International Journal of Forecasting 36 (2020) 217–223

Contents lists available at ScienceDirect

International Journal of Forecasting

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

Discussion

Responses to discussions and commentaries

Spyros Makridakis *, Evangelos Spiliotis, Vassilios Assimakopoulos

University of Nicosia, Cyprus

All of the discussions and commentaries were extremely useful, providing us with valuable feedback about the M4 Competition while also proposing interesting ways of improving future forecasting competitions. We would therefore like to thank all of the contributors for taking the time to prepare their discussions and commentaries, as well as for sharing their insightful remarks. Having been in the academic world for a long time, we were expecting a great deal of criticism and disagreement when designing M4. Therefore, we were pleasantly surprised that most of the papers praised the M competitions, and the M4 in particular, expressing few concerns overall, but those few extremely useful.

Note that the invited papers and commentaries were written by academicians, researchers and practitioners who were asked to express their views on the M Competitions in general and the M4 in particular, including possible problems and concerns that should be taken into consideration in order to improve future forecasting competitions. Most of the invited papers/commentaries focus on the special issue of the M4 Competition, but others expand their analysis to additional dimensions.

Our responses to the discussions and commentaries are organized into six categories, corresponding to the five major areas of criticism/discussion and a sixth one containing various other points of concern. The responses have been arranged according to the points that we perceived to be most important. We then conclude our response by answering the hypothetical question, "*what would we have done differently if we had known of these disagreements beforehand?*", to demonstrate how such criticisms could have affected our original choices about the M4 Competition.

1. Is the M4 data set representative of reality?

This is a hard question to answer and the commentators are right to raise it. In fact, we considered this question in depth even before announcing M4, and presented our thoughts and conclusions in a recent paper written before the end of the M4 (Spiliotis, Kouloumos, Assimakopoulos, & Makridakis, 2020). Clearly, the data selected for M4 may not be representative of meteorological, hydrological, energy or web traffic forecasting, but they do cover five major domains of business forecasting (micro, industry, macro, finance and demographic), and also reflect six different frequencies (yearly, quarterly, monthly, weekly, daily and hourly) that are meaningful for such applications. Thus, as we will argue below, the large number of series included in each domain/frequency, as well as their diversity, makes them representative of what we can consider as "business reality", and therefore useful for supporting relevant decisions about the forecasting accuracy related to such decisions.

We should begin by making it clear that the purpose of the M competitions is not to determine the optimal forecasting solution for a specific forecasting situation. For instance, the best-performing method in M4, determined based on its average accuracy across 100,000 different and mostly uncorrelated series, may not be the optimal choice for predicting hourly or minute-by-minute web traffic. This explains why the top-performing participants for each domain and frequency were all invited to describe their methods for inclusion in this special issue. Each forecasting situation has its own special characteristics and particular needs, meaning that no forecasting competition can cover all such specific characteristics/needs adequately. However, the M competitions can be effective in determining which alternatives are more likely to work best for the situation considered, based on what worked well for similar series that were included in the competition's rich and diverse dataset. As a consequence,

* Corresponding author.
E-mail address: makridakis.s@unic.ac.cy (S. Makridakis).

https://doi.org/10.1016/j.ijforecast.2019.05.002

we believe that the correct way of exploiting the results of the M competitions is to determine which series have the same characteristics as the ones being considered for some specific forecasting situation and decide which methods to utilize based on its performance on this particular subset of series.

As the commentators correctly pointed out, the M4 data are not representative of the weekly and hourly frequencies, as there are not enough data in these two particular categories to be able to generalize to all weekly/hourly ones. Similarly, the M4 series are likely to differ from those of tourism, weather and transport, as, by its design, M4 includes multiple domains and frequencies in order to be representative of the "business reality". Clearly, a dataset that is truly representative of reality must, by definition, differ from specific subcategories of data, while including complete data from as many categories as possible that represent the preferred reality. For example, the "other" data in M4 include retail, transport, energy, and weather series. Thus, this subset might not be representative of the above-mentioned categories globally, but locally it may provide useful information about which forecasting approach might work well for each case. The same applies to the comment made by Fry and Brundage (2020) that the M4 data are generally "more neat and tidy" than the data used in the 2017 Kaggle competition on Web Traffic Time Series Forecasting (WTTSF). This is true, but a careful examination of the two datasets shows that the WTTSF data are practically just a subset of the M4 data (Fry & Brundage, 2020, Figure 2): series similar to WTTSF can be found in M4, just in different proportions. Thus, identifying the most accurate methods for the "less tidy" data of M4 could also help us to determine what might possibly work well for the WTTSF series.

