**MANAGEMENT SCIENCE**
Vol. 00, No. 0, Xxxxx 0000, pp. 000–000
ISSN 0025-1909 | EISSN 1526-5501 | 00 | 0000 | 0001

**INFORMS**
DOI 10.1287/xxxx.0000.0000


# Forecast Selection and Representativeness

Fotios Petropoulos
School of Management, University of Bath, UK, f.petropoulos@bath.ac.uk

Enno Siemsen
Wisconsin School of Business, University of Wisconsin-Madison, USA, esiemsen@wisc.edu

Effective approaches to forecast model selection are crucial to improve forecast accuracy and to facilitate the use of forecasts for decision-making processes. Information criteria or cross-validation are common approaches of forecast model selection. Both methods compare forecasts with the respective actual realizations. However, no existing selection method assesses out-of-sample forecasts before the actual values become available – a technique used in human judgment in this context. Research in judgmental model selection emphasizes that human judgment can be superior to statistical selection procedures in evaluating the quality of forecasting models. We therefore propose a new way of statistical model selection based on these insights from human judgment. Our approach relies on an asynchronous comparison of forecasts and actual values, allowing for an ex-ante evaluation of forecasts via representativeness. We tested this criterion on numerous time series. Results from our analyses provide evidence that forecast performance can be improved when models are selected based on their representativeness.

*Key words*: forecasting, model selection, model combination, information criteria, representativeness, empirical evaluation

## 1. Introduction

Selecting the right forecasting model is a challenge. In a world of fat tail distributions (Taleb 2008, Makridakis et al. 2010), the best forecasting model can be elusive. Although a single model might perform well in homogeneous settings across a large set of time series, the heterogeneity and dynamism characterizing the market environments of many firms require different forecasting models. Fildes and Petropoulos (2015) showed that if we always chose the best performing model among several widely used ones, the forecast error would decrease by as much as 30%. Such a reduction in error would in turn translate into large savings in inventory, increased customer service, and a significant decrease in waste. This insight has led to he quest for the "horses for courses" in forecasting (Petropoulos et al. 2014), and emphasized the importance of developing effective selection criteria.

Criteria for model selection help in picking a single ("best") model for a specific time series; they also support combining forecasts across models. Forecast combinations are beneficial in improving forecasting performance (Timmermann 2006, Lichtendahl et al. 2013) and in decreasing the variance

in forecasts (Hibon and Evgeniou 2005). Although simple combinations can work well (Genre et al. 2013), smart combinations of forecasts, in which combination weights depend on some outcome, can significantly increase performance (Kolassa 2011, Kourentzes et al. 2019). The ability to accurately assess whether a model is suitable for forecasting a specific series facilitates such smart combinations.

Existing algorithms for model selection and combination are based on information criteria, assessment of past forecasting performance, and an analysis of time series features. One common element among these approaches is that they look backward and focus on observed past values. No existing selection criterion assesses the forecasts produced for the out-of-sample periods.

Motivated by research on human judgment in forecast model selection, we propose a novel selection criterion. This new criterion reflects not only on past model fit, but also looks forward by assessing the out-of-sample forecasts in terms of their representativeness compared with past actuals. We will define this approach, argue on its theoretical differences over existing selection approaches, and perform a large empirical evaluation of this approach. Our results suggest that representativeness should be the new paradigm in forecast model selection and combination.

The next section surveys the relevant academic literature, followed by the motivation for our study and the research gap. In section 3, we define this new criterion for forecast selection, differentiate it from existing approaches, and reason when and why it may work better. Section 4 presents the empirical results of the study. Section 5 discusses the theoretical and practical significance of the results. We conclude in Section 6.

## 2. Background Research and Gap

### 2.1. Statistical Model Selection

Notable families of forecasting models include exponential smoothing (ETS – Hyndman et al. 2002, Gardner 2006) and the autoregressive integrated moving average (ARIMA – Box et al. 2008) models. Exponential smoothing is based on three time series components – error, trend, and seasonality – that can appear in different forms (such as additive or multiplicative) and that are conceptualized as relatively stable or changing. The combination of these components and forms allows for 30 possible ETS models. Autoregressive integrated moving average models are based on autoregressive and moving average components. In theory, the ARIMA family consists of an infinite number of possible models because the order of the autoregressive and moving average terms is without upper bounds. In practice, only small orders of autoregressive and moving average terms should be considered to avoid model overfitting.

Regardless which forecasting models are considered, selecting the right one is important. For instance, if the task is to forecast a series that exhibits a strong seasonal signal, a non-seasonal model can result in poor forecasts. Although aggregate (cross-sectional) model selection has been suggested,

especially when dealing with sets of homogeneous series (Fildes 1989), individual (per series) model selection is the preferred option that results in superior performance (Fildes and Petropoulos 2015). Moreover, as patterns of time series evolve, the selected forecasting models should be updated. Given that firms need to forecast for many markets and products, this selection process among different forecasting models is often an automated "machine" task.

A prevalent statistical selection approach is based on information criteria (IC – Gardner 2006, Hyndman et al. 2008). This approach offers a balance between performance (goodness-of-fit) and model complexity. Information criteria are derived from the likelihood of the models, which is calculated given a specific cost function, then penalized based on the size of the model (number of parameters involved) and, sometimes, on the number of available historical observations. The general form of an information criterion can be expressed as IC = performance measure + penalty, in which performance is measured with in-sample fit.

In time-series forecasting, Goodrich (1990) was the first to propose the use of IC as the basis for selection between models. His rationale for using IC was based on avoiding complex models (and, thus, overfitting). Information criteria are expected to work well for stable and relatively short series that would render hold-out evaluation difficult (Ord et al. 2017). Gardner (2006) argued that another advantage of IC is their ability to select between models with different types of seasonality (additive versus multiplicative).

Popular IC include the Bayesian Information Criterion (BIC, developed by Schwarz 1978), the Akaike's Information Criterion (AIC – Akaike 1974) and the bias corrected (for small sample sizes) Akaike's Information Criterion ($\text{AIC}_c$ – Sugiura 1978). The AIC is defined as $\text{AIC} = -2\log(L) + 2k$, in which $L$ is the likelihood of the model and $k$ is the number of the associated parameters. The log-likelihood of a model can be written in terms of the residuals sum of squares (RSS) as $\log(L) = -\frac{n}{2}\ln(\frac{\text{RSS}}{n}) + C$, in which $C$ is a data-dependent constant and $n$ is the number of observations. In this study, we focused on the $\text{AIC}_c$, for which the penalty applied depends only on the size of the model ($k$) and the length of the in-sample data ($n$). $\text{AIC}_c$ is defined as

$$\text{AIC}_c = \underbrace{-2\log(L)}_{\text{performance measure}} + \underbrace{2k + \frac{2k(k+1)}{n-k-1}}_{\text{penalty}} = \underbrace{n\ln\frac{\text{RSS}}{n}}_{\text{performance measure}} + \underbrace{2k + \frac{2k(k+1)}{n-k-1}}_{\text{penalty}} - 2C. \quad (1)$$

We focus on $\text{AIC}_c$ in our research for two reasons. First, the $\text{AIC}_c$ is the default setting for model selection in statistical forecasting packages, such as the *smooth* and the *forecast* packages for the R statistical software (Hyndman et al. 2019, Svetunkov 2019). Second, the $\text{AIC}_c$ may have some advantages in small samples, and its value converges to that of the AIC for large samples. Therefore, Kolassa (2011) suggests always using the $\text{AIC}_c$.

Validation is another approach to model selection, based on evaluating past performance (Fildes and Petropoulos 2015). The available in-sample data are divided into training and validation sets. The training set is used to fit models. The forecasts from these fitted models are then evaluated against the validation set. If the validation set is large enough, the process can be repeated by adding additional observations at the end of the training set, refitting models, and re-evaluating the forecasts; i.e., cross-validation (CV) for time-series data. The process is similar to the fixed and rolling-origin evaluation approaches (Tashman 2000). Validation and CV are prevalent methods within the field of machine learning. Neither of these two approaches penalizes model complexity.

Cross-validation generally outperforms validation in model selection because the single validation window can include unusual observations or temporal pattern breaks. Averaging the performance of models across many origins is more robust (Tashman 2000). Both approaches work with either one-step-ahead or multiple-steps-ahead evaluations. Using more than one step-ahead forecasts in the evaluation will result in better performance (Fildes and Petropoulos 2015). Furthermore, it is a good strategy to match the evaluation horizon with the required out-of-sample horizon and to match the cost function with the performance indicator that will be used when the forecast is used. Montero-Manso et al. (2020) proposed a well-performing approach for time-series forecasting based on CV as part of their submission for the M4 forecasting competition (Makridakis et al. 2020). Other noteworthy approaches for statistical model selection include rules based on the variances of differences series (Gardner and McKenzie 1988), rules based on expertise and domain knowledge (Adya et al. 2001), discriminant analysis (Shah 1997), descriptive statistics (Meade 2000), time-series features that use regression models on past forecasting performance (Petropoulos et al. 2014), and meta-learning (Talagala et al. 2018).

### 2.2. Insights from Judgment

As Kahneman and Tversky stated, "judgments of likelihood essentially coincide with judgments of similarity but are quite unlike the estimates of base rates" (1973, p. 239). This insight would lead one to expect that humans performing judgmental model selection will base their selection on the similarity between the out-of-sample forecasts and the in-sample historical values. A stream of research examines whether the way people create forecasts is based on such similarity (Harvey 1995). A series of judgmental forecasts produced by a human forecaster tends to resemble the series of data provided. This is odd, because the data should contain more noise than the forecasts. However, forecasters tend to reproduce the noise in a time series in their forecasts, rather than filter it out. As the noise in the data increases, so does the noise in the series of judgmental forecasts.

Human judgment contains noise, which in turns deteriorates performance (Kahneman et al. 2006). The observation that the noise of the time series influences this noise in judgmental forecasts is

important. Similar effects exist in related decision making, such as inventory orders, in which the noise of the underlying time series influences the consistency of ordering decisions (Lee and Siemsen 2017). Harvey et al. (1997) examined many alternative explanations for this phenomenon, such as an inability to perceive patterns, or a desire to mask such patterns with noise. After several experiments, he settled on representativeness as the best explanation: People add noise to their series of forecasts to make each forecast typical of the data underlying their forecasts. This conclusion would suggest that the representativeness of the forecasts is indeed an intuitive criterion for human decision makers; forecasters prefer forecasts that are representative.

What can we learn from this insight about the efficacy of human judgment in model selection? Three recent studies investigated this question (Petropoulos et al. 2018, Han et al. 2019, De Baets and Harvey 2020). Petropoulos et al. (2018) performed a behavioral experiment to determine whether humans can effectively select between statistically produced forecasts. Participants in the experiment had to select between forecasts from four widely-used exponential smoothing models. They received a graph that displayed both the historical actual values (monthly frequency) as well as the one-year-ahead forecasts. The performance of the judgmentally selected forecasts was compared with a selection through AIC. Results suggested that on average, humans select models as well as AIC does.

Another important finding by Petropoulos et al. (2018) was the importance of the user interface. The authors demonstrated that an interface based on decomposition and selection of applicable series patterns (trend/seasonality) is superior to simply selecting between alternative forecasts. This result was confirmed by Han et al. (2019), who also demonstrated that the decomposition interface requires less cognitive load and working memory. The study by De Baets and Harvey (2020) provided evidence that the ability of people to judgmentally select the best model depends on the relative performance between models and the noise in the series. When performance is similar and noise is high, selecting a model is more difficult. Apart from the noise, the strength and the direction of the trend are also important factors that influence the efficacy of judgmental model selection (Han et al. 2019).

The most relevant finding of Petropoulos et al. (2018) for our study referred to the frequency with which humans versus AIC select the best and the worst models. Specifically, selection by AIC outperformed humans in identifying ex-ante the model performing the best ex-post. However, humans outperformed AIC in rejecting the worst models. We believe this ability of humans to effectively avoid bad forecasts was due to the graphical tool they were provided: It put them in a position to perform a mental extrapolation and indirectly compare the out-of-sample forecasts with the historical data – in other words, to assess representativeness. This in turn allowed them to reject forecasts that seemed unreasonable. In a relevant commentary, Goodwin (2019) hypothesized that "the task draws the forecaster's attention to the big picture of how the data series has behaved over its entire history [...] availing [the forecaster] of all the useful information that lies in the full data history".

