International Journal of Forecasting 36 (2020) 54–74

Contents lists available at ScienceDirect

International Journal of Forecasting

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

The M4 Competition: 100,000 time series and 61 forecasting methods

Spyros Makridakis [a,*], Evangelos Spiliotis [b], Vassilios Assimakopoulos [b]

[a] Institute for the Future, University of Nicosia, Cyprus
[b] Forecasting & Strategy Unit, School of Electrical and Computer Engineering, National Technical University of Athens, Greece

ARTICLE INFO

Keywords:
Forecasting competitions
M competitions
Forecasting accuracy
Prediction intervals
Time series methods
Machine learning methods
Benchmarking methods
Practice of forecasting

ABSTRACT

The M4 Competition follows on from the three previous M competitions, the purpose of which was to learn from empirical evidence both how to improve the forecasting accuracy and how such learning could be used to advance the theory and practice of forecasting. The aim of M4 was to replicate and extend the three previous competitions by: (a) significantly increasing the number of series, (b) expanding the number of forecasting methods, and (c) including prediction intervals in the evaluation process as well as point forecasts. This paper covers all aspects of M4 in detail, including its organization and running, the presentation of its results, the top-performing methods overall and by categories, its major findings and their implications, and the computational requirements of the various methods. Finally, it summarizes its main conclusions and states the expectation that its series will become a testing ground for the evaluation of new methods and the improvement of the practice of forecasting, while also suggesting some ways forward for the field.

1. Introduction

This paper provides a detailed description of the most recent M Competition, the M4. It presents its results, discusses its findings and states its conclusions. In addition, it examines its implications for the theory and practice of forecasting and outlines some ways forward. Hyndman's (2020) excellent history of forecasting competitions elucidates their benefits and their contribution to the field of forecasting, so we do not need to add anything more, other than to agree with his conclusions and thank him for his considerable support in all stages of the M4, as well as in the design and realization of this special issue.

A major innovation of the M4 Competition was to predict/hypothesize on its findings more than two months before its completion. Our ten predictions/hypotheses were a clear statement of our expectations of such findings, rather than rationalizing its results in a post-hoc reasoning. We have now evaluated these ten predictions/hypotheses in a separate paper (Makridakis, Spiliotis & Assimakopoulos, 2020), and we are pleased to say that we were entirely correct in at least six out of the ten predictions/hypotheses made beforehand. Furthermore, we have explained where we went wrong, where we were partially correct, and where additional information is required to confirm our claims.

Another innovation of M4 was the appointment of a guest editor, Fotios Petropoulos, to supervise this special issue and assure objectivity and fairness. While certain decisions on the issue, such as its structure and length, and the number and content of invited papers and commentaries, were agreed upon between him, Rob J. Hyndman and Spyros Makridakis, the final decision about this and all other papers submitted remained solely with the guest editor. His editorial, Petropoulos and Makridakis (2020) describes the criteria used for inviting papers, the

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

https://doi.org/10.1016/j.ijforecast.2019.04.014

reviewing process and the deadlines set. In addition, it has been agreed that this paper, together with all others, would be put through a regular reviewing process before being published. Makridakis and Petropoulos (2020) conclude this special issue, while Makridakis, Hyndman and Petropoulos (2020) summarize the state of the art of forecasting in social sciences, discuss its achievements and limitations, and deliberate on its future prospects.

The present paper consists of three main parts, along with five appendices that are available online as supplementary material. The first part describes the background, organization and running of the M4 Competition. The second part presents the results of the competition, including its major findings and their implications, as well as the top performing methods, both overall and across the various subcategories, including data frequencies and application domains. In addition, it comments on the poor accuracy of the pure machine learning (ML) methods that were submitted to M4 and provides summarizing graphs and tables of various performance measures, including a graph showing the computational time required for applying a method versus its forecasting accuracy. The last part then summarizes the paper and highlights the conclusions of the competition, stating our suggestion that the 100,000 series of the M4 should become a testing ground for guiding further theoretical and practical developments by being a huge sample of time series on which the performances of new forecasting methods can be assessed. Finally, the five appendices provide comprehensive tables and graphs of the overall findings, as well as of the various subcategories, including information regarding the benchmarks used and the forecasting performances achieved by the participating methods for each forecasting horizon, frequency and domain, in terms of both point forecasts (PFs) and prediction intervals (PIs).

2. The background

Forecasting competitions have influenced the field of forecasting greatly over the years, providing a solid basis for assessing different extrapolation approaches and learning empirically how to advance forecasting theory and practice (Hyndman, 2020). Each forecasting competition has introduced some new features or data, while trying to either address possible limitations of previous ones, or focus on specific fields of application such as energy and tourism. M4 likewise involved some new features, which can be summarized as follows: (i) the introduction of high-frequency data (weekly, daily and hourly) along with low-frequency data (yearly, quarterly and monthly); (ii) the consideration of PIs as well as PFs; (iii) an emphasis on the reproducibility of the results; and, finally, (iv) the incorporation of a vast number of diverse series and benchmarks. Apart from introducing these features, though, M4 was motivated by the fact that new competitions also tend to bring to light innovative methods which can be tested against existing ones on completely unknown sets of series. This is of major importance given that M3, the previous M Competition, has been being used as a standard benchmark for comparison for almost two decades, meaning that newly proposed methods could eventually over-fit its published test sample. From our point of view, M4 has inspired the development of cutting-edge methods, as well as providing a new, larger, and therefore more difficult to over-fit dataset, thus assisting researchers and practitioners to explore and re-evaluate best practices in forecasting in greater detail.

The M4 Competition was initially announced at the beginning of November, 2017, first on the University of Nicosia's website (www.unic.ac.cy) and then on both the IIF blog (www.forecasters.org) and that of Rob J. Hyndman (www.robjhyndman.com). In addition, invitation emails were sent to all those who had participated in the previous M Competitions (Makridakis, Andersen, Carbone, Fildes, Hibon, et al., 1982; Makridakis, Chatfield, Hibon, Lawrence, Mills, et al., 1993; Makridakis & Hibon, 2000), the tourism forecasting competition (Athanasopoulos, Hyndman, Song, & Wu, 2011), the NN3 competition (Crone, Hibon, & Nikolopoulos, 2011), and recent International Symposium on Forecasting (ISF) events, as well as to those who had published relevant articles in the *International Journal of Forecasting* and other well-known forecasting, neural network and machine learning journals (*Foresight: The International Journal of Applied Forecasting*, *Journal of Forecasting*, *Technological Forecasting and Social Change*, *Expert Systems with Applications*, *Neural Networks* and *Neurocomputing*). The data were made available on 31st December, originally on the M4 website (www.m4.unic.ac.cy) and later via the *M4comp2018* R package (Montero-Manso, Netto, & Talagala, 2018) and the M4 GitHub repository (www.github.com/M4Competition). The competition ended at midnight, May 31st, 2018. A short paper with the initial results of the M4 was published in the *International Journal of Forecasting* on June 20th, 2018 (Makridakis, Spiliotis, & Assimakopoulos, 2018b), which was made available as an open access article by the generous contribution of Elsevier during the ISF conference in Boulder, Colorado, USA.

Like the previous three M Competitions, M4 was an open competition with the aim of ensuring fairness and objectivity. Moreover, in order to promote replicability in forecasting research and facilitate future work in the field (Makridakis, Assimakopoulos, & Spiliotis, 2018), the participants were encouraged to submit the code of their method, as well as a detailed description of it, to the M4 GitHub repository. Participants who claimed proprietary rights (e.g. software vendors) did not have to make their code available publicly, but provided a version of their program to the organizers of the competition to allow them to evaluate its reproducibility and measure its computational complexity. Table 7 lists the submitted methods and shows which have been replicated successfully, as well as the extent of the replication.

The rules of the competition, prizes and additional details were all made available on the M4 website. In order for participants to be eligible for the prizes, they had to provide PFs and, optionally, PIs for all 100,000 series of the competition shown in Table 1. The dataset was subdivided into six data frequencies and six application domains, and the PFs and PIs were evaluated for each subcategory, in addition to computing overall averages. Thus, the various tables in the main text, as well

Table 1
Number of M4 series per data frequency and domain.

Time interval between successive observations	Micro	Industry	Macro	Finance	Demographic	Other	Total
Yearly	6,538	3,716	3,903	6,519	1,088	1,236	**23,000**
Quarterly	6,020	4,637	5,315	5,305	1,858	865	**24,000**
Monthly	10,975	10,017	10,016	10,987	5,728	277	**48,000**
Weekly	112	6	41	164	24	12	**359**
Daily	1,476	422	127	1,559	10	633	**4,227**
Hourly	0	0	0	0	0	414	**414**
Total	**25,121**	**18,798**	**19,402**	**24,534**	**8,708**	**3,437**	**100,000**

as in the appendices, show the overall accuracy of PFs (Appendix A) and the overall precision of PIs (Appendix B) for the different data frequencies and forecasting horizons, while Appendix C lists the best performing methods for the various frequencies and domains for both PFs and PIs. Appendix D shows the percentage of time series for which Comb (the simple arithmetic average of Single, Holt and Damped exponential smoothing) was more accurate than the other methods for PFs (Fig. D.1), as well as the percentage of time series for which Naïve 1 was more accurate for PIs (Fig. D2). The benchmarks used in the competition are described analytically in Appendix E.

It should be noted that the 100,000 time series used in the M4 were selected from a database (called ForeDeCk) compiled at the National Technical University of Athens (NTUA) that contains 900,000 continuous time series, built from multiple, diverse and publicly accessible sources. ForeDeCk emphasizes business forecasting applications, including series from relevant domains such as industries, services, tourism, imports & exports, demographics, education, labor & wage, government, households, bonds, stocks, insurances, loans, real estate, transportation, and natural resources & environment (Spiliotis, Kouloumos, Assimakopoulos & Makridakis, 2020).

The M4 dataset was created on December 28th, 2017, when Professor Makridakis chose a seed number to randomly select the sample of 100,000 time series to be used in the M4. The selected series were then scaled to prevent negative observations and values lower than 10, thus avoiding possible problems when calculating various error measures. The scaling was performed by simply adding a constant to the series so that their minimum value was equal to 10 (29 occurrences across the whole dataset). In addition, any information that could possibly lead to the identification of the original series was removed so as to ensure the objectivity of the results. This included the starting dates of the series, which did not become available to the participants until the M4 had ended.

Note that the number of series considered for each frequency and domain was determined based mainly on how likely a company or an organization is to generate forecasts for the applications implied by them, as well as how important they are in terms of operation and strategic planning. For instance, when dealing with business forecasting, monthly forecasts are required more frequently than quarterly or yearly ones. Similarly, micro and financial data are more likely to be used to support decision making than demographic data. Details about how M4 was constructed can be found in the study by Spiliotis, Kouloumos et al. (2020), along with some guidelines that are intended to facilitate the design of future forecasting competitions.

It should also be mentioned that, as in the previous M Competitions, low-volume and intermittent time series were not considered for the M4 dataset. The reasoning behind this choice was, first, the continuity of M4 with past M Competitions; second, the many methodological problems that zero values would raise; and third, the significantly different nature of models that would be required for extrapolating non-continuous series. Thus, the authors would like to clarify that the findings of M4 refer to continuous business series, meaning that some of them may not apply for low-volume or intermittent series.

Another notable innovation of the M4 Competition was the introduction of various benchmarks, both statistical and ML. The field of forecasting has progressed a great deal since the original M Competition, which concluded that "more complex or statistically sophisticated methods are not necessarily more accurate than simpler methods", and over time, new methods have been proposed that have clearly proven to be more accurate than simpler ones. In this regard, the organizers of the competition decided to include ten benchmark methods for two reasons. First, to evaluate the improvement of the M4 submissions over simple/standard approaches, and, second, to be able to identify the causes of improvements by comparing each submission directly with different methods of well-known properties. For instance, Naïve 2 (Makridakis, Wheelwright, & Hyndman, 1998) captures seasonality only, Single Exponential Smoothing (DSES) captures the level (Gardner, 1985), Holt (Gardner, 2006) extrapolates using a linear trend, and Damped (Gardner, 2006), as its name implies, dampens the linear trend. Thus, comparing the submitted methods to the benchmarks better enables one to detect the factors that lead to improved forecasts, such as ways of handling seasonality or trend optimally. Table 2 lists the ten benchmarks used in M4, as well as two additional standards for comparison: ETS (exponential smoothing; Hyndman, Koehler, Snyder, & Grose, 2002) and (auto) ARIMA (Hyndman & Khandakar, 2008), which were included due to their extensive utilization in forecasting studies over the last few years, as well as in this competition. Of these benchmarks, we decided to use Comb as the single benchmark against which to compare the forecasting accuracies

Table 2
The benchmarks and standards for comparison of the M4 Competition.