Another way of investigating the question of whether forecasting competition data represent reality is to provide evidence that reality is characterized by characteristics that are significantly different from those in their datasets. "*The main problem with providing such evidence is that obtaining a complete picture of the 'real world' is impossible in practice, due to its unlimited applications and types of data involved*" (Spiliotis et al., 2020, p. 3). At the same time, the theoretical work of Kang, Hyndman, and Smith-Miles (2017) proposed a method for visualizing time series data collections in a two-dimensional instance space, and by so doing, demonstrated that the M3 data were not distributed evenly over the entire range of the instance space. However, this in no way indicated whether or not the M3 data truly represented reality, since there is absolutely no reason to expect real data to be distributed evenly over such a space. Similarly, although Spiliotis et al. (2020) concluded that the data from the M3 and M4 competitions were quite similar in principle, this does not necessarily mean that both, or neither, of these represent reality adequately. However, it is evident that M4 covers a much wider range of data, offering more opportunities for forecasters to explore forecasting accuracy issues because of its large number of diverse time series. Thus, utilizing the M4 dataset over those of M3 or M1 is certainly an improvement and a step in the right direction, providing more useful information for the empirical evaluation of the available forecasting approaches.

Another concern was the distribution of the series across the various domains and frequencies. The M competitions started with data taken from the business and economic environment, and we decided to remain consistent with M4. In the business world, one of the major forecasting activities is preparing budgets. The process starts in September or October of each year and covers the entire following one. These monthly forecasts are then aggregated to quarterly ones for stock market and other purposes, and then re-aggregated again to yearly ones for long-term strategic planning purposes (hence our judgmental decision to have many more monthly series, followed by quarterly and yearly ones). However, we do accept that times have changed and, as we have been told repeatedly, the dataset's structure should be modified so to include more high frequency series (daily, hourly or even sub-hourly). We would therefore consider incorporating more high frequency series in future forecasting competitions, as well as including new domains of data to address, such as web and social media applications, among others.

2. Are the error measures that are used to evaluate forecasting accuracy the most appropriate ones?

This is a persistent concern that has been raised in all previous M competitions. It seems that forecasters all have their own favorite measures and are therefore arguing in support of their use. We completely understand their concerns, as some accuracy measures have been proven to display better mathematical/statistical properties, while others are more intuitive and practical to use.

However, before arguing about the measures selected for M4, we must first question whether the main conclusions of the competition would have been any different if alternative measures had been used. It is true that different measures might have resulted in different rankings, but the large number of series included in the M4 dataset means that such differences would have been negligible for all practical purposes, and we do not believe that they would have affected the main findings of the study significantly. For instance, the Spearman's rank correlation coefficient of the two measures utilized, the symmetric mean absolute error (sMAPE) and the mean absolute scaled error (MASE), is 0.90, indicating a close relationship between the rankings of these two measures, despite them being very different in nature. Moreover, the top six methods, which were found to perform significantly better than the rest of the submissions, would still be sharing the same range of ranks, regardless of the measure used. The same is true for the two winning submissions, which displayed the most accurate forecasts in terms of both MASE and sMAPE.

Another alternative would have been to consider more measures, such as medians and geometric means. That was done for M1, which included numerous error metrics, but this added little (or no) extra value to its conclusions. We believe that the size and dimensions of the

M4 Competition would have made such an alternative completely impractical, complicating the interpretation of the results while making determining the winner more problematic. In our view, all error measures have various clear advantages and drawbacks that must be weighed properly in order to decide which is the most appropriate for assessing the participating methods within a forecasting competition's context. We made our choice with a full awareness of such drawbacks, but realizing that some of them could not be avoided completely and that there would be objections to whatever error measures we selected.