There is evidence that human forecasters are influenced by representativeness in their preference for some forecasts over others, and that this criterion – while also leading to some unnecessary noise if forecasters are allowed to influence forecasts – may allow them to select a good forecasting model among a set of different models. Thus, in the right context, representativeness appears as a fast and frugal heuristic (Gigerenzer and Todd 1999). Using this criterion in a selection algorithm is thus akin to using insights from human judgment in the creation of an algorithm.

There is extensive research that examines how algorithms and decision makers can coexist – a question of increased importance in the age of analytics. One of the guiding principles of the Watson artificial intelligence (AI) project at IBM is that the purpose of AI is to augment, not to replace, human decision makers.[1] Starting with the work of Blattberg and Hoch (1990), researchers long pointed to combination mechanisms between human judgment and algorithms as superior to either approach. The resulting literature stream is summarized in Arvan et al. (2019). Combination methods included averaging algorithmic forecasts with judgmental forecasts, allowing judgmental adjustments to forecasts, or creating an algorithmic correction of a judgmental forecast. Recent work has also begun to examine how best to elicit information from human forecasters to feed into algorithms (Flicker 2018, Rouba et al. 2021, Brau et al. 2021). Another line of research explores how decision making can benefit if an algorithm transfers exceptional cases to a human decision maker (Fügener et al. 2019). More relevant to our work, a further line of thought uses the insights from human judgment to design better algorithms. For example, van Donselaar et al. (2010) showed how insights from store manager orders can be used to improve inventory-ordering software. Along the same lines, we propose to develop an algorithm that can outperform existing algorithms by using the insight that human decision makers base their effective model selection on representativeness.

### 2.3. The Research Gap

Current statistical criteria to forecast model selection focus on past performance and the ability of forecasts to fit in-sample data. To the best of our knowledge, no statistical selection approach considers the representativeness of the out-of-sample forecasts – that is, whether such forecasts could possibly match with reality given the past observed data. In this study, we therefore propose a new statistical approach for forecast selection based on representativeness and evaluate it against several state-of-the-art statistical alternatives. Our motivation for this approach is the good forecasting performance of judgmental model selection and the potential of humans to avoid bad forecasts (Petropoulos et al. 2018). Counter to the trend of eliminating human judgment from supply chain processes (Lyall et al. 2018), our research contributes to the growing stream of research that takes a slightly different point of view: Much can learned from human judgment in the design of algorithms.

[1] See https://www.ibm.com/blogs/policy/trust-principles/

# 3. Selecting via representativeness

## 3.1. The representativeness of the forecasts

Selecting models based on information criteria or validation/CV relies on a comparison of *concurrent* pairs of actual and forecasted values. We argue that an *asynchronous* comparison is also feasible, by capturing the representativeness of the out-of-sample forecasts with the in-sample data. We regard *forecast representativeness* as the correspondence between the patterns of the out-of-sample forecasts and the actual values. For example, if the actual data are strongly seasonal, then a representative set of forecasts exhibits a similar seasonal pattern. If the actual data exhibit a positive linear trend, then a representative set of forecasts behaves similarly. In this study, we focus on measuring representativeness in terms of point forecasts, that is, how close the out-of-sample point forecasts are compared with past actual values. We apply appropriate data transformations and scaling to make such comparisons valid. Given that distribution forecasts are becoming more important, we also discuss how representativeness extends to distribution forecasts in section 5.

Figure 1 provides an illustrative example of representativeness. Without a loss of generality, let us assume that we have a monthly series, $y$, of $n = 48$ in-sample observations (four years) that exhibits trend and seasonality. Let us also assume that we wish to produce forecasts, $f$, for the next year ($h = 12$). We then produce forecasts using two models. The first captures only the trend in the data. The second model captures both the trend and the seasonality. The actual data and the point forecasts from these two models are depicted in the first panel of figure 1 with black, red, and blue, respectively. It should be noted that matching the periodicity of the data, $s$, with the forecast horizon, $h$, simplifies the presentation but is not necessary to measure forecast representativeness.

Assessing representativeness requires measuring the similarity between the point forecasts from each model and the past in-sample data. Given that we have forecasts for the next year and historical observations for the past four years, we consider four non-overlapping in-sample windows for this comparison. Specifically, we compare each of these in-sample windows to the forecasts from each model. To render the comparison meaningful, we first apply a Box-Cox transformation to stabilize the variances and convert multiplicative seasonal patterns into additive ones. The Box-Cox transformation is applied using the same parameter value (i.e., $\lambda$) for both the actual in-sample data as well as the forecasts. The parameter, $\lambda$, was automatically selected using Guerrero's method (Guerrero 1993) on the actual data. We then scale the Box-Cox transformed data by subtracting the mean of an in-sample window and dividing by the standard deviation of the same in-sample window. The forecasts are scaled as many times as the number of windows by subtracting their mean and dividing with the standard deviation of the respective in-sample window. Note that we use the variance of each in-sample window as reference, with the forecasts being appropriately adjusted for every window.

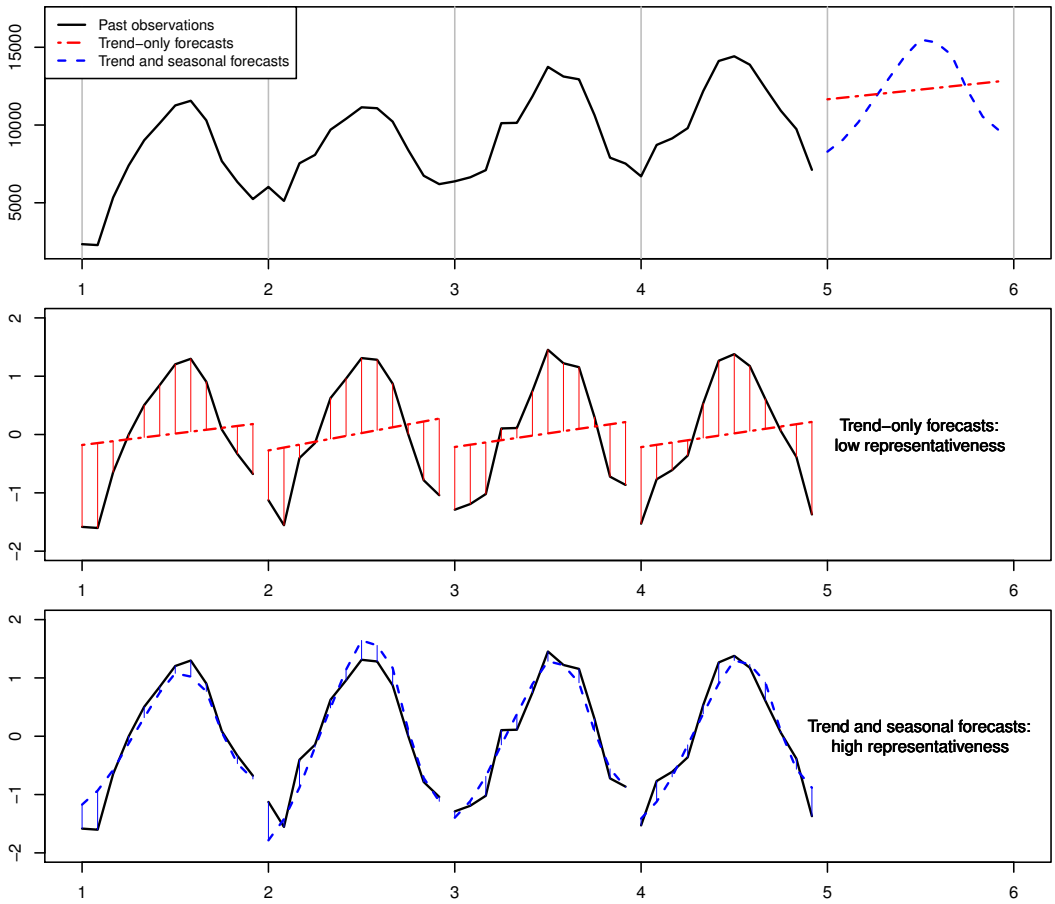


**Figure 1** **A toy example of forecast representativeness.**

We present the results of the above window-splitting and transformation-scaling process in the second and third panels of figure 1. We measure representativeness as the closeness of the values of the point forecasts with the past data – or, the *representativeness gap* as the distance between the forecasts and the past data. We use the sum of the absolute deviations ($\mathcal{L}_1$ norm) as our distance measure. For our toy example (figure 1), it is clear that the trend and seasonal forecasts are by far more representative that the trend-only forecasts, and should be preferred.

Representativeness can also be measured for non-seasonal data (such as yearly data). In such cases, the length of the non-overlapping in-sample windows should match the forecasting horizon. Generally, we set the length of the in-sample windows to be $p = \lceil h/s \rceil s$, with $s$ being the periodicity of the series (length of seasonal cycle), and $\lceil \cdot \rceil$ the ceiling function. The number of available windows is $\lfloor n/p \rfloor$, in which $\lfloor \cdot \rfloor$ is the floor function. As a result, the first $n - \lfloor n/p \rfloor$ observations are not considered in representativeness. In the cases of multiple seasonal cycles, $s$ can simply refer to the longest cycle. If the number of the past available observations is larger than the forecasting horizon ($n > h$), as in figure 1, we can measure representativeness over several non-overlapping windows. In case of $n < h$, we can measure representativeness by comparing a subset of the forecast horizons.

Data patterns (such as trends and seasonality) change over time. This implies that assessing representativeness would benefit from applying a higher weight to more recent windows of data. This weighting can be controlled through a *discount factor*, $\delta \in [0,1]$. Higher $\delta$ values result in higher discount rates. When $\delta = 0$, all data windows are considered with equal weights. A value of $\delta > 0$ imitates the way humans assess the forecasts by anchoring on more recent past values. When $\delta = 1$, only the most recent time window is compared with the forecasts. We suggest $\delta = 0.5$ as a standard parameter value, meaning that the ratio of the weights between two consecutive windows is 2 (half-life). A sensitivity analysis on this parameter is discussed in section 4.3.

In summary, we propose that, once both vectors have been transformed and scaled, the gap in the point-forecast representativeness is measured as the weighted sum of the distances of the out-of-sample forecasts with non-overlapping windows of the observed values. More formally,

$$\text{representativeness gap} = \sum_{i=1}^{\lfloor n/p \rfloor} (1-\delta)^{i-1} \left\| \check{y}_{[i]}, \check{f} \right\|_1 \tag{2}$$

in which $\|\cdot,\cdot\|_1$ is the $\mathcal{L}_1$ norm (sum of the absolute differences) between two vectors; $y_{[i]}$ is a window of $y$ such that $y_{[1]}$ refers to the first $h$ of the $p$-most recent observations, $y_{[2]}$ refers to the first $h$ observations of the second most recent window of $p$ periods that does not overlap with $y_{[1]}$, and so on and so forth; and the vectors $\check{y}_{[i]}$ and $\check{f}$ are conversions of $y_{[i]}$ and $f$ so that transformation and scaling have been applied as detailed above.

The representativeness gap can be measured for different sets of forecasts produced by different models or methods. Lower values are better, suggesting higher representativeness. In the case of $p = 1$, the forecasts from all models are equally representative. This could be the case when forecasting yearly data with $s = 1$, and $h = 1$. To distinguish between models in such cases, a longer forecasting horizon ($h > 1$) can be used for the selection stage.