Methods			Description
Point Forecasts (PFs)	**Statistical benchmarks**	Naïve 1	A random walk model, assuming that future values will be the same as that of the last known observation.
		Naïve S	Forecasts are equal to the last known observation of the same period.
		Naïve 2	Like Naïve 1 but the data are seasonally adjusted, if needed, by applying a classical multiplicative decomposition. A 90% autocorrelation test is performed to decide whether the data are seasonal.
		SES	Exponentially smoothing the data and extrapolating assuming no trend. Seasonal adjustments are considered as per Naïve 2.
		Holt	Exponentially smoothing the data and extrapolating assuming a linear trend. Seasonal adjustments are considered as per Naïve 2.
		Damped	Exponentially smoothing the data and extrapolating assuming a damped trend. Seasonal adjustments are considered as per Naïve 2.
		Theta	As applied to the M3 Competition using two Theta lines, $\vartheta_1 = 0$ and $\vartheta_2 = 2$, with the first one being extrapolated using linear regression and the second one using SES. The forecasts are then combined using equal weights. Seasonal adjustments are considered as per Naïve 2.
		Comb	The simple arithmetic average of SES, Holt and Damped exponential smoothing (*used as the single benchmark for evaluating all other methods*).
	ML benchmarks	MLP	A perceptron of a very basic architecture and parameterization. Some preprocessing like detrending and deseasonalization is applied beforehand to facilitate extrapolation.
		RNN	A recurrent network of a very basic architecture and parameterization. Some preprocessing like detrending and deseasonalization is applied beforehand to facilitate extrapolation.
	Standards for comparison	ETS	Automatically provides the best exponential smoothing model, indicated through information criteria.
		ARIMA	An automatic selection of possible ARIMA models is performed and the best one is chosen using appropriate selection criteria.
Prediction intervals (PIs)		*Naïve 1*	A random walk model, assuming that future values will be the same as that of the last known one (*used as the single benchmark for the PIs*).
		ETS	Automatically provides the best exponential smoothing model, indicated through information criteria.
		ARIMA	A stepwise selection over possible models is performed and the best ARIMA model is returned using appropriate criteria.

of all other methods. The reason for this is that Comb is simple to compute, usually more accurate than the three individual methods being averaged, and intuitive; furthermore, it was also used in M2 and M3, thus enabling comparisons among the three competitions. For the case of the PIs, we decided to use three benchmarks, namely Naïve 1, ETS and ARIMA, with Naïve 1 being utilized as the single benchmark against which to compare all other methods.

Note that the benchmarks introduced included two pure ML ones: a perceptron and a recurrent network. This was done to further emphasize the objective of the competition in regard to the evaluation of ML approaches in forecasting applications, as well as to encourage the participation of multiple ML methods. The two benchmarks were networks of very basic architectures, trained individually for each series and utilizing typical preprocessing options in order to avoid limiting the participants and to trigger innovative solutions.

The code for generating and reproducing the forecasts of the benchmarks and the standards for comparison listed above became available from the M4 GitHub repository at the start of the competition. The statistical benchmarks were all estimated using the v8.2 of the *forecast* package for R (Hyndman, 2017), while the ML ones were developed in Python using the Scikit v0.19.1, Keras v2.0.9 and TensorFlow v1.4.0 libraries. An exception was made for the case of the Theta method (Assimakopoulos & Nikolopoulos, 2000), which was provided as a separate R function (note that none of the benchmarks were eligible for prizes). Finally, two of the methods submitted, those of Spiliotis & Assimakopoulos and Legaki & Koutsouri, were not eligible for prizes either, because their authors were related to the organizing team (both methods were variations of the original Theta with the aim of improving its accuracy). Detailed information about the benchmarks can be found in Appendix E.

3. Organizing and running the M4 competition

The M4 Competition began on January 1st, 2018, when its dataset became available for download from the M4 website. The rules and detailed information about the competition were likewise posted on this site (M4 Team, 2018). The dataset could be downloaded from the M4 site, the GitHub or as the *M4comp2018* R package by anyone wishing to participate.

3.1. Training and test set

As was noted in Section 2, the M4 dataset consists of 100,000 time series of yearly, quarterly, monthly and other (weekly, daily and hourly) data, which are divided into training and test sets. The training set was made available to the participants at the beginning of the competition, while the test set was kept secret till its end,

when it was released and used by the organizers for evaluating the submissions. The minimum numbers of observations in the training test are 13 for yearly, 16 for quarterly, 42 for monthly, 80 for weekly, 93 for daily and 700 for hourly series. It is worth mentioning that M4 consists of much longer series than M3 on average, thus offering more opportunities for complicated methods that require large amounts of data for proper training.

As in the previous M Competition, the participants were asked to produce the following numbers of forecasts beyond the available data that they had been given: six for yearly, eight for quarterly and 18 for monthly series. In addition, we requested 13 forecasts for the weekly series and 14 and 48 forecasts respectively for the daily and hourly ones. The forecasting horizons were determined based on the nature of the decisions that each frequency of data is most likely to support within a company or organization. For instance, yearly data are used typically to support long-term decisions on a strategic level, for between one and five years ahead. On the other hand, quarterly and monthly forecasts are typically used for budgeting purposes, varying from a few months to two years ahead. Finally, high frequency data are usually used for supporting operations at a short-term level, varying from a few hours to a few weeks ahead.

3.2. Performance measures for point forecasts (PFs)

There are many measures available in the forecasting literature for evaluating the performances of forecasting methods (Hyndman & Koehler, 2006; Kim & Kim, 2016). In the previous M Competitions, several of these measures were used without any clear agreement as to the advantages and drawbacks of each (Goodwin & Lawton, 1999). Given such a lack of agreement, we decided to use the average of two of the most popular accuracy measures, referred to as the overall weighted average (OWA), believing that this would help us to achieve a higher level of objectivity. These measures were the symmetric mean absolute percentage error (sMAPE; Makridakis, 1993) and the mean absolute scaled error (MASE; Hyndman & Koehler, 2006).

The first measure, which has been selected in the past for evaluating the submissions of the M3 Competition, uses percentage errors that are scale independent, intuitive to understand and part of an everyday vocabulary (e.g. the value of the Apple stock increased by 1.5% yesterday). The second measure aims to correct some potential problems of the first and to provide an alternative with better mathematical properties (Franses, 2016). For instance, the proposed MASE has a defined mean and a finite variance, is scale-independent and can be computed for a single forecast horizon, being less than one if it arises from a better forecast than the average one-step Naïve S forecast computed in-sample, and vice-versa.

The formulae for computing the two measures are as follows:

$$sMAPE = \frac{2}{h}\sum_{t=n+1}^{n+h}\frac{\left|Y_t - \widehat{Y}_t\right|}{|Y_t| + \left|\widehat{Y}_t\right|} * 100(\%)$$

$$MASE = \frac{1}{h}\frac{\sum_{t=n+1}^{n+h}\left|Y_t - \widehat{Y}_t\right|}{\frac{1}{n-m}\sum_{t=m+1}^{n}|Y_t - Y_{t-m}|},$$

where Y_t is the value of the time series at point t, $\widehat{Y}_t$ the estimated forecast, h the forecasting horizon, n the number of the data points available in-sample, and m the time interval between successive observations considered by the organizers for each data frequency, i.e., 12 for monthly, four for quarterly, 24 for hourly and one for yearly, weekly and daily data. The reasoning behind this choice is that there is no obvious seasonality for the case of the weekly, daily and hourly data. For example, not all years consist of 52 weeks, daily data may include five, six or seven observations per week, depending on the application considered, and hourly data may display double (7 days $\times$ 24 h) or even triple (7 days $\times$ 24 h $\times$ 12 months) seasonality. The proposed m values simplify the assumptions made regarding the seasonality present and match the M4 with the M3, given that the "other" data of the latter were also regarded as not being seasonal, despite being of weekly and daily frequencies. The organizers announced the m values at the beginning of the competition, though they highlighted the fact that these values were assumed only for estimating the MASE, with that participants being free to consider any other alternative for generating their forecasts.

Note that the MASE used in M4 differs from the one proposed by Hyndman and Koehler (2006) in that originally m was set equal to one regardless of the frequency of the data. Thus, the benchmark (denominator) that was proposed originally for scaling the absolute error of the examined method (numerator) was the in-sample absolute error of the Naïve 1 method (random walk). We decided to use Naïve S instead of Naïve 1 because we believed that this would be a much more indicative benchmark for seasonal series, as it provided a more reasonable scaling option.

Undoubtedly, Naïve 2 could be another alternative with similar properties. However, Naïve S was preferred over Naïve 2 because there is no unique way of either defining which series are seasonal or estimating the seasonal indexes. Thus, using Naïve 2 as a benchmark for MASE could make the replication of the results a much more difficult and complicated process. In contrast, Naïve S is straight-forward and easy to compute, and requires no assumptions or additional information.

We compute the OWA of sMAPE and MASE by first dividing their total value by the corresponding value of Naïve 2 to obtain the relative sMAPE and the relative MASE, respectively, and then computing their simple arithmetic mean. Thus, if Method X displays a MASE of 1.6 and an sMAPE of 12.5% across the 100,000 series of M4, while Naïve 2 displays a MASE of 1.9 and an sMAPE of 13.7%, the relative MASE and sMAPE of Method X would be equal to $1.6/1.9 = 0.84$ and $12.5/13.7 = 0.91$, respectively, resulting in an OWA of $(0.84 + 0.91)/2 = 0.88$, which indicates that, on average, the method examined is about 12% more accurate than Naïve 2, taking into account both MASE and sMAPE. Detailed numerical examples of the computation of the OWA are provided by the M4 Team (2018).

Note that sMAPE and MASE are first estimated for each series by averaging the error computed for each forecasting horizon, then averaged again across all time series to compute the average value for the entire dataset. On the other hand, OWA is computed only once at the end of the evaluation process for the whole sample of series. Thus, although OWA is relative in nature, it is more indicative and robust than typical relative measures and measures based on relative errors.

It should also be mentioned that Naïve 2 was preferred over Naïve S for estimating the OWA, despite being more complicated in its computation, because Naïve 2 is very popular for time series forecasting, is typically more accurate than Naïve S, has been used repeatedly as a benchmark in many past forecasting studies, and has been estimated for previous M Competitions, thus enabling direct comparisons.

Note that the estimation of sMAPE and MASE differs from those originally adopted in the previous M Competitions, where all of the errors computed for each series and forecasting horizon were averaged together. For instance, the M3 Competition involved a total of 37,014 errors, 3870 referring to yearly (six forecasts across 645 series), 6048 to quarterly (eight forecasts across 756 series), 25,704 to monthly (18 forecasts across 1428 series) and 1392 to other (eight forecasts across 174 series) data. Clearly, subsets that involve more series and longer forecasting horizons will have larger impacts on the accuracy estimate. This is exactly why M4 begins by averaging the errors at a series level, thus weighting all of the series in the dataset equally.

We will not pretend that the selected measures were the most appropriate to use, and we would definitely expect different opinions, or perhaps even strong objections to their selection. Indeed, the literature is full of adequate measures that could have been used alternatively for evaluating the results of the competition (Armstrong & Collopy, 1992; Chen, Twycross, & Garibaldi, 2017; Davydenko & Fildes, 2013; Hyndman & Koehler, 2006; Kolassa, 2016). However, we do believe that the large sample of series used in M4 mitigates the effect that the utilization of different error measures would have had for determining the final ranks of the participating methods. An additional reason for using sMAPE is its continuity with the previous M Competitions, especially after mitigating its major shortcomings by excluding negative and small positive values, still considering its intuitive, everyday interpretation/understanding (for example, Section 4 compares M4 to previous M Competitions based on the reported sMAPE values). On the other hand, MASE is used widely in the contemporary forecasting literature, being independent of the scale of the data, less sensitive to outliers, and infinite or undefined only when all historical observations are equal, which is impossible in practice.

