Editorial

The M5 competition: Conclusions

Spyros Makridakis
Institute for the Future (IFF), University of Nicosia, Nicosia, Cyprus

Fotios Petropoulos
School of Management, University of Bath, United Kingdom

Evangelos Spiliotis
Forecasting and Strategy Unit, School of Electrical and Computer Engineering, National Technical University of Athens, Greece

1. Introduction

There have been exactly 40 years since 1982 when the findings of the first M competition (M or M1) were published (Makridakis et al., 1982). At that time, when computers were not much faster than today's calculators and all communications were done by telephone and regular mail, the results of the seven participants who took part in M1 were fiercely contested by the statistical establishment of the time. Moving to 2020, the fifth installment of the Makridakis competitions, M5, attracted more than 8,200 contestants, achieving almost 100,000 submissions and publishing its results, findings, and key insights in this special issue of the *International Journal of Forecasting (IJF)*, including methodological and discussion papers, as well as short notes and editorials. Most importantly, the conclusions of the M5, as well as those of the previous M competitions, are accepted as essential knowledge, guiding forecasting theory and practice. In these 40 years, the field of forecasting has gone from a cold winter to a blossoming spring, transforming itself from a collection of subfields to a unified discipline of academics and practitioners working harmoniously together to advance it. It is the purpose of this concluding paper to outline the changes of the field over the last 40 years, as witnessed by the five M competitions, and then discuss its future.

2. The major findings of the M competitions

Each M competition provided some unique, empirical evidence for guiding objectively the theory and practice of forecasting that are summarized next.

2.1 The M competition, 1982: M1 is considered the first true forecasting competition in the sense that anyone could submit forecasts and compare the accuracy of their methods over other approaches (Hyndman, 2020). Its findings, which were based on 1,001 time series, were identical to those of the Makridakis and Hibon (1979) study and can be summarized as follows:

- Statistically sophisticated or complex methods do not necessarily produce more accurate forecasts than simple ones.

- The accuracy of a combination of methods outperforms, on average, the accuracy of the individual methods being combined. Moreover, combinations perform well compared to single methods.
- Forecasting accuracy depends on the length of the forecasting horizon, while the relative performance of different methods depends on the measure used for the evaluation.

These findings, and especially the first two ones, were contrary to those espoused by the dominant approach of that time, the Box-Jenkins methodology to ARIMA models, advocating a true model for each series that had to be identified judgmentally and then used to forecast once its residuals were assured to be random.

2.2 The M2 competition, 1993: The M2 competition (Makridakis et al., 1993) owes its origin to the belief that the only reason for the superiority of the simple methods over the sophisticated ones in M1 was that no human judgment could be applied to the statistical forecasts. This assumption, however, had to be tested on real time data so that judgment could be applied to the statistical forecasts, making M2 the first live forecasting competition from the 18 major ones conducted until now (Makridakis et al., 2021). This innovative design requirement was met by providing the participants with the budget data of four companies (Honeywell, Squibb, Aussedat-Rey, and a car company, not willing to disclose its name) and asking them to make 15 monthly forecasts at the end of October over two consecutive years. The five forecasters who participated could, therefore, modify their statistical forecasts to incorporate their judgmental inputs about the economy, industry, and firm from which they could ask and receive additional information about its operations and plans. The results indicated that, on average, human judgment deteriorated the accuracy of the statistical forecasts. Moreover, they suggested that simple methods performed better than statistically sophisticated ones, thus reconfirming the findings of M1.

2.3 The M3 competition, 2000: The M3 competition (Makridakis & Hibon, 2000) validated the findings of the previous two M competitions, with the only difference that the number of series was increased to 3,003, thus providing a larger, more representative sample to test the accuracy of various forecasting methods. More importantly, this revalidation was made given the involvement of more participants that employed advanced - for that time - forecasting approaches, including expert systems and some premature machine learning (ML) methods, among others. Finally, M3 identified a new forecasting method, Theta (Assimakopoulos & Nikolopoulos, 2000), that proved to be the most accurate one, despite being relatively simple. Overall, we can argue that a long winter has been prevailing in the field of forecasting since it was established by Robert Brown in 1959.

