On the Impact of Forecast Reliability on Downstream Decision-Making

Jente Van Belle[a,c], Wouter Verbeke[a], Nikolaos Kourentzes[b]

[a]*Information Systems Engineering Research Group (LIRIS), Faculty of Economics and Business, KU Leuven, 3000 Leuven, Belgium*
jente.vanbelle@kuleuven.be
wouter.verbeke@kuleuven.be
[b]*Independent, Stockholm, Sweden*
nikolaos@kourentzes.com
[c]*Presenting author*

Abstract

While assessing forecast accuracy is of key importance in the evaluation of time series forecasting models, the consistency of their performance over time has remained largely underexplored in the literature. We refer to this property as temporal forecast reliability. A common evaluation approach is the rolling-origin setup, in which forecast traces containing one- to h-step-ahead forecasts are generated from multiple forecasting origins. By aggregating performance metrics across these origins, the evaluation captures how methods adapt to varying conditions and accounts for multiple test periods. This is typically limited to mean performance across forecast traces and thus overlooks the consistency of performance over time: occasional unreliable forecasts may be masked by these summary statistics. For example, an analysis of performance over time may reveal one method to have stationary and homoscedastic errors, while another might have periods of highly volatile or nonstationary performance. We formalize this concept of temporal forecast reliability and demonstrate its importance as an additional evaluation criterion. Specifically, we show that when forecasts are used as inputs to decision-making processes, differences in forecast reliability can affect downstream decision-making performance due to a nonlinear propagation of forecast errors. As operational decision-making becomes increasingly automated, we argue that forecast reliability should be treated as a key evaluation dimension rather than an afterthought.

Keywords: Time series, Rolling-origin evaluation, Knapsack problem, Semivariance.

1. Introduction

Time series forecasts usually serve as inputs to decision-making processes. Hence, they add value through their impact on the quality of the solution to a decision-making problem. Given recent evidence that forecast accuracy does not necessarily correlate well with decision-making quality (Kourentzes et al., 2020; Yardley and Petropoulos, 2021), there is growing interest in designing forecasting methods and processes that incorporate additional considerations relevant to decision-makers.

Existing work in this area can be grouped into two categories. A first line of research follows a utility-based approach, in which forecast accuracy is replaced by problem-specific metrics

that capture how forecasts will be used to support decision-making during model evaluation (see, e.g., Petropoulos et al., 2019), model optimization (see, e.g., Kourentzes et al., 2020), or model combination (see, e.g., Ghazanfariharandi and Mieth, 2025). A second line of research follows the traditional predict-then-optimize rationale (i.e., a standard loss function is used to optimize the forecasting model) but takes into account complementary evaluation criteria during model selection (see, e.g., Pritularga and Kourentzes, 2024), model optimization (see, e.g., Van Belle et al., 2023, 2024), or model combination (see, e.g., Godahewa et al., 2025).

With this paper, we contribute to the second line of research. Specifically, we focus on one of the additional evaluation criteria proposed by Spavound and Kourentzes (2022): forecast reliability. Spavound and Kourentzes (2022) conceptually describe forecast reliability as being about minimizing the likelihood of extreme errors. In this work, we formalize one type of forecast reliability, namely temporal forecast reliability, which concerns the consistency of accuracy metrics over time, and we assess its impact on downstream decision-making performance. We distinguish temporal forecast reliability from cross-sectional forecast reliability, which concerns the consistency of performance across different time series and is not the focus of this paper. We investigate whether our proposed definition of temporal forecast reliability can be used in addition to forecast accuracy results to guide model selection towards forecasting models that consistently perform well in terms of decision-making performance.

This paper is structured as follows. In Section 2, we look at the differences between utility-based approaches and approaches that consider complementary evaluation criteria in addition to forecast accuracy to better align forecasting with decision-making in more detail. We then review existing work on the complementary evaluation criterion central to our study and provide an overview of commonly adopted forecast evaluation schemes, with a focus on the best practice: the rolling-origin evaluation. This rolling-origin evaluation approach forms the basis of our proposed definition of temporal forecast reliability, which is introduced in Section 3. To assess whether temporal forecast reliability affects downstream decision-making, we consider in Section 4 a knapsack problem in which part of the problem parameters must be forecasted from historical data. In Section 5, we briefly discuss how forecast reliability is conceptually distinct from forecast accuracy and how it should be weighed against it. Finally, in Section 6, we summarize the findings of this paper and outline directions for future research.

2. Literature Review

2.1. Forecasting for Decision-Making

Plenty of evidence exist that better forecast accuracy does not necessarily result in better decision-making performance. For instance, Kourentzes et al. (2020) and Theodorou et al. (2025), find that the most accurate forecasts do not necessarily lead to the best inventory performance. To better align forecasting and downstream decision-making, existing work focuses on either a utility-based approach, or complements forecast accuracy with additional evaluation criteria.

Following the utility-based approach, decision-making is explicitly placed at the center by directly evaluating and selecting forecasts based on this utility-based perspective, or even directly optimizing forecasts from this perspective. For instance, Sagaert et al. (2019) set up an inventory simulation to evaluate the gains of incorporating external macroeconomic information in demand forecasts directly in terms of inventory performance. Kourentzes et al. (2020) take this one step further by relying on simulation optimization to optimize simple exponential smoothing forecasts

directly for inventory performance. Such utility-based optimization is also referred to in the literature as predict-and-optimize/predict+optimize (Mandi et al., 2020), decision-focused learning (Mandi et al., 2024) and value-oriented forecasting (Carriere and Kariniotakis, 2019).