3.3. Performance measures for prediction intervals (PIs)

The M4 Competition adopted a 95% prediction interval (PI) for estimating the uncertainty around the point forecasts. This confidence level was selected because it is one of the most commonly used levels in the business world, as it is neither too tight (e.g. 99%) nor too wide (e.g. 90%) for the majority of economic and financial forecasting applications. Additional levels were not considered because that would have increased the complexity of the competition, requiring more estimations for a huge sample of series and therefore discouraging participation.

The performances of the generated PIs were evaluated using the Mean Scaled Interval Score (MSIS) of Gneiting and Raftery (2007), as follows:

$$MSIS = \frac{1}{h} \times \frac{\sum_{t=n+1}^{n+h} (U_t - L_t) + \frac{2}{a}(L_t - Y_t)1Y_t < L_t + \frac{2}{a}(Y_t - U_t)1Y_t > U_t}{\frac{1}{n-m}\sum_{t=m+1}^{n} |Y_t - Y_{t-m}|},$$

where L_t and U_t are the lower and upper bounds of the prediction intervals, Y_t are the future observations of the series, a is the significance level and **1** is the indicator function (being 1 if Y_t is within the postulated interval and 0 otherwise). Since the forecasters were asked to generate 95% prediction intervals, a is set to 0.05.

The following algorithm illustrates how the MSIS is estimated in practice and highlights its logic when it is used for comparing the precisions of the intervals generated by two different forecasting methods:

- A penalty is calculated for each method at the points at which the future values are outside the specified bounds. This captures the coverage rate of each method.
- The width of the prediction interval is added to the penalty, if any, to get the interval score (IS). In this respect, the methods of larger intervals are penalized over those of smaller ones, regardless of the coverage rate achieved.
- The ISs estimated at the individual points are averaged to get the mean interval score (MIS).
- MIS is scaled by dividing its value by the mean absolute seasonal difference of the series, as is done for the case of the MASE used in M4, in order to make the measure scale-independent.

Since MSIS uses a complex formula to evaluate the coverage rate and the width of the submitted intervals concurrently, this study also uses the absolute coverage difference (ACD) as a supplementary measure of the precision of PIs; however, it is not involved in determining the most precise PIs of the competition, only in examining and understating them. ACD is simply the absolute difference between the average coverage of the method and the target set (here 0.95). Thus, if the future values across the 100,000 time series of the competition are outside the bounds specified by a method an average of 2% of the time (coverage of 98%), the ACD will be $|0.98 - 0.95| = 0.03$.

3.4. The prizes

As has been mentioned already, in order for a participant to be eligible for a prize, PFs had to be provided for all 100,000 series of the competition. Submitting PIs, as well as the code for reproducing the results, was optional

Table 3
The six prizes of the M4 Competition.

Prize name	Description	Amount
1st prize	Best performing method according to OWA	9000€
2nd prize	Second-best performing method according to OWA	4000€
3rd prize	Third-best performing method according to OWA	2000€
Prediction intervals prize	Best performing method according to MSIS	5000€
The UBER student prize	Best performing method among student competitors according to OWA	5000€
The Amazon prize	The best reproducible forecasting method according to OWA	2000€

but highly recommended. Note that there were no restrictions preventing a given participant from collecting more than one of the prizes listed in Table 3.

The amount of 20,000€ was generously provided by the University of Nicosia. In addition, the global transportation technology company Uber generously awarded a special Student Prize of 5000€ to the student with the most accurate forecasting method, and Amazon generously awarded 2000€ for the best reproducible forecasting method.

3.5. Participants and valid submissions

There were 248 participants who were registered to participate in the M4 Competition, from over 46 countries. Of those, only 49 individuals/teams submitted valid PFs for all 100,000 time series. There were even fewer PI submissions, with only 20 forecasters providing valid PIs for all 100,000 alongside their PFs. There were three major reasons for not having more valid submissions:

- The enormous number of time series discouraged many participants from completing the task on time. Even some of those who were using simple statistical methods found the task of predicting such a huge number of series overwhelming. From our experience, we believe that six months was an adequate amount of time for the participants to evaluate their alternatives and submit their forecasts, provided of course that the time was managed well. In this regard, our suggestion for future competitions would be for the organizers to approach potential participants as soon as possible and keep them engaged on the task in order to avoid them wasting valuable time.
- Many participants realized that their methods were no more accurate than the benchmarks, and therefore decided not to submit their forecasts because they were concerned about their overall ranking.
- We suspect that forecasting software vendors and major software firms using statistical and ML methods avoided submitting their forecasts due to concerns about the negative impact of their ranking in case their method was not at the top.

Most of the participants (28) were from academic institutions, while the rest were individuals (11) or from companies/organizations (10). Moreover, the majority of participants (27) utilized combinations of methods, either statistical or ML, many used pure statistical methods (17), one developed a hybrid approach and only four used pure ML ones, probably due to the computational requirements and the complexity involved in forecasting 100,000 time series. Note that we use the term "pure" here to refer to submissions that were based solely on a specific type of method (either ML or statistical), and did not use the other one in any kind of combination or optimization scheme.

The various tables presented in the remainder of this paper include a total of 61 methods for the PFs. These comprise the 49 methods submitted, the 10 benchmarks and the two standards for comparison. In addition, there were 23 methods for the PIs, namely the 20 submissions, the single benchmark and the two standards for comparison.

4. The results of the M4

It is clear that there are plenty of results to report in order to properly evaluate the performances of the various methods for the six data frequencies and the six application domains, covering the accuracy of the PFs and the precision of the PIs. This reporting is done in detail in the first four appendices of this paper (A–D), but this section presents the seven major findings of the M4 and discusses their implications for the field of forecasting. In addition, it provides two summary tables of the results, with Table 4 ranking the various methods according to the accuracy of the PFs (numerical accuracy in terms of sMAPE, MASE and OWA) and Table 5 ranking them according to the precision of the PIs (numerical precision in terms of MSIS and ACD). Three columns of Table 4 show the improvement of each of the submitted methods over the single Comb benchmark for the PFs and two columns of Table 5 show the improvement over the Naïve 1 for the PIs.

4.1. The seven major findings of the M4 competition

The M4 offers a wealth of information and many important findings. We present the seven that we consider to be the most important below, and allow the commentators of the paper to decide whether they agree with us or not. Hopefully, future papers exploring the M4 dataset, which is available to everyone, will add to these findings and to our understanding of the M4 Competition.

Finding 1: ***The improved numerical accuracy of combining.*** The most important finding of the M4 Competition was that all of the top-performing methods, in terms of both PFs and PIs, were combinations of mostly statistical methods, with such combinations being more

Table 4
The performances of the 61 methods (49 submitted, 10 benchmarks and 2 standards for comparison) in terms of the accuracy of PFs.

Author(s)	Affiliation	sMAPE	MASE	OWA	Rank			% improvement of method over the benchmark			Type of forecasting approach				
					sMAPE	MASE	OWA	sMAPE	MASE	OWA	Hybrid	Statistical	ML	Combination	Benchmark
Smyl	**Uber Technologies**	11.374	1.536	0.821	1	1	1	9.4	7.7	8.6	✔				
Montero-Manso, et al.	**University of A Coruña & Monash University**	11.720	1.551	0.838	3	3	2	6.6	6.7	6.7				✔	
Pawlikowski, et al.	**ProLogistica Soft**	11.845	1.547	0.841	5	2	3	5.7	6.9	6.3				✔	
Jaganathan & Prakash	**Individual**	11.695	1.571	0.842	2	6	4	6.8	5.5	6.2				✔	
Fiorucci & Louzada	**University of Brasilia & University of São Paulo**	11.836	1.554	0.843	4	4	5	5.7	6.6	6.1				✔	
Petropoulos & Svetunkov	**University of Bath & Lancaster University**	11.887	1.565	0.848	6	5	6	5.3	5.9	5.6				✔	
Shaub	**Harvard Extension School**	12.020	1.595	0.860	9	7	7	4.3	4.1	4.2				✔	
Legaki & Koutsouri	**National Technical University of Athens**	11.986	1.601	0.861	8	8	8	4.5	3.7	4.1		✔			
Doornik, et al.	**University of Oxford**	11.924	1.627	0.865	7	10	9	5.0	2.2	3.6				✔	
Pedregal, et al.	**University of Castilla-La Mancha**	12.114	1.614	0.869	10	9	10	3.5	2.9	3.2				✔	
Spiliotis & Assimakopoulos	**National Technical University of Athens**	12.148	1.628	0.874	12	11	11	3.2	2.1	2.7		✔			
Roubinchtein	**Washington State Employment Security Department**	12.183	1.633	0.876	13	13	12	3.0	1.8	2.4				✔	
Ibrahim	**Georgia Institute of Technology**	12.198	1.644	0.880	14	14	13	2.8	1.1	2.0		✔			
Tartu M4 seminar	**University of Tartu**	12.496	1.633	0.888	19	12	14	0.5	1.8	1.1				✔	
Waheeb	**Universiti Tun Hussein Onn Malaysia**	12.146	1.706	0.894	11	24	15	3.3	−2.6	0.4				✔	
Darin & Stellwagen	**Business Forecast Systems (Forecast Pro)**	12.279	1.693	0.895	15	22	16	2.2	−1.8	0.3		✔			
Dantas & Cyrino Oliveira	**Pontifical Catholic University of Rio de Janeiro**	12.553	1.657	0.896	21	15	17	0.0	0.4	0.2				✔	
Theta - Benchmark		12.309	1.696	0.897	16	23	18	2.0	−2.0	0.0		✔			✔
Comb - Benchmark		12.555	1.663	0.898	23	16	19	0.0	0.0	0.0				✔	✔
ARIMA - Standard for comparison		12.669	1.666	0.903	26	17	20	−0.9	−0.2	−0.5		✔			✔
Nikzad	**Individual**	12.370	1.724	0.907	17	27	21	1.5	−3.7	−1.0				✔	
Damped - Benchmark		12.661	1.683	0.907	25	20	22	−0.8	−1.2	−1.0		✔			✔
ETS - Standard for comparison		12.726	1.680	0.908	27	18	23	−1.4	−1.0	−1.2		✔			✔
Segura-Heras, et al.	**Universidad Miguel Hernández & Universitat de Valencia**	12.507	1.717	0.910	20	26	24	0.4	−3.3	−1.4				✔	
Trotta	**Individual**	12.894	1.682	0.915	29	19	25	−2.7	−1.1	−1.9			✔		
Chen & Francis	**Fordham University**	12.554	1.730	0.915	22	28	26	0.0	−4.0	−2.0				✔	
Svetunkov, et al.	**Lancaster University & University of Newcastle**	12.464	1.745	0.916	18	30	27	0.7	−4.9	−2.0				✔	
Talagala, et al.	**Monash University**	12.902	1.687	0.917	30	21	28	−2.8	−1.4	−2.1		✔			
Sui & Rengifo	**Fordham University**	12.855	1.743	0.930	28	29	29	−2.4	−4.8	−3.6				✔	
Kharaghani	**Individual**	13.063	1.716	0.930	31	25	30	−4.1	−3.2	−3.6				✔	
Smart Forecast	**Smart Cube (Smart Forecast)**	13.214	1.788	0.955	33	32	31	−5.2	−7.5	−6.3				✔	

(*continued on next page*)

Table 4 (*continued*).

