International Journal of Forecasting 36 (2020) 161–166

Contents lists available at ScienceDirect

International Journal of Forecasting

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

Discussion

The value added by machine learning approaches in forecasting

Michael Gilliland

SAS, United States

A B S T R A C T

This discussion reflects on the results of the M4 forecasting competition, and in particular, the impact of machine learning (ML) methods. Unlike the M3, which included only one ML method (an automatic artificial neural network that performed poorly), M4's 49 participants included eight that used either pure ML approaches, or ML in conjunction with statistical methods. The six pure (or combination of pure) ML methods again fared poorly, with all of them falling below the Comb benchmark that combined three simple time series methods. However, utilizing ML either in combination with statistical methods (and for selecting weightings) or in a hybrid model with exponential smoothing not only exceeded the benchmark, but performed at the top. While these promising results by no means prove ML to be a panacea, they do challenge the notion that complex methods do not add value to the forecasting process.

1. Introduction

When Robert G. Brown (1956) published *Exponential smoothing for predicting demand*, he opened the modern era of time series forecasting methods. While computational power, the availability of data, and method sophistication have all increased to levels that would have been unfathomable in the 1950s, we have not observed a commensurate increase in our ability to forecast the future accurately. Over half a century of research, and the three previous M forecasting competitions, have led skeptics to raise the legitimate question of whether technological and methodological advances have delivered any value at all for forecasting.

At one extreme, in a special issue of *Journal of Business Research* on the topic of simple versus complex methods for forecasting, Green and Armstrong (2015) found that, "Remarkably, no matter what type of forecasting method is used, complexity harms accuracy" (p. 1684). Their conclusion was drawn from a review of 32 papers reporting on 97 simple versus complex comparisons. However, this starkly negative view of complexity is not unchallenged.

The more mainstream opinion has been being expressed in various forms for at least 40 years. In their 1978 presentation before the Royal Statistical Society, Makridakis and Hibon (1979) called for more research as to "why, under certain circumstances, simpler methods do as well as or better than sophisticated ones" (p. 116). They later characterized a result of the first M competition – an attempt to conduct such research – as that "Statistically sophisticated or complex methods do not necessarily provide more accurate forecasts than simpler ones" (Makridakis & Hibon, 2000, p. 452). Closer to the present day, Morlidge's (2014a) review of the M competitions up to that point concluded that "sophistication is no guarantee of performance" (p. 36).

A slightly more positive (although still guarded) opinion of sophisticated methods was expressed by Crone, Hibon, and Nikolopoulos (2011) in their analysis of the NN3 forecasting competition (in which neural network and other "complex" methods were applied to a disguised sample of the M3 data). These authors found reason for optimism: "Overall, we hope that the success of complex algorithms on such a well-established dataset will at least rekindle the discussion of innovative, sophisticated algorithms for time series extrapolation in forecasting, econometrics and statistics" (p. 657). Furthermore, such methods "can perform competitively relative to established statistical methods in time series prediction, but still cannot outperform them. However, … we can no longer assume that they are inferior" (p. 657).

E-mail address: mvgilliland@gmail.com.

https://doi.org/10.1016/j.ijforecast.2019.04.016
0169-2070/

Table 1
Performances of the top two methods and selected benchmarks in terms of PF accuracy.

				Improvement over Comb		
Method	sMAPE	MASE	OWA	sMAPE	MASE	OWA
Smyl	11.374	1.536	0.821	9.4%	7.7%	8.6%
Montero-Manso et al.	11.720	1.551	0.838	6.6%	6.7%	6.7%
Comb	12.555	1.663	0.898	–	–	–
Naïve2	13.564	1.912	1.000	−8.0%	−15.0%	−11.4%
Naïve1	14.208	2.044	1.058	−13.2%	−22.9%	−17.9%

Table 2
Performances of the top two methods and the naïve benchmark in terms of PI precision.