Relying on these utility-based solutions has an important consequence: there is no "single source of truth" with regard to the forecast anymore, as different decision-problem parameters might yield different forecasts. This can be undesirable, for example, when the forecasts are used as inputs to multiple, different, decision-making problems (e.g., sales forecasts might be used as inputs to decision-making in both an operations and a financial context). Moreover, these approaches focus on tightly integrating forecasting and algorithmic decision-making, making the assumption that the operational information (e.g., decision-problem parameters) required to set up simulations and parameterize the decision problem is readily available. However, many decisions are not algorithmic, and the operational information needed to fully characterize all relevant decision problems is not always readily available, for example due to organizational silos (Athanasopoulos and Kourentzes, 2023). In such cases, one must resort to the more classical two-stage predict-then-optimize approach, in which forecasting and decision-making are dealt with separately and independently. Finally, from a process perspective, integrated utility-based metrics can complicate the management and development of forecasting capabilities and forecasting experts, as any quantitative feedback depends on decisions or actions that lie beyond the forecaster's control.

This is where the second category of solutions fits in. These solutions follow the classical predict-then-optimize rationale, but implicitly acknowledge that decision-making is the ultimate goal. They do so by complementing forecast accuracy with additional quality criteria for model selection or optimization, thereby weakening the independence between forecasting and decision-making while still retaining a single source of truth. That is, additional metrics that are indirectly related to decision-making performance are considered during evaluation or optimization. Spavound and Kourentzes (2022) have proposed four additional quality dimensions based on what practitioners have expressed: (i) forecast stability, (ii) forecast intelligibility, (iii) forecast alignment with user needs, and (iv) forecast reliability.

Forecast stability refers to the variability in forecasts for a specific period caused by updating the forecast when new observations become available. Since, in a setting in which forecasts are updated on a rolling basis, more stable forecasts can be expected to lead to fewer and smaller revisions to plans and to greater trust in the forecasting system—potentially resulting in fewer unwarranted judgmental adjustments that often lead to a decrease in forecast accuracy—taking this dimension into account in model selection or optimization would plausibly have a positive indirect impact on decision-making performance (Van Belle et al., 2023). The argument about greater trust also underlies the rationale for forecast intelligibility, which Spavound and Kourentzes (2022) characterize as the need for a certain degree of explainability and a functional understanding of a forecast and of a forecasting model's key elements, respectively. Forecast alignment with user needs refers to building and evaluating forecasting models using metrics that correlate well with decision outcomes. This criterion is closely related to the first category of solutions, with the key difference being that they propose using such metrics to complement traditional forecast accuracy metrics, rather than to substitute for them. Finally, forecast reliability concerns minimizing the likelihood of extreme errors. It is clear that these considerations require broader organizational engagement and go beyond strict model-building activities.

2.2. The Concept of Forecast Reliability

To capture the general idea of forecast reliability, Spavound and Kourentzes (2022) formulate it as follows: "A method with a low MAPE that carries risks of outlying errors may be less reliable than one with a higher average error but lower error variance." This intentionally general statement can be interpreted in multiple ways: in terms of (i) the variability in performance of a method across time series, (ii) the variability in raw forecast errors being summarized into a specific error metric, and (iii) the variability of an error metric across time.

Although formal definitions of forecast reliability, or metrics to evaluate it, have not yet been established, the well-established practice of using a Multiple Comparisons with the Best (MCB) test (Koning et al., 2005) to assess whether differences in mean forecast accuracy across time series are statistically significant does take into account the variability in a method's performance across time series. This is because the MCB test is based on the average rank of each method across all time series in a dataset. Intervals are constructed around these average ranks, and methods whose intervals do not overlap can be considered to result in statistically significant differences in forecast accuracy. The MCB test is a post-hoc test that builds on the Friedman test and is a special case of the Nemenyi test (Hollander et al., 2014). There is also work that considers the other types of variability mentioned above. For instance, with respect to the second type of variability, Barrow and Kourentzes (2016) investigate the impact of combining multiple forecasts on the distribution of forecast errors and find that forecast combinations, on average, result in more consistent behavior between in-sample and out-of-sample error distributions and that these distributions are closer to the normal distribution than those of the base forecasts, which are desirable properties for safety stock calculations. Finally, with respect to the variability of an error metric over time, the work of Pritularga et al. (2021) on forecast reconciliation not only evaluates the impact of different strategies for estimating the reconciliation weights matrix on the average of various forecast accuracy metrics, but also explicitly reports the variability of these metrics across different test periods. They find that some strategies lead to considerable improvements in average accuracy, but these come at the cost of increased variability across test periods, whereas other strategies yield more modest average accuracy gains but with more consistent improvements across different test periods. They argue that users may prefer strategies that yield smaller average improvements but more consistent accuracy gains to support decision-making.

This latter work is closest in spirit to the type of variability on which we will focus. In Section 3, we build on similar reasoning to formalize temporal forecast reliability, while in Section 4 we seek empirical evidence that more consistent accuracy performance is beneficial for decision-making. Importantly, as in Pritularga et al. (2021), our definition of temporal forecast reliability also requires multiple test periods. The risk of relying on a single test period was already highlighted by Pack (1990) and Fildes et al. (1998), who examined the variability of error metrics across multiple test periods and showed that the accuracy-based ranking of forecasting methods can change appreciably as the test period varies. As explained in the next section, the use of multiple test periods is commonly achieved via a rolling-origin evaluation setup.

2.3. Forecast Evaluation Schemes

Much has been written about the choice of forecast accuracy metrics (see, e.g., Hyndman and Koehler, 2006; Kourentzes and Athanasopoulos, 2019; Kolassa, 2020), but there has long been a consensus that these metrics should be based on holdout samples rather than on measures of in-sample fit. The seminal work of Tashman (2000) provided guidelines for the implementation of

such out-of-sample evaluation. To improve the "efficiency and reliability" of these evaluations, the author proposes the use of multiple test periods through rolling-origin evaluation—where the forecasting origin is the final time in the data used for model fitting, that is, the point from which the forecasts are generated—with updating of model parameters. In a rolling-origin evaluation, the forecasting origin is successively updated and forecast traces are produced that contain one- to h-step-ahead forecasts from each new origin. Forecast accuracy metrics are then averaged across these multiple test periods, thereby capturing how methods adapt to varying inputs and mitigating sensitivity to special events unique to any single test period (e.g., a specific phase of the business cycle).