Author(s)	Affiliation	sMAPE	MASE	OWA	Rank			% improvement of method over the benchmark			Type of forecasting approach				
					sMAPE	MASE	OWA	sMAPE	MASE	OWA	Hybrid	Statistical	ML	Combination	Benchmark
Wainwright, et al.	**Oracle Corporation (Crystal Ball)**	13.336	1.798	0.962	34	34	32	−6.2	−8.1	−7.2		✔			
Holt - Benchmark		13.775	1.772	0.971	42	31	33	−9.7	-6.6	−8.2		✔			✔
SES - Benchmark		13.087	1.885	0.975	32	36	34	−4.2	-13.3	−8.6		✔			✔
Valle dos Santos, et al.	**Individual**	13.820	1.789	0.977	43	33	35	−10.1	−7.6	−8.9				✔	
Reilly	**Automatic Forecasting Systems, Inc. (AutoBox)**	13.756	1.873	0.997	41	35	36	−9.6	−12.6	−11.0		✔			
Naïve 2 - Benchmark		13.564	1.912	1.000	36	38	37	−8.0	-15.0	−11.4		✔			✔
Iqbal, et al.	**Wells Fargo Securities**	14.312	1.892	1.022	46	37	38	−14.0	−13.8	−13.9		✔			
Fritschi	**Individual**	13.530	2.069	1.040	35	46	39	−7.8	−24.4	−15.8		✔			
Bontempi	**Université Libre de Bruxelles**	13.990	2.023	1.045	44	41	40	−11.4	−21.7	−16.4				✔	
Naïve 1 - Benchmark		14.208	2.044	1.058	45	42	41	−13.2	-22.9	−17.9		✔			✔
Bandara, et al.	**Monash University**	12.653	2.334	1.077	24	49	42	−0.8	−40.4	−20.0				✔	
Naïve S - Benchmark		14.657	2.057	1.078	49	44	43	−16.7	-23.7	−20.1		✔			✔
Patelis	**Individual**	14.430	2.098	1.081	47	47	44	−14.9	−26.2	−20.4				✔	
Kyriakides & Artusi	**University of Nicosia**	14.889	2.068	1.090	50	45	45	−18.6	−24.4	−21.4				✔	
Viole & Vinod	**Fordham University**	15.392	2.015	1.094	51	40	46	−22.6	−21.2	−21.9				✔	
Chirikhin & Ryabko	**Novosibirsk State University**	16.468	1.957	1.119	54	39	47	−31.2	−17.7	−24.6				✔	
Alves Santos Junior	**Individual**	16.638	2.056	1.151	55	43	48	−32.5	−23.7	−28.2			✔		
Mohamed	**ServiceNow**	15.901	2.310	1.190	53	48	49	−26.7	−38.9	−32.6				✔	
Clark	**Individual**	15.881	2.583	1.261	52	50	50	−26.5	−55.3	−40.5		✔			
Selamlar	**Dokuz Eylul University**	13.587	3.305	1.365	39	54	51	−8.2	−98.8	−52.1		✔			
Taylan	**Yapı Kredi Invest**	13.577	3.322	1.369	38	55	52	−8.1	−99.8	−52.5		✔			
Yapar, et al.	**Dokuz Eylul University**	13.573	3.371	1.382	37	57	53	−8.1	−102.7	−53.9		✔			
Yilmaz	**Dokuz Eylul University**	13.627	3.676	1.464	40	59	54	−8.5	−121.1	−63.1		✔			
Çetin	**Dokuz Eylul University**	14.489	3.576	1.469	48	58	55	−15.4	−115.0	−63.7		✔			
Mukhopadhyay	**University of Texas**	18.469	3.059	1.481	56	52	56	−47.1	−84.0	−65.0			✔		
RNN - Benchmark		21.152	2.685	1.482	59	51	57	−68.5	-61.4	−65.1			✔		✔
Pełka	**Czestochowa University of Technology**	19.573	3.341	1.595	58	56	58	−55.9	−100.9	−77.7			✔		
MLP - Benchmark		21.653	3.225	1.642	60	53	59	−72.5	-93.9	−82.9			✔		✔
Dudek	**Czestochowa University of Technology**	26.137	6.110	2.561	61	60	60	−108.2	−267.4	−185.3		✔			
Sirotin	**Siberian State University of Telecommunications and Information Sciences**	19.136	14.081	4.387	57	61	61	−52.4	−746.7	−388.8		✔			

Table 5
The performances of the 23 methods (20 submitted and 3 benchmarks) in terms of the precision of PIs.

Author(s)	Affiliation	MSIS	ACD	Rank		Cover rate (percentage of the time that future values were within the PIs)	% improvement over the benchmark		Type of forecasting approach				
				MSIS	ACD		MSIS	ACD	Hybrid	Statistical	ML	Combination	Benchmark
Smyl	**Uber Technologies**	12.230	0.002	1	1	94.78%	49.2%	97.4%	✔				
Montero-Manso, et al.	**University of A Coruña & Monash University**	14.334	0.010	2	2	95.96%	40.4%	88.8%				✔	
Doornik, et al.	**University of Oxford**	15.183	0.043	3	5	90.70%	36.9%	50.0%				✔	
ETS - Standard for comparison		15.679	0.037	4	4	91.27%	34.8%	56.6%		✔			✔
Fiorucci & Louzada	**University of Brasilia & University of São Paulo**	15.695	0.065	5	9	88.52%	34.8%	24.6%				✔	
Petropoulos & Svetunkov	**University of Bath & Lancaster University**	15.981	0.072	6	10	87.81%	33.6%	16.4%				✔	
Roubinchtein	**Washington State Employment Security Department**	16.505	0.061	7	8	88.93%	31.4%	29.4%				✔	
Talagala, et al.	**Monash University**	18.427	0.085	8	11	86.48%	23.4%	0.9%		✔			
ARIMA - Standard for comparison		18.681	0.092	9	14	85.80%	22.3%	−7.0%		✔			✔
Ibrahim	**Georgia Institute of Technology**	20.202	0.094	10	15	85.62%	16.0%	−9.1%		✔			
Iqbal, et al.	**Wells Fargo Securities**	22.001	0.086	11	12	86.41%	8.5%	0.1%		✔			
Reilly	**Automatic Forecasting Systems, Inc. (AutoBox)**	22.367	0.121	12	21	82.87%	7.0%	−41.1%		✔			
Wainwright, et al.	**Oracle Corporation (Crystal Ball)**	22.674	0.120	13	20	82.99%	5.7%	−39.6%		✔			
Segura-Heras, et al.	**Universidad Miguel Hernández & Universitat de Valencia**	22.717	0.049	14	6	90.10%	5.6%	43.1%				✔	
Naïve 1 - Benchmark		24.055	0.086	15	13	86.40%	0.0%	0.0%		✔			✔
Trotta	**Individual**	24.989	0.016	16	3	93.39%	−3.9%	81.3%			✔		
Alves Santos Junior	**Individual**	27.043	0.124	17	22	82.61%	−12.4%	−44.1%			✔		
Svetunkov, et al.	**Lancaster University & University of Newcastle**	28.139	0.056	18	7	89.40%	−17.0%	34.8%				✔	
Selamlar	**Dokuz Eylul University**	84.155	0.113	19	18	83.74%	−249.9%	−30.9%		✔			
Taylan	**Yapı Kredi Invest**	84.403	0.115	20	19	83.53%	−250.9%	−33.4%		✔			
Yapar, et al.	**Dokuz Eylul University**	86.644	0.108	21	16	84.19%	−260.2%	−25.7%		✔			
Çetin	**Dokuz Eylul University**	90.388	0.153	22	23	79.68%	−275.8%	−78.1%		✔			
Yilmaz	**Dokuz Eylul University**	98.478	0.111	23	17	83.85%	−309.4%	−29.6%		✔			

accurate numerically than either pure statistical or pure ML methods. More specifically, there was only one pure statistical method among the first ten most-accurate ones, as measured by PFs, while all of the ten least-accurate methods were either pure statistical or pure ML. Similarly, for PIs, there was only one pure statistical method among the first five most-accurate ones, while all of the five least-accurate methods were pure statistical.

Combining has long been considered as a useful practice in the forecasting literature (Bates & Granger, 1969; Clemen, 1989; Makridakis & Winkler, 1983). Over the years, the forecast combination puzzle (Claeskens, Magnus, Vasnev, & Wang, 2016), i.e., the fact that optimal weights often perform poorly in applications, has been examined both theoretically and empirically (Chan & Pauwels, 2018; Smith & Wallis, 2009), and many alternatives have been proposed in order to exploit the benefits of combining. Jose and Winkler (2008) used trimming and winsorizing to deal with the sensitivity of the mean operator to extreme values, and therefore its vulnerability to increasing the risk of significant errors. Kolassa (2011) used weights derived from Akaike and other information criteria to combine exponential smoothing point and interval forecasts. Assimakopoulos and Nikolopoulos (2000) proposed a framework for decomposing data into multiple Theta lines, each of which emphasizes a different time series characteristic, and which can then be extrapolated individually and combined to enhance the forecasting accuracy (Fiorucci, Pellegrini, Louzada, Petropoulos, & Koehler, 2016). The concept of temporal aggregation can also help in strengthening or attenuating the signals of different time series components (Kourentzes, Petropoulos, & Trapero, 2014), while also mitigating the modelling uncertainty through the combination that it implies (Athanasopoulos, Hyndman, Kourentzes, & Petropoulos, 2017). Bootstrapping with aggregation (bagging) has been proven to improve the forecasting performance by handling the data, model and parameter uncertainty simultaneously (Petropoulos, Hyndman, & Bergmeir, 2018). In this regard, the results of M4 re-confirm the benefits of combining and contribute to the literature by introducing new effective approaches for the optimal combination of forecasts.

The value of combining can be seen in Table 6, which lists the sMAPE, MASE and OWA of Single, Holt and Damped exponential smoothing, as well as those of their simple arithmetic average (Comb) across the entire M4 dataset. In this table, Comb provides more accurate PFs than the three individual methods being combined. In addition, it can be observed that its accuracy in M4 is almost identical to that of Theta, the most accurate method of the M3 Competition. Moreover, given that Theta is ranked 18th in the M4, we can conclude that the improvements achieved by combining these three exponential smoothing (ES) methods improved the forecasting accuracy significantly.

The higher numerical accuracy of combining, coupled with the poor performances of pure statistical/ML methods, confirms the findings of the previous three M Competitions, as well as those of other competitions/empirical studies. It implies that 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.

Table 6
Accuracy of the Single, Holt and Damped exponential smoothing PFs, as well as those of Comb, Theta, ETS and ARIMA, across the M4 dataset.

Method	sMAPE	MASE	OWA	Rank
SES	13.087	1.885	0.975	34
Holt	13.775	1.772	0.971	33
Damped	12.661	1.683	0.907	22
Comb	12.555	1.663	0.898	19
Theta	12.309	1.696	0.897	18
ETS	12.726	1.680	0.908	23
ARIMA	12.669	1.666	0.903	20

It is also worth mentioning that combining forecasting models, even simple ones like Comb, leads to better PFs than trying to select a "best" model for each series, as is done by methods like ETS and ARIMA, for instance. This becomes evident for the PIs as well, given that both ETS and ARIMA are outperformed by the top-performing combination schemes submitted to the competition. In addition, some single models provide more accurate PFs than complex selection approaches. For example, Damped, which is a particular case of ES, performs slightly better than ETS (being 0.1% and 0.5% better according to OWA and sMAPE, respectively, but 0.2% worse according to MASE), even though the latter selects the most appropriate ES model from a wide range of options. Furthermore, Theta provides more accurate PFs than Comb, ARIMA and ETS, despite utilizing a simple approach to forecasting. This demonstrates that the various sources of uncertainty that are included in the data, models and parameters, make model selection a challenging and difficult task, and that specific strategies, rules, tools and tests should be considered to mitigate their effect and enable improvements in forecasting accuracy (Petropoulos et al., 2018). The same is true for model combinations, where the pool of models should be considered carefully, along with their weighting scheme.

