

What is time series forecasting?

A time series forecast extends a series from its own past. In the M4 competition, a hybrid and combinations led; pure methods trailed.

Author	Published	Reading time	Words
Sinan Isoglu, MBA (Quantic) Commercial growth leader, lecturer and doctoral researcher	18 October 2026	16 min	3,513

Management summary Time series forecasting predicts the future values of one series from that series' own past, by carrying its level, trend and seasonal pattern forward. This concept guide defines it, sets out the method families (naive and seasonal naive benchmarks, exponential smoothing, ARIMA, combinations, machine learning and hybrids), reports what the M4 competition found when 61 methods forecast 100,000 series, and explains why Hyndman and Koehler propose scaled errors, such as the mean absolute scaled error, for comparing accuracy across series instead of errors in units or percentages. The sources are one forecasting competition on continuous series, which left out intermittent demand, and one methods article on accuracy measures.

Contents	§ 1	What is time series forecasting?	2
	§ 2	Which methods extend a series from its own past?	3
	§ 3	What did 100,000 series show about these methods?	4
	Figure 1	Accuracy of 14 of the 61 M4 methods, relative to Naïve 2	5
	§ 4	Why can forecast errors not be averaged across series as they stand?	7
	§ 5	What should a forecast comparison fix before the first score?	9
	Figure 2	A forecast comparison record	9
	§ 6	Which of these results would not survive a short or intermittent series?	10
		References	12

What is time series forecasting?

§ 1 of 6

It is forecasting a series from its own past. Hyndman and Koehler (2006)¹ set out to “discuss and compare measures of accuracy of univariate time series forecasts”, forecasts of a single series, and their four example methods, from “the historical mean using data up to the most recent observation” to Holt’s method, use nothing but the series itself. In the M4 competition, Makridakis et al. (2020)² gave the participants the history of each series and held back its end: “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”. Before release, “any information that could possibly lead to the identification of the original series was removed so as to ensure the objectivity of the results”, starting dates included. The authors list as an open question “the possible improvement in PF and PI performances that can be achieved by expanding time series forecasting to include explanatory/exogenous variables”.

¹ Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4), 679–688.

² Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2020). The M4 Competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, 36(1), 54–74.

Makridakis et al.² describe their benchmark methods by the part of a series’ pattern that each one carries forward: Naïve 2 “captures seasonality only”, single exponential smoothing “captures the level”, Holt’s method “extrapolates using a linear trend”, and the damped method “dampens the linear trend”. In this guide’s words, the level is where the series stands now, the trend is the direction it is moving in, and seasonality is a pattern that repeats over a fixed period, such as the months of a year.

In this guide, a time series forecast is a forecast of the future values of one series made from past values of that series, by a method that states how it carries the level, the trend and any seasonal pattern forward, and judged on values the method did not see. The method may also learn from other series of the same kind: the three most accurate M4 methods “introduced information from multiple series (aggregated by data frequency)” to choose how to forecast each series or how to weight a combination (p. 14 of the in-press version). A forecast that also uses other variables, such as prices, promotions or the open opportunities in a sales pipeline, goes beyond this definition; M4 left such variables out, and so does this guide. The examples here

are revenue and demand series. The database M4 drew on “emphasizes business forecasting applications”, and its 100,000 series come from six domains: micro, industry, macro, finance, demographic and other (Table 1, p. 3 of the in-press version).

Which methods extend a series from its own past?

§ 2 of 6

The simplest are the naive benchmarks. In M4, Naïve 1 is “A random walk model, assuming that future values will be the same as that of the last known observation.” Naïve S, the seasonal naive method, sets each forecast equal to “the last known observation of the same period”, so next March is forecast as last March. Naïve 2 is “Like Naïve 1 but the data are seasonally adjusted, if needed, by applying a classical multiplicative decomposition”, with a 90 per cent autocorrelation test deciding whether a series is seasonal. On this guide’s reading, they are yardsticks, not recommendations: a more complex method is worth its cost only if it beats them on data it has not seen.