			Improvement over Naïve1	
Method	MSIS	ACD	MSIS	ACD
Smyl	12.230	0.002	49.2%	97.4%
Montero-Manso, et al.	14.334	0.010	40.4%	88.8%
Naïve1	24.055	0.086	–	–

Table 3
Running times of selected methods and benchmarks.

Author	sMAPE	MASE	OWA	Runtime (mins)	Runtime vs. Comb
Smyl	11.374	1.536	0.821	8056.0	242×
Montero-Manso, et al.	11.720	1.551	0.838	46108.3	1388×
Legaki & Koutsouri	11.986	1.601	0.861	25.0	0.75×
Theta	12.309	1.696	0.897	12.7	0.38×
Comb	12.555	1.663	0.898	33.2	–
ARIMA	12.669	1.666	0.903	3030.9	91×
Naïve2	13.564	1.912	1.000	2.9	0.09×
Naïve1	14.208	2.044	1.058	0.2	0.01×

The results of the M4 affirm this optimism. Unlike some of the poorly-substantiated claims in the machine learning literature (see Makridakis, Assimakopoulos & Spiliotis, 2018), M4 provides evidence of progress in the appropriate use of ML methods for improving forecasting. The remainder of this discussion will focus on the value added by ML approaches relative to traditional statistical forecasting methods, consider some curiosities in the results, and examine why the practice of business forecasting continues to fall well short of its accuracy potential.

2. Results of the M4 forecasting competition

A full presentation of the M4 results, with details regarding particular methods and definitions of the performance metrics, can be found elsewhere in this issue (Makridakis, Spiliotis, & Assimakopoulos, 2020). Table 1 provides a summary of the point forecast (PF) results for the two top-performing competitors and selected benchmarks.

Table 2 provides a summary of the prediction interval (PI) results for the two top-performing competitors and the benchmark.

It should be noted that the methods of Smyl (2020) and Montero-Manso, Talagala, Hyndman, and Athanasopoulos (2020) (hereafter M-M) incorporate machine learning along with traditional time series forecasting techniques, and so are of particular interest for this discussion. (Six other "pure" ML approaches (either individual or combination ML models) fared relatively poorly: all fell below the Comb benchmark (a simple average of SES, Damped, and Holt), and all but Trotta's ML ensemble fell below the Naïve1 and Naïve2 benchmarks.)

The Smyl and M-M methods are both described fully by their authors in this issue. The methods are not simple, either conceptually or computationally. Smyl is characterized as a "hybrid" approach, mixing exponential smoothing with a recurrent neural network forecasting engine. M-M is a combination of seven statistical methods and one ML method, with weights calculated by a ML algorithm. Both methods use information from multiple series to predict individual ones. So, where do the M4 results lead us, given that the top-rated performances are achieved by complex methods that utilize both statistical and ML approaches? Is it time to discard the belief that complex methods are no better than simple methods for forecasting?

In a limited sense, the answer seems to be yes. Previous studies have found that complex methods can perform better under particular circumstances, and M4 now shows that improvements over simple benchmarks are possible over a broad range of series types and frequencies, in terms of both PFs and PIs. As Makridakis et al. (2020) conclude, "More complex methods can possibly lead to greater forecast accuracy".

3. Accuracy improvement – at what cost?

Granting the potential of ML to improve upon traditional forecasting methods, this raises questions as to the cost of such improvements.

The organizers of the M4 sought to replicate the participating methods on a standard hardware configuration, with the running time for each method serving as a proxy for its complexity. Although the full replication work is not yet complete at the time of this writing, the organizers found that, "on average, the accuracy improved as the computational time increased, signifying that processing power was exploited beneficially to achieve more accurate forecasts" (Makridakis et al., 2020). Table 3 shows the running times of a selection of participating methods and benchmarks.

The method of Legaki and Koutsouri (hereafter L&K) is the top-ranking statistical method that is not a combination model. The top ranked method in M3 was Theta, which served as a benchmark in M4. The relevance of both will be discussed below.