2.4 The M4 competition, 2018: The M4 competition (Makridakis et al., 2020) revised the first, major finding of the M1, signifying the end of the long forecasting winter. This outcome can be attributed to two key factors. First, advances in ML, typically exploited in areas such as image recognition and signal processing, were effectively applied for time series forecasting and promising algorithms, such as recurrent neural networks (NNs) and decision trees, were proposed as accurate alternatives to simple forecasting methods. Second, the number of series in the M4 was increased to 100,000, allowing participants to develop global models (Januschowski et al., 2020) that learn from multiple series simultaneously how to forecast the target ones (Semenoglou et al., 2020). As a result, the winning submissions provided substantially more accurate point forecasts than the benchmarks while estimating uncertainty (95% prediction intervals) with an amazing degree of precision. Nevertheless, on average, ML methods continued to underperform, while combinations of statistical forecasts continued to dominate rankings. Consequently, the M4 competition identified ML methods and hybrids as promising alternatives to statistical predictions and reconfirmed the value of combining.

2.5 The M5 competition, 2020: While M1, M3, and M4 covered a large number of data frequencies and domains, providing insights on the methods that are expected to work well in various forecasting applications, the data of the M5 was representative of the retail industry, involving daily, hierarchical series of product sales. Moreover, the M5 data exhibited intermittency at the lower hierarchical levels, thus requiring special processing and modeling. Furthermore, M5 data included exogenous variables, such as information about prices and special events, that participants could exploit to further improve the performance of their solutions. While M4 showed the potential of ML methods, M5 established their value over simple ones as the great majority of the top 50 performing methods, both in the "Accuracy" and the "Uncertainty" challenge, employed a particular tree-based forecasting method, LightGBM, advanced, deep NNs, or combinations of those. These methods achieved significant improvements in terms of accuracy and uncertainty over the statistical benchmarks considered by the organizers and some participants that were diminishing, however, at lower hierarchical levels and at the tails of the uncertainty distributions. Overall, it became apparent that complex, global ML methods are more appropriate for forecasting in retail sales settings, especially when useful explanatory variables are available to incorporate and techniques that improve the generalization of such methods are effectively applied, both to minimize the risk of overfitting and improve the learning process, either in terms of feature and hyper-parameter selection or in terms of data availability. This includes efficient combining, proper cross-validation, and data augmentation, among others.

2.6 The value of the M competitions: The data and results of the M competitions are open and freely available[1] for anyone to download and use to verify their correctness and perform additional research. Moreover, their findings are disseminated through academic and professional publications, including dedicated special issues of the *IJF* for the M3, M4, and M5 competitions. In the M competitions, there is an emphasis on learning and advancing the theory and practice of forecasting by challenging existing best practices and identifying possible new ones. In addition, according to Seaman & Bowman (this issue) "*the Makridakis competitions are not only extremely valuable in their own right; they form a historical record of how the technology has evolved over the years.*" We expect to continue the tradition of the M competitions to future ones.

3. The future of forecasting competitions

There have been eighteen major forecasting competitions until now (Makridakis et al., 2020, 2021), with all of them, except one, M2, having similar attributes. The evaluation in these competitions was done by concealing the latest part of the data, asking participants to submit forecasts for the unknown observations, and using the concealed data to measure

[1] The data of the M competitions can be found at GitHub (https://github.com/Mcompetitions) and the website of the International Institute of Forecasters (https://forecasters.org/resources/time-series-data/).

the performance of the submissions made. Furthermore, all submissions were numerical, as no judgment could be incorporated to data whose characteristics and period were either completely or partially unknown. In addition, in most cases the evaluation was made by considering a single forecasting origin instead of multiple ones. Finally, all of the competitions required point forecasts, while six asked for additional uncertainty estimates.