Exponential smoothing is a family. Simple exponential smoothing (SES) means “Exponentially smoothing the data and extrapolating assuming no trend”; Holt’s method assumes a linear trend and the damped method “a damped trend”; for all three in M4, “Seasonal adjustments are considered as per Naïve 2.” ETS chooses among exponential smoothing models: it “Automatically provides the best exponential smoothing model, indicated through information criteria.” The Theta method, “the most accurate method of the M3 Competition”, is a single method too; Hyndman and Koehler (2006)¹ cite a demonstration that it “is equivalent to simple exponential smoothing (SES) with drift where the drift is half the value of the slope of a linear regression fitted to the data”.

ARIMA models come from the Box-Jenkins methodology, which, in Makridakis et al.²’s account, “claimed that an optimal model existed for each time series” and that it could be found by identifying, estimating and testing a model. In M4, ARIMA forecasting was automatic: “An automatic selection of possible ARIMA models is performed and the best one is chosen using appropriate selection criteria.” The M4 report describes ARIMA by how a model is chosen, not by its equations, and this guide does the same. Both automatic procedures choose by an information criterion computed on the fitting data: “ETS and ARIMA select the simplest model that is still good at producing accurate short-term forecasts”, and the authors add that such a 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”.

A forecast combination averages the forecasts of several methods. M4’s main benchmark, Comb, is “The simple arithmetic average of SES, Holt and Damped exponential smoothing”. M4 also ran two machine learning benchmarks, a perceptron (MLP) and a recurrent network (RNN), each “of a very basic architecture and parameterization” and “trained individually for each series”. A hybrid joins the two traditions: the winning M4 method “mixes ES formulas with a recurrent neural network (RNN) forecasting engine”, and its smoothing parameters “are fitted concurrently with the RNN weights by the same gradient descent method”.

What did 100,000 series show about these methods?

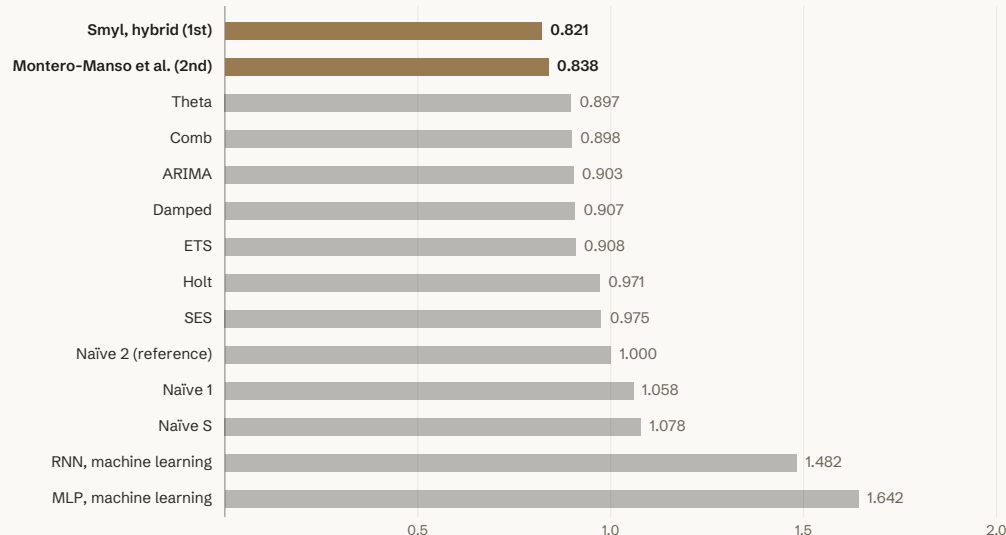
§ 3 of 6

M4 ran from January to May 2018. It asked for forecasts of 100,000 series: 23,000 yearly, 24,000 quarterly, 48,000 monthly and 5,000 weekly, daily or hourly (Table 1, p. 3 of the in-press version). Participants forecast “six for yearly, eight for quarterly and 18 for monthly series”, and 13, 14 and 48 steps for the weekly, daily and hourly ones, and the organisers scored the forecasts against the test set they had kept back. Of 248 registered participants, 49 submitted valid forecasts for every series; with ten benchmarks and two standards for comparison, ETS and ARIMA, the results tables rank 61 methods.

Accuracy was scored with two measures, which the next section explains: the symmetric mean absolute percentage error (sMAPE) and the mean absolute scaled error (MASE). Each was averaged over a series’ forecast horizons first and over series second, “thus weighting all of the series in the dataset equally”. The ranking used the overall weighted average (OWA) of the two: “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.” An OWA of 1 is Naïve 2’s accuracy and lower is better; in the authors’ example, an OWA of 0.88 means a method is “about 12% more accurate than Naïve 2, taking into account both MASE and sMAPE”. They add: “We will not pretend that the selected measures were the most appropriate to use”. Figure 1 shows 14 of the 61 methods.