The key takeaway from the runtime analysis is that benchmarks are computationally efficient, as expected.

Even though Comb (at 33.2 min) took over 150 times as long as the most efficient method (Naïve1 at 0.2 min), this difference is not of practical importance. Both should be fast enough for even large-scale forecasting situations (such as we might find at a large retailer that requires millions of store/SKU forecasts each week).

What is of practical importance is that ARIMA (2.1 days) took nearly 100 times as long as Comb, and Smyl (5.6 days) and M-M (32.0 days) took nearly 250 and nearly 1400 times as long. Both would require considerable performance tuning to make them suitable for large-scale forecasting situations.

While excessive computational time does not disqualify Smyl, M-M, or the other top performers from practical applications, it may – for the moment – limit their use to only the most important, high-value forecasts. However, it is very likely that further research and experience using these approaches (including parallelizing the algorithms on inexpensive hardware) will reduce the running time dramatically. It is possible for efficiencies in algorithm coding and data movement, and the use of distributed processing, to deliver order-of-magnitude reductions.

Also, while ML (and even classical methods) can be costly to fit, they are cheap to evaluate. Thus, the practical implications of excessive fit times could be allayed through a combination of parallelization and scheduling. For example, if daily forecasts are needed, models could be re-learned weekly, then used to forecast on a daily basis. There may be no need to refit models for every new forecast.

4. Is the PF improvement meaningful?

In addition to considering the costs of the PF and PI improvements using the top-ranked methods, we must also consider their practical importance, given the sizes of the improvements.

A more accurate forecast, by itself, delivers no practical value to an organization. Its value is created through improved business decisions, which lead to actions that improve business results.

Computer-generated forecasts are routinely distrusted and ignored by management. (For a discussion of the "algorithm aversion" versus "algorithm appreciation" problem, see Logg, Minson, & Moore, 2019.) They are *distrusted* when they have proved not to be sufficiently accurate, or are produced through a mechanism ("black box") that cannot be explained. They are *ignored* when they do not forecast the kind of future that management wishes to see happen. Thus, rather than a forecast doing what it should – expressing a "best guess" at what is really going to happen – management often prefers forecasts to be aspirational – expressing the goal or target they wish to achieve. (It is called *evangelical forecasting* when a powerful and charismatic leader provides the forecasts, unencumbered by what the data may be saying.) Unfortunately, taking operational and financial decisions based on aspirational forecasts can have negative consequences (such as excess inventory, and revenue shortfalls).

A few percentage points of improved PF accuracy are unlikely to be of much practical importance. This is not enough improvement for an untrusted forecasting process to become trusted, or to result in significantly different decisions or actions.

The Smyl and M-M reductions in sMAPE over Comb, which itself is 13.2% better than Naïve1, do merit attention. These improvements are especially interesting because they occur over a broad range of forecasting situations. However, Smyl's overall sMAPE still drops a mere 1.2 percentage points below Comb (11.374% vs. 12.555%) and just 2.8 percentage points below Naïve1 (14.208%). The forecasting literature is full of new methods that purport to demonstrate slight accuracy improvements, at least over some limited range of forecasting situations. But is eking out a few extra percentage points of accuracy the ultimate solution to the business forecasting problem? We'll return to this question in Section 9.

5. Is the PI improvement meaningful?

There has long been easy money to be made by betting on future actuals falling outside the calculated prediction intervals. Previous methods for computing PIs generally underestimated the future forecast errors, often dramatically, leading to unrealistically narrow bands. (This observation led Makridakis, Hogarth, & Gaba, 2009, to quip that calculated PI bands should be doubled in width.)

Perhaps the most surprising outcome of M4 is the success of the Smyl and M-M methods in specifying PIs correctly. Although Athanasopoulos, Hyndman, Song, and Wu (2011) reported instances of accurate PIs in a tourism forecasting competition, Makridakis, Spiliotis and Assimakopoulos (2018) state of Smyl and M-M that "These are the first methods we are aware of that have done so, rather than underestimating the uncertainty considerably" (p. 803).