### 3.2. The REP criterion

We propose a new criterion for selecting between forecast models based on representativeness, denoted as REP. REP consists of a concurrent part that refers to the in-sample performance and an asynchronous part that measures the representativeness gap. This two-part criterion is similar to IC; however, the penalty no longer relies on model complexity but on the representativeness gap between the out-of-sample forecasts and the actuals. Just as IC would reject large models to avoid overfitting, REP will reject models that produce less representative forecasts. Our proposition is

$$\text{REP} = \text{performance gap} + \text{representativeness gap}. \tag{3}$$

We measure the performance gap in Equation (3) through the distance between the observed in-sample data, $y$, and the in-sample fitted forecasts of each model, as denoted by $g$. The differences are

calculated on the Box-Cox transformed and scaled vectors ($\check{y}$ and $\check{g}$) so that none of the quantities "performance gap" and "representativeness gap" in Equation (3) dominates the other. Perfectly matching the scales of "performance gap" and "representativeness gap" is possible, but not required. Information criteria have a similar asymmetric scaling; for example, in AIC the complexity penalty is fixed while maximum likelihood is increasing in the sample size. Further note that if $\delta > 0$, the representativeness gap will not increase as much as the performance gap does for a larger $n$.

Combining the two components of Equation (3), we get

$$\text{REP} = \underbrace{\|\check{y}, \check{g}\|_1}_{\text{performance gap}} + \underbrace{\sum_{i=1}^{\lfloor n/p \rfloor} (1-\delta)^{i-1} \left\|\check{y}_{[i]}, \check{f}\right\|_1}_{\text{representativeness gap}}. \tag{4}$$

The REP values from different forecasting models can be compared to select the one with the lowest REP. Alternatively, the REP values can form the basis for estimating weights for model combinations, with lower REP values translating into higher contribution weights (see Kolassa 2011, who proposed weights from IC).

### 3.3. Analytical insights

Consider a time series that exhibits a deterministic trend and/or seasonality. If there is little noise in the data, then we would expect any reasonable selection criterion to point toward the correctly specified model. However, as the signal-to-noise ratio decreases, the selection might change. Consider IC, specifically. An increase in the noise relative to the signal will result in more similar likelihoods for models with correctly specified and misspecified forms. Selection then becomes less driven by fit than by differences in the complexity between models. Information criteria will tend to select the simpler (i.e., less parameters) model, which may lead to the selection of a misspecified model. To illustrate this point, suppose that the data generating process can be described by $y_t = c + t\beta + m_t\gamma + \epsilon_t$. The term $m_t$ is a vector of binary values, indicating an increase in $y_t$ in every other period, and the term $t\beta$ indicates a linear trend. In essence, the above process describes a signal that consists of a constant, a deterministic trend, a deterministic seasonality with periodicity equal to 2, and an independent noise process that follows a normal distribution, $\epsilon \sim N(0, \sigma^2)$.

Further, suppose that we use two different forecasting models, $F_1$ and $F_2$. $F_1$ is misspecified by forecasting a level and trend only, but $F_2$ is correctly specified by forecasting level, trend, and a seasonal change with according periodicity. Both forecasting models perfectly estimate their parameters without error, resulting in

$$F_1 : \hat{c}_1 = c + \frac{\gamma}{2},\ \hat{\beta}_1 = \beta, \qquad \text{and} \qquad F_2 : \hat{c}_2 = c,\ \hat{\beta}_2 = \beta,\ \hat{\gamma}_2 = \gamma.$$

The expectation of the $\text{AIC}_c$ for model $F_1$ is $\mathbb{E}[\text{AIC}_c^1] = n\ln\left(\sigma^2 + \frac{\gamma^2}{4}\right) + 4 + \frac{12}{n-3} - 2C$. For model $F_2$, $\mathbb{E}[\text{AIC}_c^2] = n\ln(\sigma^2) + 6 + \frac{24}{n-4} - 2C$ (see electronic Appendix A, Case 3, for derivations). Selecting by $\text{AIC}_c$ will lead to the correct model on average only if $\mathbb{E}[\text{AIC}_c^1] > \mathbb{E}[\text{AIC}_c^2]$, or

$$\ln\left(1 + \frac{\gamma^2}{4\sigma^2}\right) > \frac{2(n-1)}{(n-3)(n-4)}.$$

By looking at the left side of this inequality, we can see a clear signal-to-noise ratio: As the noise increases, then the tendency to select a misspecified model will increase as well. The right side of the equation also shows a dependence on the sample size: As the sample size increases, then the tendency to select a correct model also increases.

We can repeat this analysis for REP instead of $\text{AIC}_c$. We should note that as a result of the assumption of perfect knowledge and the stability of the parameters (i.e., nothing is updated), the forecasts for future periods will be the same as the forecasts of past periods. Also assuming $\mathcal{L}_2$ (sum of the squared differences) for REP and setting $\delta = 0$ (i.e., no discounting of the past in-sample windows because the time series is stable), this means that REP simply compares $\text{RSS}_1$ to $\text{RSS}_2$. Because $\text{RSS}_2$ is always smaller than $\text{RSS}_1$, REP will always pick the correct model. The same insight holds for selecting via CV.

In the electronic Appendix A, Case 4, we consider a case in which $F_1$ is misspecified in such a way that it forecasts only level and seasonality (but not trend). We also consider two simpler cases (Cases 1 and 2), in which the data generating process consists of level and seasonality or level and trend, with the misspecified model able to forecast only the level. This simple analysis illustrates that IC tend to select the wrong model when (a) a pattern exists and (b) the signal-to-noise ratio of the underlying data is low. Selection based on representativeness would not suffer from this issue. In the electronic Appendix B, we conduct a simulation exercise that tests these theoretical insights.

### 3.4. Theoretical comparisons and practical examples

The previous subsection focused on the complexity penalty and the signal-to-noise ratio as an explanation for a performance advantage of REP. This subsection provides additional arguments for why we expect REP to work well in practice. We support our arguments with a series of stylized examples: Figure 2 presents nine cases based on real-life time series from the quarterly frequency of M1, M3, and M4 forecasting competitions (Makridakis et al. 1982, Makridakis and Hibon 2000, Makridakis et al. 2020). Exponential smoothing models were fitted using the observed actual data (depicted in black) and the "best" model is selected via $\text{AIC}_c$ (red), CV (blue), and REP (green) criteria. The forecasts selected by using these criteria are contrasted against the actual realizations (gray). The series identifiers are presented as labels in the graphs. For example, "M1 - Q98" refers to the $98^{\text{th}}$ quarterly series of the M1 competition.

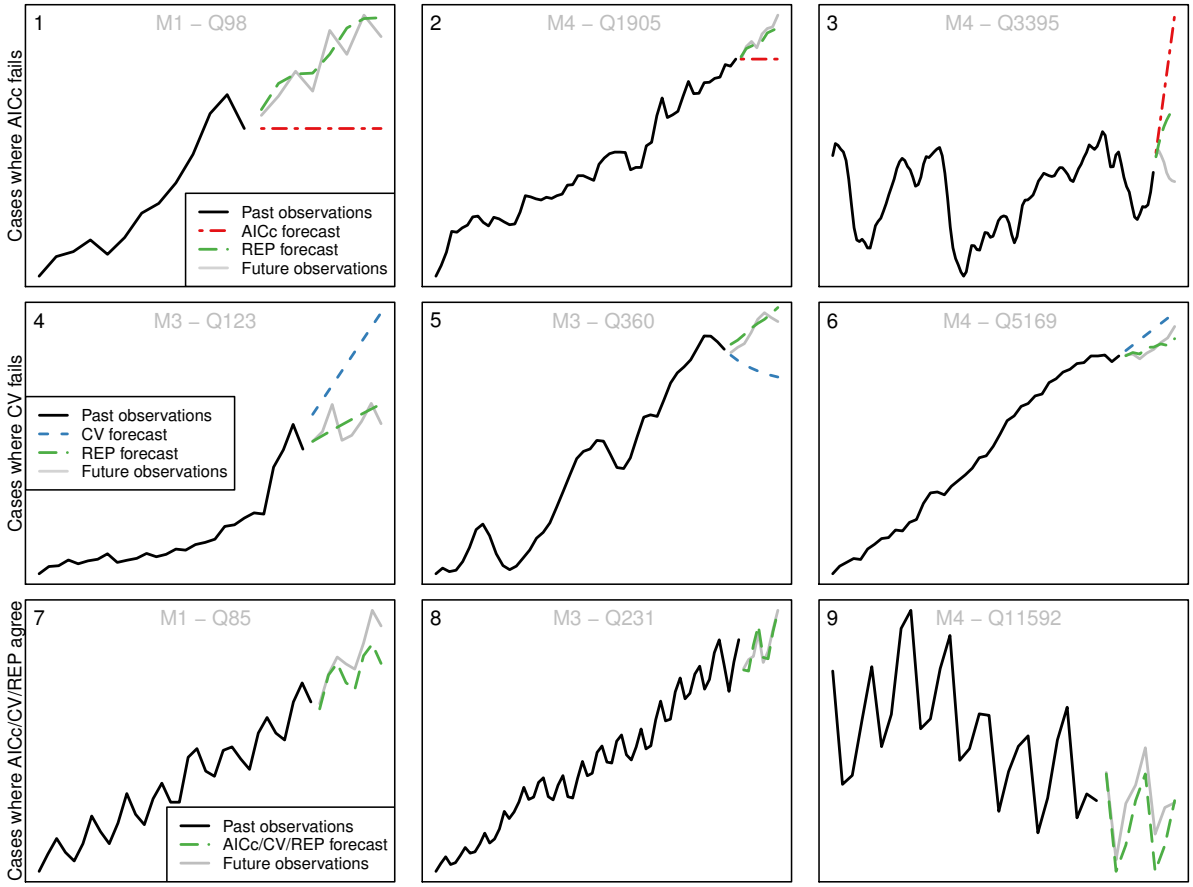


**Figure 2 Examples of cases in which the AIC$_c$ or the CV forecast is not representative.**

In cases 1 to 3, the forecasts selected by AIC$_c$ are not representative. In cases 4 to 6, the forecasts selected by CV are not representative. All three criteria are able to select representative forecasts in cases 7 to 9. The series identifiers are presented as labels in the graphs. For example, "M1 - Q98" refers to the 98$^{\text{th}}$ quarterly series of the M1 competition.

The main difference between REP and the other selection criteria is the use of the out-of-sample forecasts. REP is the only selection criterion that, in conjunction with the in-sample fit of the model, considers the out-of-sample forecasts and assesses them (via representativeness) before the future actual values occur. We expect that REP will reject models that produce forecasts with low representativeness, even if the respective models have produced low IC values or high past performance. In essence, REP does exactly what teachers of statistics advise their students to do: "Check the output of your models."

In cases 1 and 2 of figure 2, the trend in the data is clear. REP chooses trended forecasts whereas AIC$_c$ selects a model without a trend. The poor model choice using IC is a result of the penalty applied for complexity. This penality can become particularly excessive when the series are short (case 1) or when the signal-to-noise ratio is high (case 2). Similarly, in cases 4 and 5, although CV correctly opts for a trended model, its forecast are unrepresentative of the actuals. In case 4, the

trend is an exaggeration of the past and highly influenced by recent values; in case 5, the forecasts have a negative trend, although the global trend of the data is positive.

REP also provides a balance between in-sample fit and out-of-sample representativeness. Simply focusing on the in-sample fit and maximizing the probability that a model follows the true data generation process may lead to over-fitting. Even if over-fitting in terms of model structure is handled in IC by applying a penalty term for the size of the model, the method does not account for over-fitting the model parameters. In case 3 of figure 2, selection via $AIC_c$ fails because of the high values of the smoothing parameters of the respective "optimal" exponential smoothing model. Although the REP selection is not ideal either, it is still better than the $AIC_c$ selection.

In addition, REP assesses the representativeness of multiple-steps-ahead forecasts while IC focuses on one-step-ahead evaluations. Assuming normally distributed forecast errors, minimizing the AIC is asymptotically equivalent to minimizing the one-step-ahead forecast for the mean squared error. Multiple-steps-ahead evaluation tends to produce better results than one-step-ahead evaluation in the case of CV for model selection (Fildes and Petropoulos 2015). For example, the one-step-ahead forecasts for cases 2 and 3 of figure 2 is almost identical for $AIC_c$ and REP (the starting points of the green and red lines coincide); however, the forecasts for the later horizons differ significantly.

