

Research Paper



No single winner: Forecast combination versus neural and tree-based learners on short monthly time series

Heri Nurdiyanto^{1*}, Abid Khan²

¹Department of Industrial Engineering, Faculty of Engineering, Universitas Negeri Yogyakarta, Yogyakarta, Indonesia.

²Faculty of Management Sciences, University of Malakand, Chakdara, Pakistan.

Article Info

Article History:

Received: 05 July 2026

Revised: 03 September 2026

Accepted: 22 September 2026

Published: 22 October 2026

Keywords:

Time Series Forecasting

Forecast Combination

Neural Networks

Gradient Boosting

Exponential Smoothing

Rolling-Origin Evaluation

Mean Absolute Scaled Error



ABSTRACT

Machine learning (ML) methods are increasingly applied to business and environmental forecasting, yet many practical series are short and dominated by trend and seasonality, conditions under which classical statistical methods have performed strongly in large forecasting competitions. This study compares nine forecasting methods on six public monthly time series of 108–240 observations, covering airline traffic, retail sales, crime, climate and solar activity. The methods comprise naive and seasonal naive benchmarks, exponential smoothing (ETS), the Theta method, a seasonal ARIMA model, an equal-weight combination of the three statistical models (Comb), a multilayer perceptron (MLP), an MLP applied to seasonally adjusted data (MLP-DS) and gradient boosting (GB). Each method produced 12-month forecasts from 25 rolling origins per series, giving 1,350 multi-step forecasts evaluated with the mean absolute scaled error (MASE) and the symmetric mean absolute percentage error. The combination achieved the lowest average MASE (1.076), the best mean rank (3.75 of 9) and the smallest worst-case regret (1.42 times the best method on any series), and it improved on the seasonal naive benchmark at 74.7% of forecast origins. The ML methods were markedly heterogeneous: they were the most accurate family on the airline series (MLP-DS, MASE 0.562) and on the non-seasonal sunspot series (GB, 0.837), but among the least accurate on the car sales and robberies series, where their MASE exceeded that of the combination by 0.46–0.98. Seasonal adjustment improved the MLP on four of the six series. Differences between methods were not significant at the series level (Friedman $p = 0.113$), reflecting the small number of series. For short monthly series, a simple combination of statistical models is therefore the most reliable default, whereas ML methods should be adopted only after series-specific validation.

Corresponding Author:

Heri Nurdiyanto

Department of Industrial Engineering, Faculty of Engineering, Universitas Negeri Yogyakarta, Yogyakarta, Indonesia.

Email: herinurdiyanto@uny.ac.id

Copyright © 2026 The Author(s). This is an open access article distributed under the Creative Commons Attribution License, (<http://creativecommons.org/licenses/by/4.0/>) which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

1. INTRODUCTION

Forecasting monthly demand, traffic, sales or environmental quantities one year ahead is among the most common predictive tasks in business and public administration. Over the past decade, machine learning (ML) methods such as neural networks and gradient-boosted trees have been promoted as general-purpose replacements for classical statistical forecasting models, and deep architectures have achieved notable successes on large collections of related series [1], [2], [3]. At the same time, large empirical comparisons have repeatedly shown that simple statistical methods and their combinations are difficult to beat on individual, relatively short series [4], [5], [6], [5] found that, on monthly series from the M3 competition, popular ML methods were less accurate than exponential smoothing and the Theta method, while the M4 competition was won by a hybrid of exponential smoothing and recurrent networks [1] and confirmed the strength of forecast combinations [6].

Practitioners who must forecast a handful of short series, rather than thousands of related ones, therefore face a genuine choice. Should they adopt an ML model, a single statistical model, or a combination of statistical models? Competition results are averaged over very many series and may hide large differences between individual series, and many published comparisons of ML and statistical forecasting evaluate a single split or a single horizon, which inflates the apparent superiority of whichever method happens to win on that split [7], [8]. What matters to a practitioner is not only average accuracy but also the risk of a poor choice on the particular series at hand.

This paper addresses that question with a transparent rolling-origin benchmark on six public monthly series with contrasting characteristics. The contributions are as follows:

A controlled comparison of nine methods, namely two benchmarks, three statistical models, their combination and three ML learners, over 25 forecast origins and horizons of 1–12 months on each of six series, evaluated with scale-free error measures.

An assessment of both average accuracy and worst-case regret, showing that the statistical combination is the most reliable default while ML methods are best on some series and among the worst on others.

Evidence that seasonal adjustment before neural-network training, as recommended by Zhang and Qi [9], improves accuracy on most but not all series, and a characterization of the series on which ML methods did and did not help.

An explicit account of the statistical uncertainty of such comparisons, including the lack of series-level significance with a small number of series.

Section 2 reviews related work, Section 3 describes the data, methods and evaluation protocol, Section 4 presents the results, Section 5 discusses them, and Section 6 concludes the paper.

2. RELATED WORK

2.1 Statistical Forecasting Methods

Exponential smoothing, introduced by Holt [10] and extended to seasonal data by Winters [11], remains one of the most widely used forecasting families; its state-space formulation [12] enables automatic model selection by information criteria, and its development is reviewed by Gardner [13]. The Theta method [14], which performed best among the methods of the M3 competition, was shown by Hyndman and Billah [15] to be equivalent to simple exponential smoothing with drift. Seasonal

autoregressive integrated moving-average (SARIMA) models, and in particular the “airline model” with one regular and one seasonal difference, are standard for trending seasonal series [16]. Since Bates and Granger [17], combining forecasts from several models has been known to reduce error variance, and a large body of evidence shows that simple equal-weight combinations are hard to beat [6], [18].