For the sake of brevity, we will not present here the reasons behind our selection of the M4 accuracy measures, i.e., the overall weighted average (OWA) of the sMAPE and MASE. These are described analytically in the main M4 paper (Makridakis, Spiliotis, & Assimakopoulos, 2020). Instead, we will merely emphasize that M4 was a completely open competition, with both the time series data and the submitted forecasts being available publicly, meaning that anyone who wishes to test alternative error measures and challenge the findings of the competition can do so easily. Interestingly, although some of the disadvantages of the selected measures were pointed out, few suggestions were actually made about candidates that would have been more appropriate. We therefore encourage relevant studies to test the implications of alternative error measures empirically in order to understand how the results are affected and to arrive at some consensus about what measures should be used in future competitions.

There were two additional points of criticism regarding this topic. The first was that, from a theoretical point of view, it makes no sense to combine two error measures, as models can be optimized to minimize either one or the other, but not both simultaneously. This is a valid criticism; however, the combination of the two measures results in a clearly defined cost function that can be minimized if necessary. In fact, this type of minimization of the OWA was utilized by many participants, including the winning submission and at least two additional ones from those ranked among the top five positions. This proves what Kolassa (2020) correctly stated, namely that it makes more sense to optimize a model based on the measure that will actually be used for its evaluation, as it will lead to more accurate results. At the same time, forecasters did not have to submit two different forecasts for the two different measures used for determining the OWA, something that would have been an interesting exercise, but clearly impractical for the competition, adding additional work and discouraging some forecasters from participating.

The second point of criticism was the lack of inclusion of multiple prediction intervals, or even full predictive densities. We agree that such an experimental design could have resulted in additional valuable insights, but only given two assumptions, both of which would be challenging to fulfill in practice: (i) the participants were willing to perform such a time-intensive and complex task, and (ii) the true future distribution of the data was known. Thus, we believe that exhaustively evaluating density forecasts and analyzing the reasons behind the success or failure of each are complex, time-consuming tasks that should be considered carefully in future forecasting competitions in order to ensure their successful implementation if truly committed participants can be found. Grushka-Cockayne and Jose (2020) provided such a preliminary assessment and highlighted some possible issues related to the estimation of prediction intervals that require further study, such as the lack of calibration, while also proposing some interesting solutions.

3. Are machine learning (ML) methods less accurate than statistical ones?

Before answering this question, we must clarify that our conclusion regarding the insufficiency of ML methods was extracted based on the accuracies reported by the methods that were self-classified as "pure" ML. Moreover, we must emphasize that we did not claim that ML methods are poor choices for extrapolating any possible type of series (this is also relevant to the first response of this paper): ML approaches can indeed provide accurate forecasts in applications where numerous, correlated and high-frequency data are available. However, the M4 Competition does not examine such cases, instead focusing mostly on business series, typically with limited sample sizes, that are also characterized by non-stationarity. The extrapolation of such series is a completely different task from typical ML applications, meaning that different tools, methods and pre-processing techniques are needed in order for ML methods to be implemented effectively. We consider that this is exactly where the M4 Competition contributed the most: it helped us to identify some special ML mechanisms that can make a huge difference in forecasting and indicated how to apply them effectively in practice. For instance, the winning submission of Smyl revealed an efficient pre-processing scheme that can be used to deal with non-stationary time series capturing, among others, both global and local components of the series that can be utilized to extract and combine information either from a single series, or at the level of multiple ones. Similarly, the second-best method of Montero-Manso and colleagues revealed the importance of "cross-learning" and demonstrated how useful information can be extracted from multiple series even if, seemingly, these are not directly correlated.