A single-origin or fixed-origin evaluation, however, is still widely adopted. This is, for instance, usually the case in forecasting competitions such as the influential M-competitions (Makridakis and Hibon, 2000; Makridakis et al., 2020, 2022). This choice facilitates the practical organization of such competitions, since a rolling-origin evaluation would require either participants to submit executable code so that organizers can generate the forecasts themselves (instead of submitting forecasts directly), or participants to be granted access to the test-period data in order to allow them to successively update the forecasting origin—thereby introducing the risk that participants may use the test data for model fitting. To mitigate the limitations of the single-origin approach, forecasting competitions typically base their evaluation on average forecast accuracy, computed by averaging accuracy metrics across a large and diverse set of time series drawn from different data sources and with different starting and ending dates. However, it is important to emphasize that a large collection of time series does not automatically guarantee diversity in forecasting situations, particularly when the series cover largely overlapping calendar periods. In general, heterogeneity across series and heterogeneity across time can stress the capabilities of a forecasting method, but do so in very different ways.

3. Temporal Forecast Reliability

Although rolling-origin evaluation is generally not adopted in forecasting competitions, it is the de facto standard evaluation approach in real-world forecasting applications. Yet, performance aggregation across the different test periods is typically limited to the mean performance, thereby overlooking the variability of the forecast accuracy metric across test periods—i.e., the third type of variability discussed in Section 2.2. This summary statistic may mask occasional unreliable forecasts (i.e., for specific test periods), since methods with similar average performance may differ substantially in their temporal consistency, with some performing consistently over time and the other exhibiting greater volatility.

As described in Section 2.2, one exception is the work of Pritularga et al. (2021), in which the logarithm of the geometric standard deviation is used to measure variability in the geometric mean of relative accuracy measurements across multiple test periods. This approach captures both upside and downside variability. While this provides useful information on the consistency of a method over time, we argue that the directional nature of accuracy measurements should be taken into account. Specifically, in a "lower is better" setting, the primary concern of forecasters and decision-makers lies in the risk associated with the upper tail of the accuracy metric distribution.

We therefore argue that this type of temporal variability in performance—i.e., the risk of poor performance for some test periods—should be avoided, since variability in forecast performance across origins (i) might propagate to decision-making performance, and (ii) may erode trust in

the forecasting method used. Regarding the former, occasional periods of relatively low forecast accuracy may have a disproportionately large negative impact on downstream decision-making performance. Moreover, such instability can induce higher variability in relative decision-making performance metrics, making absolute performance more dependent on decision-problem-instance-specific parameters.

To operationalize the measurement of risk in the upper tail of the accuracy metric distribution, we draw inspiration from risk measures commonly used in financial and insurance risk management to quantify so-called downside risk, which measures only the variability of returns below a reference value: lower partial moments (McNeil et al., 2015), with the lower semivariance (Markowitz, 1959) as a special case. Two variants of the lower semivariance can be defined. First, the below-target semivariance, denoted by $\sigma^2_{L,\text{target}}$, is defined as the dispersion of returns $R \in \mathbb{R}$ below a target level $b \in \mathbb{R}$:

$$\sigma^2_{L,\text{target}} = \mathbb{E}\Big[\big(\min\big(R-b,\,0\big)\big)^2\Big]. \tag{1}$$

If $b = \mathbb{E}[R]$, this expression reduces to the below-mean semivariance. In a "lower is better" setting, we can use the upper semivariance variants of the above definitions. Note that the below-target semivariance can equivalently be written as

$$\sigma^2_{L,\text{target}} = \mathbb{E}\Big[(R-b)^2 \mid R < b\Big]\,\mathbb{P}(R<b), \tag{2}$$

where $\mathbb{P}(R<b)$ can be interpreted as the frequency of below-target returns and the conditional expectation captures the severity of such returns when they occur. Hence, the semivariance weights the severity of below-target returns by their empirical frequency.

To formally define temporal forecast reliability, let o_i $(i = 1,\ldots,n)$ denote the i-th forecasting origin used in a rolling-origin evaluation of forecasting method k $(k = 1,\ldots,m)$. The forecasting origins may correspond to successive time periods, but can also be spaced using a step size larger than one. For each origin o_i, we collect one- to h-step-ahead errors $e_{o_i,j}$ $(j = 1,\ldots,h)$. Using an accuracy metric M (with a "lower is better" convention) to summarize performance per origin o_i, we obtain a collection of accuracy measurements $(M_{o_1}, M_{o_2}, \ldots, M_{o_n})$:

$$\begin{bmatrix} e_{o_1,1} & e_{o_1,2} & \cdots & e_{o_1,h} \\ e_{o_2,1} & e_{o_2,2} & \cdots & e_{o_2,h} \\ \vdots & \vdots & \ddots & \vdots \\ e_{o_n,1} & e_{o_n,2} & \cdots & e_{o_n,h} \end{bmatrix} \quad \begin{matrix} \longrightarrow \\ \longrightarrow \\ \\ \longrightarrow \end{matrix} \quad \begin{bmatrix} M_{o_1} \\ M_{o_2} \\ \vdots \\ M_{o_n} \end{bmatrix} \quad \begin{matrix} \\ \\ \\ \\ \downarrow \\ \bar{M} \end{matrix},$$

where $\bar{M}$ denotes the mean performance across the different test periods.

