
compared	0.005	0.025	0.165	0.250	0.500	0.750	0.835	0.975	0.995	
ARIMA	0.021	0.072	0.278	0.349	0.426	0.339	0.265	0.070	0.022	0.205
Winning team	**0.011**	0.050	0.197	0.249	0.316	0.267	0.217	0.062	0.019	0.154
Best team per level (BEST)	**0.011**	**0.047**	0.197	0.249	0.299	0.267	0.216	0.057	**0.015**	0.151
Combination of best teams (COMB)	0.014	0.049	0.183	0.235	0.302	0.257	0.212	0.057	0.017	0.147
Combination of top five teams (COMB-Top5)	0.016	0.050	**0.183**	**0.232**	**0.295**	**0.250**	**0.204**	**0.055**	0.016	**0.145**
Improvement of winner over ARIMA	**46.3%**	31.2%	29.1%	28.6%	25.9%	21.1%	18.1%	11.8%	13.7%	24.6%
Improvement of BEST over ARIMA	**46.3%**	**34.5%**	29.1%	28.6%	29.8%	21.2%	18.3%	18.7%	**29.5%**	26.2%
Improvement of COMB over ARIMA	35.4%	32.4%	34.0%	32.5%	29.1%	24.1%	20.1%	18.0%	22.4%	28.0%
Improvement of COMB-Top5 over ARIMA	26.0%	31.0%	**34.1%**	**33.5%**	**30.7%**	**26.2%**	**22.8%**	**20.6%**	27.6%	**29.3%**

starting at 46.3% at quantile 0.005 and then decreasing, in a similar, linear way in Table 5, to 25.9%, 11.8%, and 13.7% at quantiles 0.500, 0.975, and 0.995, respectively. The reasons for such behavior by the winning method between the lower and higher quantiles are not obvious and need further investigation. However, we find that when the "best" method is selected at each quantile, the overall performance is increased by an additional 2% over the winning method. Moreover, although the improvements are small for the left side of the distribution, they are notable for the median and the right tail of the distribution, where an additional improvement of 5.2%, 7.9%, and 18.3% is achieved for quantiles 0.500, 0.975, and 0.995, respectively. The COMB approach leads to even greater improvements over the winning method of approximately 4.4% which, in contrast to BEST, are observed across all parts of the distribution, particularly its middle and right tail, with the only exception being quantile 0.005. The COMB-Top5 approach has a slightly better overall performance than COMB, but its effect is more visible in the middle of the distribution.

Based on the above, we conclude that "horses for courses" and simple combinations of precise methods display great potential for improving uncertainty estimation. The benefits of these approaches are expected to be greater when applied to identify best practices at a cross-sectional level, where the characteristics of the series may significantly differ. However, they are also substantial when employed to select the most precise method for each quantile, where different assumptions about how the realized values are distributed may be required. As demonstrated by the findings of the M5 "Uncertainty" competition, such selections and combinations are possible when effective CV strategies and augmentation techniques are utilized. It remains to be seen how forecasters will explore these practices in the future.

5.2. Limitations and the uniqueness of the "Uncertainty" challenge

We are not going to discuss the primary limitations of the "Uncertainty" challenge in detail as they are similar to those of the "Accuracy" one (Makridakis et al., 2020a); (i) the M5 data set included the hierarchical unit sales of a single company, Walmart, meaning that its results and findings may not apply to other companies whose sales could display different characteristics, be driven by different pricing and promotion strategies, and be organized in a different structure; (ii) improvements in uncertainty estimation were not directly linked to Walmart's underlying operational costs, thus missing how the narrower and better calibrated PIs of the winning teams affect holding costs and service levels in practice; (iii) a limited number of teams provided a description of their methods and the code required for reproducing their results, thus limiting our focus on the winning submissions and not allowing us to determine the reasons that other teams failed to perform equally well. Nevertheless, we would add to these limitations the issues related to the benchmarks selected in the "Uncertainty" challenge, which, compared to the "Accuracy" one, were more straightforward, more limited in number, and in most cases theoretically inappropriate for forecasting intermittent demand data, i.e., the series at the lowest three cross-sectional levels of the competition. Therefore, similarly to Spiliotis et al. (2021), future studies should consider more appropriate benchmarks to analyze the results of the competition further and challenge its findings.