4. Is it possible to draw a distinction between statistical and ML methods?

This was a common concern among the commentators, whose criticisms ranged from *"both terms are ill defined"* to *"the distinction is probably of a tribal nature"*. Their major concern was that such a distinction *"limits the insights into the appropriateness and effectiveness of forecasting methods"*, while restricting the benefits *"from cross-pollination between the ML and statistics communities"*. We agree that the distinction made in M4 was unfortunate and that, though there are "*distinct ML and statistics communities in forecasting"*, they must break their silos and learn how to communicate effectively with each

other. Therefore, the important point now is for the wider forecasting community, working both on ML and statistics, to agree on some classifications that are meaningful and can help us to convey the reality of forecasting more appropriately. Undoubtedly, the achievement of this goal will require both communities to cooperate and exchange fruitful ideas, merging their theories, tools, algorithms and methods, so that the accuracy and applicability of the two major approaches to forecasting can be improved in the future.

For example, Januschowski et al. (2020) mention that the terms "statistics" and "ML" were being used as far back as 2001, when there were attempts to establish a clear distinction between the two terms. However, the fact that no agreement has yet emerged attests to the difficulty of doing so and shows that it may take some time before a common language is accepted. At the same time, it is clear that, at their core, both statistics and ML methods are extrapolation techniques, focusing on predictive modeling and the common task framework (CTF), with the aim of evaluating forecasting methods objectively. Thus, Januschowski and colleagues discuss several objective and subjective dimensions according to which such methods can be meaningfully classified and applied.

Having read the proposals of the commentators, it seems that the term "model driven" could be more appropriate for describing statistical methods and the term "data driven" for ML ones (Januschowski et al., 2020), respectively indicating that the data generating process is defined a priori and learned from the data. Alternatively, the forecasting community could consider the terms "structured" and "unstructured", as proposed by Barker (2020), which is very similar in nature to the classification described above. Another very interesting classification is the distinction between "local" and "global" models (Fry & Brundage, 2020), as the results of M4 clearly show that combining information from multiple series, while also considering the particularities of the individual ones, is the most promising solution for developing a highly accurate forecasting method.

5. Is the decomposition method that is used to deseasonalize the data and reseasonalize the forecasts the most suitable one?

This is a valid question with two answers. First, the three previous M Competitions used a similar approach for seasonally adjusting the data and implementing the forecasting methods that were unable to model seasonality directly, i.e., that do not incorporate a seasonal component. In fact, this is considered standard practice in time series forecasting and is used widely in the literature. Second, the data in the M1 Competition were deseasonalized/reseasonalized using two versions of the Census II decomposition method in addition to the approach employed in M4 (Makridakis, Andersen, Carbone, Fildes, Hibon, et al., 1982). The comparison showed that neither of these Census II alternatives produced forecasts that were more accurate than the simple decomposition approach used at that time, and later in M2, M3 and M4. However, the commentator is correct in suggesting that the adjustment process should be described analytically so that interested researchers can carry out such analyses on other datasets. Such a description is actually provided in Appendix E of the main M4 paper (Makridakis et al., 2020), indicating when seasonal adjustments are considered and how classical multiplicative decomposition is applied through the *decompose()* function of the *stats* package for R, an open and well-documented solution for anyone interested in its utilization. Moreover, the code for implementing the M4 benchmarks, including the seasonally adjusted ones, is available from the M4 GitHub repository (www.github.com/M4Competition). We also agree with the commentator that it would be interesting to compare the forecasts that were provided originally by ETS and ARIMA models with those produced by the same models using seasonally adjusted data as an input, in order to allow us to determine whether there are any significant differences in forecasting accuracy between the two alternatives. Hopefully, researchers will follow up and implement these leads.

6. Miscellaneous concerns

i. One commentator pointed out that two aspects of the competition were not made clear to all participants. The first refers to the submission of multiple-step-ahead forecasts instead of one-step-ahead, and the second to the utilization of multivariate time series analyses for selecting the most appropriate forecasting methods for various series. We would like to emphasize that the rules of the competition, as well as the guidelines, were posted well in advance of its launch, and no one received any inside or private information in addition to the general public announcement put on the M4 website (https://www.mcompetitions.unic.ac.cy/deadlines-evaluation/). The competition's guide explains clearly that different multiple-step-ahead forecasts had to be submitted depending on the frequency of the series, and it was not stated anywhere that participants were not allowed to use multivariate time series approaches. On the contrary, the participants were encouraged to use whichever forecasting method they regarded as the most appropriate, and many of them decided to apply such multivariate methods.