2.2 Machine Learning For Time Series Forecasting

Neural networks have been applied to forecasting for several decades [19]. Zhang and Qi [9] showed that multilayer perceptrons (MLPs) struggle to model seasonality directly and that deseasonalising and detrending the data before training substantially improves their accuracy. Multi-step forecasting with ML can be organised recursively or by direct strategies that train one output per horizon, and the choice affects accuracy [20]. Gradient boosting [21], [22] dominated the M5 competition on hierarchical retail data [23], and specialized deep architectures such as N-BEATS [2] and hybrid models [1] achieved strong results on large datasets. [3] Found that recurrent networks can be competitive but require careful preprocessing and are sensitive to series characteristics.

2.3 Evaluation of Forecasting Methods

Tashman [7] argued for rolling-origin evaluation over multiple origins rather than a single hold-out window. Hyndman and Koehler [24] recommended the mean absolute scaled error (MASE), which scales errors by the in-sample error of a naive benchmark and is therefore comparable across series of different scales. Formal comparison of forecast accuracy uses the Diebold–Mariano test [25] for pairs of methods and rank-based procedures for multiple methods and series [26], [27], [5] emphasized that comparisons should include simple benchmarks and should report results over many origins and horizons, a practice followed in this study.

3. METHODOLOGY

3.1 Data

Six public monthly series with contrasting characteristics were selected Table 1 and Figure 1: international airline passengers [28], monthly car sales [28], monthly sales of writing paper [28], monthly armed robberies [28], the monthly mean of daily minimum temperatures [28], and the most recent 240 months of the monthly mean sunspot number [29]. The first five series were obtained through the Time Series Data Library [28]; the daily temperature record was aggregated to monthly means. The series contain 108–240 observations. To characterize them, trend strength FT and seasonal strength FS were computed from a robust STL decomposition [30] of the training portion (logarithms were taken for strictly positive series), following [31]: $FT = \max(0, 1 - \text{Var}(R)/\text{Var}(T + R))$ and $FS = \max(0, 1 - \text{Var}(R)/\text{Var}(S + R))$, where T, S and R are the trend, seasonal and remainder components. The six series thus range from strongly trending and seasonal (airline passengers) to strongly seasonal without trend (minimum temperature) and to trending or cyclical but essentially non-seasonal (robberies and sunspots).

Table 1. Characteristics of the Six Monthly Series. Trend and Seasonal Strength Were Computed on the Data Preceding the Evaluation Region

Series	Observations (N)	Domain	Trend Strength FT	Seasonal Strength FS	Positive-Valued
Airline passengers	144	Transport	0.99	0.95	Yes
Car sales	108	Retail	0.79	0.89	Yes
Writing paper sales	147	Retail	0.69	0.94	Yes
Robberies	118	Crime	0.93	0.24	Yes
Minimum temperature	120	Climate	0.00	0.92	Yes
Sunspots	240	Solar physics	0.88	0.12	Yes

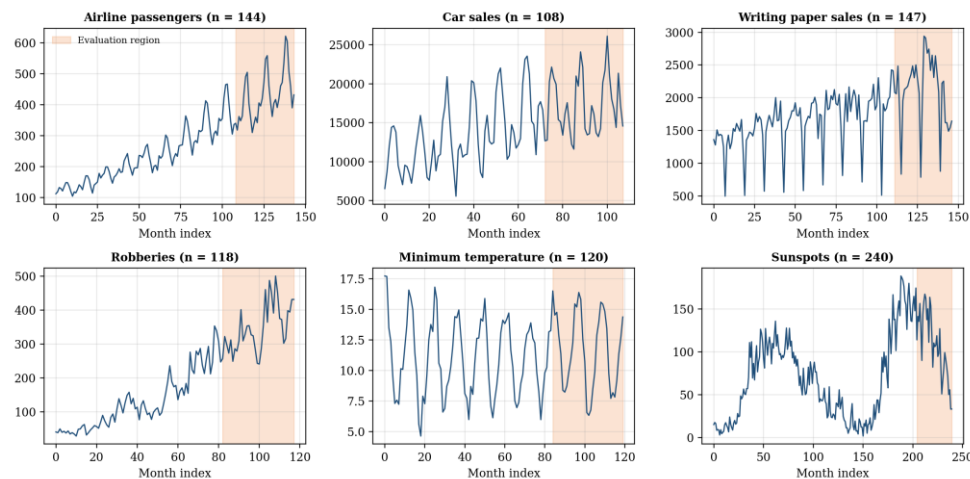


Figure 1. The Six Monthly Series. The Shaded Region Marks the Final 36 Months, Which contain all Forecast Targets of the Rolling-Origin Evaluation

3.2 Forecasting Methods

Nine methods were compared Table 2. The naive method repeats the last observation and the seasonal naive (sNaive) method repeats the observation from the same month of the previous year [16]. ETS fits Holt–Winters exponential smoothing models [10], [11] with no trend or an additive damped trend and additive or multiplicative seasonality, selecting among them by the corrected Akaike information criterion. Theta is the standard Theta method with automatic deseasonalisation [14]. SARIMA is the airline model ARIMA [1](0,1,1)12 fitted to logarithms of the data [32]. Comb is the equal-weight average of the ETS, Theta and SARIMA forecasts [17], [18]. The statistical models were implemented with statsmodels [33].