Finally, the standard definitions of IC and CV suggest that the comparison of the in-sample forecasts with past values is performed using equal weights across all observations. Representativeness divides the in-sample period into non-overlapping windows and assigns a higher importance to the windows nearer the forecast origin by discounting more distant windows. Exponentially weighted IC have been previously proposed (Taylor 2008) because of their better performance compared with the standard IC. For example, case 6 of figure 2 shows selection with CV may fail if performance on past data is treated equally and when the patterns have changed in the most recent observations. On the other hand, REP is able to select the better model by focusing on the most recent data.

Cases 7 to 9 present three examples in which all three selection criteria, $AIC_c$, CV, and REP, agree and are correct in picking the best performing model. This happens with a strong signal (pattern) in the data. As established in the previous subsection, we do not expect REP to have an advantage if the signal-to-noise ratio is high.

## 4. Empirical Evaluation

### 4.1. Design

In this section, we use a set of exponential smoothing models to compare the selection performance of REP with $AIC_c$ and CV. Exponential smoothing models are widely used in practice (Weller and Crone 2012), have robust forecasting performance in a range of data categories, and compute quickly. We ran our analysis in R, using the ets() function of the *forecast* package (Hyndman et al. 2019).

By default, the set of exponential smoothing models we consider does not include multiplicative trend models or models with additive error and multiplicative seasonal components. The resulting set includes 15 models, six of which are suitable for non-seasonal series. When modelling time series with frequencies ($s$) higher than 24, we used the es() function of the *smooth* package instead, matching the model space to that of the ets() function. Our analysis in sections 4.2 and 4.3 focuses on the exponential smoothing family. We applied REP on the ARIMA family as well as to a custom set of forecasts that included the Theta method (Assimakopoulos and Nikolopoulos 2000) in section 4.4.

For each criterion, $\text{AIC}_c$, CV, and REP, we consider both model selection (in which a single model that minimizes the respective criterion is selected) and model combination (in which forecasts of different models are combined with weights that reflect the values of the various selection criteria). Burnham and Anderson (2002) and Kolassa (2011) suggested the use of a normalized exponential (softmax) function to assign weights for combination. These weights provide the probabilities that a model is optimal (correctly specified). Assuming Cr is a vector that holds the values of a selection criterion (such as $\text{AIC}_c$ or REP), the combination weights can be derived as

$$w_i^{\text{Cr}} = \frac{\exp\left(-\frac{1}{2}\Delta_i^{\text{Cr}}\right)}{\sum_{j=1}^{M}\exp\left(-\frac{1}{2}\Delta_j^{\text{Cr}}\right)}, \tag{5}$$

where $M$ is the number of models available, $\exp(\cdot)$ is the exponential function and $\Delta_i^{\text{Cr}} = \text{Cr}_i - \min(\text{Cr})$, with $\min(\cdot)$ returning the minimum value of a vector. We replaced Cr in turn with $\text{AIC}_c$, CV, and REP. These combination weights were applied to both the point forecasts as well as to the prediction intervals obtained from each model. For the results presented in sections 4.2 and 4.3, only models from the exponential smoothing family were combined. We additionally benchmarked against an equal-weighted combination across all available models in a given pool, as denoted by EQW.

We used the yearly, quarterly, monthly, weekly, daily, and hourly data from the M, M3, and M4 forecasting competitions (Makridakis et al. 1982, Makridakis and Hibon 2000, Makridakis et al. 2020). In total, we considered 103,830 time series from a variety of fields. Each series consisted of in-sample and out-of-sample (test) sets of observations. We estimated models on the in-sample data and evaluated performance on the out-of-sample/test data. Although the length of the in-sample information varied across series, the length of the required forecast horizon (which matches the length of the out-of-sample set) was fixed for each frequency. Table 2 of the electronic Appendix C provides more details on the data.

We pooled series from several forecasting competitions for three reasons: First, by pooling the series from the M, M3, and M4 forecasting competitions, we achieved a larger and more diverse set of data. Note that the insights presented in section 4.2 generally hold for each individual competition data pool. Second, our results should not be compared directly with the results of any of the participating

methods in these competitions, because our evaluation takes place after the release of the test data. Third, our intention is not to propose a new model but to propose a general approach to model selection and combination that can be applied when multiple models are considered.

In measuring the values of the CV criterion, we set the length of the initial in-sample to two seasonal cycles when data might contain seasonality (i.e., 24 for monthly data) or eight observations for non-seasonal data (yearly). We repeatedly produced (by rolling the origin one period at a time) and evaluated forecasts using the mean absolute error. The only exceptions were the cases of weekly and hourly data, in which we limited the number of forecast origins to $h$ because of the excessive computational cost required otherwise. The forecast horizon for CV was set equal to $h$ or lower, if too few data were available. In measuring the out-of-sample performance of model selection and combination by $\text{AIC}_c$, CV, and REP, we adopted a fixed-origin evaluation approach with horizon $h$ (Tashman 2000). We produced forecasts for each series once by using all available in-sample data and all applicable exponential smoothing models. We evaluate forecast performance with seven performance indicators: two for measuring the accuracy of the point forecasts, one for evaluating the performance of the prediction intervals, three measures closely linked with the utility of the forecasts, and one for measuring the bias of the forecasts.

We calculated the symmetric Mean Absolute Percentage Error (sMAPE – see for example: Makridakis and Hibon 2000) and the Mean Absolute Scaled Error (MASE – Hyndman and Koehler 2006) to measure the accuracy of the forecasts. Although sMAPE suffers from asymmetry with regard to positive and negative forecast errors (Goodwin and Lawton 1999), it is a standard in many forecasting competitions. The MASE is the current gold standard measure for comparing the accuracy of point forecasts (Franses 2016). Assuming that for each series we produce $h$-steps-ahead forecasts using $n$ available observations in the in-sample set, we can define the sMAPE and MASE as

$$\text{sMAPE} = \frac{200}{h} \sum_{t=n+1}^{n+h} \frac{|y_t - f_t|}{|y_t| + |f_t|}, \qquad \text{and} \qquad \text{MASE} = \frac{1}{h} \frac{\sum_{t=n+1}^{n+h} |y_t - f_t|}{\frac{1}{n-s} \sum_{i=s+1}^{n} |y_i - y_{i-s}|},$$

in which $y_t$ is the actual observation at time period $t$, $f_t$ is the respective forecast, and $s$ is the length of the seasonal cycle ($s = 12$ for monthly data). The values of sMAPE and MASE can be averaged across series because both measures are scale independent. Lower values of the sMAPE and MASE suggest that the forecasts are closer to the actual data, thus indicating better accuracy.

We calculated the interval score (Gneiting and Raftery 2007) to evaluate prediction intervals. This utility function is intuitively appealing because it simultaneously considers the width of the prediction intervals as well as its hit-rate (coverage). To be able to average the values of the interval scores

across series of different levels, we use the Mean Scaled Interval Score (MSIS – Makridakis et al. 2020). For one series and $h$-steps-ahead forecasts, the MSIS can be calculated as

$$\text{MSIS} = \frac{1}{h} \frac{\sum_{t=n+1}^{n+h} \left( u_t - l_t + \frac{2}{\alpha}(l_t - y_t)\mathbb{1}\{y_t < l_t\} + \frac{2}{\alpha}(y_t - u_t)\mathbb{1}\{y_t > u_t\} \right)}{\frac{1}{n-s}\sum_{i=s+1}^{n} |y_t - y_{t-s}|},$$

in which $u_t$ and $l_t$ are the upper and lower prediction intervals for period $t$, $\alpha$ is the significance level so that $(1-\alpha)\times 100\%$ is the confidence level, and $\mathbb{1}\{\cdot\}$ is an indicator function that returns a value of 1 if the condition is true, 0 otherwise. Similar to the sMAPE and the MASE, the values of the MSIS are scale-independent and can be averaged across multiple series. In this paper, we evaluated the prediction intervals at a 95% confidence level which corresponds to $\alpha = 0.05$.

We also calculated three measures that relate to the use of forecasts for decision making (Svetunkov and Petropoulos 2018). The gains in supply chain efficiency from improved forecasting are context specific, but these additional measures allow us to approximate potential efficiency gains. These measures are the *coverage* (percentage of times that the actual future values lie within the prediction intervals), *upper coverage* (percentage of times that the actual future values do not exceed the upper prediction interval; a proxy to measuring achieved service level), and *spread* of the intervals (scaled by the in-sample mean); a proxy for the required safety stock. Finally, we calculate the average signed error scaled by the in-sample mean as a measure of bias:

$$\text{Bias} = \frac{1}{h} \frac{\sum_{t=n+1}^{n+h} (y_t - f_t)}{\frac{1}{n}\sum_{i=1}^{n} y_i}.$$

We used $\mathcal{L}_1$ (sum of absolute differences) as the distance measure for the performance and representativeness gaps of REP. This measure is aligned with our out-of-sample accuracy measures. Further, the $\mathcal{L}_1$ norm leads to slightly better performance than the $\mathcal{L}_2$ norm.

Some models can produce unrealistically wide prediction intervals. In these cases, we applied Tukey's fences approach to exclude outliers. Specifically, we removed models in which the upper (or the lower) prediction interval of the furthest horizon exceeds $\text{Q}_3 + 1.5(\text{Q}_3 - \text{Q}_1)$ (or is lower than $\text{Q}_1 - 1.5(\text{Q}_3 - \text{Q}_1)$), in which $\text{Q}_1$ and $\text{Q}_3$ are the first and third quartiles of the respective values across models. A similar outlier removal treatment was applied by de Oliveira et al. (2021). In estimating the combination weights, we also excluded models with outlier criteria values. Finally, the maximum likelihood could not be estimated for some series and models. We excluded those models, so that we could render the comparisons across criteria as fairly as possible.

### 4.2. Results

Table 1 reports the performance of each criterion for each data frequency and measure. Each measure is calculated per series, and then averaged across series by using the arithmetic mean. The best selection and combination approach is boldfaced for each frequency and measure.