Instead, we will emphasize the distinctiveness of the "Uncertainty" challenge and its contribution to the forecasting literature. To our knowledge, three published studies deal with uncertainty competitions. The first two are the Global Energy Forecasting Competitions (GEFComs) of 2014 (Hong et al., 2016) and 2017 (Hong et al., 2019), which required the production of load, price, wind, and solar probabilistic forecasts and hierarchical probabilistic load forecasts, respectively. The third refers to the M4 competition (Makridakis et al., 2020c) which required the prediction of the 95% central prediction intervals for a large number of diverse series. In brief, GEFCom2014 involved 15 series, attracted 581 contestants, and required the submissions of 99 quantiles submitted in a rolling fashion. GEFCom2017 involved 183 delivery point meters series (aggregated into 16 groups and two control zones), attracted 177 teams, and required the submissions of nine indicative quantiles. Finally, M4 involved 100,000 series, attracted 49 teams, and required the submissions of two quantiles. As seen, GEFComs have focused on the particular domain of energy forecasting and required the prediction of the complete uncertainty distributions. On the other hand, M4 was more generic in terms of the forecasting applications covered by its data. It included significantly more series but only focused on the tails of the uncertainty distribution. More importantly, all three competitions attracted a limited number of participants, mainly academics, researchers, and students.

The M5 "Uncertainty" challenge has become the first forecasting competition to focus on the prediction of the entire uncertainty distributions in the domain of retail sales forecasting by considering nine indicative quantiles, while also (i) attracting a great number of participants (892 teams), (ii) involving a large data set of 42,840 series that are organized hierarchically and allow for comparisons of statistical significance, and (iii) providing complete information of the methods used and the benchmarks considered for evaluating their relative performance. Furthermore, given that most M5 participants had little or no knowledge about forecasting, as they were mostly practitioners in computer science, the "Uncertainty" challenge introduced many data scientists to the area of probabilistic forecasting, which, unfortunately, remains largely unexplored by the ML community. We hope that the M5 will promote the cross-pollination between the ML and statistics communities, allow them to learn from each other, and facilitate their communication to advance the theory and practice of forecasting (Makridakis et al., 2020b).

5.3. Directions for future research

The directions for future research are similar to those described for the "Accuracy" challenge, i.e. (i) focusing on the further exploration and development of ML forecasting methods, (ii) determining ways for efficiently running such methods in everyday operations, (iii) replicating the findings of the M5 for other retail sales data sets, (iv) exploiting "cross-learning" and "horses-for-courses", and (v) investing in powerful CV and oversampling strategies to mitigate overfitting. Therefore, we will not discuss them further in this section but note that the research done in probabilistic forecasting is still limited, requiring new benchmarks and methods for estimating and evaluating uncertainty precisely.

However, we will emphasize the need for future research to sensitize academics and educate practitioners on the usefulness of uncertainty and the need to comprehend its risk implications and take actions to be prepared to face such risks (Taleb et al., 2020). Unfortunately, practitioners expect forecasting to reduce future uncertainty by providing accurate predictions like those in hard sciences. However, this is a great misconception. A major purpose of forecasting is not to reduce uncertainty but reveal its full extent and implications by estimating it as precisely as possible, i.e., measuring past volatility and assuming that future uncertainty will be similar to the past, as long as patterns and relationships remain constant. The challenge for the forecasting field is how to persuade practitioners of the reality that all forecasts are uncertain and that this uncertainty cannot be ignored, as doing so could lead to catastrophic consequences. Moreover, it aims to illustrate that uncertainty does not always behave in a usual manner. Instead, as the example of COVID-19 illustrates, it can create huge, fat-tailed uncertainty with catastrophic consequences (Pinson & Makridakis, 2020).

6. Conclusions

The main conclusions of the "Accuracy" competition also hold for the "Uncertainty" one. Therefore, we will limit ourselves to emphasize the promising performance of ML methods, particularly the LightGBM that was used by the great majority of the top 50 contestants, and discuss its implications for the theory and practice of forecasting.