We define the above-mean temporal forecast reliability for method k as

$$\text{TFR} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\big(\max(M_{o_i}-\bar{M},0)\big)^2}. \tag{3}$$

We further define the above-target temporal forecast reliability, denoted by $\text{TFR}_{\min}$, where $\bar{M}$ is replaced by the best mean performance across all methods $k = 1,\ldots,m$: $\bar{M}_{\min} = \min\{\bar{M}_1, \bar{M}_2, \ldots, \bar{M}_m\}$.

The use of $\bar{M}$ implicitly assumes that the calculated metrics are stationary across the different test periods. This condition is easier to meet when M is based on a metric that accounts for the forecast difficulty of each period, which is often captured by the performance of the random walk. Therefore, we recommend using metrics that are scaled in this way.

To emphasize the severity of poor performance when it occurs, relying on the decomposition of the semivariance in Eq. (2), we define the severity-only variants as

$$\mathrm{TFR}^{\mathrm{sev}}_{\mathrm{target}} = \sqrt{\frac{1}{n_+}\sum_{i=1}^{n}\big(\max(M_{o_i} - M_{\mathrm{target}}, 0)\big)^2}, \quad n_+ = \sum_{i=1}^{n}\mathbf{1}\{M_{o_i} > M_{\mathrm{target}}\}. \tag{4}$$

If the impact of occasional very poor forecasting performance on decision-making is expected to be very high, these severity-only temporal forecast reliability metrics might be preferred. The metrics TFR and $\mathrm{TFR}^{\mathrm{sev}}$ provide information that can be used in addition to the mean performance $\bar{M}$, while $\mathrm{TFR}_{\mathrm{min}}$ and $\mathrm{TFR}^{\mathrm{sev}}_{\mathrm{min}}$ directly provide an integrated view across all methods.

In the finance literature, metrics have been proposed that explicitly account for the upside potential of returns above a target level relative to the downside risk of returns below a target level; see, for example, the upside-potential ratio proposed by Sortino et al. (1999). Although test periods in which better-than-mean performance is observed are also desirable in the context of forecasting for decision-making, such metrics are less well aligned with the above motivation for temporal consistency in performance—namely, that poor forecasting performance often propagates nonlinearly to downstream decision-making and erodes trust—and are therefore not included here.

4. Simulation Experiment

4.1. Problem Setup and Data

The decision problem we consider is a knapsack problem (Kellerer et al., 2004) in which part of the problem parameters must be forecasted from historical data. Specifically, at the beginning of each day, we must decide when to use C units of "free" stored energy instead of purchasing electricity at time-varying market prices. Each day is divided into 48 time slots of 30 minutes, indexed by $i = 1, \ldots, 48$. Each slot corresponds to a knapsack item $x_i \in \{0,1\}$, with a known scheduled time-varying electricity consumption $w_i \in \mathbb{R}^+$ and a unit electricity price $p_i \in \mathbb{R}^+$.

The time-varying electricity prices p_i are not known at the time when the energy usage schedule must be determined (i.e., when the combinatorial optimization problem is solved). To deal with this problem structure, we adopt a sequential approach: first, we generate forecasts for the electricity prices p_i, denoted $\hat{p}_i$; second, we solve the resulting 0-1 knapsack problem using these forecasted prices as inputs to maximize the expected cost savings. That is, time slots are selected such that the sum of the expected saved electricity costs is maximized, subject to the capacity constraint C.

Mathematically, the problem is formulated as

$$\begin{aligned} &\text{maximize} \sum_{i=1}^{48} \hat{p}_i w_i x_i \\ &\text{subject to} \sum_{i=1}^{48} w_i x_i \leq C, \quad x_i \in \{0,1\}. \end{aligned} \tag{5}$$

To solve this problem, we use the `mknapsack()` function from the `adagio` R package (Borchers, 2023), which reformulates the knapsack problem as an integer linear program and solves it with the `lp_solve` Mixed Integer Linear Programming solver (Csárdi and Berkelaar, 2024). While this approach typically yields high-quality solutions and supports continuous (positive) problem parameters, we replace the solution with a greedy solution when the latter achieves a higher objective value. The greedy approach sorts the time slots in descending order of forecasted electricity prices $\hat{p}_i$ and then iterates over them, selecting a time slot ($x_i = 1$) if the remaining capacity allows, and skipping it ($x_i = 0$) otherwise.

For our experiment, we simulate knapsack problem instances. For the electricity prices p_i, we use historical electricity price data from the Irish Single Electricity Market Operator (Mandi et al., 2020), recorded at 30-minute intervals from 00:00 on 1 November 2011 to 31 December 2013. The scheduled electricity consumption w_i for each time slot is sampled from a normal distribution, $w_i \sim N(2,2)$, with negative values replaced by zero, and the day-dependent capacity C is sampled from a Poisson distribution, $C \sim P(20)$. Using this setup, we simulate 50 knapsack problem instances for each day of 2013, the final year in the electricity price dataset.

4.2. Forecasting Methods

The goal of this simulation experiment is to assess whether temporal forecast reliability affects downstream decision-making. To this end, for each day of 2013, we generate one- to 48-step-ahead forecasts for the electricity prices p_i $(i = 1,\ldots,48)$ using different forecasting methods:

- Nv1: the naïve method (random walk), i.e., the last observed value is used as the forecast for all time slots;
- NvD: the seasonal naïve method with a seasonal period of 48, i.e., the forecast for each horizon is equal to the price at the same time slot on the previous day;
- NvW: the seasonal naïve method with a seasonal period of 336 (48×7), i.e., the forecast for each horizon is equal to the price at the same time slot on the same day of the previous week;
- AR: a linear regression using autoregressive lags of the target variable. The series has no level or seasonal unit roots, and therefore we provide a large selection of lags to capture local dynamics and seasonal cycles (lags: 1 to 47, 48, 96, 144, 192, 240, 288 [1 to 6 days], 336, 672, 1008, 1344 [1 to 4 weeks], 17472 [1 year]). The selection of lags is refined by maximizing the AIC with a backward selection strategy. We use autoregressive lags as they are more economical in inputs than typical seasonal features and allow for stochastic components;
- Rdg: the same inputs as AR, but with a Ridge penalty for the estimation;
- RF: the same inputs as AR, but using a random forest as the learning algorithm (`randomForest` R package, Liaw and Wiener, 2002);
- ELM: the same inputs as AR, but using an extreme learning machine as the learning algorithm. The model uses a hidden layer with 100 neurons, and the linear output layer relies on stepwise selection to estimate their contributions (`nnfor` R package, Kourentzes, 2023).