**Table 1 The average forecasting performance of each criterion for each frequency and measure.**

| | | Selection | | | Combination | | | |
|---|---|---|---|---|---|---|---|---|
| Frequency | Measure | $AIC_c$ | CV | REP | EQW | $AIC_c$ | CV | REP |
| Yearly | sMAPE | 15.319 | 14.847 | **14.081** | 14.830 | 15.127 | 14.833 | **13.938** |
| 23826 series | MASE | 3.405 | 3.307 | **3.125** | 3.245 | 3.351 | 3.300 | **3.101** |
| | MSIS | **34.822** | 39.679 | 35.034 | 30.949 | 33.164 | 37.459 | **30.379** |
| | Coverage | **0.838** | 0.788 | 0.822 | **0.874** | 0.850 | 0.805 | 0.863 |
| | Upper Coverage | 0.898 | 0.882 | **0.904** | 0.903 | 0.904 | 0.889 | **0.913** |
| | Spread | 1.323 | **1.046** | 1.151 | 1.169 | 1.324 | **1.078** | 1.124 |
| | Bias | 0.091 | 0.070 | **0.053** | 0.125 | 0.092 | **0.073** | 0.076 |
| Quarterly | sMAPE | 10.292 | 10.113 | **10.055** | 10.159 | 10.148 | 10.067 | **9.882** |
| 24959 series | MASE | 1.163 | 1.167 | **1.147** | 1.175 | 1.148 | 1.159 | **1.134** |
| | MSIS | **9.547** | 10.055 | 9.800 | 9.119 | 9.350 | 9.603 | **9.032** |
| | Coverage | **0.929** | 0.913 | 0.925 | **0.948** | 0.934 | 0.923 | 0.941 |
| | Upper Coverage | **0.953** | 0.943 | **0.953** | **0.961** | 0.955 | 0.949 | 0.960 |
| | Spread | 0.888 | **0.823** | 0.839 | 0.893 | 0.893 | **0.830** | 0.841 |
| | Bias | 0.012 | 0.019 | **0.008** | 0.022 | **0.013** | 0.019 | 0.014 |
| Monthly | sMAPE | 13.417 | 13.055 | **12.976** | 13.033 | 13.296 | 13.023 | **12.730** |
| 50045 series | MASE | 0.941 | 0.921 | **0.918** | 0.948 | 0.933 | 0.916 | **0.906** |
| | MSIS | **8.144** | 8.399 | 8.311 | 8.214 | 8.054 | 8.174 | **7.956** |
| | Coverage | **0.931** | 0.920 | 0.926 | **0.949** | 0.935 | 0.927 | 0.938 |
| | Upper Coverage | **0.960** | 0.956 | 0.959 | **0.969** | 0.962 | 0.960 | 0.964 |
| | Spread | 0.824 | 0.780 | **0.777** | 0.854 | 0.826 | **0.773** | 0.787 |
| | Bias | -0.005 | **0.000** | -0.004 | **0.000** | -0.005 | **0.000** | -0.001 |
| Weekly | sMAPE | 7.233 | 7.758 | **7.184** | **7.034** | 7.267 | 7.515 | 7.078 |
| 359 series | MASE | 0.511 | 0.515 | **0.507** | **0.489** | 0.513 | 0.508 | 0.497 |
| | MSIS | 3.964 | 3.996 | **3.958** | **3.767** | 3.955 | 3.864 | 3.782 |
| | Coverage | **0.955** | 0.928 | 0.939 | **0.969** | 0.955 | 0.943 | 0.954 |
| | Upper Coverage | **0.985** | 0.966 | 0.974 | 0.981 | **0.985** | 0.974 | 0.981 |
| | Spread | 0.616 | **0.519** | 0.567 | 0.591 | 0.615 | **0.528** | 0.574 |
| | Bias | 0.003 | 0.005 | **-0.002** | 0.012 | 0.004 | 0.004 | **0.001** |
| Daily | sMAPE | 3.109 | 3.052 | **3.033** | **3.005** | 3.105 | 3.075 | 3.015 |
| 4227 series | MASE | 1.246 | **1.142** | 1.156 | 1.155 | 1.243 | **1.147** | 1.148 |
| | MSIS | 16.309 | 10.760 | **10.456** | 270.539 | 16.297 | **10.219** | 10.232 |
| | Coverage | **0.951** | 0.947 | 0.948 | **0.955** | 0.952 | 0.952 | 0.953 |
| | Upper Coverage | **0.973** | 0.970 | 0.971 | **0.974** | 0.973 | 0.971 | **0.974** |
| | Spread | 0.716 | 0.234 | **0.224** | 11.504 | 0.715 | 0.229 | **0.225** |
| | Bias | **0.001** | 0.010 | 0.009 | 0.008 | **0.001** | 0.011 | 0.009 |
| Hourly | sMAPE | 13.655 | **12.981** | 13.224 | 17.530 | 13.640 | 13.385 | **13.227** |
| 414 series | MASE | 0.903 | **0.834** | 0.858 | 1.436 | 0.902 | 0.915 | **0.849** |
| | MSIS | 9.521 | **7.825** | 9.246 | 80.486 | 9.502 | **8.831** | 9.322 |
| | Coverage | 0.902 | 0.883 | **0.908** | **0.968** | 0.902 | 0.921 | 0.929 |
| | Upper Coverage | 0.952 | 0.945 | **0.956** | **0.988** | 0.952 | 0.967 | 0.969 |
| | Spread | 1.386 | **0.807** | 1.348 | 1.835 | 1.385 | **0.943** | 1.374 |
| | Bias | 0.011 | 0.012 | **0.009** | 0.009 | 0.011 | -0.009 | **0.005** |

Focusing on selection, we observed that selecting by representativeness leads to improved forecasting accuracy for most frequencies and measures. This is especially true for the yearly frequency in which the forecast error (in terms of MASE) for REP is 8.2% and 5.5% lower, respectively, than for $AIC_c$ and CV. REP is close, but still better, than CV in the monthly frequency. The relative rankings between the three criteria are in-line with our simulation findings (subsection 3.3). The only exceptions are for hourly frequency, in which REP ranks second after CV, and in daily frequency, in which REP is slightly better than CV in terms of sMAPE but slightly worse in terms of MASE.

REP shows strong performance with prediction intervals as well (MSIS). Specifically, REP performs better on average than CV in five out of six data frequencies. REP also outperforms $AIC_c$ in the weekly, daily, and hourly frequencies. $AIC_c$ offers the best (upper) interval coverage, but this performance comes at a price in terms of interval spread. Exactly the opposite situation occurs with CV, which results in the lowest coverage but tight prediction intervals. REP offers balanced coverage versus spread that translates to a balance between achieved service level and holding costs. For example, REP outperforms $AIC_c$'s upper coverage in the yearly frequency (90.4% versus 89.8%) with significantly lower interval spread (1.151 versus 1.323). Similarly, REP outperforms CV in (upper) coverage and spread in the monthly and daily frequencies. Finally, selection with REP results in the least biased forecasts in four of six data frequencies. The combination approaches outperformed the selection approaches on most measures. Overall, forecast combinations with weights based on REP offer better results compared with combinations based on $AIC_c$ or CV.

To test the statistical significance of these results, we performed multiple comparisons from the best (MCB – Koning et al. 2005). Figure 3 presents the results of the MCB test based on MASE. The first six panels (one for each data frequency) show the results for selecting one model with the last six panels showing the results from combining across models. Within each panel the approaches are ranked from the worst (top row) to the best (bottom row) mean ranks.

In terms of selection, REP significantly outperforms $AIC_c$ and CV in the yearly, quarterly, and monthly data. The three approaches perform on par with the weekly data. In the daily and hourly data, CV is better than the other two selection approaches. In terms of combinations, we observed that REP is ranked first in five data frequencies, and statistically better in four of them than all other combination approaches. In the daily data, combinations via REP and CV perform similarly, but both have significantly worse mean ranks than either $AIC_c$ or EQW.

We next explored the frequency with which REP selects a less or more complex model compared with the two other approaches, $AIC_c$ and CV. These results are presented in table 2. We observed that REP generally selects a more complex model than those selected by $AIC_c$ and CV. Although the additional complexity results from increased number of components or parameters in the selected exponential smoothing models (trend component, seasonal component, dampening factor for the

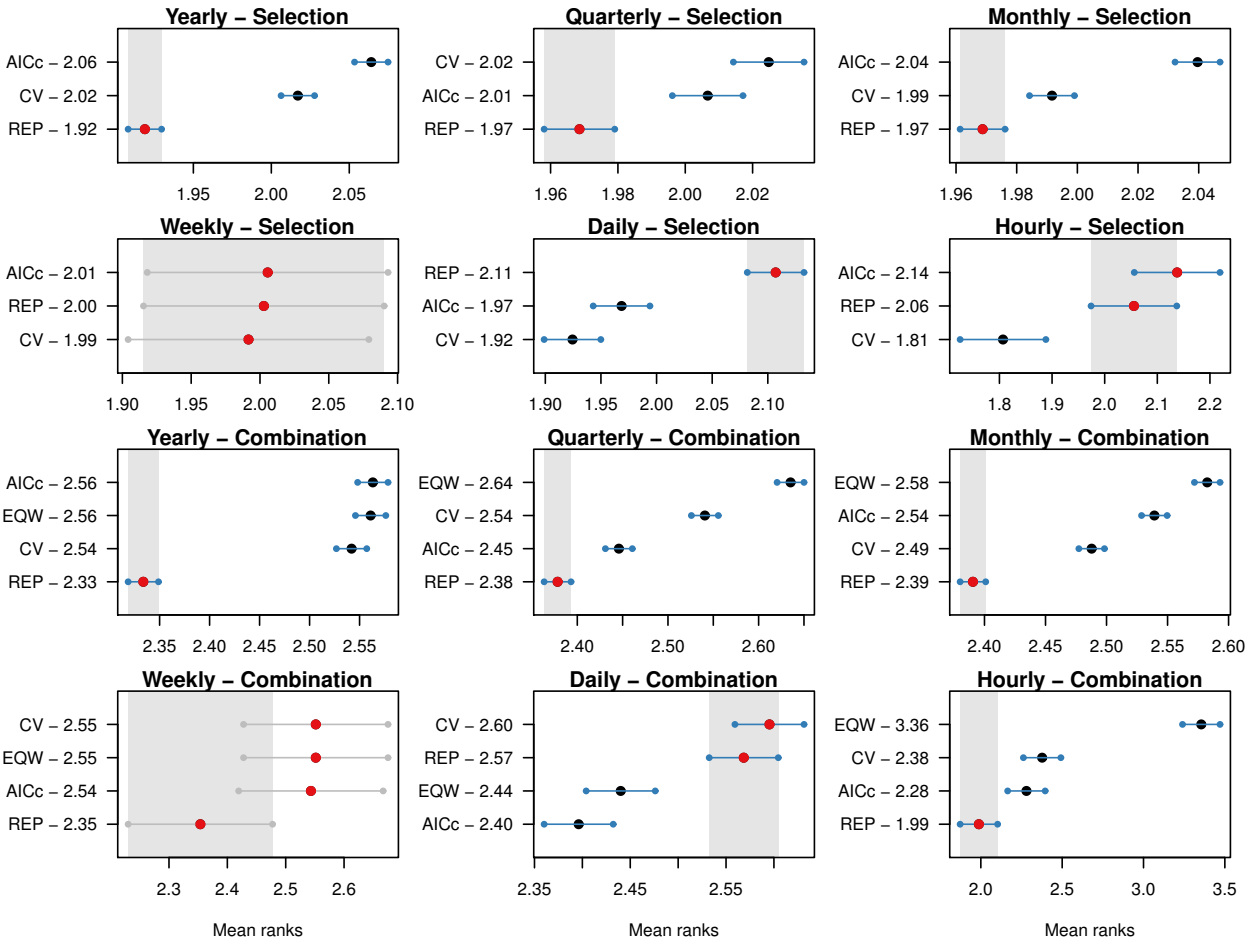


**Figure 3 Multiple comparisons from the best (MCB) for the MASE.**

When the confidence intervals of two approaches overlap, their mean ranks are not significantly different. These intervals are presented with blue when at least two approaches are significantly different; otherwise they are presented in gray. The gray area reflects the confidence intervals of the REP approach (selected or combined). The mean ranks of the statistically indifferent approaches to REP are depicted with red dots. The mean ranks of the approaches that are significantly different from REP are presented with black dots.

trend component, or a combination of these), REP did not opt significantly more for one component over another. However, there are cases in which REP will opt for a less complex model than CV will, for instance in a third of the weekly and hourly time series. The main takeaway is that $AIC_c$ indeed penalizes model complexity too heavily and inappropriately leads to less complex models.

### 4.3. Analysis

To better understand why REP worked well in our previous analysis, we empirically explored the theoretical differences between the selection criteria articulated in Sections 3.3 and 3.4. Note that the analysis in this subsection focuses on selecting the best model using each criterion rather than combining across models based on equation (5).

**Table 2 The REP frequency selection of less/more complex models compared to $AIC_c$ and CV.**

| | Less complex than | | More complex than | |
|---|---|---|---|---|
| Frequency | $AIC_c$ | CV | $AIC_c$ | CV |
| Yearly | 7.4% | 21.4% | 38.2% | 29.0% |
| Quarterly | 10.2% | 18.7% | 61.5% | 56.5% |
| Monthly | 14.6% | 15.5% | 55.8% | 58.5% |
| Weekly | 10.3% | 33.4% | 53.8% | 42.1% |
| Daily | 4.8% | 9.3% | 79.0% | 73.4% |
| Hourly | 1.4% | 36.2% | 6.8% | 7.0% |

**4.3.1. Level of noise versus length of series.** In section 3.3, we demonstrated that a low signal-to-noise ratio will result in REP outperforming $AIC_c$, especially for short series. Because we do not know the data generation process for any real-life time series, we empirically estimate the noise of each series relatively to its signal. To do so, we performed a Seasonal and Trend decomposition using Loess (STL) decomposition on the Box-Cox transformed data; for yearly data, we simply applied the Loess method and calculated the remainder, assuming an additive relationship. After separating the series into its three components (trend, seasonality, and remainder) we reconstructed the series without the remainder. In essence, we split the series into a "signal" (consisting of the trend and seasonal components) and the remainder. We then calculated the ratio of the absolute value of the remainder of each observation and the "signal" for the same observation, and averaged this ratio across all observations for a series.