Software vendors have not adopted the presentation of PIs around their point forecasts universally. However, even when PIs are calculated and presented in forecasting software, they largely go unused by industry practitioners, and not solely because of their imprecision, as there are a number of psychological issues to contend with. For example, Yaniv and Foster (1995) hypothesize that "… the vagueness … of judgmental estimation under uncertainty involves a trade-off between two conflicting objectives: accuracy and informativeness" (p. 424). They found that people may prefer a narrower interval that does not include the true value to a wider interval that does. Du, Budescu, Shelly, and Omer (2011) found that people will tolerate intervals up to a certain width – that some degree of uncertainty is unavoidable – but that wider intervals lose credibility.

Goodwin (2014) provides a good summary of the psychological issues around PIs. While the exceptional PI performance of Smyl and M-M might spur more adoptions of PIs by practitioners, it may not. The quality of the calibration of the PIs does not appear to be related to the motivation to use them. As Goodwin states, "An interval forecast may accurately reflect the uncertainty, but it is likely to be spurned by decision makers if it is too wide and judged to be uninformative" (p. 5). The challenge remains to educate practitioners on the interpretation and use of PIs, as well as of even more valuable uncertainty indicators such as full predictive densities.

Table 4
Performances of the Legaki & Koutsouri, Theta, and Comb methods across M3 and M4.

	sMAPE		Improvement over Comb		Runtime (mins)
Method	M3	M4	M3	M4	M4
Legaki & Koutsouri	NA	11.986	NA	4.5%	25.0
Theta	13.0	12.309	3.8%	2.0%	12.7
Comb	13.5	12.555	–	–	33.2

6. An indirect affirmation of middle-out forecasting

M4 provided new examples of the utilization of information from multiple time series for predicting individual series. Industry practitioners have long realized the benefits of this approach – for time series arranged in a hierarchy – by performing "middle-out" forecasting.

Middle-out forecasting is used when the lowest-level (most granular) time series (e.g. an item at a given location) are intermittent or otherwise too noisy to capture anything beyond the average level in the model. Moving up the hierarchy (e.g., to the item- or brand-level time series) allows the trend and seasonality to be modeled, and then apportioned back down to the most granular level.

Other, more sophisticated ways of performing hierarchical forecasting have been proposed, further highlighting the value of information that is provided in multiple time series. These include an optimal forecast reconciliation method (Wickramasuriya, Athanasopoulos, & Hyndman, 2018) that guarantees coherent forecasts (which add up exactly to the forecasts of the aggregated series) that are at least as good as the (pre-reconciliation) base forecasts.

Several of the M4 participants, including Smyl and M-M, exploited information from multiple (albeit, non-hierarchical) series. Unfortunately, the M4 results do not allow us to isolate the performance gains that can be attributed to this approach directly.

7. Curiosities in the results

Datasets often contain outliers, or at least "curiosities" that merit further investigation, and the M4 results do not disappoint in this respect.

7.1. Performance of the Legaki & Koutsouri statistical model

Legaki & Koutsouri (L&K) is the top-performing pure statistical method that does not use a combination of models (it is actually a variation of the Theta method using a Box–Cox transformation). It ranked 8th overall for PFs, but did not provide PIs. In terms of sMAPE, L&K delivered a 4.5% improvement over the Comb benchmark, versus the 9.4% improvement by Smyl (see Table 4). Although L&K's performance was exceeded in M4 by seven methods that combined statistical and/or ML models, its improvement over Comb would have been good enough to beat the Theta model that "won" M3 (Theta had a 3.8% improvement over Comb in M3).

The importance of L&K's results is likely to be overshadowed by the focus on the ML and combination methods. However, the fact that it demonstrated meaningful improvements over the benchmarks with an extremely short processing time merits investigation. L&K's running time is on a par with those of Theta and Comb, the times of which were at least two orders of magnitude lower than those of the higher-performing combination methods that have been replicated as of the time of this writing.