Finding 2: ***The superiority of a hybrid approach that utilizes both statistical and ML features.*** The biggest surprise of the M4 Competition was a new, innovative method that was submitted by Slawek Smyl, a data scientist at Uber Technologies, which mixes ES formulas with a recurrent neural network (RNN) forecasting engine. Smyl clarifies that his method does not constitute a simple ensemble of exponential smoothing and neural networks. Instead, the models are truly hybrid algorithms in which all parameters, like the initial ES seasonality and smoothing coefficients, are fitted concurrently with the RNN weights by the same gradient descent method. The improvement of this method over that of Comb was close to an impressive 10%, showing a clear advancement in the field of forecasting by exploiting the advantages of both statistical and ML methods, while also avoiding their shortcomings. A full description of this hybrid method is provided by the author (Smyl, 2020), who explains

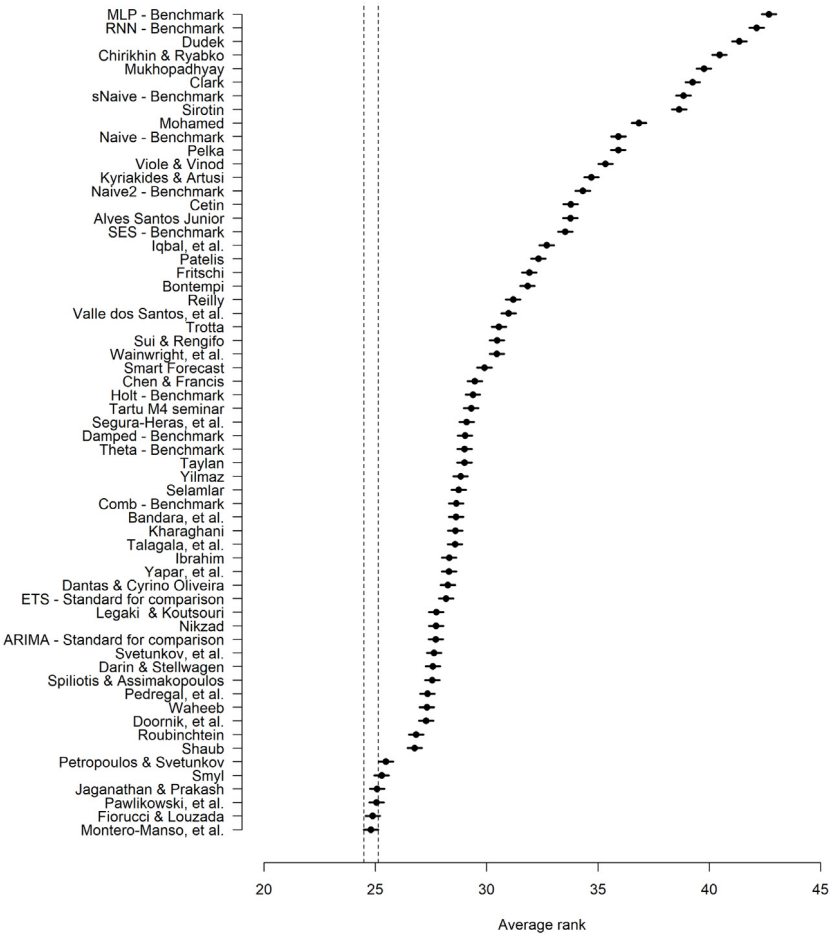

Fig. 1. Average ranks and 95% confidence intervals of 61 methods over all M4 series: multiple comparisons with the best (sMAPE used for ranking the methods) as proposed by Koning, Franses, Hibon, and Stekler (2005).

in detail its success based on the following two main components:

- The hierarchical nature of the selection process including criteria (parameters) for each time series and frequency aggregates of the series.
- Extending the machinery of RNNs with formulas taken from, or at least inspired by, the "battle tested" statistical models of Holt and Holt-Winters exponential smoothing methods.

Finding 3: ***The significant differences between the six top-performing methods and the rest in terms of PFs.*** The second most accurate method in terms of PFs was a combination of seven statistical methods and a ML one, with the weights that were utilized for the averaging being calculated by a ML algorithm that was trained to minimize the forecasting errors through holdout tests for each individual series. This method (Montero-Manso, Athanasopoulos, Hyndman, & Talagala, 2020) was submitted jointly by Spain's University of A Coruña and Australia's Monash University, where Montero-Manso is a Ph.D. student. The third to sixth methods were also combinations of statistical ones with no statistically significant differences in their accuracies, displaying OWA differences from the winning method of less than 0.05.

These differences are presented in Fig. 1, where the multiple comparisons with the best (MCB) test is applied to the rankings, based on sMAPE, of all methods submitted. The figure shows those with average ranks across all series that are not statistically different from those of others (Koning et al., 2005). In particular, if the intervals of two methods do not overlap, this indicates a statistically different performance; thus, methods that do not overlap with the dashed lines of Fig. 1 are significantly worse than the best, and vice versa. As has been observed, the methods of Montero-Manso and co-authors, Fiorucci and Louzada (2020), Jaganathan and Prakash (2020), Pawlikowski, Chorowska, and Yanchuk (2020) and Smyl do not differ statistically, while the method of Petropoulos and Svetunkov (2020) is very close to the others, and also displays a clear gap from the remaining ones. Note that similar conclusions are drawn and the same six methods are confirmed as being the dominant ones when MASE is

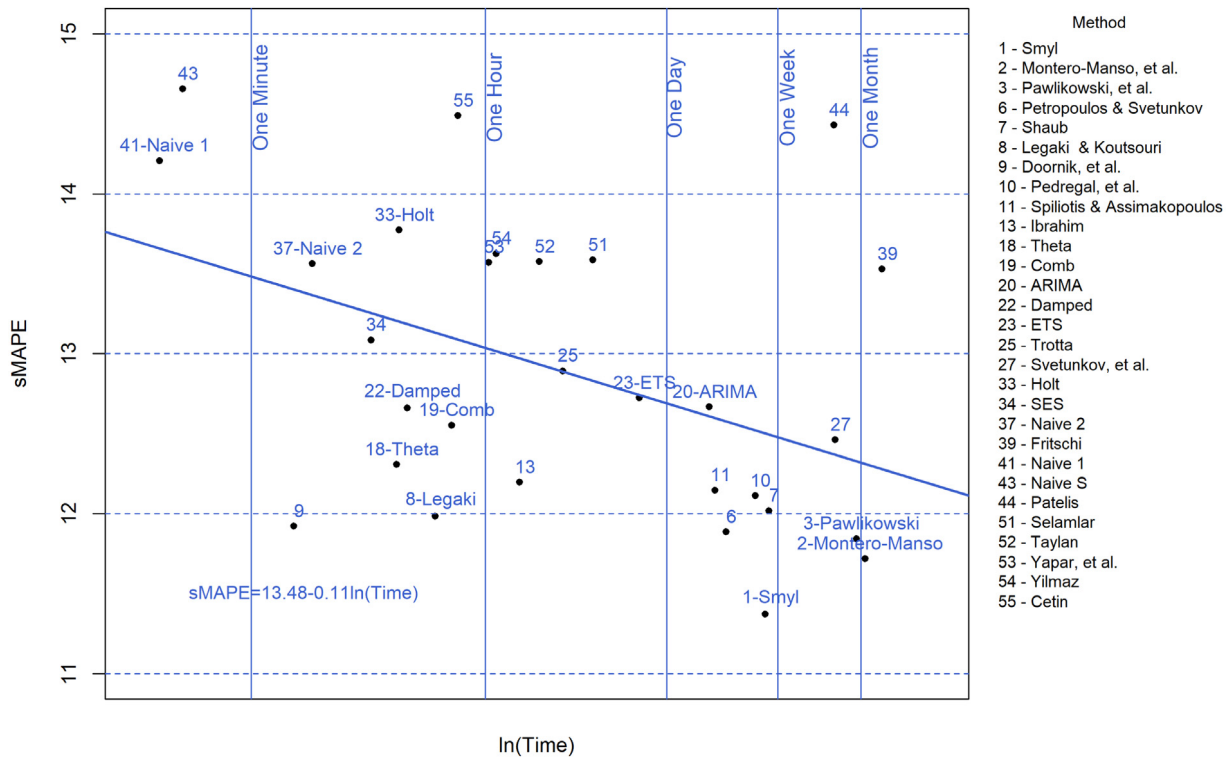

Fig. 2. The accuracy achieved (sMAPE) when replicating the computational time required to obtain PFs for all 100,000 time series by each method.

used for conducting the MCB test instead of sMAPE, given that the Spearman's rank correlation coefficient of the two measures is 0.90.

The small differences in forecasting accuracy that are reported for the six top-performing methods, as measured by PFs, point to advancements in the field of forecasting and the diffusion of knowledge to several researchers around the world who are able to apply such knowledge successfully to improve the accuracy of their predictions. We would expect that a further evaluation of the results of the M4 would allow additional improvements in accuracy to be achieved by studying the hybrid method, as well as the various forms of combinations of statistical and/or ML methods that have managed to exploit their advantages while avoiding their shortcomings.

Note that the M4 dataset consists of a huge number of series of different domains, frequencies, types and characteristics. Thus, we are confident that the diversity of the M4 series ensures the validity of the inferences reported above based on the MCB results, for both slow- and fast-moving data. However, it would be interesting in the future to examine the differences in the submitted forecasting methods further for subsets of particular characteristics, in order to identify the best practice for each case. This is exactly why the top performing methods for each frequency and domain were invited to submit papers to this special issue, presenting their forecasting approaches in detail.

Finding 4: *The improved precision of the PIs.* The most accurate and second most accurate methods of the M4 in terms of PFs also achieved amazing levels of success in specifying the 95% PIs precisely. These are the first methods that we are aware of that have done so, rather than underestimating the uncertainty considerably, as has been the case in the past (Makridakis, Hibon, Lusk, & Belhadjali, 1987), as well as with the remaining methods of the M4. Table 5 shows the astonishing precision of these two methods, which extends to all forecasting horizons and domains (see Appendix B), in addition to the numbers shown in this table. Therefore, we would expect to learn a great deal from Smyl and Montero-Manso and co-authors on how they achieved such phenomenal precision in specifying PIs. At the same time, the underestimation of uncertainty by the remaining methods must be addressed and corrected by learning from the best two methods.

Finding 5: *More complex methods can possibly lead to a greater forecasting accuracy.* The previous three M Competitions, and possibly other studies (Crone et al., 2011; Nikolopoulos & Petropoulos, 2018), found that the use of more computational time to apply more sophisticated and complex methods did not seem to improve the forecasting accuracy. As can been seen in Fig. 2 and Table 7, this has been reversed in the M4, where, on average, the accuracy of PFs improved as the computational time increased, signifying that processing power was exploited beneficially to achieve more accurate forecasts. There are some notable exceptions from pure ML methods and some results from pure statistical ones that are less accurate than expected.

Table 7
The methods for which forecasters provided code for replicating their submissions.

Method	sMAPE	MASE	OWA	Rank	Software	Replicability (%)	Replicability	Running time (min)
Smyl	11.374	1.536	0.821	1	C++	98.5	Full replicable	8056.0
Montero-Manso, et al.	11.720	1.551	0.838	2	R	99.5	Full replicable	46108.3
Pawlikowski, et al.	11.845	1.547	0.841	3	R	99.6	Full replicable	39654.8
Petropoulos & Svetunkov	11.887	1.565	0.848	6	R	99.5	Full replicable	4049.5
Shaub	12.020	1.595	0.860	7	R	100.0	Full replicable	8575.0
Legaki & Koutsouri	11.986	1.601	0.861	8	R	100.0	Full replicable	25.0
Doornik, et al.	11.924	1.627	0.865	9	OX 8	100.0	Full replicable	2.1
Pedregal, et al.	12.114	1.614	0.869	10	R	100.0	Full replicable	6742.6
Spiliotis & Assimakopoulos	12.148	1.628	0.874	11	R	100.0	Full replicable	3335.9
Roubinchtein	12.183	1.633	0.876	12	R	–	Not replicable	–
Ibrahim	12.198	1.644	0.880	13	R	100.0	Full replicable	109.6
Tartu M4 seminar	12.496	1.633	0.888	14	Python	–	Not replicable	–
Waheeb	12.146	1.706	0.894	15	R	–	Not replicable	–
Darin & Stellwagen	12.279	1.693	0.895	16	ForecastPro	–	Not replicable	–
Dantas & Cyrino Oliveira	12.553	1.657	0.896	17	R	Unknown	Unknown	> 2 months
Theta - Benchmark	12.309	1.696	0.897	18	R	100.0	Full replicable	12.7
Comb - Benchmark	12.555	1.663	0.898	19	R	100.0	Full replicable	33.2
ARIMA - Standard for comparison	12.669	1.666	0.903	20	R	100.0	Full replicable	3030.9
Nikzad	12.370	1.724	0.907	21	R	–	Not replicable	–
Damped - Benchmark	12.661	1.683	0.907	22	R	100.0	Full replicable	15.3
ETS - Standard for comparison	12.726	1.680	0.908	23	R	100.0	Full replicable	888.8
Segura-Heras, et al.	12.507	1.717	0.910	24	.exe	–	Not replicable	–
Trotta	12.894	1.682	0.915	25	Python	97.9	Replicable	232.1
Svetunkov, et al.	12.464	1.745	0.916	27	R	99.6	Full replicable	27432.3
Holt - Benchmark	13.775	1.772	0.971	33	R	100.0	Full replicable	13.3
SES - Benchmark	13.087	1.885	0.975	34	R	100.0	Full replicable	8.1
Naïve 2 - Benchmark	13.564	1.912	1.000	37	R	100.0	Full replicable	2.9
Iqbal, et al.	14.312	1.892	1.022	38	SAS	Not tested	Unknown	–
Fritschi	13.530	2.069	1.040	39	R	100.0	Full replicable	62242.9
Bontempi	13.990	2.023	1.045	40	R	–	Not replicable	–
Naïve 1- Benchmark	14.208	2.044	1.058	41	R	100.0	Full replicable	0.2
Bandara, et al.	12.653	2.334	1.077	42	R & Python	–	Not replicable	–
Naïve S - Benchmark	14.657	2.057	1.078	43	R	100.0	Full replicable	0.3
Patelis	14.430	2.098	1.081	44	VB.net	–	Not replicable	26719.2
Chirikhin & Ryabko	16.468	1.957	1.119	47	Python	–	Not replicable	–
Selamlar	13.587	3.305	1.365	51	R/.exe	100.0	Full replicable	393.5
Taylan	13.577	3.322	1.369	52	.exe	100.0	Full replicable	154.8
Yapar, et al.	13.573	3.371	1.382	53	.exe	100.0	Full replicable	63.6
Yilmaz	13.627	3.676	1.464	54	.exe	100.0	Full replicable	72.5
Çetin	14.489	3.576	1.469	55	.exe	100.0	Full replicable	37.2
Mukhopadhyay	18.469	3.059	1.481	56	SAS	Not tested	Unknown	–
RNN - Benchmark	21.152	2.685	1.482	57	Python	100.0	Full replicable	64857.10
Pełka	19.573	3.341	1.595	58	Matlab	Not tested	Unknown	–
MLP - Benchmark	21.653	3.225	1.642	59	Python	100.0	Full replicable	1484.37
Dudek	26.137	6.110	2.561	60	Matlab	Not tested	Unknown	–
Sirotin	19.136	14.081	4.387	61	Python	Not tested	Unknown	–