The M2 competition has been the only live competition to involve multiple forecasting origins, following the budgeting process of having to produce forecasts for the next fifteen months and then evaluating their accuracy once the actual data becomes available. Doing so allowed judgmental inputs to the numerical forecasts as the participants could adjust the statistical forecasts of their models to incorporate personal and publicly available information about the economy, industry, and firms. In addition, since the M2 was done on two rolling origins (budget years), it allowed the participants to receive feedback and possibly improve their predictions for the second year. Although M2 has received the least attention of all M competitions, it has been a unique, innovative one in terms of design, closely simulating common business forecasting practices. Its main disadvantage has been the limited number of series (29 in total) and evaluation rounds (two in total).

The M6 duathlon competition: The concealed-data, single-origin competitions are the easiest to organize and run. For this reason, they are also the most popular. Live, multi-origin competitions are on the other end of the difficulty scale as they require more time and effort, both to organize and complete. Organizers must, therefore, strike a balance between having as realistic of a competition as possible, while not discouraging participants because of the hard work and the long period of time required to stay committed in the competition. In this respect, the M6[2] competition will last for an entire year and require twelve monthly submissions for a participant to be eligible for the global prizes. However, it will also award participants that decide to compete in particular quarters. In addition, M6 will be a duathlon, requiring participants to accurately predict, on a scale from 1 to 5, the rank of 100 assets (50 stocks and 50 ETFs) in terms of percentage returns and, at the same time, to make profitable investment decisions by selecting among these 100 assets the ones that will produce the highest returns, using appropriate weights of investment with short and long positions.

There are two major research questions to be empirically investigated by the M6. The first is the Efficient Market Hypothesis (EMH). According to the yearly Morningstar's "Active/Passive Barometer", active, professionally managed investment funds do not beat, on average, random stock selections. However, legendary investors like Warren Buffett, Peter Lynch, and George Soros, among others, and celebrated firms like Bridgewater Associates, Renaissance Technologies, and DE Shaw have achieved phenomenal returns that are impossible to justify by mere chance or be explained by the EMH, rendering the latter hypothesis into a paradox. The second is to determine possible links between forecasting accuracy/uncertainty and investment returns by investigating if above average financial returns are achieved by one or a combination of the following:

- The ability to forecast more accurately than others overall market returns, or those of individual stocks/ETFs.

[2] More information about the M6 competition, including its detailed guidelines, can be found at https://mofc.unic.ac.cy/the-m6-competition/.

- The ability to properly appreciate overall market uncertainty or that of individual stocks/ETFs.
- The identification of effective ways to assess forecasting accuracy/uncertainty when making investment decisions and selecting particular stocks/ETFs.
- The exploitation of judgment in making investment decisions.
- The employment of a consistent investment strategy.
- Other possible factors like judgmental biases that negatively affect the assessment of accuracy/uncertainty.

4. The future of forecasting

In a study comparing various statistical and ML methods published in 2018, Makridakis et al. (2018) found that the worst statistical model was more accurate than the best ML one. However, this conclusion was drawn due to training the ML models in a local, series-by-series fashion and has been reversed with the introduction of global ML methods that were still in their infancy in forecasting settings in 2016 when the study was originally conducted. This reversal was proven with the top two performing methods of the M4 competition and the excellent performance of the LightGBM method, as well as some deep learning NNs, in both the “Accuracy” and “Uncertainty” challenges of the M5 competition. For instance, in the M5 “Accuracy” challenge, the top performing ML method was, overall, 22.4% more accurate than the best statistical benchmark (exponential smoothing), while in the M5 “Uncertainty” challenge, the top performing ML method was, overall, 24.6% more accurate than the top statistical benchmark (ARIMA).