All methods rely on iterative forecasting, where the next step-ahead forecast is iteratively used as an autoregressive input to obtain the subsequent forecast, and so on. Given the substantial training sample, we estimate the methods once and do not updated during the evaluation.

4.3. Forecasting and Decision-Making Performance Evaluation

We evaluate the one- to 48-step-ahead electricity price forecasts for each day in the test set in terms of accuracy using a variant of the root mean squared scaled error (RMSSE):

$$\text{RMSSE} = \sqrt{\frac{\frac{1}{48}\sum_{i=1}^{48}(p_i - \hat{p}_i)^2}{\frac{1}{48-1}\sum_{i=2}^{48}(p_i - p_{i-1})^2}}, \tag{6}$$

where the out-of-sample actual prices are used in the denominator to scale the forecast errors in order to account for time-varying intrinsic uncertainty in the electricity prices. Using the same scaling approach, we also compute the absolute mean scaled error (|MScaledE|) for each test period to assess the magnitude of forecast bias.

To evaluate the quality of the knapsack solutions obtained using the forecasted prices as inputs instead of the actual electricity prices, we calculate the regret, defined as the difference between the savings that would have been achieved if the actual prices were known and the savings realized by using the solution based on the forecasted prices. Let $\hat{\mathbf{x}}$ denote the solution for a specific knapsack problem instance obtained by solving it using the forecasted electricity prices $\hat{p}_i$, and let $\mathbf{x}^*$ denote the solution for the same problem instance under the actual prices p_i. The realized cost savings of an energy usage schedule $\mathbf{x}$, denoted by Savings($\mathbf{x}$), are given by $\sum_{i=1}^{48} p_i w_i x_i$. The relative regret (RelRegret) is then defined as

$$\text{RelRegret} = \frac{\text{Savings}(\mathbf{x}^*) - \text{Savings}(\hat{\mathbf{x}})}{\text{Savings}(\mathbf{x}^*)}. \tag{7}$$

4.4. Results

To assess the impact of temporal forecast reliability on downstream decision-making, we investigate the relationship between the distributions across test periods of forecast accuracy (RMSSE) and forecast bias magnitude (|MScaledE|) on the one hand, and decision-making performance (RelRegret) on the other. Boxplots of these distributions are shown in Figure 1. For the RelRegret distribution, we compute RelRegret for each day in the test set by averaging over 50 different knapsack problem instances with different problem parameters but identical electricity price forecasts (see Section 4.1). This averaging mitigates the impact of instance-specific parameters and allows us to better assess the relationship between forecast performance and decision-making performance, as the link between poor forecast accuracy (high RMSSE) for a given test period and the RelRegret for a specific knapsack problem instance may be weak when large forecast errors occur at time slots with low scheduled consumption.

From Figure 1, we observe a clear relationship between average accuracy and RelRegret, with the best-performing forecasting methods, AR and Rdg, achieving the lowest average RelRegret. This is also reflected in the average accuracy and RelRegret values reported in Table 1. When considering average accuracy and bias alone, one might be largely indifferent between AR and Rdg. However, by additionally taking into account the temporal forecast reliability metrics, this conclusion clearly changes in favor of Rdg. The more consistent accuracy performance of Rdg appears to have a positive effect on the RelRegret distribution, characterized by a slightly lower average RelRegret and reduced upside variability. This is further reflected in a lower $\text{CTE}^{0.75}_{\text{RelRegret}}$, which is the empirical conditional tail expectation at the 75% level (i.e., calculated over the upper 25% of the RelRegret measurements).

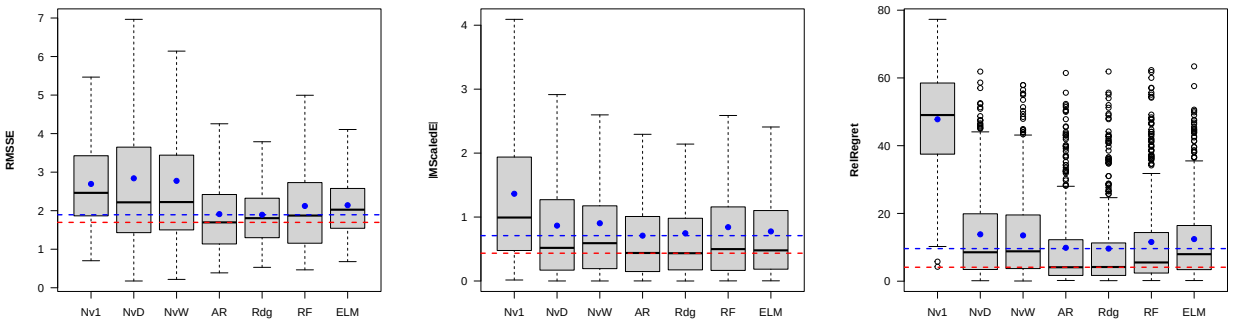

Figure 1: Boxplots showing the distributions of accuracy (RMSSE), bias (|MScaledE|), and relative regret (RelRegret). Blue dots indicate the means of the distributions. The red and blue dashed lines, centered on the minimum medians and means, respectively, serve as reference lines. Outliers are not shown for accuracy and bias for visual clarity.