7.2. Death knell for ARIMA?

The performance of ARIMA has improved relative to benchmarks since the first M competition, so that now, in M4, it is essentially as good as Comb (Table 5). The problem is that ARIMA's running time is nearly 100 times as long as that of Comb. (It should be noted that ARIMA's replicated runtime on the M4 data is surprisingly long. Automated ARIMA modeling has been a staple of forecasting software for a number of years, and I suspect that a significant reduction in runtime would be possible, although not to the level of the other simple benchmarks.)

If it is not quite time to retire ARIMA from the forecaster's toolbox, at least there seems to be ample evidence that its role should be limited to niche situations where its performance is less likely to trail those of more computationally efficient benchmarks. For example, Chatfield (2007) stated, "I would only recommend ARIMA modeling for a series showing short-term correlation where the variation is not dominated by trend and seasonality, provided the forecaster has the technical expertise to understand how to carry out the method" (p. 5).

7.3. Poor performance of some combination models

Combination models dominate the top of the rankings, and there is a wealth of research (as well as common sense reasons) showing why combining models improves the accuracy. As is stated by Makridakis et al. (2020), "no single method can capture the time series patterns adequately, whereas a combination of methods, each of which captures a different component of such patterns, is more accurate because it cancels the errors of the individual models through averaging".

However, M4 also saw some combination models perform poorly, considerably below the Comb benchmark, and there may be valuable lessons from exploring these results. Possible explanations include a poor selection of models to be combined, or improper weighting. Akin to the "simple vs. complex" issue in forecast modeling, where simple models often perform better, the simple average often performs best for forecast combination. Claeskens, Magnus, Vasnev, and Wang (2016) provide theoretical reasons why weightings other than a simple average may degrade the forecast accuracy.

Table 5
ARIMA performance relative to selected M competition benchmarks.

	MAPE		sMAPE		Runtime	Runtime
Method	M1	M2	M3	M4	M4	vs. Comb
Naïve2	17.8	13.3	15.5	13.6	2.9	0.09×
SES	16.8	11.9	14.3	13.1	8.1	0.24×
Damped	NA	12.8	13.7	12.7	15.3	0.46×
Comb	NA	11.7	13.5	12.6	33.2	–
ARIMA	18.0	16.0	14.0	12.7	3030.9	91×
FVA vs Naïve2	−1.1%	−20.3%	9.7%	6.6%		
FVA vs Comb	NA	−36.8%	−3.7%	−0.8%		

8. Other applications of ML methods in forecasting

If nothing else, M4 shows that there is legitimate hope for the application of machine learning methods for forecasting.

Further examples of work in progress include projects that utilize ML to guide manual overrides to statistical forecasts, with such work being conducted independently at Kellogg's and in SAS research and development. Manual overrides to system-generated forecasts are common, reaching 80% in some organizations (Fildes, Goodwin, Lawrence, & Nikolopoulos, 2009). This of course consumes considerable management time in reviewing and adjusting potentially thousands of individual forecasts. Both Kellogg's (Pineda & Stevens, 2018) and SAS R&D (Valsaraj, Gallagher, & Chase, 2018) have reported favorable results in both improving the quality of overrides and reducing the number of manual overrides being performed. Although their positive initial results have not yet been proven to rigorous academic standards, the work is still in its early stages. Further empirical investigation will show the efficacy (or failure) of these approaches.

9. The shocking disappointment of real-life business forecasting

It is not a revelation that the practice of business forecasting falls well short of the potential exhibited in academic research and in competitions like M4. In an early IJF editorial, Chatfield (1986) bemoaned the state of forecasting practice and called on statisticians to find ways of communicating the better use of existing methods to practitioners. Lawrence's (2000) IJF editorial some years later, "What does it take to achieve adoption in sales forecasting?" begins by citing Chatfield, then highlights the lack of progress toward Chatfield's goal, deploring the lack of research devoted to studying the causes. "If we don't support and encourage this line of research it is possible that in another thirteen years there will be another editorial lamenting the lack of progress in getting business to make use of the available forecasting techniques" (p. 148).