Does this finding suggest the domination of global ML methods and the eventual abandonment of the local, statistical ones? To answer this question, we shall pay attention to the distribution of the overall improvements of these methods across the various hierarchical levels and validate whether these are uniform or not. In the “Accuracy” challenge, the improvement at the top level was about 53%, dropping to about 31% at middle levels, and reaching 3% at the lowest one. Similarly, in the “Uncertainty” challenge, the improvement starts at 53% at the top level, drops to around 30% at the middle levels and falls to 12% at the lowest one. It seems that simple and fast to run statistical methods can still compete with the global ML ones at the lowest hierarchical levels, where randomness and intermittency becomes dominant. In a similar fashion, the improvement of the top performing ML method in the “Uncertainty” challenge drops from 26% at quantile 0.5 (median) to 14% at quantile 0.995, indicating that statistical methods remain competitive when it comes to forecasting the right tail of the uncertainty distributions.

There are, however, some additional, thought-provoking factors to consider. Out of the 2,666 participating teams in the “Accuracy” challenge, less than half (48.4%) managed to outperform the Naive benchmark, only 35.8% outperformed the sNaive benchmark, and just 7.5% beat the top performing statistical benchmark. The results are better for the “Uncertainty” challenge where, from the 769 participating teams, 86.2% managed to outperform the Naive benchmark, 62% outperformed the sNaive benchmark, and 22.6% beat the top performing statistical benchmark.

What may these statistics suggest exactly? Although the performance of the top performing teams in both challenges is excellent, there is a clear deterioration in terms of improvements as we consider more teams, particularly in the “Accuracy” challenge where 92.5% of the participants did not manage to beat, on average, the top statistical benchmark. This may mean that there is a trade-off between using simple, ready-to-use, and computationally fast methods like exponential smoothing versus ML ones, given that the latter typically require more time and effort to be sufficiently tuned, both in terms of inputs and hyper-parameters. The results of the M5 competition demonstrate that global ML methods can indeed outperform simple ones, but also that this is expected to be the case only when sufficient effort is put on the design and training of the ML solutions.

An equally interesting observation is that the top position winner of the “Accuracy” challenge, YeonJun In, was a senior undergraduate student at a Korean university at the time the competition was conducted and with practically no previous experience in forecasting. Nevertheless, his submission managed to beat those of his 7,091 competitors, some of them being Kaggle masters and grandmasters.

What does this tell us about the experience and expertise required to produce accurate forecasts? What about the skills modern forecasters should have? Clearly, the top winner of the “Accuracy” challenge was a forecasting novice but skillful in experimenting with ML algorithms like LightGBM, a method that has shown lots of potential in recent Kaggle competitions. The fact that he was a forecasting novice did not seem to be a disadvantage given that ML methods can identify patterns and data relationships on their own and provide accurate predictions without the forecaster being required to provide specific instructions on how to perform such a task. It suffices for them to fine tune the ML model used to automatically provide the forecasts. Yet, there are reasons to believe that the maturity of the available ML models in dealing with particular forecasting tasks could affect the above conclusion. For instance, most publicly available ML models currently focus on point forecasts and, as a result, a limited number of advanced ML approaches can be used in practice to precisely estimate uncertainty. This observation may partially justify why, in contrast to the “Accuracy” challenge, the top performing submissions in the “Uncertainty” challenge were all made by professional forecasters and practitioners with strong backgrounds in statistics, forecasting, and computer science.

It seems that the role of the forecaster may be changing to someone who knows how to skillfully handle ML models without necessarily having particular experience and expertise in forecasting and, undoubtedly, to someone who conducts valuable research on such unstructured algorithms to further improve their performance. This is a big difference from using statistical methods, where knowledge about the properties of each method is typically required to select and train the most appropriate one and achieve accurate results. Such a difference can be confirmed by comparing the winning methods of the M4 provided by, Slawek Smyl, an experienced forecaster who developed his own hybrid model with that of YeonJun In who won without knowing much about forecasting.