Note: In addition to each method's accuracy (sMAPE, MASE, OWA and Rank), the table presents the software used for generating its forecasts, its running time and its replicability rate.

Note that a total of four Amazon servers of the following characteristics have been used for a total of six months for replicating the results of the competition (the cost of which was generously paid by the University of Nicosia): 8 cores, 16 GB RAM, 500 GB, HDD, Linux Ubuntu 16.10. The replicability of each method was found by estimating the difference in the sMAPE between the original and the reproduced by the organizers forecasts. We believe that, since many methods involve random initializations, any method that displays a replicability rate higher than 98% should be considered as fully replicable, while any below that are considered as partially replicable (the number was selected arbitrarily, believing that other alternatives would be either too strict or too flexible). The methods for which the submitted code was impossible to replicate, due to either problems in trying to run the submitted code or unreasonable execution times, are also reported in Table 7. Note that, before classifying a method as not being replicable, we either tried to fix the errors returned by the code ourselves, or contacted the original author(s) for support. Finally, given the effort required to test all of the submitted methods effectively, we decided to exclude from the analysis any method which performed worse than the Naïve 2 method and was implemented using packages that we were not familiar with.

Finding 6: ***Using information from multiple series to predict individual ones.*** For the first time, the top three performing methods of the M4, as measured by PFs, introduced information from multiple series (aggregated by data frequency) in order to decide on the most effective way of forecasting and/or selecting the weights for combining the various statistical/ML methods considered. It must be investigated whether such information was

part of the reason why these three methods managed to perform better that the remaining ones, and if so, to what extent. For instance, Trotta also introduced a pure ML method which exploited information from multiple series to extrapolate the individual ones, but it was not among the top performers. However, the performance of his method was significantly better than those of the rest of the pure ML methods, highlighting the potential of this "cross-learning" approach.

Finding 7: ***The poor performances of the submitted pure ML methods.*** A total of five ML methods (four pure and one combination) were submitted in M4. All of them were less accurate than the Comb benchmark in terms of PFs, and only one was more accurate than the Naïve 2 method. These results are in complete agreement with a paper we published in PLOS ONE (Makridakis, Spiliotis, & Assimakopoulos, 2018a) which proved empirically the poor performances of standard ML methods, as well as with the findings of past forecasting competitions dedicated to NN and computational intelligence methods (Crone et al., 2011). We believe that once such a performance is accepted and the reasons behind it are understood, the accuracy of pure ML methods will be improved substantially, as there is no reason for such complex methods to be less accurate than simple statistical benchmarks, at least once their shortcoming of over-fitting is corrected. The reasoning behind the failure of the submitted ML methods, based on their description and results, could be the limited usage of cross-learning, i.e., using information from multiple series to predict individual ones, together with the limited sample sizes of the individual series, the effect of over-fitting and the non-stationarity of the data. On the other hand, given the paucity of entries in the ML category, more information might still be needed to make it safe to generalize this finding. Thus, the experts in the field are encouraged to experiment further with their ML approaches and report their performances in order to help the forecasting community learn more about their actual potential and possible limitations.

4.2. Discussing the findings and their implications

The above findings confirm

- The poor performances of pure statistical and ML methods;
- The superior accomplishment of the hybrid approach;
- The improved accuracy of combinations of statistical and/or ML methods;
- The effective usage of processing power to improve the accuracy of PFs and the precision of PIs;
- The beneficial usage of information from aggregates of series to improve the accuracy of individual ones; and finally,
- The amazing precision of the PIs of the top two methods and the dreadful performances of the remaining ones.

From all this, we can conclude that the way forward is to exploit the advantages of both statistical and ML methods while avoiding their drawbacks, probably by combining them and extending such combinations with hybrid forecasting frameworks. Moreover, more work is needed to obtain a better understanding of the most effective way of choosing which methods to combine and how to estimate the optimal weights of the methods being combined. We can speculate that the success of the method of Montero-Manso and collaborators is due, at least to some extent, to its ability to determine the optimal weights for combining the various methods being used by means of ML algorithms and a significant amount of computing power, as the optimization was done for each series.

The seven major findings presented in this section are not the only possible ones. Our hope and expectation is that, now that all of the data, PFs and PIs are available readily in the public domain, researchers will study them and discover additional findings or challenge some of our conclusions. The 3,003 time series of the M3 Competition have been used widely, leading to additional findings and the testing of new methods. We would envisage that the 100,000 series of the M4 will provide an even more fertile data ground for further research and new findings.

4.2.1. Accuracy across the four M competitions

Table 8 shows the numerical accuracies of PFs (MAPE or sMAPE) from four standard forecasting methods across the four M Competitions, plus the most accurate method in each one. By using relative measures (Naïve 2 is used in order to scale the forecasting errors and enable direct comparisons between the individual datasets), and assuming that the main principles of the time series do not differ significantly across the datasets (Spiliotis, Kouloumos et al., 2020), some revealing observations can be made from this table.

- M1 and M2 saw the accuracy of DSES, an extremely simple method that requires little computational time, being higher than that of the statistically sophisticated and highly popular ARIMA models, referred to at that time as the Box-Jenkins methodology for forecasting (Box & Jenkins, 1970). This methodology claimed that an optimal model existed for each time series, and that it could be identified by using their suggested approach to identify, estimate and test the adequacy of such a model. However, this assumed that the best model fit to past data would also assure the most accurate forecasts for the future. As a consequence, it took a very long time for theoretical statisticians to accept the results of empirical studies like those of M1, M2 and M3 and realize that goodness of fit was not necessarily related to post-sample accuracy. In fact, often the opposite is true (see for instance Fig. 3 of Makridakis et al., 2018a). Such was also the case for combinations of methods, which have repeatedly been proven empirically to be more accurate than the theoretically correct model that is fitted to historical data.
- We observe that ARIMA became more accurate than DSES by a small margin (∼2%) in M3, while that margin increased (∼3%) in M4. It seems that the

Table 8
The accuracies of four standard forecasting methods across the four M Competitions, plus the best one in each competition.

Methods	MAPE		sMAPE	
	M1: 1982	M2: 1993	M3: 2000	M4: 2018
Naïve 2	17.8 *(1.00)*	13.3 *(1.00)*	15.5 *(1.00)*	13.6 *(1.00)*
DSES	16.8 *(0.94)*	11.9 *(0.89)*	14.3 *(0.92)*	13.1 *(0.96)*
Damped	*NA*	12.8 *(0.96)*	13.7 *(0.88)*	12.7 *(0.93)*
Comb	*NA*	11.7 *(0.88)*	13.5 *(0.87)*	12.6 *(0.93)*
ARIMA	18.0 *(1.01)*	16.0 *(1.20)*	14.0 *(0.90)*	12.7*(0.93)*
Best method	Parzen (ARARMA)	Comb	Theta	Hybrid
	15.4 *(0.87)*	11.0 *(0.83)*	13.0 *(0.84)*	11.4 *(0.84)*
% improvement of the best method over the Naïve 2	13.5	17.3	15.9	16.2
% improvement of the best method over the Comb	*NA*	0.0	3.8	9.4
Number of series	111	23	3,003	100,000

Note: The parentheses provide the relative MAPE/sMAPE, considering the Naïve 2 method as a benchmark.

accuracy of the ARIMA models improved over time as those using it identified its shortcomings, and realized and mitigated its major problems, such as over-fitting. Undoubtedly, further research must be conducted to validate this claim, but we would expect that something similar would occur with pure ML methods in the future by likewise minimizing their drawbacks.

- It is interesting that the ARIMA models are more accurate numerically than the ETS models in the M4. Even though the improvement is small, and not statistically significant according to Fig. 1, this contradicts previous studies, while also confirming the accuracy improvement of the former over time. Moreover, since both ARIMA and ETS are automatic model selection algorithms with multiple settings, it would be interesting to investigate the sensitivity of their performances when their parameters are changed, and to determine whether the improvement reported is due to the superiority of the ARIMA models over the ES ones, or to the default settings used for selecting and training the individual models. For instance, it could be that the efficient order selection mechanism introduced by Hyndman and Khandakar (2008) has contributed significantly to the improvements reported over time, especially relative to previous Box-Jenkins implementations.
- The accuracies of Damped and Comb in M3 and M4 in terms of sMAPE are very similar both between themselves and across the two competitions, indicating some kind of stability in their relative and total performances. However, their improvements over the Naïve 2 benchmark are 12% and 7%, respectively, demonstrating that, despite being similar in terms of forecasting performances, their improvements over standard alternatives might be influenced by the time series characteristics of the dataset under examination, such as the seasonality and trend.
- The numerical accuracies of Damped, Comb and ARIMA are very similar in M4, but the accuracy of ARIMA is still the same as that of Damped (a simple method) while Comb is a little more accurate than ARIMA. On the other hand, Fig. 1 indicates that hardly any statistically significant differences among the methods can be found. This implies that there is probably no such thing as a "best" model, and that theoretical postulates do not seem to apply in practice, as there is no consistency between goodness-of-fit and post-sample accuracy.
- Each competition has a new best-performing method. In the M2, the best method was the Comb; in M3, it was Theta; and in M4, it was the hybrid method of Smyl. The accuracy improvements of these latter methods over that of Comb is 3.5% for the M3, while the corresponding improvement for M4 is much greater (9.4%). This indicates an improvement over time that increases with Smyl's method, probably due to its exploitation of the advantages of both statistical and ML features and its avoidance of their disadvantages.
- What is worth noting is the improvement of the best method over Naïve 2, which is amazingly consistent over the four M Competitions, indicating some kind of stability over a period spanning more than three and a half decades.

4.2.2. Forecasting accuracy versus computational time

Fig. 2 shows the average accuracies (sMAPE) reported for the PFs of the reproducible methods of M4 (see Table 7) for all 100,000 time series versus the total computational time required for their estimation in logarithmic scale. It can be observed that, on average, there is a clear negative relationship between the two variables, indicating that, as the computational time required increases, i.e., the model sophistication rises, the post-sample sMAPE becomes smaller, that is, the accuracy improves. This average relationship (for the number of methods included) is equal to

$$sMAPE = \underset{(0.29)}{13.48} - \underset{(0.04)}{0.11} \log(Computational\ Time),$$

where the parentheses below the coefficients provide the standard errors of the parameters to prove their statistical significance.

The equation indicates that an increase in computational time leads on average to a decrease in sMAPE; more specifically, if the computational time increases by 1%, the sMAPE will decrease by 0.0011 units on average. Similar relationships can be observed for the MASE and OWA measures, and therefore they are not reported here for reasons of brevity. We merely note that the

Spearman's rank correlation coefficients between sMAPE and these measures are 0.90 and 0.98, respectively. Note also that the two methods (MLP and RNN) that displayed extremely high sMAPE values were excluded from the present analysis as outliers, in order to ensure that the scale of the figure and the conclusions of the analysis were meaningful.