More generally, we observe that higher "bad" variability in RMSSE, as captured by the temporal forecast reliability metrics, propagates to decision-making as higher "bad" variability in RelRegret, as reflected in $\text{CTE}^{0.75}_{\text{RelRegret}}$. For example, although ELM performs slightly worse than RF in terms of average accuracy and average RelRegret, the heavier upper tail of the RMSSE distribution for RF—reflected in higher values for the TFR metrics—translates into a heavier upper tail of the RelRegret distribution. This results in a $\text{CTE}^{0.75}_{\text{RelRegret}}$ of 32.2 for RF, compared to 30.9 for ELM. Depending on the forecast user's risk profile, ELM may therefore be preferred over RF to mitigate the probability of incurring large absolute losses (i.e., forgone cost savings). For the severity-only temporal reliability variants, this propagation is less consistent, with the lower $\text{TFR}^{\text{sev}}_{\text{min}}$ of ELM compared to AR not being reflected in the cost savings metrics.

Finally, Nv1 exhibits high forecast bias and poor decision-making performance, with substantially worse RelRegret than NvD and NvW despite comparable accuracy. This weak connection between forecast and decision-making performance for Nv1 results from the fact that, by construction, Nv1 does not differentiate between time slots based on price and therefore fails to provide informative signals for solving the knapsack problems to optimality.

	Accuracy	**Bias**	**Temporal reliability RMSSE**				**Cost savings**	
	$\overline{\text{RMSSE}}$	$\overline{\lvert\text{MScaledE}\rvert}$	TFR	TFR_{min}	TFR^{sev}	$\text{TFR}^{\text{sev}}_{\text{min}}$	$\overline{\text{RelRegret}}$	$\text{CTE}^{0.75}_{\text{RelRegret}}$
Nv1	2.694	1.364	0.962	1.420	1.479	1.645	47.8	65.5
NvD	2.840	0.866	1.988	2.412	3.340	3.153	13.8	34.9
NvW	2.774	0.903	1.657	2.085	2.839	2.726	13.5	33.2
AR	*1.911*	**0.709**	0.836	*0.844*	1.293	1.297	*9.9*	*28.3*
Rdg	**1.894**	*0.747*	**0.695**	**0.695**	**1.032**	**1.032**	**9.6**	**27.5**
RF	2.122	0.842	1.008	1.122	1.566	1.604	11.6	32.2
ELM	2.142	0.775	*0.739*	0.865	*1.084*	*1.138*	12.4	30.9

Table 1: Average accuracy and bias, together with temporal reliability metrics for RMSSE measurements; and average relative cost savings and the empirical conditional tail expectation at the 75% level of the RelRegret measurements. Lower is better. Bold indicates the best result per column, and italics indicate the second-best.

5. Discussion

Ultimately, the objective is to support an analyst in selecting the forecast that provides the most value for the supported decisions. In the theoretical case, where a forecasting model would match the underlying data-generating process, maximizing accuracy would be sufficient, as the errors would be well behaved (matching the assumed innovation distribution). In practice, however, this is not feasible, with the quality of the forecasts depending on the function form difference from the unknown data-generating process. This typically motivates the use of multiple evaluation metrics, particularly given the limited correlation between forecast accuracy and decision quality observed in many applications (see Section 2.1). Athanasopoulos and Kourentzes (2023) propose adopting a multi-objective framing to accommodate alternative metrics, and identify forecasts that are globally dominant (outperforming alternatives across all metrics) or partially dominant (outperforming on some metrics). Alternatively, when analysts can identify a weighting function over the different metrics, the multi-objective loss can be linearized into a single metric that can be optimized.

Figure 2 follows this logic and plots the average accuracy ($\overline{\text{RMSSE}}$) and temporal forecast reliability (TFR) reported in Table 1. The size of the points is inversely proportional to the cost savings ($\overline{\text{RelRegret}}$), such that larger points are preferable. The Rdg forecasts are dominant; however, identifying the second-best method—or more generally ranking the remaining forecasting methods—is non-trivial, as they are only partially dominant. For instance, ELM is more reliable than AR and RF, with a substantial accuracy difference from AR but only a marginal one from RF. Similarly, among the naïve methods, Nv1 dominates NvW and NvD in terms of accuracy and reliability, but the latter provide more valuable forecasts as measured by $\overline{\text{RelRegret}}$. The same can be said about the partial dominance of Nv1 over RF.

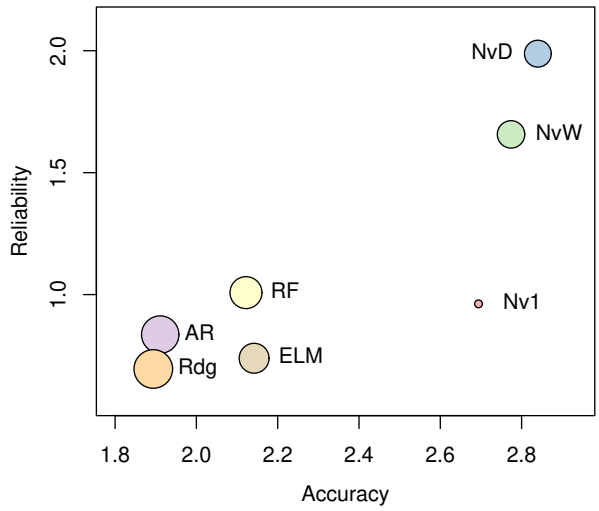

Figure 2: Accuracy ($\overline{\text{RMSSE}}$) versus reliability (TFR), with decision-making value indicated by point size (inversely proportional to $\overline{\text{RelRegret}}$; larger points denote more valuable forecasts).