4.1 Statistical and ML forecasting methods

ML has expanded the repertoire of forecasting possibilities to include some powerful alternatives to statistical methods that can significantly improve forecasting accuracy and precisely estimate uncertainty, at least at the higher hierarchical levels. In practice, ML ended the long forecasting winter by introducing generalized, global models, capable of identifying and exploiting data similarities across related series, cross-validation strategies that improve the overall accuracy while avoiding overfitting, and approaches that estimate uncertainty realistically. The major advantage of ML methods over statistical ones is their superior performance on average and the minimal interventions they require for producing the forecasts. Their disadvantage is that they need considerable time to be trained and be properly implemented, providing also little explanations on the way the forecasts are made, thus being a "black box" approach.

According to Kolassa (this issue), ML methods are overrated in terms of their practical value. It is a fact that only 7.5% of the participating teams in the M5 "Accuracy" challenge managed to beat automatic exponential smoothing and that the top method outperformed such benchmark only on 58.5% of the series in terms of MSE, being certainly statistically significant but not overwhelmingly. It seems that there is a substantial difference between the performance of these methods when it comes to forecasting the top or middle hierarchical levels versus the bottom ones. What does Kolassa suggest? "*If you are tempted to invest heavily in data scientists and expect them to work wonders, make sure to compare their methods to simple benchmarks that you can probably implement at a fraction of the cost of an ML pipeline.*" His advice seems valid at the lower levels of the hierarchy, characterized by high randomness and intermittency, where improvements in accuracy/uncertainty over the statistical benchmarks are minimal. At the same time, the average improvements of ML methods over such statistical benchmarks are impressive for the higher hierarchical levels and hard to be ignored.

In our view, ML methods are here to stay whatever their current disadvantages or weaknesses. First, ML methods require less human effort to produce forecasts for multiple series as their predictions are made mechanically by powerful and generalizable algorithms. Second, ML methods are complementary to statistical ones, allowing using them only when their improvements exceed development and running costs. Third, the speed and memory of computers is increasing at a fast pace while running costs are continuously dropping, making the use of ML methods more economical. Fourth, as with all new methods, existing limitations will be identified and addressed while novel improvements will be introduced, thus further reducing costs and simplifying the process of making predictions.

Most importantly, ML methods are connected to AI and bound to follow its progress that would lead to further improvements in forecasting accuracy and uncertainty. After all, the utilization of modern predictive ML methods is less than half a decade old with high expectations of significant improvements in the future, coming from AI scientists working outside the forecasting field. Statistical methods, on the other hand, are unlikely to improve much unless some unexpected breakthrough occurs.

4.2 The maturity of the forecasting field

The forecasting field has changed a great deal since Robert Brown's (1959) *Exponential Smoothing for Predicting Demand* book was published, establishing in effect the field, and so has the role of the forecaster. Brown was a practitioner who introduced a class of exponential smoothing methods that could be used to predict demand. These methods were simple enough to be applied using the systems available at that time but also powerful enough to provide reasonably accurate forecasts. Nevertheless, they were initially looked down by the academic community for lacking statistical sophistication, ignoring the opportunity firms had to benefit from their utilization.

Instead, the field moved “towards” the introduction of statistically sophisticated methods, including large scale econometric models, cultivating with the Box and Jenkins (1970) book that provided a complex methodology to time series forecasting that was adored by the academic community. This approach, however, encouraged overfitting, emphasized the need of judgment to identify a true model for each time series, and discouraged combinations of forecasts. The overall inappropriateness of the Box-Jenkins methodology was confirmed by the results of the M competitions where it was found that its accuracy was not necessarily better than Brown’s exponential smoothing and was also worse than other simple methods, like Theta, and combinations of simple statistical models.