Fast forward to 2018, and nearly every M4 contestant beats Naïve1 (the "no change" model), with top contestants outperforming it handily (with roughly a 20% reduction in sMAPE, and a 25% reduction in MASE). However, Morlidge's (2014b) study of eight consumer and industrial businesses (with over 17,500 products) reported that 52% of their forecasts were less accurate than Naïve1. "This result distressingly suggests that, on average, a company's product forecasts do not improve upon naïve projections" (p. 29). How have we reached this state of affairs?

The fact that even good methods will underperform simple benchmarks some of the time is no surprise. Morlidge's (2014a) review of M3 (focusing on the 334 one-month ahead forecasts of industry data) found that all 24 of the methods performed worse than Naïve1 more than 30% of the time. Similarly, in M4, both Smyl and M-M performed worse than Comb over 40% of the time. (Of course, Comb is a much higher-performing benchmark than Naïve1.)

Typically, business forecasting is not an objective search for the truth, for an "unbiased best guess". Instead, a business forecast may express the wants, wishes, and personal agendas of forecasting process participants, rather than their belief as to what is really going to happen. In addition, these participants, too often, are failing to look critically at their process and its less-than-stellar results. (In a survey of forecasters by Fildes and Goodwin (2007), 25% failed to indicate that they used any error measure at all!) "What is most important for improving forecast performance is not optimizing the forecasting method, but rather avoiding methods that are patently inadequate" (Morlidge, 2014a, p. 39).

The M4 has demonstrated that many approaches, even simple and inexpensive ones, can outperform naïve models. So why does real-life forecasting remain so vexing? The solution may not lie in new algorithms, even when they are demonstrably better, because they don't appear to be being used. Instead, taking a "defensive" approach to business forecasting – such as using forecast value added (FVA) analysis to identify and avoid bad practices – may have a much more positive impact on process results (Gilliland, 2013, 2017).

The sad conclusion that we can draw from M4 is that real life business forecasting continues to be a shocking disappointment. As Chatfield and Lawrence have pleaded before, research on forecasting practice must be encouraged to continue.

10. Conclusions

On average, the Smyl and M-M methods demonstrate superior performance on PFs, and unprecedented performance on PIs. However, averages hide a lot of relevant information (e.g., Savage, 2009).

Given the amount of effort and cost involved in their application, there is no immediate need for practitioners

to apply the new ML-enhanced approaches in every situation: there are plenty of fast and inexpensive statistical methods that are "good enough" for most practical purposes – especially large-scale forecasting situations where thousands or even millions of forecasts are needed, and time and cost become significant considerations.

However, M4 has shown that sophisticated methods, when properly applied, can improve the accuracy. So perhaps it is time to temper the "common wisdom" – which I myself have adhered to – that sophisticated methods are of little use for forecasting.

Also, the M4 results re-affirm many years of research on the value of combination methods. Perhaps we have reached the point where practitioners (and forecasting software vendors) should cease their pursuit of the perfect individual forecasting model, and instead let the "wisdom of the crowds" (i.e., an ensemble of imperfect models) do their forecasting for them.

References

Athanasopoulos, G., Hyndman, R. J., Song, H., & Wu, D. C. (2011). The tourism forecasting competition. *International Journal of Forecasting*, *27*, 822–844.

Brown, R. G. (1956). *Exponential smoothing for predicting demand*. MA: Cambridge.

Chatfield, C. (1986). Simple is best? *International Journal of Forecasting*, *2*, 401–402.

Chatfield, C. (2007). Confessions of a pragmatic forecaster. *Foresight*, *6*, 3–9.

Claeskens, G., Magnus, J. R., Vasnev, A. L., & Wang, W. (2016). The forecast combination puzzle: A simple theoretical explanation. *International Journal of Forecasting*, *32*, 754–762.