We are interested in the percentage of cases in which REP outperformed $AIC_c$. After excluding the instances in which both criteria selected the same model (see also table 2), we split the series with regard to the level of noise and the length, as depicted in table 3. For splitting to short, medium, and long lengths and low, moderate, and high noise levels, we considered the sample quantiles that corresponded to probabilities 0.333 and 0.667. We did not split the hourly series into different length buckets because $AIC_c$ and REP rarely disagree here. When the series is short, a positive relationship occurs for yearly, quarterly and monthly data between the level of noise and the percentage of cases in which REP outperforms $AIC_c$. The percentage of cases in which REP outperforms $AIC_c$ decreases as the lengths of the series increases. These empirical findings are aligned with our analytical insights in section 3.3.

**4.3.2. Balancing information from in-sample and out-of-sample forecasts.** In section 3.4, we argued that the good performance of REP is further driven by a combination of three factors. First, REP offers a balance between the in-sample model fit and the out-of-sample forecasts. REP is the only approach to take into account (through representativeness) the out-of-sample forecasts, whereas $AIC_c$ and CV focus only on the in-sample forecasts (in terms of model fit or rolling origin evaluation, respectively). Second, REP assesses representativeness by giving more weight to more

**Table 3 Percentages of series in which REP outperforms AIC$_c$ for different levels of noise and series lengths.**

| | | Length | | | | | Length | | |
|---|---|---|---|---|---|---|---|---|---|
| Frequency | Noise | Short | Medium | Long | Frequency | Noise | Short | Medium | Long |
| Yearly | Low | 57.5% | 57.3% | 54.2% | Quarterly | Low | 52.1% | 50.2% | 52.3% |
| | Moderate | 58.4% | 58.6% | 55.5% | | Moderate | 52.8% | 51.4% | 50.9% |
| | High | 60.0% | 54.8% | 54.9% | | High | 53.6% | 50.5% | 48.1% |
| Monthly | Low | 55.7% | 51.2% | 51.8% | Weekly | Low | 35.7% | 44.8% | 50.0% |
| | Moderate | 54.7% | 51.6% | 52.3% | | Moderate | 59.3% | 44.4% | 53.6% |
| | High | 55.9% | 51.0% | 50.8% | | High | 45.9% | 60.6% | 66.7% |
| Daily | Low | 49.8% | 42.0% | 40.1% | Hourly | Low | – | 66.7% | – |
| | Moderate | 51.0% | 39.6% | 34.7% | | Moderate | – | 70.0% | – |
| | High | 49.2% | 46.4% | 51.6% | | High | – | 61.9% | – |

recent data through the discount factor, $\delta$. Third, REP assesses representativeness through multiple forecast horizons. Although CV evaluates the (in-sample) forecasts over multiple horizons, AIC$_c$ focuses on the one-step-ahead in-sample forecast error. We empirically examined the first two factors in this subsection. We consider the effect of the forecast horizon in the next subsection.

To explore the importance of the balance of in-sample and out-of-sample forecasts, we devised two additional definitions of REP (REP$_{in}$ and REP$_{out}$) that focus on the in-sample fit and the out-of-sample forecast representativeness:

$$\text{REP}_{in} = \|\check{y}, \check{g}\|_1, \qquad \text{and} \qquad \text{REP}_{out} = \sum_{i=1}^{\lfloor n/p \rfloor} (1-\delta)^{i-1} \left\|\check{y}_{[i]}, \check{f}\right\|_1.$$

Selecting by REP$_{in}$ is equivalent to selecting by the one-step-ahead mean absolute in-sample error. Selecting by REP$_{out}$ requires only the out-of-sample forecasts. We simultaneously investigated the effects of the discount factor, $\delta$, for REP and REP$_{out}$ by considering different values in $[0 \dots 1]$. The results from our analysis are depicted in figure 4.

We observed that REP is superior to REP$_{out}$ for the yearly, quarterly, monthly, and daily frequencies. Selecting by REP$_{out}$ would result in better performance for the weekly and hourly frequencies. A value of $\delta$ in the region between 0.4 and 0.6 seems to work well for lower data frequencies (yearly, quarterly, and monthly). Smaller discounts (lower $\delta$ values) should be considered for higher data frequencies (weekly, daily, and hourly data). Longer series are expected for such frequencies.

With the exception of daily data, the performance of REP$_{in}$ is worse than that of REP. It is also noteworthy that the performance of REP$_{in}$ closely resembles that of AIC$_c$ for the quarterly frequency, and REP$_{in}$ achieves better accuracy on the monthly and daily data than AIC$_c$, even if the former does not apply any penalties for model complexity. Still, AIC$_c$ performs much better than REP$_{in}$ on the yearly, weekly, and hourly series.

**4.3.3. The effect of the forecast horizon.** To explore the importance of evaluating multiple-steps-ahead when selecting between forecasting models, we calculated the forecast accuracy (by means

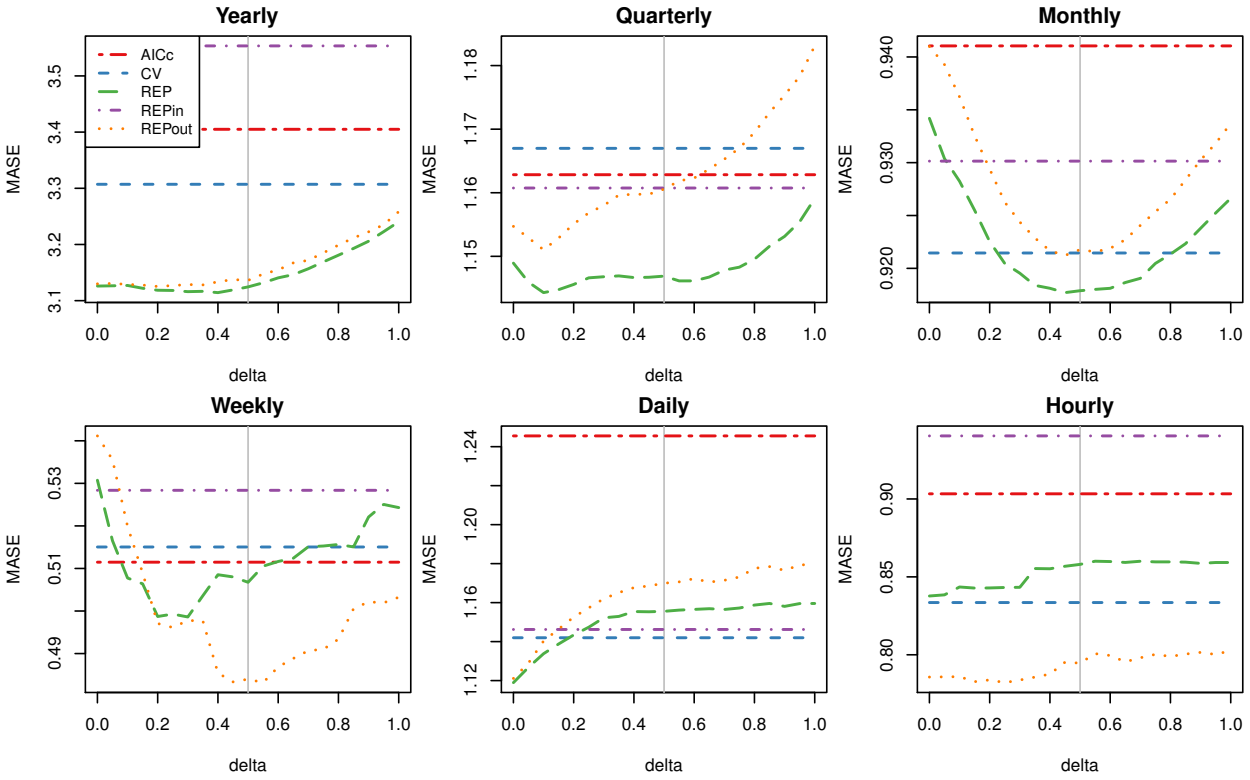


**Figure 4** **The effects of in-sample performance measurement, out-of-sample representativeness, and the $\delta$ discount factor.**

of average MASE) for one-step-ahead forecasts ($h = 1$) as well as for short, medium, and long forecast horizons. Short, medium, and long are defined by data frequency. Table 4 provides a summary of our analysis. The strong performance of REP is evident across all planning horizons. The performance gap between REP and $\text{AIC}_c$ increases for longer horizons. The performance differences are smaller for one-step-ahead forecasts, especially for the quarterly, monthly, and hourly frequencies. We expected this result because $\text{AIC}_c$ focuses on one-step-ahead forecasts.

**Table 4** **The average MASE performance of each approach for different planning horizons.**

| Frequency | Horizon | $\text{AIC}_c$ | CV | REP | Frequency | Horizon | $\text{AIC}_c$ | CV | REP |
|---|---|---|---|---|---|---|---|---|---|
| Yearly | One-Step (1) | 1.511 | 1.514 | 1.469 | Quarterly | One-Step (1) | 0.600 | 0.606 | 0.589 |
| | Short (1-2) | 1.909 | 1.893 | 1.823 | | Short (1-2) | 0.694 | 0.701 | 0.684 |
| | Medium (3-4) | 3.420 | 3.325 | 3.142 | | Medium (3-5) | 1.094 | 1.101 | 1.080 |
| | Long (5-6) | 4.886 | 4.703 | 4.409 | | Long (6-8) | 1.544 | 1.543 | 1.523 |
| Monthly | One-Step (1) | 0.454 | 0.454 | 0.453 | Weekly | One-Step (1) | 0.416 | 0.405 | 0.425 |
| | Short (1-6) | 0.649 | 0.633 | 0.627 | | Short (1-4) | 0.410 | 0.402 | 0.424 |
| | Medium (7-12) | 0.959 | 0.947 | 0.949 | | Medium (5-9) | 0.577 | 0.571 | 0.573 |
| | Long (13-18) | 1.215 | 1.184 | 1.178 | | Long (10-13) | 0.531 | 0.559 | 0.506 |
| Daily | One-Step (1) | 0.428 | 0.399 | 0.400 | Hourly | One-Step (1) | 0.241 | 0.275 | 0.237 |
| | Short (1-4) | 0.760 | 0.663 | 0.668 | | Short (1-16) | 0.818 | 0.778 | 0.792 |
| | Medium (5-9) | 1.220 | 1.115 | 1.125 | | Medium (17-32) | 0.794 | 0.730 | 0.754 |
| | Long (10-14) | 1.660 | 1.552 | 1.576 | | Long (33-48) | 1.098 | 0.993 | 1.028 |

### 4.4. Selecting and combining models from different classes

In the previous sections, we explored the performance of REP against $AIC_c$ and CV when selecting among or combining different models within the exponential smoothing family. REP is not limited to a single class of models. It can be used to select between models of any other family (such as ARIMA), or to select between forecasts across different model families.

We first applied REP, CV, and $AIC_c$ to select the "best" ARIMA model, or combine across ARIMA models. We followed Hyndman and Khandakar (2008) and implemented a search process. First, we applied non-seasonal and seasonal differences on the data based on the result of unit root tests, and tested simple ARIMA models. Then, the search expanded in a step-wise fashion in which the ARIMA model parameters (non-seasonal and seasonal autoregressive and moving average orders) were increased or decreased by one and the constant was included or excluded. If a better model was identified given a criterion ($AIC_c$, CV, or REP), then the search continued. This process was implemented for the IC within the auto.arima() of the *forecast* package for the R statistical software. We chose similar default hyperparameters to the auto.arima() function in terms of constraining the maximum orders of the models to avoid overfitting. Given that this is a step-wise non-exhaustive search, the application of different criteria ($AIC_c$, CV, REP) resulted in different search paths. This would be also the case if one changed the default information criterion ($AIC_c$) to another one (AIC or BIC). Accordingly, EQW is not meaningful in this case.

We limited the number of origins considered by the CV approach for validation purposes to $2h$ ($h$ for the weekly and hourly frequencies). We made this choice for two reasons. First, decreasing the length of the available data results in rejection of many ARIMA models because their fits cannot be estimated. In turn, this leads to a very poor performance for the CV criterion. Second, the computational cost for ARIMA is significant.

