Finding Wisdom in the Crowd: Adaptive Accuracy-Diversity Trimming for Forecast Combinations

Xiaoqian Wang[1], Yanfei Kang[2] and Feng Li[3*]

[1]*Center for Forecasting Science, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China*

[2]*School of Economics and Management, Beihang University, Beijing 100191, China*

[3]*Guanghua School of Management, Peking University, Beijing 100871, China*

Abstract

Forecast combinations support production and inventory decisions, but inaccurate or redundant candidates can dilute crowd wisdom. We develop a modular forecast-trimming framework based on robustness, accuracy, and diversity, introducing autoAD and autoRAD. Both adaptively balance error magnitude and complementarity, with autoRAD additionally applying robustness screening. Relative Diversity (RelDiv) provides a scale-normalised diagnostic of the initial pool. Across 103,826 M-competition series, autoAD performs comparably to accuracy-only trimming overall, with significant MASE-rank gains for monthly series and moderate-RelDiv pools. On 5,000 Royal Air Force spare-parts series, it significantly outperforms accuracy-only trimming in rank at all three horizons. Allowing a zero diversity weight lets autoAD revert to accuracy-based selection when validation evidence does not support complementarity. Together with the weaker performance of alignment-only trimming, these findings support diversity as a selective addition to accuracy. In the spare-parts application, autoAD achieves interval scores comparable to the full-pool and accuracy-only benchmarks, while autoRAD attains the lowest scores considered. Both substantially reduce pool size while maintaining similar service levels, with modest inventory cost savings under moderate backorder penalties. These savings disappear when penalties are high. The framework provides a practical approach to constructing parsimonious forecasting crowds for operational decision support.

Practitioner Summary

Practitioners often have many plausible forecasting models but limited guidance on which ones should enter a forecast combination. Including every model can retain inaccurate or redundant information, while selecting models only by past accuracy may discard forecasts that provide useful complementary information. This study develops a practical trimming framework that considers forecast stability, accuracy, and error complementarity. Its adaptive approach treats accuracy as the primary criterion and uses diversity only when validation data indicate that it adds value.

Tests on more than 100,000 benchmark series and 5,000 Royal Air Force spare-parts demand series show that diversity is most useful as a selective complement to accuracy rather than as a stand-alone criterion. In the spare-parts application, the adaptive methods substantially reduce the number of forecasts while maintaining similar service levels and prediction-interval quality. They can also produce modest inventory cost savings when

E-mail addresses: xiaoqian.wang@amss.ac.cn (Xiaoqian Wang), yanfeikang@buaa.edu.cn (Yanfei Kang), feng.li@gsm.pku.edu.cn (Feng Li, Corresponding author).

backorder penalties are moderate, although these gains disappear when shortages are very costly. For practitioners, the key implication is to take a deliberate approach to choosing which forecasts to retain: start with accuracy-based selection, allow diversity to influence the pool only when supported by validation evidence, and assess the retained pool against the service targets and shortage costs relevant to the application. This provides a practical way to simplify forecasting systems without materially weakening predictive or operational performance.

Word count: 8213

Keywords: Forecast combination, Forecast trimming, Demand forecasting, Supply chain forecasting, Forecast diversity, Intermittent demand

1 Introduction

Demand forecasts inform production planning, inventory replenishment, resource allocation, and maintenance scheduling (Petropoulos et al. 2022). Modern production systems generate forecasts for large numbers of products and components from an expanding set of statistical and machine-learning models (Wang & Petropoulos 2016, Garred et al. 2026). Forecast combination avoids committing each series to a single model. Combining several plausible forecasts can improve accuracy and reduce exposure to model, parameter, and data uncertainty (Bates & Granger 1969, Petropoulos et al. 2018, Li et al. 2023, Wang et al. 2023), particularly when no method dominates across products, horizons, or operating conditions.

The wisdom of this forecasting crowd, however, depends on who is included. Combination performance is determined by both the aggregation rule and the composition of the forecast pool. Flexible weighting schemes can exploit differences among candidates, yet estimation error may offset their theoretical advantages (Hibon & Evgeniou 2005, Smith & Wallis 2009, Geweke & Amisano 2011, Claeskens et al. 2016). This trade-off helps explain the forecast combination puzzle: simple averaging often matches or outperforms more elaborate weighting methods (Stock & Watson 2004). With simple weights, careful pool construction becomes especially important because unstable, inaccurate, or redundant forecasts can dilute the information in the combination.

Forecast trimming, also known as forecast subset selection or combination selection, retains an informative subset before aggregation. Its premise is that a selected subset can outperform the full pool (Wang et al. 2023). This shift from estimating weights to deciding which forecasts merit inclusion is particularly relevant to production forecasting. Alongside potential accuracy gains, a smaller deployed pool may simplify forecast attribution, governance, and communication with operational decision makers.