If methods are below the trend line, this indicate that their accuracy is better than the average based on their computational time, and vice versa. For instance, the methods of Trotta, ETS and ARIMA are very close to the average line, those of Doornik, Legaki & Koutsouri, Smyl, Theta and Ibrahim are more accurate than the average time they utilize, and those of Fritschi and Cetin show the opposite performance. Such an analysis can provide useful insights in regard to the utilization of simple yet effective forecasting methods, keeping in mind that a lower accuracy might not always matter in many applications of businesses and organizations when predictions for large numbers of items are required, such as in inventory forecasting (Nikolopoulos & Petropoulos, 2018). Given the information provided in Fig. 2 and the precise relationship between sMAPE and the average computational time needed to run each method, it would be possible to calculate the cost of utilizing each method and to be able to balance such a cost against the expected accuracy.

It should be noted that no such relationship was present in the previous M Competitions, indicating that the M4 methods utilized processing power effectively to improve the forecasting accuracy. Hopefully this relationship will continue to improve in the future, increasing the accuracy even further as computers become faster and more affordable.

It should be understood that the computational times estimated for the methods considered may depend strongly on the efficiency of the code utilized by the authors. For instance, approaches exploiting multiple CPUs might require less computational time than much simpler ones implemented in a non-parallel fashion. Thus, although computational time can be used as a proxy for model complexity, it does not guarantee that time-intensive methods are also more complex. This was one of the main reasons why efficiency, although of paramount importance in real-life forecasting applications, was not used as a criterion for evaluating the submitted methods (the aim of the competition was to learn how to improve the forecasting accuracy, not to assess the programming skills of the forecasters). However, this does not mean that such information is not important for practical purposes, and additional measures such as the actual computational cost (money required for renting a server and running the method) could be introduced in future studies in order to capture the efficiency of the proposed forecasting methods better and to reach conclusions regarding their actual sophistication.

5. Conclusions and next steps

What have we learned from the M4 Competition and how can such learning be used to improve the theory and practice of forecasting? Answering these two questions is of fundamental scientific importance, and, although much further work must and hopefully will be performed to analyze/digest the results of the M4 competition, some preliminary conclusions are attempted next.

We must first emphasize again that 100,000 time series is a formidable sample size, and being able to evaluate 49 submissions from top forecasters and 12 benchmarks/standard methods is a colossal achievement, allowing us to draw definite conclusions and to have a high level of confidence that such conclusions are valid and can be applied to advance the theory and improve the practice of forecasting, at least as far as generic forecasting algorithms are concerned. Thus, going beyond the seven major findings discussed above, let us summarize what we believe we have learned and ways in which this knowledge can be exploited/applied successfully, keeping in mind that our conclusions will be challenged and possibly re-evaluated in the future, as more researchers and practitioners examine the diverse dataset of M4 in greater detail.

The two indispensable forecasting tasks: What we have been calling PFs are insufficient and often dangerous to utilize for decision making on their own without using their complement, the PIs, to specify the uncertainty around the PFs clearly and unambiguously (Bermúdez, Segura, & Vercher, 2010; Ord, Koehler, & Snyder, 1997). PIs express the conviction that only "prophets", who unfortunately do not exist anymore, can predict the future with perfect certainty. Hitherto, most forecasting methods have specified overly narrow PIs, thus underestimating the uncertainty (Makridakis et al., 1987). Two of the M4 methods, which happened to be the best and second-best estimated PIs, with an amazing degree of precision, contribute significantly to the alleviation of the underestimation problem once the approach behind their success is understood clearly and applied widely. Determining inventory levels correctly (Petropoulos, Wang, & Disney, 2019; Prak & Teunter, 2019) and assessing the risk of investments realistically (Ardia, Bluteau, Boudt, & Catania, 2018; Moore, Boehm, & Banerji, 1994) are just two of the numerous applications that require the estimation of the uncertainty and the identification of robust forecasting approaches (Spiliotis, Nikolopoulos & Assimakopoulos, 2019), providing a major, practical contribution to decision and policy makers.

The new generation of global forecasters: The main contributors (except in two cases) of the top sixteen methods that had accuracies which exceeded that of Comb, as measured by PFs, were mostly young forecasters, often newcomers in the field. Moreover, they originated from nine different countries (USA, Spain, Australia, Poland, Brazil, UK, Greece, Estonia, and Malaysia) in four continents. We are pleased that forecasting displays such a widespread utilization around the world and that new forecasters are contributing to the field with innovative and cutting-edge forecasting approaches. It is also worth mentioning that, with the exception of the best hybrid method, the performances of the remaining five top methods were close to each other, with all of them utilizing combinations of

similar forecasting methods. However, the element that distinguished the most accurate methods from the rest was the way in which they combined the PFs of the individual methods being utilized. For instance, the three top-performing methods exploited the processing power of computers to estimate the weights of combining optimally, in some cases using advanced ideas borrowed from ML algorithms. Furthermore, they all used information from the entire sample of series, based mainly on the similarities in data frequencies, to predict individual ones more accurately, thus opening some promising avenues for improving the forecasting accuracy further.

Continuing to advance the theory of forecasting: The old, mistaken belief was that there is an optimal model for each real-life time series. M4 and the previous competitions/empirical studies have questioned such a belief and opened new directions for future work, placing an emphasis on which models produced good forecasts, rather than on their mathematical properties or popularity, and separating the task of forecasting from time series analysis or computer science. M4 confirms that there are no limits to innovative ways of combining: it is not just averaging the predictions of individual methods. The best example is Smyl's hybrid approach, which blended exponential smoothing and ML concepts into a single model of remarkable accuracy. The next example was Montero-Manso et al.'s idea of using a ML method to estimate the weighting scheme for combining their various methods, while also exploiting information from multiple series in order to predict individual ones more accurately. In our view, as the results of M4 are evaluated further by researchers, additional ways of combining will be experimented with by exploiting the advantages of various approaches, while also avoiding their drawbacks. We believe strongly that such ideas will open fruitful directions for future theoretical research to better understand the reasons involved.

Expanding/improving the practice of forecasting: Forecasting has multiple uses, from predicting the demand for hundreds of thousands of inventory items (Syntetos, Nikolopoulos, & Boylan, 2010) to forecasting long term trends and their implications for strategic planning and strategy formulation (Jakubovskis, 2017). The 100,000 M4 series cover most types of forecasting situations: from hourly data to yearly ones, and from micro applications to macro and demographic ones. A positive development for practitioners and small organizations is the fact that the forecasting software for replicating the forecasts of the majority of the participating methods is available readily, for free, to anyone who is interested in using them in their organization. The M4 GitHub repository provides the code for replicating the results of the competition, most of which are written for open-access programming languages like R and Python. Moreover, Table 7 shows the computational times required for running the various forecasting methods, thus providing useful guidelines for deciding which method to use, given their accuracy in terms of PFs and computational requirements that can be adopted to users' specific needs. In addition, the ten benchmark methods listed in Table 2 are an excellent, easy way to start in applying forecasting, as they are all ready to use and easy to implement through the *forecast* package for R. A good illustration of the application of forecasting is that of a group of students at the University of Tartu in Estonia, Ingel, Shahroudi, Kangsepp, Tattar, Komisarenko, and Kull (2020), who, despite not having any forecasting experience prior to attending a forecasting seminar, submitted their PFs guided by their professor and, surprisingly, managed to rank 14th overall and second in daily data, beating many far more experienced forecasters.

Pure ML forecasting methods: expectations and reality: The pure ML forecasting methods in the M4 definitely performed below expectations. As was mentioned earlier, all ML methods submitted were less accurate than the Comb benchmark, and only one was more accurate than the Naïve 2. This poor performance was not consistent with the findings of other ML studies (Hamzaçebi, Akay, & Kutay, 2009; Salaken, Khosravi, Nguyen, & Nahavandi, 2017; Zhang, Patuwo, & Hu, 1998) that have been published in well-known ML and neural network journals, which have found excellent levels of accuracy. At the same time, it confirms the finding of Makridakis et al. (2018a) that the most accurate ML method was less accurate than the worst statistical ones. We believe that there is an urgent need for the results of important ML studies claiming superior forecasting performance to be replicated/reproduced, and therefore call for such studies to make their data and forecasting algorithms publically available, which is not currently the case (see Makridakis et al., 2018). This would allow the literature to separate hype from reality and introduce an objective basis for properly assessing the forecasting accuracies of pure ML methods, proving that the reported improvements are real and the methods more accurate than simple benchmarks. Revised guidelines for authors and reviewers, large open-access datasets, adequate descriptions of methods, software availability, clearly defined error measures, the use of fair benchmarks and a mitigation of advocacy and bias are some of the proposed ways forward.

Best model versus combining against single methods: A consistent finding of forecasting competitions and other empirical studies (Bates & Granger, 1969; Clemen, 1989; Smith & Wallis, 2009) has been a higher accuracy of combining than of single methods, and this has been confirmed in M4, in terms of both PFs and PIs. The difficulty of separating the pattern from the noise means that the selection of a "best" model for each series is a difficult task that could lead to over-fitting if some of the noise is included in the pattern, leading to an inability to identify the most appropriate model. On the other hand, combining several forecasting models/methods cancels random errors, generally resulting in more accurate forecasts (Chan & Pauwels, 2018). However, combining does not guarantee that their forecasts will outperform single models. For instance, this is the case in Table 6 for Comb, which, despite providing better PFs than DSES, Holt and Damped, is outperformed by Theta, a single, simple method. In addition, Damped is more accurate than ETS, despite the latter selecting the optimal model out of 32

exponential smoothing ones. This means that a single method, Damped, one of the 32, can be more accurate than the optimal ETS selection.

The reasoning behind this could be the fact that "optimal" is quite a relative term, and is related closely to the objectives of the forecaster. For instance, ETS and ARIMA select the "best" model by minimizing an information criterion, such as the AIC, the parameters of which are defined typically by minimizing the in-sample one-step-ahead squared error. In other words, ETS and ARIMA select the simplest model that is still good at producing accurate short-term forecasts. As a result, it becomes evident that the constructed model could be far from optimal if the forecasting horizon considered is much greater than one period and the measure used for evaluating the forecasts differs from the one used for training. This indicates that further research is needed in the field of automatic forecasting algorithms, but that such research could make them much more competitive in the future.

Black Swans and fat tails: The first three M Competitions, and almost all major empirical studies comparing the performances of forecasting methods, have been concerned mainly with PFs, and little or not at all with PIs and uncertainty. The studies of Athanasopoulos et al. (2011) and Hong, Pinson, Fan, Zareipour, Troccoli, and Hyndman (2016) are two interesting exceptions that emphasized both point forecast accuracy and forecast interval coverage and probabilities. M4 contributed further in that direction, evaluating the PIs of 23 methods across a formidable sample of series. However, no forecasting competitions have ever examined Black Swans (Makridakis & Taleb, 2009), Gray Rhinos (Wucker, 2016), or fat, non-normally distributed data (Coleman & Swanson, 2007; Taleb, 2009). Although this was outside the scope of the M4 Competition, we strongly believe that another special competition is needed to cover such phenomena and analyze their implications for risk management and decision making (Makridakis, Hogarth, & Gaba, 2009).

The M Competitions, including its predecessor, the Makridakis and Hibon (1979) study, began 39 years ago and have continued through to the M4, which is the largest and most ambitious to date. We would like to believe that M4 will become the most beneficial forecasting competition for both the theory and practice of forecasting. As has been stated, we believe that its 100,000 series will become the standard testing ground for evaluating theoretical developments, testing new methods and determining the value that they add. Most importantly, we would expect to learn a great deal about PFs and PIs by analyzing the large volume of results that have been made available to everyone, and thus to manage to advance the theory of forecasting. On the side of the practice of forecasting, the two largest benefits will first come from Fig. 2, which displays sMAPE values versus the computational time required to run each method. Such a graph, together with those of Appendices A, B and C, allows practitioners to select the most appropriate method(s) for their forecasting needs, considering the cost of running each method and its expected accuracy. The second advantage will come from being able to utilize the ready-to-use code that is available in the M4 GitHub repository, enabling practitioners, companies, organizations and researchers to run the method of their choice. One thing that remains to be determined is the possible improvement in PF and PI performances that can be achieved by expanding time series forecasting to include explanatory/exogenous variables. This element could be explored in future M Competitions, thus expanding time series forecasting competitions in a new and ambitious direction.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ijforecast.2019.04.014.