We present the results for the different sampling frequencies in table 5. Similar to our insights from exponential smoothing, we observed that REP results in better performance compared with $AIC_c$ and CV. Differences are larger for the yearly, weekly, and hourly frequencies. It is noteworthy that the average performance gap between selection and combination is relatively small. Four hourly series were excluded from this analysis because the resulting MASE value for $AIC_c$ selection and combination was very large.

We also investigated the performance of REP in selecting (and combining) models from the following popular pool: Theta model (Assimakopoulos and Nikolopoulos 2000), exponential smoothing (ETS: Hyndman et al. 2008), and ARIMA (Hyndman and Khandakar 2008). We used the functions thetaf(), ets() and auto.arima() of the *forecast* package for the R statistical software (when $s > 24$, we used the es() function of the *smooth package* instead of the ets() function). The functions ets()/es() and auto.arima() automatically and by default select the "best" exponential smoothing and ARIMA

**Table 5 The average forecasting accuracy (in MASE) of $AIC_c$, CV, and REP in selecting and combining across ARIMA models.**

| | Selection | | | Combination | | |
|---|---|---|---|---|---|---|
| Frequency | $AIC_c$ | CV | REP | $AIC_c$ | CV | REP |
| Yearly | 3.384 | 3.357 | **3.237** | 3.352 | 3.471 | **3.243** |
| Quarterly | 1.170 | 1.192 | **1.158** | 1.161 | 1.190 | **1.148** |
| Monthly | 0.928 | 0.948 | **0.919** | 0.919 | 0.945 | **0.914** |
| Weekly | 0.539 | 0.558 | **0.493** | 0.538 | 0.558 | **0.488** |
| Daily | **1.154** | 1.156 | 1.159 | 1.152 | 1.153 | **1.151** |
| Hourly | 0.764 | 0.787 | **0.738** | 0.769 | 0.782 | **0.745** |

models via an information criterion. Selection and combination across the three sets of forecasts (forecasts from the Theta model, the "best" ETS model, and the "best" ARIMA model) was done by using the REP criterion and compared with the CV criterion. Similar to the results of table 5, we limited the validation period for CV to $2h$ forecast origins ($h$ for the weekly and hourly frequencies). We did not benchmark against $AIC_c$ because information criteria values are not comparable across model families and packages.

We show the average MASE in table 6, together with the individual average performance of each model (Theta, ETS, and ARIMA). We observed that both CV and REP performed better than any of the individual models when selecting forecasts but especially when combining them. REP outperformed CV in four of the six frequencies. The performance differences between the two criteria are small for the weekly and the daily data. However, CV is better than REP in the hourly frequency.

**Table 6 The average forecasting accuracy (in MASE) of CV and REP in selecting and combining across three models: Theta, ETS, and ARIMA.**

| | Individual methods | | | Selection | | Combination | | |
|---|---|---|---|---|---|---|---|---|
| Frequency | Theta | ETS | ARIMA | CV | REP | EQW | CV | REP |
| Yearly | 3.365 | 3.431 | 3.389 | 3.277 | **3.116** | 3.143 | 3.267 | **3.062** |
| Quarterly | 1.232 | 1.165 | 1.171 | 1.175 | **1.146** | 1.130 | 1.170 | **1.127** |
| Monthly | 0.969 | 0.947 | 0.931 | 0.921 | **0.914** | 0.906 | 0.917 | **0.905** |
| Weekly | 0.546 | 0.508 | 0.532 | 0.465 | **0.453** | 0.487 | 0.460 | **0.452** |
| Daily | 1.153 | 1.239 | 1.204 | **1.152** | 1.158 | 1.172 | **1.150** | 1.151 |
| Hourly | 0.949 | 0.888 | 0.750 | **0.747** | 0.797 | 0.774 | **0.763** | 0.792 |

We did not apply any pruning techniques (as discussed at the last paragraph of 4.1) because we chose to retain the output of the *forecast* and *smooth* packages as is. This explains the small differences in the average MASE values between the columns ETS and ARIMA in table 6 and the columns Selection with $AIC_c$ in tables 1 and 5. Overall, tables 5 and 6 demonstrate that the superiority of REP over other selection criteria is not limited within the class of the exponential smoothing models but can be generalized across other classes of forecasting models. Further, REP can work on top of existing automatic algorithms in choosing models across many different forecasting classes.

## 5. Discussion

We conceptualized and tested representativeness as an approach to forecast selection. We were motivated by the good performance of judgmental model selection in this context and the ability of humans to avoid models with unreasonable forecasts (Petropoulos et al. 2018). Using the time series that formed the basis of our empirical investigation in section 4, we performed an analysis to check the frequency with which the proposed REP criterion selects the best and worst models compared with the frequency of selection based on $AIC_c$ and CV. For this purpose, we split the models into three groups (top, middle, and bottom thirds) based on their out-of-sample performance as measured by MASE. Table 7 presents the respective relative frequencies. We observed that forecast selection by representativeness, like judgmental model selection, selects the worst models less frequently than selection with IC. At the same time, REP is also able to identify the best models more frequently than either $AIC_c$ or CV.

**Table 7 The relative frequencies with which $AIC_c$, CV, and REP select the best versus the worst models.**

| Group | $AIC_c$ | CV | REP |
|---|---|---|---|
| Top 1/3 Performance | 41.0% | 41.4% | 44.6% |
| Middle 1/3 Performance | 33.1% | 35.7% | 31.8% |
| Bottom 1/3 Performance | 25.8% | 22.8% | 23.6% |

Although REP can also lead to benefits in interval estimation, REP has a clear advantage when point forecasts are used; in practice, many firms are still using point forecasts and do not rely on prediction intervals. A recent survey by one of the authors indicates that 36% of responding firms still rely exclusively on point forecasts, without even calculating prediction intervals, best case/worst case scenarios, or probability distributions. This implies that many firms could benefit by selecting their forecasting models based on REP.

One advantage of the REP approach is that it requires only a small amount of additional computational time because the calculations for obtaining the REP value of each model are trivial. The performance improvements reported in section 4.2 do not come with a significant additional computational cost. REP is only 3% slower than IC at the monthly frequency. In contrast, CV is multiple times more computationally expensive than IC, depending on the number of origins considered. This is an important advantage of REP in the era of big data because companies nowadays must produce forecasts for tens of thousands of stock-keeping units over thousands of locations overnight (Nikolopoulos and Petropoulos 2018) and performance gains of new approaches should be evaluated against the additional cost involved (Gilliland 2019) regardless of the scalability of the infrastructure.

Another advantage of the proposed approach is that its $REP_{out}$ variant, which performs competitively (see subsection 4.3.2), does not require access to in-sample forecasts (as is the case with IC)

or past forecasting performance (as is the case with validation or CV approaches). In fact, selecting between forecasts using $\text{REP}_{out}$ requires only the historical data and the forecasts. Because of this feature, it can be easily applied across forecasts of different families, including judgmental forecasts. Several authors have already devised ex-ante criteria on when to apply univariate versus causal versus judgmental methods to produce forecasts based, among others, on data availability and factors affecting the environment (see, for instance Hyndman and Athanasopoulos 2018, chapter 4). However, this is the first approach, to the best of our knowledge, that allows selection between statistical or judgmental forecasts ex-post (once the forecasts have been produced).

In some cases, as shown in table 2, REP will end up selecting more complex models than either $\text{AIC}_c$ or CV. Following Kang et al. (2017) and Spiliotis et al. (2020), we analyzed the strength of the trend and seasonality of each series and observed that REP selects models that include such components (trend and/or seasonality) when the respective signals are strong. This is in-line with the simulation results in section 3.3 in which we saw that, unlike $\text{AIC}_c$, REP does not automatically levy harsh penalties on complex models. But even when REP unnecessarily opts for a more complex, suboptimal model, this choice may not have an adverse effect on forecasting performance. This insight follows previous research on forecasting under suboptimality (Nikolopoulos and Petropoulos 2018).

It is important to note that we have tested REP conditional on the parameters of each model having been specified. For example, the forecasts produced by Single Exponential Smoothing (SES), one of the exponential smoothing family models, were based on the optimal values of its two parameters: the initial level and the smoothing parameter for the level. This optimization is usually done by fitting multiple SES models with different parameters and choosing the one that minimizes the one-step-ahead MSE or any other in-sample criterion. We believe that REP could also be used toward identifying the optimal sets of parameters within each model.

We focused on the representativeness of the point forecasts in our research. It is also possible to measure representativeness using prediction intervals instead of the point forecasts. Prediction interval representativeness can be measured, for example, in terms of interval scores (Gneiting and Raftery 2007), after transformations and scaling have been applied to the lower/upper prediction intervals. To retain the (a)symmetry of such intervals, the scaling should be applied using the point forecasts as a reference for the mean and the respective in-sample window as a reference for the variance. In case we are able to produce multiple forecast quantiles, then proper scoring rules (Grushka-Cockayne et al. 2017) could also be used to measure the representativeness of the prediction intervals. This is an important avenue for future research.

The results of this study are relevant to both forecasting practitioners as well as software developers. The former can enhance their forecasting processes and improve their approaches on how to select between different models and forecasts. The latter can incorporate the ideas described in this

paper into their forecasting support systems. In both cases, we believe our proposition provides a robust framework for accurate selection between forecasting models that outperforms existing state-of-the-art solutions. The benefits derived from its improved forecast accuracy can be amplified if inventory evaluation is performed (Petropoulos et al. 2019). As such, the suggested change in the existing processes and software will add value to the forecasting functions used in practice. Based on our empirical results in this paper, we recommend that the discount factor, $\delta$, is set to a maximum of 0.5, with lower values preferable for higher frequency data. With regard to the distance measurement, we suggest the use of $\mathcal{L}_1$ for both the performance gap and representativeness gap of REP because the results based on $\mathcal{L}_2$ were generally worse. Moreover, we recommend measurement of representativeness over at least a full seasonal cycle.

The efficacy of the proposed forecast selection criterion has been tested on a very large set of real series (more than 100,000 series). The good performance of REP, in terms of the accuracy of the point forecasts, is evident. However, as with any empirical study, the results presented here are limited to the data sets used. Because 95% of the time series used in this study were recorded in yearly, quarterly, or monthly frequencies, we regard the results and insights obtained for these three data frequencies as robust. We believe further testing should be done, primarily focusing on larger collections of higher frequency data such as weekly, daily, and hourly. Further, examining hierarchical series and series with high intermittency is important.

## 6. Conclusions

In this study, we proposed a new way to select statistical models for forecasting. Our approach is the first to take into account how well out-of-sample forecasts represent the historical data. We achieved this through an asynchronous comparison of forecasts and past actuals. Our criterion, REP, significantly outperforms existing approaches for automatic model selection – namely IC and time series cross-validation – in terms of forecast accuracy. In addition, it performs well in performance indicators associated with the utility of these forecasts. Furthermore, model combinations using information from REP showed significant performance improvements compared with similar combinations for other criteria.

The construction of our criterion is similar to IC in that it consists of two parts: how well the model fits the historical data and a penalty. The penalty in IC is related to the complexity of the model, whereas the penalty of the new approach is related to the representativeness of the forecasts. Although our proposition does not differentiate between simple and complex models, it assesses whether the resulting forecasts are a natural continuation of the historical data by applying large penalties to models that produce unrepresentative sets of forecasts. In fact, REP avoids the worst models more often than IC, while also selecting the best models more frequently.

We showed analytically and empirically that a penalty that is solely based on complexity, as is the case in IC, will lead to incorrect model selections in cases of low signal-to-noise ratios and short series. When uncertainty is greatest (low signal-to-noise ratios, short series, and long horizons) REP should be the preferred option for model selection.

REP was originally motivated by the results of judgmental selection of forecasting models (Petropoulos et al. 2018). Like REP, humans can avoid the worst models by implicitly judging their future paths. However, individual judgmental selections did not always provide better accuracy compared with IC. Our study offers more evidence that algorithms can be infused with insights from human judgment toward an improved solution that outperforms both existing approaches and humans (Goldberg 1970). We offer an algorithm that is grounded in psychology, not statistics, and can regularly beat the two leading statistical approaches for model selection.