The figure highlights the additional information provided by the reliability dimension. Although Rdg is preferable in terms of both accuracy and reliability, a different set of forecasting methods could easily complicate the ranking. Methods that perform well across both dimensions generally provide greater decision-making value, as reflected in lower $\overline{\text{RelRegret}}$. Resolving the

choice of a single forecasting method, however, requires weighing the different evaluation dimensions. To facilitate this, we introduce the notions of metrics and meta-metrics. Metrics, such as accuracy and bias, are scores that have inherent value, as they directly indicate how well a method captures the properties of the target series. The connections between these metrics can often be investigated analytically, for instance, through the decomposition of forecast errors into bias and variance. In contrast, meta-metrics derive their value from the application that the forecasts support. The value of reliability, for example, depends on the decision context: applications requiring high risk aversion (see, e.g., Kourentzes and Svetunkov, 2026) will place a high value on forecast reliability. Other complementary metrics also fall under meta-metrics, such as forecast stability (Van Belle et al., 2023), where, for example, the characteristics of a supply chain may make more stable forecasts preferable. Quantifying the value of meta-metrics requires careful consideration of the application context, potentially with input from stakeholders. This is the key distinction between meta-metrics and metrics. Once the importance of the different metrics and meta-metrics is specified, the multi-objective cost of a method (see Figure 2) can be linearized, making it possible to construct a simple ranking that reflects both data and organizational considerations.

6. Conclusions

In this work, we propose a quantitative characterization of temporal forecast reliability, which has so far been treated in the literature as an abstract notion. We show that incorporating reliability as an additional dimension of forecast evaluation can help identify forecasts that add value to decision-making. Naturally, the importance of reliability is application-dependent and falls under the proposed class of meta-metrics. Meta-metrics derive their value from organizational and stakeholder considerations, in contrast to metrics, which depend on how well forecasts capture the properties of a target series.

Future work is needed to investigate the effectiveness of the proposed reliability meta-metric in other settings, such as inventory management. For instance, Kourentzes et al. (2020) show that forecast accuracy correlates weakly with inventory performance and resort to computationally expensive simulation optimization to generate appropriate forecasts. In such contexts, reliability could arguably help analysts identify valuable forecasts without the need for simulation-based approaches. The variety of organizational decisions mapped onto different hierarchical levels of organizational functions and forecast horizons listed by Athanasopoulos and Kourentzes (2023) provides important dimensions of analysis for future applications of forecast reliability.

The introduction of additional evaluation dimensions increases the complexity of the model selection process. We argue that the weight assigned to meta-metrics in the final scoring of alternative forecasts depends on the application context and stakeholder preferences. One can sidestep assigning explicit weights by constructing Pareto frontiers of forecasting methods (see, e.g., Caljon et al., 2025). Nonetheless, optimal practical use of meta-metrics, such as reliability, explainability, and stability, requires methodologies that help elicit their implied utility for stakeholders. Operationalizing this is an important open question that lies beyond the focus of this work.

Finally, this work is connected to the broader field of data-centric machine learning, which advocates incorporating data-centric considerations throughout the entire machine learning pipeline, extending beyond a focus on model development. The DC-Check framework proposed by Seedat et al. (2024) can serve as a guide for practitioners and researchers in eliciting data-centric considerations at the different stages of the pipeline: data, training, testing, and deployment.

Funding

Jente Van Belle gratefully acknowledges the financial support from Research Foundation – Flanders [grant number 12AZX24N].

References

Athanasopoulos, G., Kourentzes, N., 2023. On the evaluation of hierarchical forecasts. International Journal of Forecasting 39, 1502–1511.

Barrow, D.K., Kourentzes, N., 2016. Distributions of forecasting errors of forecast combinations: Implications for inventory management. International Journal of Production Economics 177, 24–33.

Borchers, H.W., 2023. adagio: Discrete and Global Optimization Routines. URL: `https://CRAN.R-project.org/package=adagio`. R package version 0.9.2.

Caljon, D., Vercauteren, J., De Vos, S., Verbeke, W., Van Belle, J., 2025. Using dynamic loss weighting to boost improvements in forecast stability. International Journal of Forecasting (In Press).

Carriere, T., Kariniotakis, G., 2019. An integrated approach for value-oriented energy forecasting and data-driven decision-making application to renewable energy trading. IEEE Transactions on Smart Grid 10, 6933–6944.

Csárdi, G., Berkelaar, M., 2024. lpSolve: Interface to 'Lp_solve' v. 5.5 to Solve Linear/Integer Programs. URL: `https://CRAN.R-project.org/package=lpSolve`. R package version 5.6.23.

Fildes, R., Hibon, M., Makridakis, S., Meade, N., 1998. Generalising about univariate forecasting methods: Further empirical evidence. International Journal of Forecasting 14, 339–358.

Ghazanfariharandi, M., Mieth, R., 2025. Value-oriented forecast combinations for unit commitment. IEEE Control Systems Letters 9, 1466–1471.

Godahewa, R., Bergmeir, C., Baz, Z.E., Zhu, C., Song, Z., García, S., Benavides, D., 2025. On forecast stability. International Journal of Forecasting 41, 1539–1558.

Hollander, M., Wolfe, D.A., Chicken, E., 2014. Nonparametric Statistical Methods. John Wiley & Sons.

Hyndman, R.J., Koehler, A.B., 2006. Another look at measures of forecast accuracy. International Journal of Forecasting 22, 679–688.

Kellerer, H., Pferschy, U., Pisinger, D., 2004. Multidimensional knapsack problems, in: Knapsack Problems. Springer, pp. 235–283.

Kolassa, S., 2020. Why the "best" point forecast depends on the error or accuracy measure. International Journal of Forecasting 36, 208–211.

Koning, A.J., Franses, P.H., Hibon, M., Stekler, H.O., 2005. The M3 competition: Statistical tests of the results. International Journal of Forecasting 21, 397–409.