This conclusion demonstrates the potential advantage of ML methods over statistical ones and the profound implications for the future of forecasting. Will these results lead to the demise of forecasters as we know them and the ascent of data scientists who take their place? The top winner of the "Accuracy" challenge was an undergraduate student with little forecasting knowledge and no experience in the domain. Yet, he effectively managed to win the competition and outperform his 7,091 competitors, including experienced Kaggle grandmasters, among others. On the other hand, the top winners of the "Uncertainty" challenge had a strong background in statistics and forecasting, and were also Kaggle masters and grandmasters.

It seems that as the performance of ML methods and the access to efficient, ready-to-use ML libraries improves, the value of knowledge and experience will become less important for developing accurate forecasting models that rely on unstructured, agnostic algorithms and require few human inputs. Thus, it could be the case that the role of the next generation of forecasters would be to introduce innovative ML methods and expand or improve existing ones to boost their intuitiveness, allow for the precise estimation of uncertainty, and enable their applicability in different forecasting domains and diverse types of data. Then, data scientists of firms will be able to implement these methods quickly and use their knowledge, experience, and programming skills to effectively adjust their settings based on the particularities of the forecasting task and put them into production. What is certain is that forecasting is experiencing significant changes and that recent advances in ML and computer science, in general, cannot be overlooked by academics and practitioners.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ijforecast.2021.10.009.

References

Barrow, D. K., & Kourentzes, N. (2016). Distributions of forecasting errors of forecast combinations: Implications for inventory management. *International Journal of Production Economics*, *177*, 24–33. http://dx.doi.org/10.1016/j.ijpe.2016.03.017.

Bojer, C. S., & Meldgaard, J. P. (2021). Kaggle forecasting competitions: An overlooked learning opportunity. *International Journal of Forecasting*, *37*(2), 587–603. http://dx.doi.org/10.1016/j.ijforecast.2020.07.007.

Fildes, R., Ma, S., & Kolassa, S. (2019). Retail forecasting: Research and practice. *International Journal of Forecasting*, http://dx.doi.org/10.1016/j.ijforecast.2019.06.004, in press.

Fildes, R., & Petropoulos, F. (2015). Simple versus complex selection rules for forecasting many time series. *Journal of Business Research*, *68*(8), 1692–1701. http://dx.doi.org/10.1016/j.jbusres.2015.03.028.

Fisher, M., & Raman, A. (1996). Reducing the cost of demand uncertainty through accurate response to early sales. *Operations Research*, *44*(1), 87–99.

Gaba, A., Popescu, D. G., & Chen, Z. (2019). Assessing uncertainty from point forecasts. *Management Science*, *65*(1), 90–106.

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

Gardner, E. S. (2006). Exponential smoothing: The state of the art-Part II. *International Journal of Forecasting*, *22*(4), 637–666. http://dx.doi.org/10.1016/j.ijforecast.2006.03.005.

Gardner Jr., E. S. (1985). Exponential smoothing: The state of the art. *Journal of Forecasting*, *4*(1), 1–28. http://dx.doi.org/10.1002/for.3980040103.

Gneiting, T. (2008). Editorial: Probabilistic forecasting. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *171*(2), 319–321. http://dx.doi.org/10.1111/j.1467-985X.2007.00522.x.

Gneiting, T. (2011). Quantiles as optimal point forecasts. *International Journal of Forecasting*, *27*(2), 197–207. http://dx.doi.org/10.1016/j.ijforecast.2009.12.015.

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, *102*(477), 359–378. http://dx.doi.org/10.1198/016214506000001437.

Grushka-Cockayne, Y., Jose, V. R. R., & Lichtendahl Jr, K. C. (2017). Ensembles of overfit and overconfident forecasts. *Management Science*, *63*(4), 1110–1130.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. http://dx.doi.org/10.1162/neco.1997.9.8.1735.

Hong, T., Pinson, P., Fan, S., Zareipour, H., Troccoli, A., & Hyndman, R. J. (2016). Probabilistic energy forecasting: Global energy forecasting competition 2014 and beyond. *International Journal of Forecasting*, *32*(3), 896–913. http://dx.doi.org/10.1016/j.ijforecast.2016.02.001.