The three ML methods use the direct multi-output strategy [20]: the input is the most recent 24 observations and the output is the next 12. Each training window was divided by the mean absolute value of its input segment, so that the models learn level-free patterns, and forecasts were rescaled by the same factor. The MLP has one hidden layer of 64 units, L2 penalty 10^{−3} and is trained with L-BFGS [34], which is well suited to small datasets; the forecast is the average of five networks with different random initializations [35]. MLP-DS applies the same network to data seasonally adjusted by classical decomposition (multiplicative for positive series) and re-applies the last estimated seasonal cycle to its forecasts, following Zhang and Qi [9]. GB trains one histogram gradient boosting regressor [21], [22] per horizon on the same windows. ML methods were implemented with scikit-learn [36]. All hyperparameters were fixed in advance and not tuned on the evaluation data.

Table 2. Forecasting Methods and Their Configurations

Method	Family	Configuration
Naive	Benchmark	Last observation repeated for all horizons
sNaive	Benchmark	Observation of the same month in the previous year
ETS	Statistical	Holt–Winters; trend: none or additive damped; seasonality: additive or multiplicative; selected by AICc
Theta	Statistical	Standard Theta method with automatic multiplicative deseasonalisation (period 12)
SARIMA	Statistical	ARIMA(1,1,1)(0,1,1)12 on log-transformed data
Comb	Combination	Equal-weight mean of ETS, Theta and SARIMA forecasts
MLP	Machine learning	24 lagged inputs → 12 outputs; one hidden layer (64 units); L2 = 10 ^{−3} ; L-BFGS; ensemble of 5 seeds; window-mean scaling
MLP-DS	Machine learning	As MLP, trained on seasonally adjusted data; last seasonal cycle re-applied

GB	Machine learning	Histogram gradient boosting, one model per horizon; 200 iterations; learning rate 0.05; min. 5 samples per leaf; same inputs as MLP
----	------------------	---

3.3 Rolling-Origin Evaluation

Forecasts were evaluated with a rolling origin [7] Figure 2. For a series of length n , the final 36 observations formed the evaluation region. Forecast origins were placed at $t_0 = n - 36, n - 35, \dots, n - 12$, giving 25 origins per series. At each origin every method was refitted on all observations before the origin and produced a forecast for the next 12 months, so that each method generated 150 twelve-step forecasts (1,800 point forecasts) and the whole study comprised 1,350 multi-step forecasts.

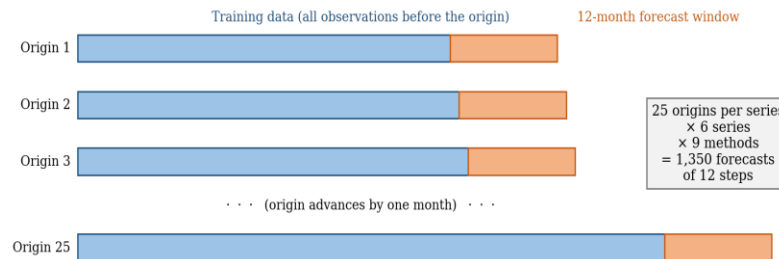


Figure 2. Rolling-Origin Evaluation. At Each of 25 Origins per Series, Every Method Is Refitted On All Data before the Origin and Forecasts the Following 12 Months

3.4 Error Measures and Statistical Analysis

Accuracy was measured by the MASE [24], which scales the absolute error at horizon h by the in-sample mean absolute error of the seasonal naive method: [1]

Where t_0 is the forecast origin and $m = 12$. A MASE below one indicates that the method is, on average, more accurate than the in-sample one-season-ahead seasonal naive forecast. The symmetric mean absolute percentage error (sMAPE), used in the M-competitions [4], [6], was reported as a secondary measure. Methods were summarised by the mean MASE over series, by the geometric mean of MASE relative to sNaive over all 150 series–origin pairs, by mean ranks, by the proportion of origins at which they beat sNaive, and by their regret, defined as the ratio of their MASE on a series to the lowest MASE achieved by any method on that series. Differences among methods were tested with the Friedman test [37] using the six series as blocks [26]. A second Friedman test with the 150 series–origin pairs as blocks is also reported, but because the forecast windows of neighbouring origins overlap by up to 11 months, it treats dependent observations as independent and is therefore anti-conservative.

4. RESULTS AND DISCUSSION

4.1 Overall accuracy

Table 3 reports the MASE of every method on every series, and Table 4 summarises the methods across series. The equal-weight combination of the three statistical models achieved the lowest mean MASE (1.076), the lowest geometric mean relative to sNaive (0.804, i.e. a 19.6% reduction in scaled error), the best mean rank over all 150 forecasts (3.75) and the highest proportion of origins at which it beat sNaive (74.7%). The individual statistical models followed closely (ETS 1.125, Theta 1.116, and SARIMA 1.208). The three ML methods had mean MASE values of 1.272 (MLP-DS), 1.315 (GB) and 1.320 (MLP), similar to or slightly worse than sNaive (1.308), although their geometric mean relative MASE (0.899–0.960) shows that they did improve on sNaive at a majority of origins. The naive method was by far the worst (2.486), confirming that seasonality dominates these series.