Crone, S. F., Hibon, M., & Nikolopoulos, K. (2011). Advances in forecasting with neural networks? Empirical evidence from the NN3 competition on time series prediction. *International Journal of Forecasting*, *27*, 635–660.

Du, N., Budescu, D., Shelly, M., & Omer, T. (2011). The appeal of vague financial forecasts. *Organizational Behavior and Human Decision Processes*, *114*, 179–189.

Fildes, R., & Goodwin, P. (2007). Against your better judgment? How organizations can improve their use of management judgment in forecasting. *Interfaces*, *37*(6), 570–576.

Fildes, R., Goodwin, P., Lawrence, M., & Nikolopoulos, K. (2009). Effective forecasting and judgmental adjustments: an empirical evaluation and strategies for improvement in supply-chain planning. *International Journal of Forecasting*, *25*, 3–23.

Gilliland, M. (2013). FVA: A reality check on forecasting practices. *Foresight*, *29*, 14–18.

Gilliland, M. (2017). Changing the paradigm for business forecasting. *Foresight*, *44*, 29–35.

Goodwin, P. (2014). Getting real about uncertainty. *Foresight*, *33*, 4–7.

Green, K., & Armstrong, F. S. (2015). Simple versus complex forecasting: the evidence. *Journal of Business Research*, *68*, 1678–1685.

Lawrence, M. (2000). What does it take to achieve adoption in sales forecasting? *International Journal of Forecasting*, *16*, 147–148.

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, *151*, 90–103.

Makridakis, S., Assimakopoulos, V., & Spiliotis, E. (2018). Objectivity, reproducibility and replicability in forecasting research. *International Journal of Forecasting*, *34*, 835–838.

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

Makridakis, S., Hogarth, R., & Gaba, A. (2009). *Dance with chance*. Oxford: Oneworld Publications.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). The m4 competition: results, findings, conclusion and way forward. *International Journal of Forecasting*, *34*, 802–808.

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.

Montero-Manso, P., Talagala, T., Hyndman, R., & Athanasopoulos, G. (2020). FFORMA: Feature-based forecast model averaging. *International Journal of Forecasting*, *36*(1), 86–92.

Morlidge, S. (2014a). Do forecasting methods reduce avoidable error? Evidence from forecasting competitions. *Foresight*, *32*(2014), 34–39.

Morlidge, S. (2014b). Forecast quality in the supply chain. *Foresight*, *33*(2014), 26–31.

Pineda, B., & Stevens, R. (2018). How machine learning boost statistical forecasting for better demand planning at Kellogg's. https://www.sas.com/en_us/events/analytics-conference/watch-live.html#formsuccess.

Savage, S. (2009). *The flaw of averages*. Hoboken, NJ: John Wiley & Sons.

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.

Valsaraj, V., Gallagher, B., & Chase, C. (2018). How demand planning will benefit from machine learning. https://www.sas.com/en_us/webinars/machine-learning-for-demand-planning.html.

Wickramasuriya, S., Athanasopoulos, G., & Hyndman, R. (2018). Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization. *Journal of the American Statistical Association*, *114*(526), 804–819.

Yaniv, I., & Foster, D. (1995). Graininess of judgment under uncertainty: An accuracy-informativeness trade-off. *Journal of Experimental Psychology: General*, *124*(4), 424–432.

Michael Gilliland is Marketing Manager for SAS forecasting software, prior to which he spent 15 years in forecasting positions in the food, consumer electronics, and apparel industries. He is author of *The Business Forecasting Deal* (2010), principal editor of *Business Forecasting: Practical Problems and Solutions* (2015), and writes The Business Forecasting Deal blog. Mike holds a BA in Philosophy from Michigan State University, and Master's degrees in Philosophy and Mathematical Sciences from Johns Hopkins University. He is interested in issues relating to forecasting process, such as worst practices and Forecast Value Added analysis, and in applying research findings for real-life improvements in business forecasting.