There was little progress in the use of systematic forecasting methods by business firms for a long time. In an article appearing in HBR, Chambers, Mullick and Smith (1971) stated: “*Our purpose here is to present an overview of this field by discussing the way a company ought to approach a forecasting problem, describing the methods available, and explaining how to match method to problem. We shall illustrate the use of the various techniques from our experience with them at Corning, and then close with our own forecast for the future of forecasting*….”, continuing, “*Although we believe forecasting is still an art, we think that some of the principles which we have learned through experience may be helpful to others.*” And, indeed, forecasting remained an art for several more decades until firms, particularly retailers, started realizing the advantages of more accurate predictions and academics became interested in improving their methods, often by exploiting advances in other fields than forecasting.

Today, the field of forecasting has matured with an emphasis on the forecasting value added rather than the sophistication of the method or its theoretical correctness. Moreover, the utilization of systematic forecasting has grown considerably, depending to a great extent on the size of the firm. Specifically, in a recent survey (Tashman et al., 2022), 51% of large firms indicated that they always or mostly use systematic forecasting while the corresponding percentage for medium firms was 40% and considerably lower for smaller ones. It is our expectation that such growth will continue as the advantages of systematic forecasting become known and as forecasting consultants provide off-the-shelf software to implement the various statistical and ML methods, providing also advice on automatically selecting the most appropriate among them. As time passes and computers become cheaper and more powerful and the benefits of systematic forecasting become better known, it is our expectation that the usage of systematic methods will further increase.

4.3 The changing role of the forecaster

The most significant change in the role of the forecaster over the years has been in the amount of human judgment required to select the most appropriate forecasting method or model. In the Box-Jenkins methodology this had to be done judgmentally and consequently verified by assuring that the residuals of the constructed model were random. Such a requirement was dropped, however, as automatic model selection algorithms were introduced, not only for the Box-Jenkins models, but also for the exponential smoothing ones (Hyndman et al., 2002). More importantly, ML methods can now be used to accurately produce forecasts for multiple series with minor human involvement. Automating model selection is changing the role of the forecaster from someone whose major task is to select the best method/model to someone who has to properly tune the algorithm that performs such a task.

Does this mean that forecasting knowledge will become outdated or that forecasters will become obsolete? The introduction of ML methods may have provided new tools that forecasters will have to become familiar with and learn how to use them to produce accurate predictions. However, statistical methods will still be useful for certain types of applications while both ML and statistical methods will be further automated allowing more time for forecasters to pursue other or additional, high value-added tasks. Here are some possibilities:

- *Judgmental adjustments*: It is well known that systematic forecasts extrapolate established patterns and data relationships to predict the future and that these forecasts must be adjusted (overridden) when structural changes are about to occur. This is a challenging, yet necessary task that must be properly done to avoid optimism and other judgmental biases (see Fildes and Goodwin, 2021) so that the adjustments made actually improve the overall forecasting performance.
- *Forecasting value added (FVA)*: Although ML methods can provide substantial improvements in forecasting performance, such improvements are not always uniform in different forecasting applications. As a result, comparing the additional costs required for using a ML method over a simple statistical one versus the improvements that such a selection would provide in terms of forecasting performance becomes vital. More importantly, these decisions should be made not only on the basis of forecasting performance improvements, but also in terms of expected utility gains that may refer to lower realized costs and better customer service, among others. This is because, although forecast error is a reasonable proxy for measuring FVA, understanding the actual benefits of utilizing a particular forecasting approach over another is even more helpful for decision making.
- *Cloud computing*: Given the widespread use of cloud computing, sharing programs and data with supply-chain partners can provide substantial benefits. Knowing the vendors' inventory levels is an example of such collaboration.
- *Uncertainty*: Uncertainty has not received the appropriate attention in either the academic community or among practitioners. Spending more time on improving the estimation of uncertainty to deal with the resulting risks more effectively can be of considerable value, improving the way that uncertainty can be faced.
- *Breaking the "black box"*: Given that most ML methods are black boxes, a better understanding of how the forecasts are computed and how uncertainty behaves would become extremely useful. Such understanding can come by comparing the forecasts produced by ML methods with those generated by statistical methods of known properties, e.g., comparing the forecasts of SIngle versus Holt's exponential smoothings.
- *Methodological innovations*: As the role of the forecaster is transitioning from that of a statistician to that of a data scientist of additional skills (linked to computer science and AI) that will eventually significantly automate and simplify the forecasting process as a whole, future forecasters will have more time to invest on research, achieving important methodological innovations and algorithmic advances that would further improve forecasting accuracy and the more realistic assessment of uncertainty.