Table 3. MASE of Each Method on Each Series (Mean $\pm$ SD Over 25 Rolling Origins, Horizons 1–12). The Lowest Value in Each Row is shown in Bold

Series	Naive	Snaive	ETS	Theta	SARIMA	Comb	MLP	MLP-DS	GB
Airline passengers	2.53 $\pm$ 0.81	1.32 $\pm$ 0.53	0.99 $\pm$ 0.34	0.90 $\pm$ 0.29	0.53 $\pm$ 0.20	0.76 $\pm$ 0.26	0.59 $\pm$ 0.28	0.56 $\pm$ 0.22	0.72 $\pm$ 0.35
Car sales	2.69 $\pm$ 0.63	0.93 $\pm$ 0.19	1.07 $\pm$ 0.27	1.19 $\pm$ 0.43	1.39 $\pm$ 0.47	1.12 $\pm$ 0.37	1.67 $\pm$ 0.32	1.81 $\pm$ 0.44	1.58 $\pm$ 0.29
Writing paper sales	3.59 $\pm$ 2.02	1.95 $\pm$ 0.50	1.77 $\pm$ 0.57	1.64 $\pm$ 0.75	1.58 $\pm$ 0.85	1.64 $\pm$ 0.73	1.81 $\pm$ 0.82	1.76 $\pm$ 0.87	1.68 $\pm$ 0.60
Robberies	1.46 $\pm$ 0.71	1.26 $\pm$ 0.39	1.14 $\pm$ 0.48	1.00 $\pm$ 0.47	1.50 $\pm$ 0.39	0.99 $\pm$ 0.39	1.92 $\pm$ 0.70	1.79 $\pm$ 0.66	1.96 $\pm$ 0.74
Minimum temperature	3.46 $\pm$ 0.76	1.08 $\pm$ 0.12	0.72 $\pm$ 0.23	0.80 $\pm$ 0.19	0.89 $\pm$ 0.19	0.77 $\pm$ 0.15	1.07 $\pm$ 0.32	0.79 $\pm$ 0.16	1.12 $\pm$ 0.37
Sunspots	1.19 $\pm$ 0.48	1.31 $\pm$ 0.46	1.07 $\pm$ 0.37	1.18 $\pm$ 0.40	1.36 $\pm$ 0.49	1.19 $\pm$ 0.42	0.86 $\pm$ 0.30	0.92 $\pm$ 0.21	0.84 $\pm$ 0.21
Mean	2.486	1.308	1.125	1.116	1.208	1.076	1.320	1.272	1.315

Table 4. Summary of Method Performance across the Six Series and 150 Forecast Origins

Method	Mean MASE	Mean SMAP E (%)	Geometric Mean MASE Relative To Snaive	Mean Rank (1–9)	Origins Beating Snaive (%)	Mean Regret	Maximum Regret	Series Won
Naive	2.486	25.22	1.779	7.79	14.0	2.94	4.82	0
sNaive	1.308	15.19	1.000	5.55	–	1.51	2.48	1
ETS	1.125	13.07	0.846	4.20	71.3	1.26	1.86	1
Theta	1.116	13.00	0.837	4.13	67.3	1.26	1.70	0
SARIMA	1.208	14.40	0.874	4.41	62.7	1.31	1.62	2
Comb	1.076	12.65	0.804	3.75	74.7	1.19	1.42	1
MLP	1.320	14.82	0.939	5.13	53.3	1.42	1.95	0
MLP-DS	1.272	14.24	0.899	4.80	58.7	1.36	1.95	0
GB	1.315	14.43	0.960	5.25	52.7	1.44	1.99	1

Regret is the ratio of a method's MASE on a series to the lowest MASE achieved by any method on that series; a value of 1 means the method was the best on that series.

At the series level, the Friedman test did not reject the hypothesis of equal performance ($\chi^2 = 12.98, df = 8, p = 0.113$); with only six series, the study has limited power to establish significance, and the ranking should be read as descriptive. When the 150 series–origin pairs were treated as blocks, the test was highly significant ($\chi^2 = 230.6, p = 2.2 \times 10^{-45}$), but, as explained in Section 3.4, this second test ignores the overlap between forecast windows and overstates the evidence.

4.2 Heterogeneity across Series

The average hides strong differences between series (Figure 3). Each of five different methods was the best on at least one series: SARIMA on airline passengers and writing paper sales, sNaive on car sales, ETS on minimum temperature, Comb on robberies and GB on sunspots. The ML methods were the most accurate family on two series. On airline passengers, which has the strongest trend and seasonality, MLP-

DS (0.562) and MLP (0.594) were second only to SARIMA (0.531) and clearly better than ETS, Theta and the combination. On sunspots, which has an 11-year cycle but no annual seasonality, GB (0.837) and MLP (0.855) outperformed every statistical model; all seasonal statistical models were misspecified on this series, and only ETS (1.070) improved on the naive method. Conversely, on car sales and robberies the ML methods were among the least accurate, exceeding the MASE of the combination by 0.46–0.98.