Hong, T., Xie, J., & Black, J. (2019). Global energy forecasting competition 2017: Hierarchical probabilistic load forecasting. *International Journal of Forecasting*, *35*(4), 1389–1399. http://dx.doi.org/10.1016/j.ijforecast.2019.02.006.

Hyndman, R. J. (2020). A brief history of forecasting competitions. *International Journal of Forecasting*, *36*(1), 7–14. http://dx.doi.org/10.1016/j.ijforecast.2019.03.015.

Hyndman, R. J., Ahmed, R. A., Athanasopoulos, G., & Shang, H. L. (2011). Optimal combination forecasts for hierarchical time series. *Computational Statistics & Data Analysis*, *55*(9), 2579–2589. http://dx.doi.org/10.1016/j.csda.2011.03.006.

Hyndman, R., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., Petropoulos, F., Razbash, S., Wang, E., & Yasmeen, F. (2020). Forecast: Forecasting functions for time series and linear models. R package version 8.12.

Hyndman, R., & Khandakar, Y. (2008). Automatic time series forecasting: the forecast package for R. *Journal of Statistical Software*, *26*(3), 1–22.

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, *22*(4), 679–688. http://dx.doi.org/10.1016/j.ijforecast.2006.03.001.

Hyndman, R. J., Koehler, A. B., Snyder, R. D., & Grose, S. (2002). A state space framework for automatic forecasting using exponential smoothing methods. *International Journal of Forecasting*, *18*(3), 439–454. http://dx.doi.org/10.1016/S0169-2070(01)00110-8.

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

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems 30* (pp. 3146–3154). Curran Associates, Inc..

Kolassa, S. (2016). Evaluating predictive count data distributions in retail sales forecasting. *International Journal of Forecasting*, *32*(3), 788–803. http://dx.doi.org/10.1016/j.ijforecast.2015.12.004.

Koning, A. J., Franses, P. H., Hibon, M., & Stekler, H. (2005). The M3 competition: Statistical tests of the results. *International Journal of Forecasting*, *21*(3), 397–409. http://dx.doi.org/10.1016/j.ijforecast.2004.10.003.

Kourentzes, N., Barrow, D., & Petropoulos, F. (2019). Another look at forecast selection and combination: Evidence from forecast pooling. *International Journal of Production Economics*, *209*, 226–235. http://dx.doi.org/10.1016/j.ijpe.2018.05.019.

Makridakis, S., Assimakopoulos, V., & Spiliotis, E. (2018). Objectivity, reproducibility and replicability in forecasting research. *International Journal of Forecasting*, *34*(4), 835–838. http://dx.doi.org/10.1016/j.ijforecast.2018.05.001.

Makridakis, S., Hibon, M., Lusk, E., & Belhadjali, M. (1987). Confidence intervals: An empirical investigation of the series in the M-competition. *International Journal of Forecasting*, *3*(3), 489–508. http://dx.doi.org/10.1016/0169-2070(87)90045-8.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). Statistical and machine learning forecasting methods: Concerns and ways forward. *PLOS ONE*, *13*(3), 1–26. http://dx.doi.org/10.1371/journal.pone.0194889.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2020). The M5 accuracy competition: Results, findings and conclusions. Working paper, available at: https://drive.google.com/drive/u/1/folders/1S6IaHDohF4qalWsx9AABsIG1filB5V0q.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2020). Responses to discussions and commentaries. *International Journal of Forecasting*, *36*(1), 217–223. http://dx.doi.org/10.1016/j.ijforecast.2019.05.002.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2020). The M4 competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, *36*(1), 54–74. http://dx.doi.org/10.1016/j.ijforecast.2019.04.014.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2021). The M5 competition: Background, organization, and implementation. *International Journal of Forecasting*, http://dx.doi.org/10.1016/j.ijforecast.2021.07.007, in press.