We discussed earlier that future research on forecasting by representativeness can include a search for optimal sets of model parameters instead of merely selecting between forecasting models. In addition, although our cost function for representativeness focused on the point forecasts, we suggested extensions of this new approach to forecast selection to include the representativeness of the prediction intervals. Future empirical work could also focus on this direction. Finally, future research could explore the effectiveness of REP in comparisons across models that also take into account exogenous variables.

## Acknowledgements

The authors would like to thank Konstantinos Nikolopoulos, Paul Goodwin, Evangelos Spiliotis, and Len Tashman for their feedback and comments. This project made use of the Balena High Performance Computing (HPC) Service at the University of Bath.

## Software

We used R statistical software (3.4.3) and the *forecast* (8.9) and *smooth* (2.5.3) packages. The MCB analysis made use of the nemenyi() function of the *tsutils* (0.9.2) package. The data are available under the packages *Mcomp* (2.8) and *M4comp2018* (0.2.0).

## References

Adya, Monica, Fred Collopy, J Scott Armstrong, Miles Kennedy. 2001. Automatic identification of time series features for rule-based forecasting. *International Journal of Forecasting* **17**(2) 143–157.

Akaike, H. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19**(6) 716–723.

Arvan, Meysam, Behnam Fahimnia, Mohsen Reisi, Enno Siemsen. 2019. Integrating human judgement into quantitative forecasting methods: A review. *Omega* **86** 237–252.

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

Blattberg, Robert, Stephen Hoch. 1990. Database models and managerial intuition: 50 % models + 50% manager. *Management Science* **36**(8) 887–899.

Box, George E P, Gwilym M Jenkins, Gregory C Reinsel. 2008. *Time Series Analysis: Forecasting and Control*. 4th ed. Wiley, New Jersey.

Brau, Rebekah, John Aloysius, Enno Siemsen. 2021. Demand planning for the digital supply chain: How to integrate human judgment and predictive analytics. *Working Paper* .

Burnham, Kenneth P, David R Anderson. 2002. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. 2nd ed. Springer, New York.

De Baets, Shari, Nigel Harvey. 2020. Using judgment to select and adjust forecasts from statistical models. *European Journal of Operational Research* **284**(3) 882–895.

de Oliveira, E M, F L Cyrino Oliveira, J Jeon. 2021. Treating and pruning: new approaches to forecasting model selection and combination using prediction intervals. *International Journal of Forecasting* **37**(2) 547–568.

Fildes, Robert. 1989. Evaluation of aggregate and individual forecast method selection rules. *Management Science* **35**(9) 1056–1065.

Fildes, Robert, Fotios Petropoulos. 2015. Simple versus complex selection rules for forecasting many time series. *Journal of Business Research* **68**(8) 1692–1701.

Flicker, Blair. 2018. Managerial insight and optimal algorithms. *Working Paper* .

Franses, Philip Hans. 2016. A note on the mean absolute scaled error. *International Journal of Forecasting* **32**(1) 20–22.

Fügener, Andreas, Jörn Grahl, Alok Gupta, Wolfgang Ketter. 2019. Cognitive challenges in human-ai collaboration: Investigating the path towards productive delegation. *Working paper* .

Gardner, Everette S. 2006. Exponential smoothing: The state of the art—part II. *International Journal of Forecasting* **22**(4) 637–666.

Gardner, Everette S, Ed McKenzie. 1988. Model identification in exponential smoothing. *The Journal of the Operational Research Society* **39**(9) 863–867.

Genre, Véronique, Geoff Kenny, Aidan Meyler, Allan Timmermann. 2013. Combining expert forecasts: Can anything beat the simple average? *International Journal of Forecasting* **29**(1) 108–121.

Gigerenzer, Gerd, Peter Todd. 1999. *Simple Heuristics that Make Us Smart*. Oxford University Press.

Gilliland, Michael. 2019. The value added by machine learning approaches in forecasting. *International Journal of Forecasting* .

Gneiting, Tilmann, Adrian E Raftery. 2007. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* **102**(477) 359–378.

Goldberg, Lewis R. 1970. Man versus model of man: A rationale, plus some evidence, for a method of improving on clinical inferences. *Psychological Bulletin* **73**(6) 422–432.

Goodrich, R. L. 1990. *Applied Statistical Forecasting.* Business Forecast Systems, Belmont, MA.

Goodwin, Paul. 2019. A surprisingly useful role for judgment. *Foresight: The International Journal of Applied Forecasting* **54** 11–12.

Goodwin, Paul, Richard Lawton. 1999. On the asymmetry of the symmetric MAPE. *International Journal of Forecasting* **15**(4) 405–408.

Grushka-Cockayne, Yael, Kenneth C Lichtendahl, Victor Richmond R Jose, Robert L Winkler. 2017. Quantile evaluation, sensitivity to bracketing, and sharing business payoffs. *Operations Research* **65**(3) 712–728.

Guerrero, Victor M. 1993. Time-series analysis supported by power transformations. *Journal of Forecasting* **12**(1) 37–48.

Han, Weiwei, Xun Wang, Fotios Petropoulos, Jing Wang. 2019. Brain imaging and forecasting: Insights from judgmental model selection. *Omega* **87** 1–9.

Harvey, Nigel. 1995. Why are judgments less consistent in less predictable task situations? *Organizational Behavior and Human Decision Processes* **63**(3) 247–263.

Harvey, Nigel, Teresa Ewart, Robert West. 1997. Effects of data noise on statistical judgement. *Thinking & Reasoning* **3**(2) 111–132.

Hibon, Michèle, Theodoros Evgeniou. 2005. To combine or not to combine: selecting among forecasts and their combinations. *International Journal of Forecasting* **21** 15–24.

Hyndman, R., G. Athanasopoulos, C. Bergmeir, G. Caceres, L. Chhay, M. O'Hara-Wild, F. Petropoulos, S. Razbash, E. Wang, F. Yasmeen, R Core Team, R. Ihaka, D. Reid, D. Shaub, Y. Tang, Z. Zhou. 2019. *forecast: Forecasting functions for time series and linear models.* R package version 8.9.

Hyndman, Rob J, George Athanasopoulos. 2018. *Forecasting: principles and practice.* 2nd ed. OTexts.

Hyndman, Rob J, Yeasmin Khandakar. 2008. Automatic time series forecasting: The forecast package for R. *Journal of Statistical Software* **27**(3) 1–22.

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

Hyndman, Rob J, Anne B Koehler, J Keith Ord, Ralph D Snyder. 2008. *Forecasting with Exponential Smoothing: The State Space Approach.* Springer Verlag, Berlin.

Hyndman, Rob J, Anne B Koehler, Ral Snyder, Simone Grose. 2002. A state space framework for automatic forecasting using exponential smoothing methods. *International Journal of Forecasting* **18**(3) 439–454.

Kahneman, Daniel, Andrew Rosenfeld, Linnea Gandhi, Tom Blaser. 2006. Noise how to overcome the high, hidden cost of inconsistent decision making. *Harvard Business Review* **94**(12) 36–43.

Kahneman, Daniel, Amos Tversky. 1973. On the psychology of prediction. *Psychological Review* **80**(4) 237–251.

Kang, Yanfei, Rob J. Hyndman, Kate Smith-Miles. 2017. Visualising forecasting algorithm performance using time series instance spaces. *International Journal of Forecasting* **33**(2) 345–358.

Kolassa, Stephan. 2011. Combining exponential smoothing forecasts using akaike weights. *International Journal of Forecasting* **27**(2) 238–251.

Koning, Alex J, Philip Hans Franses, Michele Hibon, Herman O Stekler. 2005. The M3 competition: Statistical tests of the results. *International Journal of Forecasting* **21**(3) 397–409.

Kourentzes, Nikolaos, Devon Barrow, Fotios Petropoulos. 2019. Another look at forecast selection and combination: Evidence from forecast pooling. *International Journal of Production Economics* **209** 226–235.

Lee, Yun Shin, Enno Siemsen. 2017. Task decomposition and newsvendor decision making. *Management Science* **63**(10) 3226–3245.

Lichtendahl, Kenneth C, Yael Grushka-Cockayne, Phillip E Pfeifer. 2013. The wisdom of competitive crowds. *Operations Research* **61**(6) 1383–1398.

Lyall, Allan, Pierre Mercier, Stefan Gstettner. 2018. The death of supply chain management. *Harvard Business Review* URL https://hbr.org/2018/06/the-death-of-supply-chain-management.

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

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

Makridakis, Spyros, Robin M Hogarth, Anil Gaba. 2010. *Dance With Chance: Making Luck Work for You*. Revised, expanded ed. Oneworld Publications.

Makridakis, Spyros, Evangelos Spiliotis, Vassilios Assimakopoulos. 2020. The M4 competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting* **36**(1) 54–74.

Meade, Nigel. 2000. Evidence for the selection of forecasting methods. *Journal of Forecasting* **19**(6) 515–535.

Montero-Manso, Pablo, George Athanasopoulos, Rob J. Hyndman, Thiyanga S Talagala. 2020. FFORMA: Feature-based forecast model averaging. *International Journal of Forecasting* **36**(1) 86–92.

Nikolopoulos, Konstantinos, Fotios Petropoulos. 2018. Forecasting for big data: Does suboptimality matter? *Computers & Operations Research* **98** 322–329.

Ord, Keith, Robert Fildes, Nikos Kourentzes. 2017. *Principles of Business Forecasting*. 2nd ed. Wessex, Inc.

Petropoulos, Fotios, Nikolaos Kourentzes, Konstantinos Nikolopoulos, Enno Siemsen. 2018. Judgmental selection of forecasting models. *Journal of Operations Management* **60** 34–46.

Petropoulos, Fotios, Spyros Makridakis, Vasilios Assimakopoulos, Konstantinos Nikolopoulos. 2014. 'Horses for Courses' in demand forecasting. *European Journal of Operational Research* **237** 152–163.

Petropoulos, Fotios, Xun Wang, Stephen M Disney. 2019. The inventory performance of forecasting methods: Evidence from the M3 competition data. *International Journal of Forecasting* **35**(1) 251–265.

Rouba, Ibrahim, Song-Hee Kim, Jordan Tong. 2021. Eliciting human judgement for prediction algorithms. *Management Science* .

Schwarz, Gideon. 1978. Estimating the dimension of a model. *Annals of Statistics* **6**(2) 461–464.

Shah, Chandra. 1997. Model selection in univariate time series forecasting using discriminant analysis. *International Journal of Forecasting* **13**(4) 489–500.

Spiliotis, Evangelos, Andreas Kouloumos, Vassilios Assimakopoulos, Spyros Makridakis. 2020. Are forecasting competitions data representative of the reality? *International Journal of Forecasting* **36**(1) 37–53.

Sugiura, Nariaki. 1978. Further analysts of the data by Akaike' s information criterion and the finite corrections. *Communications in Statistics - Theory and Methods* **7**(1) 13–26.

Svetunkov, I. 2019. *smooth: Forecasting Using State Space Models*. R package version 2.5.0.

Svetunkov, Ivan, Fotios Petropoulos. 2018. Old dog, new tricks: a modelling view of simple moving averages. *International Journal of Production Research* **56**(18) 6034–6047.

Talagala, Thiyanga S, Rob J Hyndman, George Athanasopoulos. 2018. Meta-learning how to forecast time series. Tech. rep., Monash University, Department of Econometrics and Business Statistics.

Taleb, Nassim Nicholas. 2008. *The Black Swan: The Impact of the Highly Improbable*. New edition ed. Penguin.

Tashman, Leonard J. 2000. Out-of-sample tests of forecasting accuracy: an analysis and review. *International Journal of Forecasting* **16**(4) 437–450.

Taylor, J W. 2008. Exponentially weighted information criteria for selecting among forecasting models. *International Journal of Forecasting* **24** 513–524.

Timmermann, Allan. 2006. Forecast combinations. *Handbook of Economic Forecasting* **1** 135–196.

van Donselaar, Karel H., Vishal Gaur, Tom van Woensel, Rob A. C. Broekmeulen, Jan Fransoo. 2010. Ordering behavior in retail stores and implications for automated replenishment. *Management Science* **56**(5) 766–784.

Weller, M., S. Crone. 2012. Supply Chain Forecasting: Best Practices & Benchmarking Study. *Lancaster Centre For Forecasting, Technical Report* .