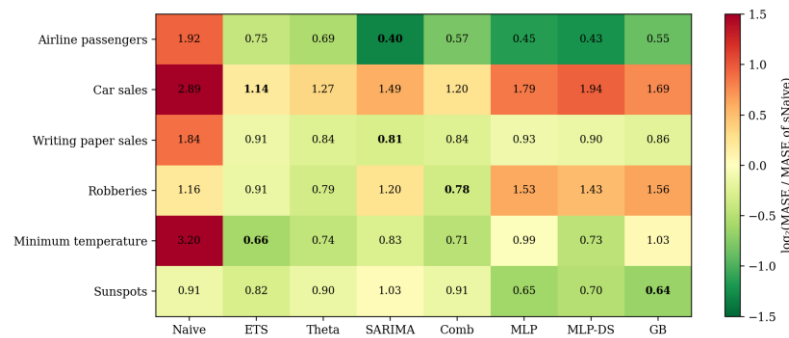


Figure 3. MASE of Each Method Relative to the Seasonal Naive Benchmark on Each Series (Values below 1 Indicate Improvement). Colors Show the $\log_2$ Ratio; the Lowest Value in Each Row among the Methods Shown is in Bold (On Car Sales, Snaive Itself Was Best)

Regret quantifies the practical consequence of this heterogeneity [Figure 4](#) and [Table 4](#). The combination was never more than 1.42 times worse than the best method on any series and had the lowest mean regret (1.19). The individual statistical models had maximum regrets of 1.62–1.86, and the ML methods of 1.95–1.99, meaning that on their worst series they were almost twice as inaccurate as the best available method. For a practitioner who must choose a method before seeing the outcome, the combination therefore offers the best protection against a poor choice.

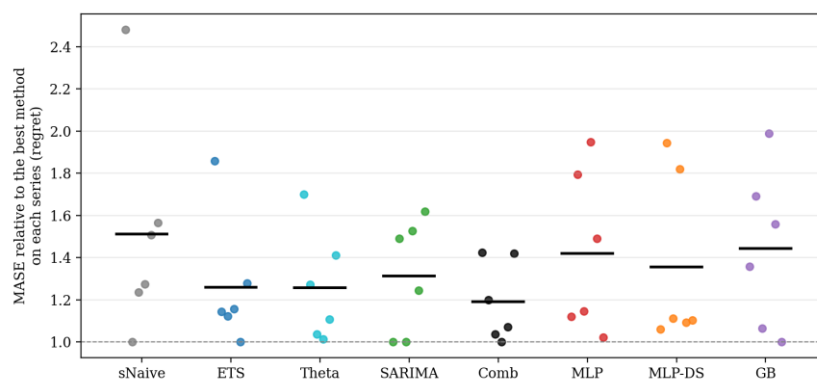


Figure 4. Regret of Each Method on Each of the Six Series, Defined As Its MASE Divided By the Lowest MASE Achieved By Any Method on That Series. Horizontal Bars Show the Mean Over Series

4.3 Paired Comparisons on Individual Series

Error! Reference source not found. compares selected pairs of methods series by series, reporting the mean MASE difference and the number of the 25 origins at which the first method was more accurate. MLP-DS was more accurate than the combination on airline passengers (19 of 25 origins) and sunspots (17 of 25), roughly equal on minimum temperature (13 of 25), and less accurate on car sales (2 of 25), robberies (3 of 25) and writing paper sales (8 of 25). Seasonal adjustment improved the MLP on four of the six series, most clearly on minimum temperature, where it reduced MASE by 0.285 and was better at 23 of 25 origins; it was harmful on the non-seasonal sunspot series and on car sales. The

combination improved on SARIMA on five series and on ETS on four, illustrating the risk-reducing effect of averaging models whose errors differ.

Table 5. Series-Level Paired Comparisons of MASE: Mean Difference (A – B) and Number of Origins (out of 25) Where a Was More Accurate

Comparison (A Vs B)	Airline Passengers	Car Sales	Writing Paper Sales	Robberies	Minimum Temperature	Sunspots
MLP-DS vs Comb	-0.193 (19)	+0.695 (2)	+0.118 (8)	+0.808 (3)	+0.016 (13)	-0.265 (17)
GB vs Comb	-0.035 (17)	+0.459 (3)	+0.043 (8)	+0.975 (3)	+0.349 (4)	-0.351 (19)
MLP-DS vs MLP	-0.032 (18)	+0.140 (8)	-0.053 (18)	-0.128 (14)	-0.285 (23)	+0.069 (8)
Comb vs ETS	-0.231 (22)	+0.052 (14)	-0.137 (20)	-0.153 (19)	+0.051 (9)	+0.119 (4)
Comb vs Theta	-0.147 (23)	-0.068 (14)	+0.001 (14)	-0.012 (12)	-0.026 (12)	+0.008 (12)
Comb vs SARIMA	+0.225 (3)	-0.271 (19)	+0.057 (7)	-0.518 (19)	-0.124 (20)	-0.166 (19)

4.4 Accuracy by Forecast Horizon

Figure 5 shows the mean scaled error by horizon, averaged over series and origins. Errors grew with the horizon for all methods. The combination was the most accurate at every horizon from one to eleven months (0.806 at one month, 1.252 at ten months) and was matched by Theta at twelve months (1.364 versus 1.375). The ML methods were least competitive at short horizons, where their one-step errors (0.971–1.122) were 20–39% higher than that of the combination, and they approached the statistical methods only at the longest horizons, where all methods converged towards the seasonal naive error. SARIMA was competitive at short horizons but deteriorated fastest, becoming worse than sNaive beyond nine months.