Montero-Manso, P., Athanasopoulos, G., Hyndman, R. J., & Talagala, T. S. (2020). FFORMA: Feature-based forecast model averaging. *International Journal of Forecasting*, *36*(1), 86–92. http://dx.doi.org/10.1016/j.ijforecast.2019.02.011.

Panagiotelis, A., Gamakumara, P., Athanasopoulos, G., & Hyndman, R. J. (2020). Probabilistic forecast reconciliation: properties, evaluation and score optimisation. In *Working paper*.

Petropoulos, F., Makridakis, S., Assimakopoulos, V., & Nikolopoulos, K. (2014). Horses for Courses in demand forecasting. *European Journal of Operational Research*, *237*(1), 152–163. http://dx.doi.org/10.1016/j.ejor.2014.02.036.

Pinson, P., & Makridakis, S. (2020). Pandemics and forecasting: The way forward through the Taleb-Ioannidis debate. *International Journal of Forecasting*, http://dx.doi.org/10.1016/j.ijforecast.2020.08.007, in press.

Rostami-Tabar, B., Babai, M. Z., Syntetos, A., & Ducq, Y. (2013). Demand forecasting by temporal aggregation. *Naval Research Logistics*, *60*(6), 479–498. http://dx.doi.org/10.1002/nav.21546.

Semenoglou, A.-A., Spiliotis, E., Makridakis, S., & Assimakopoulos, V. (2021). Investigating the accuracy of cross-learning time series forecasting methods. *International Journal of Forecasting*, *37*(3), 1072–1084. http://dx.doi.org/10.1016/j.ijforecast.2020.11.009.

Smyl, S. (2020). A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting. *International Journal of Forecasting*, *36*(1), 75–85. http://dx.doi.org/10.1016/j.ijforecast.2019.03.017.

Spiliotis, E., Kouloumos, A., Assimakopoulos, V., & Makridakis, S. (2020). Are forecasting competitions data representative of the reality? *International Journal of Forecasting*, *36*(1), 37–53. http://dx.doi.org/10.1016/j.ijforecast.2018.12.007.

Spiliotis, E., Makridakis, S., Kaltsounis, A., & Assimakopoulos, V. (2021). Product sales probabilistic forecasting: An empirical evaluation using the M5 competition data. *International Journal of Production Economics*, *240*, Article 108237. http://dx.doi.org/10.1016/j.ijpe.2021.108237.

Spiliotis, E., Makridakis, S., Semenoglou, A.-A., & Assimakopoulos, V. (2020). Comparison of statistical and machine learning methods for daily sku demand forecasting. *Operational Research: An International Journal*, http://dx.doi.org/10.1007/s12351-020-00605-2, in press.

Svetunkov, I., & Petropoulos, F. (2018). Old dog, new tricks: a modelling view of simple moving averages. *International Journal of Productions Research*, *56*(18), 6034–6047.

Syntetos, A. A., & Boylan, J. E. (2005). The accuracy of intermittent demand estimates. *International Journal of Forecasting*, *21*(2), 303–314. http://dx.doi.org/10.1016/j.ijforecast.2004.10.001.

Taleb, N. N., Bar-Yam, Y., & Cirillo, P. (2020). On single point forecasts for fat-tailed variables. *International Journal of Forecasting*, http://dx.doi.org/10.1016/j.ijforecast.2020.08.008, in press.

Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: an analysis and review. *International Journal of Forecasting*, *16*(4), 437–450. http://dx.doi.org/10.1016/S0169-2070(00)00065-0.

Theodorou, E., Wang, S., Kang, Y., Spiliotis, E., Makridakis, S., & Assimakopoulos, V. (2021). Exploring the representativeness of the M5 competition data. *International Journal of Forecasting*, http://dx.doi.org/10.1016/j.ijforecast.2021.07.006, in press.

Trapero, J. R., Cards, M., & Kourentzes, N. (2019). Quantile forecast optimal combination to enhance safety stock estimation. *International Journal of Forecasting*, *35*(1), 239–250. http://dx.doi.org/10.1016/j.ijforecast.2018.05.009.

Winkler, R. L. (1996). Scoring rules and the evaluation of probabilities. *Test*, *5*(1), 1–60.

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