References

Ardia, D., Bluteau, K., Boudt, K., & Catania, L. (2018). Forecasting risk with Markov-switching GARCH models: a large-scale performance study. *International Journal of Forecasting*, *34*(4), 733–747.

Armstrong, J. S., & Collopy, F. (1992). Error measures for generalizing about forecasting methods: empirical comparisons. *International Journal of Forecasting*, *8*(1), 69–80.

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

Athanasopoulos, G., Hyndman, R. J., Kourentzes, N., & Petropoulos, F. (2017). Forecasting with temporal hierarchies. *European Journal of Operational Research*, *262*(1), 60–74.

Athanasopoulos, G., Hyndman, R. J., Song, H., & Wu, D. C. (2011). The tourism forecasting competition. *International Journal of Forecasting*, *27*(3), 822–844.

Bates, J. M., & Granger, C. W. J. (1969). The combination of forecasts. *The Journal of the Operational Research Society*, *20*(4), 451–468.

Bermúdez, J. D., Segura, J. V., & Vercher, E. (2010). Bayesian Forecasting with the Holt-Winters model. *The Journal of the Operational Research Society*, *61*(1), 164–171.

Box, G., & Jenkins, G. (1970). *Time series analysis: forecasting and control*. San Francisco: Holden-Day.

Chan, F., & Pauwels, L. L. (2018). Some theoretical results on forecast combinations. *International Journal of Forecasting*, *34*(1), 64–74.

Chen, C., Twycross, J., & Garibaldi, J. M. (2017). A new accuracy measure based on bounded relative error for time series forecasting. *PLoS One*, *12*(3), 1–23.

Claeskens, G., Magnus, J. R., Vasnev, A. L., & Wang, W. (2016). The forecast combination puzzle: a simple theoretical explanation. *International Journal of Forecasting*, *32*(3), 754–762.

Clemen, R. (1989). Combining forecasts: a review and annotated bibliography with discussion. *International Journal of Forecasting*, *5*(4), 559–608.

Coleman, C. D., & Swanson, D. A. (2007). On MAPE-R as a measure of cross-sectional estimation and forecast accuracy. *Journal of Economic and Social Measurement*, *32*(4), 219–233.

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*(3), 635–660.

Davydenko, A., & Fildes, R. (2013). Measuring forecasting accuracy: the case of judgmental adjustments to SKU-level demand forecasts. *International Journal of Forecasting*, *29*(3), 510–522.

Fiorucci, J. A., & Louzada, F. (2020). GROEC: combination method via generalized rolling origin evaluation. *International Journal of Forecasting* *36*(1), 105–109.

Fiorucci, J. A., Pellegrini, T. R., Louzada, F., Petropoulos, F., & Koehler, A. B. (2016). Models for optimising the theta method and their relationship to state space models. *International Journal of Forecasting*, *32*(4), 1151–1161.

Franses, P. H. (2016). A note on the mean absolute scaled error. *International Journal of Forecasting*, *32*(1), 20–22.
Gardner, E. S. (1985). Exponential smoothing: the state of the art. *Journal of Forecasting*, *4*, 1–28.
Gardner, E. S. (2006). Exponential smoothing: The state of the art – Part II. *International Journal of Forecasting*, *22*(4), 637–666.
Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, *102*(477), 359–378.
Goodwin, P., & Lawton, R. (1999). On the asymmetry of the symmetric MAPE. *International Journal of Forecasting*, *15*(4), 405–408.
Hamzaçebi, C., Akay, D., & Kutay, F. (2009). Comparison of direct and iterative artificial neural network forecast approaches in multi-periodic time series forecasting. *Expert Systems with Applications*, *36*(2), 3839–3844.
Hong, T., Pinson, P., Fan, S., Zareipour, H., Troccoli, A., & Hyndman, R. J. (2016). Probabilistic energy forecasting: Global Energy Forecasting Competition 2014 and beyond. *International Journal of Forecasting*, *32*(3), 896–913.
Hyndman, R. J. (2017). *forecast: Forecasting functions for time series and linear models. R package version 8.2*.
Hyndman, R. J. (2020). A brief history of forecasting competitions. *International Journal of Forecasting* *36*(1), 7–14.
Hyndman, R., & Khandakar, Y. (2008). Automatic time series forecasting: the forecast package for R. *Journal of Statistical Software*, *26*, 1–22.
Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, *22*(4), 679–688.
Hyndman, R. J., Koehler, A. B., Snyder, R. D., & Grose, S. (2002). A state space framework for automatic forecasting using exponential smoothing methods. *International Journal of Forecasting*, *18*(3), 439–454.
Ingel, A., Shahroudi, N., Kangsepp, M., Tattar, A., Komisarenko, V., & Kull, M. (2020). Correlated daily time series and forecasting in the M4 competition. *International Journal of Forecasting*, *36*(1), 121–128.
Jaganathan, S., & Prakash, P. (2020). Combination based forecasting method: M4 competition. *International Journal of Forecasting* *36*(1), 98–104.
Jakubovskis, A. (2017). Strategic facility location, capacity acquisition, and technology choice decisions under demand uncertainty: robust vs. non-robust optimization approaches. *European Journal of Operational Research*, *260*(3), 1095–1104.
Jose, V. R. R., & Winkler, R. L. (2008). Simple robust averages of forecasts: some empirical results. *International Journal of Forecasting*, *24*(1), 163–169.
Kim, S., & Kim, H. (2016). A new metric of absolute percentage error for intermittent demand forecasts. *International Journal of Forecasting*, *32*(3), 669–679.
Kolassa, S. (2011). Combining exponential smoothing forecasts using Akaike weights. *International Journal of Forecasting*, *27*(2), 238–251.
Kolassa, S. (2016). Evaluating predictive count data distributions in retail sales forecasting. *International Journal of Forecasting*, *32*(3), 788–803.
Koning, A. J., Franses, P. H., Hibon, M., & Stekler, H. O. (2005). The M3 competition: statistical tests of the results. *International Journal of Forecasting*, *21*(3), 397–409.
Kourentzes, N., Petropoulos, F., & Trapero, J. R. (2014). Improving forecasting by estimating time series structural components across multiple frequencies. *International Journal of Forecasting*, *30*(2), 291–302.
M4 Team (2018). M4 competitor's guide: prizes and rules. See https://www.m4unic.ac.cy/wp-content/uploads/2018/03/M4-CompetitorsGuide.pdf.
Makridakis, S. (1993). Accuracy measures: theoretical and practical concerns. *International Journal of Forecasting*, *9*(4), 527–529.
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., Assimakopoulos, V., & Spiliotis, E. (2018). Objectivity, reproducibility and replicability in forecasting research. *International Journal of Forecasting*, *34*(4), 835–838.
Makridakis, S., Chatfield, C., Hibon, M., Lawrence, M., Mills, T., et al. (1993). The M2-competition: a real-time judgmentally based forecasting study. *International Journal of Forecasting*, *9*(1), 5–22.
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*(4), 451–476.
Makridakis, S., Hibon, M., Lusk, E., & Belhadjali, M. (1987). Confidence intervals: an empirical investigation of the series in the M-competition. *International Journal of Forecasting*, *3*, 489–508.
Makridakis, S., Hogarth, R. M., & Gaba, A. (2009). Forecasting and uncertainty in the economic and business world. *International Journal of Forecasting*, *25*(4), 794–812.
Makridakis, S., Hyndman, R. J., & Petropoulos, F. (2020). Forecasting in social settings: the state of the art. *International Journal of Forecasting* *36*(1), 15–28.
Makridakis, S., & Petropoulos, F. (2020). The m4 competition: Conclusions. *International Journal of Forecasting* *36*(1), 224–227.
Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018a). Statistical and machine learning forecasting methods: concerns and ways forward. *PLoS One*, *13*(3), 1–26.
Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018b). The M4 competition: results, findings, conclusions and ways forward. *International Journal of Forecasting*, *34*(4), 802–808.
Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2020). Predicting/hypothesizing the findings of the M4 competition? *International Journal of Forecasting* *36*(1), 29–36.
Makridakis, S., & Taleb, N. (2009). Decision making and planning under low levels of predictability. *International Journal of Forecasting*, *25*(4), 716–733.
Makridakis, S. G., Wheelwright, S. C., & Hyndman, R. J. (1998). *Forecasting: methods and applications* (3rd ed.). New York: Wiley.
Makridakis, S., & Winkler, R. L. (1983). Averages of forecasts: some empirical results. *Management Science*, *29*, 987–996.
Montero-Manso, P., Athanasopoulos, G., Hyndman, R. J., & Talagala, T. S. (2020). FFORMA: feature-based forecast model averaging. *International Journal of Forecasting* *36*(1), 86–92.
Montero-Manso, P., Netto, C., & Talagala, T. (2018). *M4comp2018: Data from the M4-competition. R package version 0.1.0*.
Moore, G. H., Boehm, E. A., & Banerji, A. (1994). Using economic indicators to reduce risk in stock market investments. *International Journal of Forecasting*, *10*(3), 405–417.
Nikolopoulos, K., & Petropoulos, F. (2018). Forecasting for big data: Does suboptimality matter? *Computers & Operations Research*, *98*, 322–329.
Ord, J. K., Koehler, A. B., & Snyder, R. D. (1997). Estimation and prediction for a class of dynamic nonlinear statistical models. *Journal of the American Statistical Association*, *92*(440), 1621–1629.
Pawlikowski, M., Chorowska, A., & Yanchuk, O. (2020). Weighted ensemble of statistical models. *International Journal of Forecasting* *36*(1), 93–97.
Petropoulos, F., Hyndman, R. J., & Bergmeir, C. (2018). Exploring the sources of uncertainty: Why does bagging for time series forecasting work? *European Journal of Operational Research*, *268*(2), 545–554.
Petropoulos, F., & Makridakis, S. (2020). The M4 competition: Bigger. Stronger. Better. *International Journal of Forecasting* *36*(1), 3–6.
Petropoulos, F., & Svetunkov, I. (2020). A simple combination of univariate models. *International Journal of Forecasting* *36*(1), 110–115.
Petropoulos, F., Wang, X., & Disney, S. M. (2019). The inventory performance of forecasting methods: Evidence from the M3 competition data. *International Journal of Forecasting*, *35*(1), 251–265.
Prak, D., & Teunter, R. (2019). A general method for addressing forecasting uncertainty in inventory models. *International Journal of Forecasting*, *35*(1), 224–238.

Salaken, S. M., Khosravi, A., Nguyen, T., & Nahavandi, S. (2017). Extreme learning machine based transfer learning algorithms: A survey. *Neurocomputing*, *267*, 516–524.

Smith, J., & Wallis, K. F. (2009). A simple explanation of the forecast combination puzzle. *Oxford Bulletin of Economics and Statistics*, *71*(3), 331–355.

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.

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.

Spiliotis, E., Nikolopoulos, K., & Assimakopoulos, V. (2019). Tales from tails: On the empirical distributions of forecasting errors and their implication to risk. *International Journal of Forecasting*, *35*(2), 687–698.

Syntetos, A., Nikolopoulos, K., & Boylan, J. (2010). Judging the judges through accuracy-implication metrics: The case of inventory forecasting. *International Journal of Forecasting*, *26*(1), 134–143.

Taleb, N. (2009). Errors, robustness, and the fourth quadrant. *International Journal of Forecasting*, *25*(4), 744–759.

Wucker, M. (2016). *The gray rhino: How to recognize and act on the obvious dangers we ignore*. New York: St. Martin Press.

Zhang, G., Patuwo, B. E., & Hu, M. Y. (1998). Forecasting with artificial neural networks: The state of the art. *International Journal of Forecasting*, *14*(1), 35–62.

Dr. Spyros Makridakis is a Professor at the University of Nicosia and its director of the Institute For the Future (IFF) as well as an Emeritus Professor at INSEAD. Spyros is the organizer of the M Competitions that started back in the early 1980s. He has authored, or co-authored, twenty- two books and more than 250 articles. He was the founding editor-in-chief of the *Journal of Forecasting* and the *International Journal of Forecasting*.

Evangelos Spiliotis is a Research Fellow at the Forecasting & Strategy Unit. He graduated from 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 he 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.