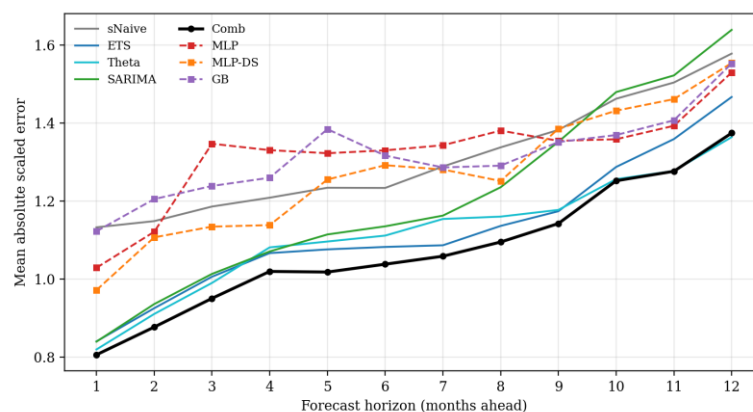


Figure 5. Mean Absolute Scaled Error By Forecast Horizon, Averaged Over The Six Series And 25 Origins. The Naive Method (Errors 1.49–3.06) Is Omitted For Readability

4.5 Discussion

4.5.1 Principal Findings

On six short monthly series, no single method was best everywhere, but an equal-weight combination of three statistical models was the most reliable choice: it had the lowest average error, the best average rank and the smallest worst-case regret. This agrees with the conclusions of the M3 and M4

competitions [4], [6], with the review of [5], and with the long-standing evidence on forecast combination [17], [18]. The present study adds a series-level view of the risk involved: the ML methods were capable of the best results in the study on two series but were nearly twice as inaccurate as the best method on others.

4.5.2 When Did Machine Learning Help?

The two series on which ML methods excelled share informative features. Airline passengers is long, smooth and multiplicative, with the strongest trend and seasonality of all series ($FT = 0.99, FS = 0.95$), so the input windows contain consistent, learnable patterns; the level-free window scaling also suits its multiplicative growth. The sunspot series is the longest (240 observations), and its dominant pattern is a quasi-periodic cycle of roughly 11 years that none of the seasonal statistical models can represent, whereas a flexible learner can extrapolate the recent cyclical shape. The series on which ML methods failed, car sales and robberies, are short (108 and 118 observations), leaving few training windows at early origins (as few as 37 for car sales), and they contain irregular level changes and noise that the ML models appear to have overfitted. These observations are consistent with earlier findings that ML forecasters need longer histories and careful preprocessing [3], [9], but with six series they remain indicative rather than conclusive.

4.5.3 Implications for Practice

Four practical recommendations follow. First, for short monthly series a simple equal-weight combination of exponential smoothing, the Theta method and a seasonal ARIMA model is a robust default that is cheap to compute and difficult to beat. Second, any proposed method, including ML, should be compared with the seasonal naive benchmark and with such a combination under rolling-origin evaluation [5], [7]; a single hold-out split would have produced very different conclusions depending on the series. Third, neural networks should be given seasonally adjusted inputs when seasonality is present [9], but adjustment should not be applied blindly to series without annual seasonality. Fourth, ML methods deserve consideration when the series is long, smooth, or dominated by patterns that standard statistical models cannot represent, as the sunspot series illustrates, and their selection should be based on series-specific validation rather than on average results.

4.5.4 Limitations

The study has several limitations. Only six series were analysed, which limits statistical power and generalisability; the series-level Friedman test was not significant, and the observations about when ML helps are hypotheses for larger studies. The ML models were deliberately simple and untuned; hyperparameter optimization, alternative lag lengths, or global models trained across many related series, as in the M4 and M5 competitions [6], [23], could improve their performance. Forecasts from neighboring origins overlap, so the 150 series-origin results are not independent. Only point forecasts were evaluated; the quality of prediction intervals, which matters for decision-making, was not assessed. Finally, the combination used equal weights over a fixed set of three models; other combinations, including hybrids of statistical and ML models, were not explored.

5. CONCLUSION

This rolling-origin study compared statistical, combined and ML forecasting methods on six short monthly series under identical conditions. An equal-weight combination of exponential smoothing, the Theta method and a seasonal ARIMA model achieved the lowest average MASE (1.076), reduced scaled error by 19.6% relative to the seasonal naive benchmark in geometric-mean terms, and was never more than 1.42 times worse than the best method on any series. Neural and tree-based learners produced the best forecasts on a long, smooth, strongly seasonal series and on a long non-seasonal cyclical series, but performed poorly on short, noisy series, with worst-case regrets close to two. Seasonal adjustment improved the neural network on most series. With only six series, the differences were not statistically significant at the series level, and the findings should be confirmed on larger collections. For practitioners

forecasting a small number of short series, the evidence supports using a statistical combination as the default and adopting ML methods only after series-specific validation. Future work will extend the benchmark to larger series collections, tuned and global ML models, hybrid combinations, and probabilistic forecasts.

Acknowledgments

The authors thank the compilers of the Time Series Data Library and the SILSO World Data Center for making the data publicly available, the developers of statsmodels and scikit-learn, and the researchers whose work is cited in this article.

Funding Information

This research has received no external funding.