ii. A point raised by one commentator was that "*the M competitions convey the message that real forecasting problems deserve very simple techniques*". In our view, this message was true with the first three competitions but not with the M4, where several sophisticated methods were found to be considerably more accurate than the benchmarks. This conclusion is summarized in the fifth finding of the M4 study, namely that "m*ore complex methods can possibly lead to a greater forecasting accuracy*". Moreover, we strongly believe that, as Gilliland (2020) correctly points out, these improvements must be not only significant, but also meaningful in terms of computational requirements. This means that, in order for a forecasting method to be practical

for large-scale implementations, its forecasts must be both quick to compute and accurate. Otherwise, much simpler and more affordable solutions, such as a simple combination of statistical methods or simple modifications of standard forecasting methods (see for instance the submission of Legaki and Koutsouri), could be a more efficient choice. The importance of this tradeoff is emphasized in the main M4 paper (Makridakis et al., 2020), where the forecasting performances of all of the replicable methods are compared to their computational requirements and relevant conclusions are drawn. To summarize, our belief is that complex methods should only be used for important, high-value forecasts when the potential benefits outperform the computational cost, implying the advisability of using simple methods for less important items.

iii. One commentator wondered whether all 100,000 time series were necessary for conducting this competition, or if a smaller number would have sufficed. Our rationale for choosing such a large number was twofold: first, because we wanted the M4 data to be as representative of reality for different domains/frequencies as possible, and second, in order to achieve results that were more likely to be statistically significant. On the opposite end, we saw no negatives to considering such a large dataset, apart from increasing the computational time required for obtaining the forecasts and, possibly, discouraging the participation of some forecasters who were not willing to commit to such a challenging task. More importantly, we should mention that the size of the dataset was probably what enabled the two top-performing methods to achieve such remarkable accuracies, by allowing their "cross-leaning" algorithms to learn effectively from multiple series. ML methods depend strongly on data availability and diversity, and M4 was the first forecasting competition to fulfill both requirements.

iv. As was pointed out by Onkal (2020), judgment was one element that was missing from the M4 Competition. Undoubtedly it was incorporated indirectly through model selection and parameterization, but it was not applicable in its pure form of adjustments. This was due mainly to the anonymous series used in the competition and their formidable number. We share the commentator's point of view, and do believe that judgment is fundamental in the forecasting process, especially in micro- and macroeconomic applications. Thus, we look forward to incorporating such a feature in a future real-time competition that will consist of fewer, known series, accompanied with contextual information and explanatory variables. However, it must be understood that the goal of M4 was not to evaluate the value added by judgment or by competitors' feedback, but rather to identify the best-performing methods across a large and diverse set of series. The two tasks are disparate in nature and, as a result, are very difficult to conduct simultaneously under the same competition's design. For example, the inclusion of explanatory variables would have required us to run the competition on a real-time basis in order for the participants to use the available information that might influence the future direction of the time series. This is exactly why M2, which investigated this aspect of forecasting, had a completely different format from M1, M3 and M4.

v. Many of the commentators discussed the superior performance of the combinations submitted in the competition, noting that M4 confirmed, at a much larger scale, the relevant conclusions of previous forecasting competitions and other empirical studies. Moreover, some of them provided explanations for the success of combining, listing, explaining and demonstrating the conditions under which forecast combinations would be expected to provide more accurate results than other alternatives. For instance, Lichtendahl and Winkler (2020) stress the importance of utilizing accurate, diverse, uncorrelated and robust methods within combination schemes in order to achieve considerable forecasting improvements. Therefore, we would like to underline once again that, like all forecasting methods, combinations do not always guarantee superior forecasts, and that other approaches, such as individual or aggregate selection, may also prove to be very effective under special circumstances. This conclusion is linked directly to the first response of this paper, arguing that data particularities are the factors that have the most influence on which method will work best in each case. No method is always superior to others, and all methods and forecasting approaches can possibly provide valuable results in some special applications. However, what became clear in M4 is that combination is probably the most effective and computationally efficient way of implementing a generalized forecasting algorithm when forecasters are dealing with numerous and diverse data that are not characterized by the same features. At the same time, as was discussed by Fry and Brundage (2020) and other commentators, smarter combination frameworks must be introduced rather than simple ensembles in order for a combination to make a real difference.