Kourentzes, N., 2023. nnfor: Time Series Forecasting with Neural Networks. URL: https://CRAN.R-project.org/package=nnfor. R package version 0.9.9.

Kourentzes, N., Athanasopoulos, G., 2019. Cross-temporal coherent forecasts for australian tourism. Annals of Tourism Research 75, 393–409.

Kourentzes, N., Svetunkov, I., 2026. Incorporating risk preferences in forecast selection. Journal of the Operational Research Society (Accepted).

Kourentzes, N., Trapero, J.R., Barrow, D.K., 2020. Optimising forecasting models for inventory planning. International Journal of Production Economics 225, 107597.

Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. R News 2, 18–22.

Makridakis, S., Hibon, M., 2000. The M3-competition: Results, conclusions and implications. International Journal of Forecasting 16, 451–476.

Makridakis, S., Spiliotis, E., Assimakopoulos, V., 2020. The M4 competition: 100,000 time series and 61 forecasting methods. International Journal of Forecasting 36, 54–74.

Makridakis, S., Spiliotis, E., Assimakopoulos, V., 2022. M5 accuracy competition: Results, findings, and conclusions. International Journal of Forecasting 38, 1346–1364.

Mandi, J., Kotary, J., Berden, S., Mulamba, M., Bucarey, V., Guns, T., Fioretto, F., 2024. Decision-focused learning: Foundations, state of the art, benchmark and future opportunities. Journal of Artificial Intelligence Research 80, 1623–1701.

Mandi, J., Stuckey, P.J., Guns, T., et al., 2020. Smart predict-and-optimize for hard combinatorial optimization problems, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 1603–1610.

Markowitz, H.M., 1959. Portfolio Selection: Efficient Diversification of Investments. Cowles Foundation Monograph 16, New York: John Wiley & Sons.

McNeil, A.J., Frey, R., Embrechts, P., 2015. Quantitative Risk Management: Concepts, Techniques and Tools - Revised Edition. Princeton University Press.

Pack, D.J., 1990. In defense of ARIMA modeling. International Journal of Forecasting 6, 211–218.

Petropoulos, F., Wang, X., Disney, S.M., 2019. The inventory performance of forecasting methods: Evidence from the M3 competition data. International Journal of Forecasting 35, 251–265.

Pritularga, K., Kourentzes, N., 2024. Forecast congruence: A quantity to align forecasts and inventory decisions. Management Science Working Paper No. 1. URL: http://dx.doi.org/10.2139/ssrn.4711817.

Pritularga, K.F., Svetunkov, I., Kourentzes, N., 2021. Stochastic coherency in forecast reconciliation. International Journal of Production Economics 240, 108221.

Sagaert, Y.R., Kourentzes, N., De Vuyst, S., Aghezzaf, E.H., Desmet, B., 2019. Incorporating macroeconomic leading indicators in tactical capacity planning. International Journal of Production Economics 209, 12–19.

Seedat, N., Imrie, F., Schaar, M.v.d., 2024. Navigating data-centric artificial intelligence with DC-check: Advances, challenges, and opportunities. IEEE Transactions on Artificial Intelligence 5, 2589–2603.

Sortino, F., Van Der Meer, R., Plantinga, A., 1999. The Dutch triangle. Journal of Portfolio Management 26, 50–57.

Spavound, S., Kourentzes, N., 2022. Making forecasts more trustworthy. Foresight: The International Journal of Applied Forecasting Q3, 21–25.

Tashman, L.J., 2000. Out-of-sample tests of forecasting accuracy: An analysis and review. International Journal of Forecasting 16, 437–450.

Theodorou, E., Spiliotis, E., Assimakopoulos, V., 2025. Forecast accuracy and inventory performance: Insights on their relationship from the m5 competition data. European Journal of Operational Research 322, 414–426.

Van Belle, J., Crevits, R., Caljon, D., Verbeke, W., 2024. Probabilistic forecasting with modified N-BEATS networks. IEEE Transactions on Neural Networks and Learning Systems 35, 18872–18885.

Van Belle, J., Crevits, R., Verbeke, W., 2023. Improving forecast stability using deep learning. International Journal of Forecasting 39, 1333–1350.

Yardley, L., Petropoulos, F., 2021. Beyond error measures to the utility and cost of the forecasts. Foresight: The International Journal of Applied Forecasting 63, 36–45.

Jente Van Belle is an FWO Junior Postdoctoral Fellow at the Faculty of Economics and Business, KU Leuven, Belgium. He received his PhD (2021) and MSc (2015) degrees in Applied Economics (Business Engineering) from Vrije Universiteit Brussel, Belgium. His research focuses on time series forecasting and prescriptive analytics and has been published in established international scientific journals, including IEEE TNNLS, IJF, and EJOR. His current research project, MBA-FORECAST, focuses on Moving Beyond Accuracy to optimize time series FORECASTs.

Wouter Verbeke is Professor of Decision Science at the Faculty of Economics and Business, KU Leuven, Belgium. His research is situated in the field of data science for business decision-making and is driven by real-life applications in operations management, pricing, fraud, and credit risk management. His work has been published in established international scientific journals such as IEEE TKDE, EJOR, and Information Sciences, and he has authored two books on fraud and profit-driven analytics. He is a member of the editorial boards of Business Information Systems Engineering and the Journal of Business Analytics.

Nikolaos Kourentzes has published extensively in forecasting, with a focus on model specification, hierarchical forecasting, and demand forecasting in supply chains, among other areas. His current research emphasizes translating forecasts into actionable decisions, with particular interest in modeling, behavioral, and organizational aspects. Known for his work in temporal hierarchies, which has received various awards and recognition, he collaborates closely with industry and organizations such as the International Monetary Fund.