Author Contribution Statement

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Heri Nurdyanto	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓
Abid Khan		✓	✓		✓	✓	✓	✓		✓	✓		✓	✓

C : Conceptualization

M : Methodology

So : Software

Va : Validation

Fo : Formal analysis

I : Investigation

R : Resources

D : Data Curation

O : Writing - Original Draft

E : Writing - Review & Editing

Vi : Visualization

Su : Supervision

P : Project administration

Fu : Funding acquisition

Conflict of Interest Statement

No conflict of interest.

Informed Consent

Not Applicable.

Ethical Approval

Not Applicable.

Data Availability

This study draws on publicly available data from the Time Series Data Library and the SILSO World Data Center. No new datasets were created. The analysis code is available from the corresponding author on reasonable request.

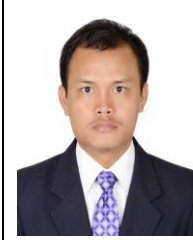
REFERENCES

- [1] S. Smyl, "A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting," *International Journal of Forecasting*, vol. 36, no. 1, pp. 75–85, 2020. doi.org/10.1016/j.ijforecast.2019.03.017
- [2] B. N. Oreshkin, D. Carпов, N. Chapados, and Y. Bengio, "N-BEATS: Neural basis expansion analysis for interpretable time series forecasting," in *Proc. 8th Int. Conf. Learning Representations*, 2020. doi.org/10.48550/arXiv.1905.10437
- [3] H. Hewamalage, C. Bergmeir, and K. Bandara, "Recurrent neural networks for time series forecasting: Current status and future directions," *International Journal of Forecasting*, vol. 37, no. 1, pp. 388–427, 2021. doi.org/10.1016/j.ijforecast.2020.06.008

- [4] S. Makridakis and M. Hibon, "The M3-competition: Results, conclusions and implications," *International Journal of Forecasting*, vol. 16, no. 4, pp. 451–476, 2000. [doi.org/10.1016/S0169-2070\(00\)00057-1](https://doi.org/10.1016/S0169-2070(00)00057-1)
- [5] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "Statistical and machine learning forecasting methods: Concerns and ways forward," *PLoS ONE*, vol. 13, no. 3, p. e0194889, 2018. doi.org/10.1371/journal.pone.0194889
- [6] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "The M4 competition: 100,000 time series and 61 forecasting methods," *International Journal of Forecasting*, vol. 36, no. 1, pp. 54–74, 2020. doi.org/10.1016/j.ijforecast.2019.04.014
- [7] L. J. Tashman, "Out-of-sample tests of forecasting accuracy: An analysis and review," *International Journal of Forecasting*, vol. 16, no. 4, pp. 437–450, 2000. [doi.org/10.1016/S0169-2070\(00\)00065-0](https://doi.org/10.1016/S0169-2070(00)00065-0)
- [8] C. Bergmeir and J. M. Benítez, "On the use of cross-validation for time series predictor evaluation," *Information Sciences*, vol. 191, pp. 192–213, 2012. doi.org/10.1016/j.ins.2011.12.028
- [9] G. P. Zhang and M. Qi, "Neural network forecasting for seasonal and trend time series," *European Journal of Operational Research*, vol. 160, no. 2, pp. 501–514, 2005. doi.org/10.1016/j.ejor.2003.08.037
- [10] C. C. Holt, "Forecasting seasonals and trends by exponentially weighted moving averages," *International Journal of Forecasting*, vol. 20, no. 1, pp. 5–10, 2004. doi.org/10.1016/j.ijforecast.2003.09.015
- [11] P. R. Winters, "Forecasting sales by exponentially weighted moving averages," *Management Science*, vol. 6, no. 3, pp. 324–342, 1960. doi.org/10.1287/mnsc.6.3.324
- [12] R. J. Hyndman, A. B. Koehler, R. D. Snyder, and S. Grose, "A state space framework for automatic forecasting using exponential smoothing methods," *International Journal of Forecasting*, vol. 18, no. 3, pp. 439–454, 2002. [doi.org/10.1016/S0169-2070\(01\)00110-8](https://doi.org/10.1016/S0169-2070(01)00110-8)
- [13] E. S. Gardner Jr., "Exponential smoothing: The state of the art — Part II," *International Journal of Forecasting*, vol. 22, no. 4, pp. 637–666, 2006. doi.org/10.1016/j.ijforecast.2006.03.005
- [14] V. Assimakopoulos and K. Nikolopoulos, "The theta model: A decomposition approach to forecasting," *International Journal of Forecasting*, vol. 16, no. 4, pp. 521–530, 2000. [doi.org/10.1016/S0169-2070\(00\)00066-2](https://doi.org/10.1016/S0169-2070(00)00066-2)
- [15] R. J. Hyndman and B. Billah, "Unmasking the Theta method," *International Journal of Forecasting*, vol. 19, no. 2, pp. 287–290, 2003. [doi.org/10.1016/S0169-2070\(01\)00143-1](https://doi.org/10.1016/S0169-2070(01)00143-1)
- [16] R. J. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*, 3rd ed. Melbourne, Australia: OTexts, 2021. otexts.com/fpp3/
- [17] J. M. Bates and C. W. J. Granger, "The combination of forecasts," *Operational Research Quarterly*, vol. 20, no. 4, pp. 451–468, 1969. doi.org/10.2307/3008764
- [18] R. T. Clemen, "Combining forecasts: A review and annotated bibliography," *International Journal of Forecasting*, vol. 5, no. 4, pp. 559–583, 1989. [doi.org/10.1016/0169-2070\(89\)90012-5](https://doi.org/10.1016/0169-2070(89)90012-5)
- [19] G. Zhang, B. E. Patuwo, and M. Y. Hu, "Forecasting with artificial neural networks: The state of the art," *International Journal of Forecasting*, vol. 14, no. 1, pp. 35–62, 1998. [doi.org/10.1016/S0169-2070\(97\)00044-7](https://doi.org/10.1016/S0169-2070(97)00044-7)
- [20] S. Ben Taieb, G. Bontempi, A. F. Atiya, and A. Sorjamaa, "A review and comparison of strategies for multi-step ahead time series forecasting based on the NN5 forecasting competition," *Expert Systems with Applications*, vol. 39, no. 8, pp. 7067–7083, 2012. doi.org/10.1016/j.eswa.2012.01.039
- [21] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001. doi.org/10.1214/aos/1013203451
- [22] G. Ke et al., "LightGBM: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems* 30, 2017, pp. 3146–3154. papers.nips.cc/paper/6907-lightgbm-a-highly-efficient-gradient-boosting-decision-tree
- [23] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "M5 accuracy competition: Results, findings, and conclusions," *International Journal of Forecasting*, vol. 38, no. 4, pp. 1346–1364, 2022. doi.org/10.1016/j.ijforecast.2021.11.013