Figure 1 Accuracy of 14 of the 61 M4 methods, relative to Naïve 2

A hybrid and a combination scored best. Theta, Comb, ARIMA, Damped and ETS lay within 0.011 of each other; the two machine learning benchmarks were the least accurate of the 14 shown.



Horizontal axis: OWA, the mean of a method's sMAPE and MASE, each divided by Naïve 2's, so Naïve 2 is 1 by construction; lower is more accurate. Comb is the plain average of SES, Holt and Damped, ARIMA and ETS select their models automatically, Naïve S is the seasonal naive method, and RNN and MLP are the two machine learning benchmarks. Table 4 (pp. 8 and 9 of the in-press version) gives no interval for OWA; of Damped, Comb and ARIMA, the authors write that hardly any difference among them is statistically significant by their test, which ranks methods by sMAPE. The figure leaves out the other 47 submissions, among them the two least accurate of all 61 methods, both pure statistical, at 2.561 and 4.387.

Source Makridakis et al. (2020), Table 4, pp. 8 and 9 of the in-press version held (pp. 1 to 21): the overall weighted average (OWA) of each method's sMAPE and MASE, each divided by Naïve 2's, over all 100,000 series. Shown are the two best submissions, all ten benchmarks and the two standards for comparison, ETS and ARIMA; the other 47 submissions are left out. Method types are those the table assigns. The authors' figures, not a re-estimation.

The authors call combining their most important finding: “all of the top-performing methods, in terms of both PFs and PIs, were combinations of mostly statistical methods”, where PFs are point forecasts and PIs prediction intervals, and “there was only one pure statistical method among the first ten most-accurate ones”, while “all of the ten least-accurate methods were either pure statistical or pure ML”; the two least accurate of all 61, at 2.561 and 4.387, were statistical submissions that Figure 1 leaves out (Table 4, p. 9 of the in-press version). The same table classes the winner as a hybrid and the next

five as combinations (p. 8 of the in-press version). Comb, the plain average, scored an OWA of 0.898 against 0.975 for SES, 0.971 for Holt and 0.907 for Damped, so “Comb provides more accurate PFs than the three individual methods being combined”. The authors also warn that “combining does not guarantee that their forecasts will outperform single models”: Theta, a single method, scored 0.897.

Among the best benchmarks and standards the differences were small: Theta 0.897, Comb 0.898, ARIMA 0.903, Damped 0.907 and ETS 0.908. Of Damped, Comb and ARIMA, the authors write that their accuracies “are very similar in M4” and that “Fig. 1 indicates that hardly any statistically significant differences among the methods can be found”. Their history of ARIMA makes a related point: it took theoretical statisticians a long time, they write, to accept from the earlier M competitions that “goodness of fit was not necessarily related to post-sample accuracy”, so a model that fits the past well need not forecast the test period well.

The most accurate method, Smyl’s hybrid of exponential smoothing and a recurrent network, scored 0.821, and the authors put its improvement over Comb “close to an impressive 10%” (9.4 per cent on SMAPE and 8.6 per cent on OWA in Table 4). Second, at 0.838, was “a combination of seven statistical methods and a ML one”, with the weights for averaging “calculated by a ML algorithm that was trained to minimize the forecasting errors through holdout tests for each individual series”. The machine learning entries did badly: “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.” The two basic machine learning benchmarks, at 1.482 and 1.642, were the least accurate methods in Figure 1. The authors’ explanation is a conjecture: the reasons “could be the limited usage of cross-learning”, that is, of information from other series (which the three best methods did use), together with short individual series, over-fitting and non-stationary data. They qualify the result themselves: “given the paucity of entries in the ML category, more information might still be needed to make it safe to generalize this finding”. The [AI demand forecasting guide](#) sets out what a forecast system around such a model has to declare beyond its accuracy.