5. The emergence of a unified data science forecasting field

The field of forecasting includes statistical and ML methods and involves academics and practitioners that work together to advance its theory and practice. With time, we expect that the field will be unified further as statistical and ML methods are complementary and will eventually be integrated into a single data science field whose purpose, at the theoretical level, will be to provide as accurate predictions and correct estimates of uncertainty as possible, while, on the practical side, to ensure maximum benefits to firms from the utilization of forecasting applications, achieved the highest FVA by minimizing costs while maximizing benefits. Theoretical improvements in forecasting performance will mainly come from advances in ML algorithms, as well as deep learning ones, originating outside the forecasting field in areas related with AI, like signal processing and image recognition. Practical improvements will mostly come from forecasting consultants and software vendors, eager to apply state-of-the-art theoretical advances to improve the forecasting benefits to firms. Furthermore, as computer costs drop and computer power increases, more effective software and services will become available, thus further increasing the automation in making predictions and changing the role of the forecaster to allow them to concentrate on new, value-adding tasks that substantially increase the benefits of forecasting. For us an increasing role of the future forecaster would be devoted to providing successful judgmental overrides to the statistical and ML predictions when patterns/relationships are changing or when new “inside” information/knowledge becomes available.

References

Assimakopoulos, V., & Nikolopoulos, K. (2000). The Theta model: a decomposition approach to forecasting. International Journal of Forecasting, 16(4), 521–530.

Box, G., & Jenkins, G. (1970). Time Series Analysis: Forecasting and Control. San Francisco: Holden-Day

Brown, R.G. (1959) Exponential Smoothing for Predicting Demand. New York, McGraw-Hill

Chambers, J.C., Mullick, S.K., & Smith, D.D. (1971) How to Choose the Right Forecasting Technique, Harvard Business Review, 65, July-August.

Fildes, R. and Goodwin, P. (2021) Stability in the inefficient use of forecasting systems: A case study in a supply chain company, International Journal of Forecasting, 37(2), 1031-1046.

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

Kolassa, S. (2021). Commentary on the M5 forecasting competition. International Journal of Forecasting.

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, 439–454.

Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Newton, J., Parzen, E., & Winkler, R. (1982). The accuracy of extrapolation (time series) methods: Results of a forecasting competition. Journal of Forecasting, 1(2), 111–153.

Makridakis, S., Chatfield, C., Hibon, M., Lawrence, M., Mills, T., Ord, K., & Simmons, L. F. (1993). The M2-competition: A real-time judgmentally based forecasting study. International Journal of Forecasting, 9(1), 5–22.

Makridakis, S., Fry, C., Petropoulos, F., & Spiliotis, E. (2021). The future of forecasting competitions: Design attributes and principles. INFORMS Journal on Data Science.

Makridakis, S., & Hibon, M. (2000). The M3-Competition: results, conclusions and implications. International Journal of Forecasting, 16(4), 451–476.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). Statistical and Machine Learning forecasting methods: Concerns and ways forward. PloS One, 13(3), e0194889.

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.

Seaman, B., & Bowman, J. (2021). Applicability of the M5 to forecasting at Walmart. International Journal of Forecasting.

Semenoglou, A.-A., Spiliotis, E., Makridakis, S., & Assimakopoulos, V. (2020). Investigating the accuracy of cross-learning time series forecasting methods. International Journal of Forecasting.

Tashman, L., Makridakis, S., Hoover, J., & Karelse, J. (2022) Business Forecasting: Impressive Advances, Lagging Implementation, Working paper, MOFC