- [24] R. J. Hyndman and A. B. Koehler, "Another look at measures of forecast accuracy," *International Journal of Forecasting*, vol. 22, no. 4, pp. 679–688, 2006. doi.org/10.1016/j.ijforecast.2006.03.001
- [25] F. X. Diebold and R. S. Mariano, "Comparing predictive accuracy," *Journal of Business & Economic Statistics*, vol. 13, no. 3, pp. 253–263, 1995. doi.org/10.1080/07350015.1995.10524599
- [26] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006. jmlr.org/papers/v7/demsar06a.html
- [27] A. J. Koning, P. H. Franses, M. Hibon, and H. O. Stekler, "The M3 competition: Statistical tests of the results," *International Journal of Forecasting*, vol. 21, no. 3, pp. 397–409, 2005. doi.org/10.1016/j.ijforecast.2004.10.003
- [28] R. J. Hyndman and Y. Yang, "tsdl: Time Series Data Library," R package version 0.1.0, 2018. pkg.yangzhuoranyang.com/tsdl/
- [29] SILSO World Data Center, "The International Sunspot Number," Royal Observatory of Belgium, Brussels. www.sidc.be/SILSO/
- [30] R. B. Cleveland, W. S. Cleveland, J. E. McRae, and I. Terpenning, "STL: A seasonal-trend decomposition procedure based on loess," *Journal of Official Statistics*, vol. 6, no. 1, pp. 3–73, 1990. www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/stl-a-seasonal-trend-decomposition-procedure-based-on-loess.pdf
- [31] X. Wang, K. Smith, and R. Hyndman, "Characteristic-based clustering for time series data," *Data Mining and Knowledge Discovery*, vol. 13, no. 3, pp. 335–364, 2006. doi.org/10.1007/s10618-005-0039-x
- [32] G. E. P. Box and D. R. Cox, "An analysis of transformations," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 26, no. 2, pp. 211–252, 1964. doi.org/10.1111/j.2517-6161.1964.tb00553.x
- [33] S. Seabold and J. Perktold, "Statsmodels: Econometric and statistical modeling with Python," in *Proc. 9th Python in Science Conf.*, 2010, pp. 92–96. doi.org/10.25080/Majors-92bf1922-011
- [34] D. C. Liu and J. Nocedal, "On the limited memory BFGS method for large scale optimization," *Mathematical Programming*, vol. 45, no. 1–3, pp. 503–528, 1989. doi.org/10.1007/BF01589116
- [35] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986. doi.org/10.1038/323533a0
- [36] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011. jmlr.org/papers/v12/pedregosa11a.html
- [37] M. Friedman, "The use of ranks to avoid the assumption of normality implicit in the analysis of variance," *Journal of the American Statistical Association*, vol. 32, no. 200, pp. 675–701, 1937. doi.org/10.1080/01621459.1937.10503522

How to Cite: Heri Nurdiyanto, Abid Khan (2026). No single winner: Forecast combination versus neural and tree-based learners on short monthly time series. *Journal of Artificial Intelligence, Machine Learning and Neural Network*, 6(2), 44-55. <https://doi.org/10.55529/jaimlenn.62.44.55>

BIOGRAPHIES OF AUTHORS

Dr. Ir. Heri Nurdyanto, S.Kom., M.T.I. is a member of the Department of Industrial Engineering, Faculty of Engineering, Universitas Negeri Yogyakarta, Yogyakarta, Indonesia. His work lies in industrial and manufacturing engineering, including automated inspection, quality control and the forecasting of demand and production quantities. He is interested in how machine learning and decision-support methods are put to work on the factory floor and in the practical constraints that industrial deployment imposes.



Dr. Abid Khan is a member of the Faculty of Management Sciences at the University of Malakand, Chakdara, Pakistan. His research interests lie in quantitative methods for management and economics, including the statistical analysis of time series and the evaluation of forecasting procedures. He is concerned with how forecasting evidence is assessed and with the conditions under which simple statistical methods remain competitive with more elaborate alternatives.