M4 also scored prediction intervals, for the 20 submissions that supplied them, Naïve 1 and the two standards, ETS and ARIMA: “The M4 Competition adopted a 95% prediction interval (PI) for estimating the uncertainty around the point forecasts.” The two best methods

were best here too. Their intervals contained 94.78 and 95.96 per cent of the future values (Table 5, p. 10 of the in-press version), while those of the other 21 methods contained between 79.68 and 93.39 per cent and so were too narrow, as most methods' intervals had been before them; the authors write that “the underestimation of uncertainty by the remaining methods must be addressed”. They call point forecasts “insufficient and often dangerous to utilize for decision making on their own” without intervals.

Why can forecast errors not be averaged across series as they stand? § 4 of 6

Because each error carries the units of its series. Hyndman and Koehler (2006)¹ call measures such as the mean absolute error (MAE) and the root mean squared error scale-dependent: “These are useful when comparing different methods applied to the same set of data, but should not be used, for example, when comparing across data sets that have different scales.” In their words, “It makes no sense to compute the MAE across series (due to their different scales).” On this guide’s reading, an error of 50 units is large for a product line that sells 100 a month and small for one that sells 10,000. If the large line misses by 500 units and the small one by 50, the mean absolute error of the two is 275, set almost entirely by the large line. The [forecast accuracy essay](#) sets out why signed and absolute errors, their distribution, the horizon and the denominator stay separate when one forecast’s errors are summarized; this section is about putting different series on one scale.

Percentage errors remove the units and are “frequently used to compare forecast performance across different data sets”, but they divide by the actual value. That makes them, the mean absolute percentage error (MAPE) among them, infinite or undefined when an actual is zero and “extremely skewed” when one is close to zero; and they “assume a meaningful zero”. The symmetric version, sMAPE, divides by the actual and the forecast together, which eases the problem without removing it: near zero, “the measure still involves division by a number close to zero”, “sMAPE and sMdAPE can take negative values”, and for the same actual a low forecast is penalised more than a high one.

Their examples make this concrete. One is three years of “monthly sales of a lubricant product sold in large containers”, a series with months of no sales. On it the MAPE was infinite or undefined for all four methods they compared, in the fitting period and in the test period, and only the MASE could be computed for every method (Table 3, p. 681).

Such series are not unusual, they stress: “intermittent demand data often contain zeros and many time series of interest to forecasters contain negative observations”. Even in a series without zeros, M3 series N0472, a relative measure that divides by the naive method’s errors, the geometric mean relative absolute error, was infinite or undefined for every method (Table 1, p. 681), because of “the occurrence of consecutive observations which take the same value, something that very often occurs with real data”.

Their proposal is to scale each error by how well the naive method did on the same series in the past. The scaled error divides a forecast error by the mean absolute error of one-step naive forecasts over the fitting period, “which is clearly independent of the scale of the data”, and the MASE is the mean of the absolute scaled errors. Below one, a method’s errors were on average smaller than the naive method’s one-step errors in the fitting period; above one, larger. Because that yardstick is a one-step forecast, “the scaled errors for multi-step forecasts are typically larger than one”, so on this guide’s reading a MASE of 1.5 at a horizon of 18 months does not show that the method lost to the naive forecast at that horizon. The measure is infinite or undefined only “when all historical observations are equal”. Hyndman and Koehler¹ propose scaled errors as “the standard approach in comparing forecast accuracy across series on different scales”, and they leave room for the older measures: “if all series are on the same scale, then the MAE may be preferred because it is simpler to explain”, and “If all data are positive and much greater than zero, the MAPE may still be preferred for reasons of simplicity.”

M4 used a variant. Its MASE scales by the seasonal naive method instead, with a season length of “12 for monthly, four for quarterly, 24 for hourly and one for yearly, weekly and daily data”; the authors note that “originally m was set equal to one regardless of the frequency of the data”. M4 also kept sMAPE out of trouble by design: the series were “scaled to prevent negative observations and values lower than 10”, by adding a constant where needed, which happened 29 times, and “low-volume and intermittent time series were not considered for the M4 dataset”. A company’s own product lines come with no such guarantee.

What should a forecast comparison fix before the first score?

§ 5 of 6

Six things, decided before any method is scored: the series, with its unit and frequency; the test period that no method may see; the horizons, chosen for the decision the forecast serves; the benchmarks; the accuracy measure and how it is scaled; and how the result will be reported, against each benchmark at each horizon. The fields follow M4's design, which kept a test set back, set its horizons "based on the nature of the decisions that each frequency of data is most likely to support" and compared each submission with benchmarks "of well-known properties", and Hyndman and Koehler's¹ case for scaled errors. The [forecast value added essay](#) applies the same discipline to a change that people make to a statistical baseline.