vi. One commentator mentioned that time series are often hierarchical and that M4 fails to capture such a reality. We agree with this comment, but note that learning how best to deal with hierarchical data was not one of the goals of the present competition, which instead contained numerous, mostly uncorrelated series. Moreover, we believe that the examination of such a reality would have required a completely different dataset and design, introducing many technical difficulties. We appreciate that many forecasters usually deal with hierarchies and we agree that conducting a large-scale competition like M4 with such data in the future would be meaningful and helpful for researchers and practitioners working in that field.

vii. Another commentator asked why forecasting competitions have been neglected outside the IIF/IJF community. IIF's community is smaller than those of many other professional associations, and possesses fewer resources to promote itself and its journals. Moreover, a considerable number of academics in forecasting are theoreticians who show little interest in empirical studies, and IJF, IIF's major journal, is devoted mostly to academics with a strong statistical background. What is surprising is that an empirical study that the authors published recently in PLOS ONE, a popular open-source journal (Makridakis, Spiliotis, & Assimakopoulos, 2018), comparing statistical and machine learning methods using a subsample of the M3 data, achieved more than 89,000 views (including more than 28,900 downloads) in a little more than a year. This proves that the forecasting community is extremely large, particularly now that it includes the field of data science, and indicates that forecasters will have to expand their communication channels in order to share their results effectively with this larger community. Hence, the problem could just be that forecasting competitions and empirical forecasting studies are not promoted properly. By way of comparison, note that the four M Competitions and their predecessor, the Makridakis and Hibon (1979) study, have received more than 3600 citations in total, which is a decent number, but more than 23 times smaller than the number of views of our PLOS ONE paper mentioned above. Our view is that more work needs to be done to promote forecasting competitions in the future, and we strongly believe that this special issue, including all these papers regarding the M4 Competition, will contribute significantly in that direction.

viii. Gilliland (2020) pointed out that the improvements that were achieved by the top-performing methods of the M4 Competition, in terms of both point forecasts and prediction intervals, are meaningful, and are large enough to attract additional applications that will open the forecasting market by persuading new users of the benefits of formalized forecasting. His view is that there are great opportunities that software vendors should exploit in order to provide realistic estimates of uncertainty, particularly in the underdeveloped area of estimating the uncertainty and specifying prediction intervals, while overcoming the resistance of users who prefer overly narrow intervals that do not include the real value over realistic ones which decision makers consider uninformative. We completely agree with Gilliland that there is a lot of work required in educating decision makers that the usual practice of underestimating uncertainty considerably must end, along with the other "shocking" disappointments of real-life business forecasting that he mentioned, such as the fact that typically 52% of business forecasts are less accurate than the Naïve method. Our belief is that such practices can be mitigated only through a substantial educational effort that emphasizes the benefits of accurate forecasting and of the precise estimation of uncertainty. Gilliland proposed using a forecast value added (FVA) analysis to identify bad practices and recognize beneficial ones. However, he emphasized that a more accurate forecast alone delivers no practical value to an organization unless it is used to make business decisions that improve the efficiency or reduce costs. Thus, the emphasis must be shifted to an FVA providing forecasts that are more accurate and a more precise estimation of uncertainty, stressing their contribution to the improvement of bottom-line results.