Figure 2 A forecast comparison record

Write down the series, test period, horizon, benchmarks, measure and the form of the result before any method is scored.

Series	Test period	Horizon	Benchmarks	Measure	Result
Which series, unit and frequency?	Which last values stay unseen?	How many steps ahead, and why?	Which naive methods and which average?	Scaled? Any zero or near-zero values?	Better than each benchmark, at each horizon?

Report the benchmark scores next to the method score, horizon by horizon.

One row per series or group of series. Fix the test period and the benchmarks before the first method is fitted. A score reported without its benchmarks cannot be read.

Source Author's decision worksheet. The test period and horizon fields follow Makridakis et al. (2020), pp. 4 and 5 of the in-press version held; the benchmark field follows Makridakis et al. (2020), pp. 3 and 4 of the in-press version; the measure field follows Hyndman and Koehler (2006), pp. 682 to 686, and Makridakis et al. (2020), p. 5 of the in-press version. The series and result fields, and the rule in the footer, are this guide's. The fields are reader inputs; no company data is supplied.

An illustration, with no company data: a manufacturer forecasts monthly unit demand for 40 spare-part lines, from a few units a month to several thousand, with three years of history. The record names each line's unit and frequency, keeps the last twelve months back as the test period, with one forecast origin per line as in M4, and sets the horizons from the decisions: three months for production planning, twelve for the budget. It compares any candidate method with the seasonal naive method, Naïve 2 and a plain average of exponential smoothing methods, because in M4 that average was more accurate than each of its three parts. It scales errors per line in the MASE way, because a mean absolute error pooled across lines would be set by the largest ones, and because some small lines have months with no demand, where a percentage error is undefined. For those lines it also records that M4 left such series out and that neither source tests methods built for intermittent demand. These are the author's examples and have not been tested.

Which of these results would not survive a short or intermittent series? § 6 of 6

The sources answer a narrower question than a revenue or demand planner asks. M4 is one competition, run once, on continuous series, and the authors say so: “the findings of M4 refer to continuous business series, meaning that some of them may not apply for low-volume or intermittent series”. Its series were stripped of their starting dates, shifted where they fell below 10, and forecast from one cut-off each, by methods that saw no other variable. The field was self-selected, since “Many participants realized that their methods were no more accurate than the benchmarks, and therefore decided not to submit their forecasts”, and the machine learning result rests on five submissions from 2018, which, the authors say, may be too few to generalize from. Later forecasting competitions are not among this guide's sources. Hyndman and Koehler (2006)¹ argue from the definitions of the measures and demonstrate on three example series and the M3 data; they do not test which measure leads to better decisions.

On this guide's reading, three things carry over with some confidence: compare any method with named naive benchmarks and a simple combination before trusting it; judge methods on a test period they did not see, not on fit to the past; and compare accuracy across series with a scaled error. Less certain is the ranking of method families for a company's own series. In M4 the best benchmarks and standards sat close together, and among Damped, Comb and ARIMA hardly any

difference was statistically significant by the authors' test; the winners combined methods or joined statistical and machine learning parts, and the least accurate methods were pure ones of either kind. How any of this holds for short revenue series, for products with months of zero demand, or for a forecast that should also use a pipeline or a price change, the sources do not say.

 End of essay

References

Page numbers given for Makridakis et al. (2020) are those of the article in press held, paginated 1 to 21; the volume, issue and pages below are from the Crossref record.

- 1 Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4), 679–688.
<https://doi.org/10.1016/j.ijforecast.2006.03.001>
- 2 Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2020). The M4 Competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, 36(1), 54–74.
<https://doi.org/10.1016/j.ijforecast.2019.04.014>

How to cite this essay

Isoglu, S. (2026, October 18). What is time series forecasting? Sinan Isoglu.
<https://isoglu.com/insights/journal/what-is-time-series-forecasting/>
Online version: <https://isoglu.com/insights/journal/what-is-time-series-forecasting/>

Sinan Isoglu, MBA (Quantic)

Commercial growth leader, lecturer and doctoral researcher
© 2026 Sinan Isoglu · isoglu.com · All rights reserved