Following our responses to the discussions and commentaries of this issue, we now conclude by answering what we would have done differently if we had received the above criticisms and disagreements while still designing the M4 Competition. Our answer is summarized as follows:

- *Predictive densities*: We decided early on that it would be impractical to estimate full prediction densities. However, we believe that including some more prediction intervals in addition to the 95% one would have been reasonable. Thus, we probably would have demanded two additional prediction intervals to better approximate the densities, such as the 90% and 68% ones.
- *High frequency data*: Understanding that many forecasters and practitioners find more value in extrapolating high frequency series than low frequency ones for their research and business purposes, we would have modified the structure of the dataset so as to include more weekly, daily and hourly data. More specifically, we would have tried to collect enough series to create a dataset in which a quarter of the total series could be characterized as high frequency ones.
- *New domains*: We would have added additional domains that would have included data from web traffic and social sites. Such new domains, coupled with high frequency data, would have covered what the commentators have been suggesting.
- *ML methods*: Understanding that some participants did not realize that ML methods could be exploited to effectively extract information from the whole M4 dataset, thus limiting their choices in the development of series-by-series models, we would have introduced two additional ML benchmarks to further inspire innovative and meaningful solutions in that promising area of modelling. The first benchmark would have utilized a simple neural network (NN) to generate forecasts using all 100,000 series as the input, while the second would have considered a deep learning NN, also using the whole M4 dataset as the input, but also exploiting a much deeper architecture in order to filter, code and extract more information from the series.
- *Inside and private information*: We might perhaps have emphasized further that there are absolutely no constraints for participating in the competition, and that all forecasters are free to choose, develop and implement their methods as they please.

References

Barker, J. (2020). Machine learning in m4: what makes a good unstructured model?. *International Journal of Forecasting*, *36*(1), 150–155.

Fry, C., & Brundage, M. (2020). The m4 forecasting competition – a practitioner's view. *International Journal of Forecasting*, *36*(1), 156–160.

Gilliland, M. (2020). The value added by machine learning approaches in forecasting. *International Journal of Forecasting*, *36*(1), 161–166.

Grushka-Cockayne, Y., & Jose, V. R. R. (2020). Combining prediction intervals in the m4 competition. *International Journal of Forecasting*, *36*(1), 178–185.

Januschowski, T., Gasthaus, J., Flunkert, V., Wang, B., Bohlke-Schneider, M., Salinas, D., et al. (2020). Criteria for classifying forecasting methods. *International Journal of Forecasting*, *36*(1), 167–177.

Kang, Y., Hyndman, R. J., & Smith-Miles, K. (2017). Visualising forecasting algorithm performance using time series instance spaces. *International Journal of Forecasting*, *33*(2), 345–358.

Kolassa, S. (2020). Why the best point forecast depends on the error or accuracy measure. *International Journal of Forecasting*, *36*(1), 208–211.

Lichtendahl, K. C., Jr., & Winkler, R. L. (2020). Why do some combinations perform better than others?. *International Journal of Forecasting*, *36*(1), 142–149.

Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., et al. (1982). The accuracy of extrapolation (time series) methods: results of a forecasting competition. *Journal of Forecasting*, *1*, 111–153.

Makridakis, S., & Hibon, M. (1979). Accuracy of forecasting: an empirical investigation. *Journal of the Royal Statistical Society, Series A*, *142*(2), 97–145.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). Statistical and machine learning forecasting methods: concerns and ways forward. *PLOS ONE*, *13*(3), 1–26.

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.

Onkal, D. (2020). M4 competition: What's next?. *International Journal of Forecasting*, *36*(1), 206–207.

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.

Spyros Makridakis is a Professor at the University of Nicosia, where he is also a director of the Institute For the Future (IFF) and Emeritus Professor at INSEAD. He has authored, or co-authored, 22 books and more than 250 articles. He was the founding editor-in-chief of the *Journal of Forecasting* and the *International Journal of Forecasting* and is therganizer of the M (Makridakis) Competitions.

Evangelos Spiliotis is a Research Fellow at the Forecasting & Strategy Unit. He graduated from the School of Electrical and Computer Engineering at the National Technical University of Athens in 2013 and got his PhD in 2017. His research interests are time series forecasting, decision support systems, optimization, statistics, energy forecasting, energy efficiency and conservation. He has conducted research and development on tools for management support in many national and European projects.

Vassilios Assimakopoulos is a professor at the School of Electrical and Computer Engineering of the National Technical University of Athens. He has worked extensively on applications of decision systems for business design and has conducted research on innovative tools for management support in an important number of projects. He specialises in various fields of strategic management, design and development of information systems, statistical and forecasting techniques using time series.