The unresolved question is what makes a forecast worth retaining. Most trimming procedures emphasise accuracy and discard candidates with poor historical performance (Granger & Jeon 2004, Mannes et al. 2014, Kourentzes et al. 2019). Yet a useful pool should also exclude unreliable candidates and preserve complementary information (Batchelor & Dua 1995, Budescu & Chen 2015, Thomson et al. 2019, Atiya 2020, Lichtendahl Jr & Winkler 2020). Robustness, accuracy, and diversity can recommend different decisions: a forecast may be accurate but redundant, diverse but inaccurate, or stable yet add little information. Two practical questions follow: how should these criteria be combined, and when does diversity improve on an accuracy-only rule?

We address these questions through a modular framework that separates robustness screening from score-based backward selection. It compares the full pool (`None`) with robustness-only

trimming (R), accuracy-only trimming (A), alignment-only trimming (D), adaptive accuracy-diversity trimming (autoAD) and its robustness-screened variant (autoRAD). The adaptive rules use a new Accuracy-Diversity Trade-off (ADT) score that balances a candidate's individual error magnitude against its error co-movement with the remaining pool. We also introduce Relative Diversity (RelDiv) to provide a scale-normalised diagnostic of error complementarity before rule-specific trimming.

This study makes three contributions. First, the common framework applies robustness, accuracy, and error complementarity separately or jointly, enabling comparisons that hold the aggregation rule fixed. Second, the ADT score and data-driven selection of its alignment weight make the accuracy-diversity trade-off pool specific. RelDiv helps examine how its value varies with initial pool conditions. Third, the evaluation links evidence from 103,826 M-competition series to 5,000 Royal Air Force (RAF) aerospace spare-parts demand series. The RAF application extends the comparison to prediction intervals and inventory costs, relating pool size and composition to operational performance.

The results show that, on the M data, autoAD performs comparably to accuracy-only trimming overall, with significant rank improvements for monthly series and moderate-RelDiv pools. It also significantly improves on A in within-series MASE ranks at all three RAF horizons. These gains contrast with the less reliable performance of alignment-only trimming, while the frequent selection of zero diversity weights in low-RelDiv pools demonstrates the adaptive rule's ability to retain accuracy-only selection. Error complementarity is therefore most valuable as a selective addition to accuracy, not as a standalone objective. The RAF extensions relate these findings to uncertainty assessment and inventory decisions. autoAD produces interval scores close to those of None and A, whereas autoRAD achieves the lowest scores across all six configurations under each interval combination rule. With unit holding cost and a backorder penalty of 19, autoAD retains 18.2% fewer models than None at a slightly lower inventory cost, and fewer models than A at an almost identical cost. These results indicate that a smaller forecasting crowd can retain predictive and operational value, although inventory outcomes depend on the shortage penalty.

The remainder of the paper is organised as follows. Section 2 reviews the related literature, and Section 3 develops the methodological framework. Section 4 presents the M-competition experiments, and Section 5 examines spare-parts forecasting and inventory performance. Section 6 discusses managerial implications, and Section 7 concludes.

2 Related literature

2.1 Forecast-pool quality: robustness, accuracy, and diversity

Forecast-pool construction is consequential in production settings because forecast quality feeds directly into planning and control. Production research has reviewed aggregation, hierarchical forecasting, and combinations across supply-chain levels (Babai et al. 2022); compared the inventory performance of selecting and combining model and expert forecasts (Wang & Petropoulos 2016); and examined the accuracy and inventory implications of intermittent-demand combinations (Li et al. 2023, Sanguri et al. 2024). Related studies develop a machine-learning model for intermittent spare-parts demand (Jiang et al. 2021) and an automatic model-selection system for supply-chain forecasting (Garred et al. 2026). This literature establishes the operational relevance of choosing among forecasts. Within this literature, we examine how accuracy and error complementarity should jointly guide subset selection for equal-weight combination.

Three principles recur in research on forecast selection and combination: exclude unreliable candidates, remove persistently inaccurate candidates, and retain candidates that contribute distinct information. We refer to these dimensions of pool quality as robustness, accuracy, and diversity. Although they are conceptually related, each targets a different failure mode and may therefore retain a different subset.

Robustness concerns the stability of a candidate's errors rather than its average error alone. Depending on the validation design, related safeguards use outlier removal, historical-performance screens, or filters for poorly behaved forecasts. Research explaining why some combinations outperform others likewise emphasises that the properties of component forecasts matter alongside aggregate performance (Lichtendahl Jr & Winkler 2020). When robustness screening is embedded in preliminary data treatment, its contribution is difficult to distinguish from that of subsequent selection. In our single-origin design, robustness has the narrower meaning of stability in absolute error across the H validation horizons, not robustness to contamination or model misspecification.

Accuracy is the dominant criterion in forecast trimming. The rationale is direct: retaining candidates with large validation errors can degrade the combined forecast. Accuracy-based methods therefore rank candidates by recent or historical performance and remove weak members. Kourentzes et al. (2019), for example, propose an automatic pooling heuristic that orders forecasts by past performance and stops after a marked deterioration in accuracy. The same principle appears in the wisdom-of-crowds literature: a smaller group of demonstrably better judges can outperform the undifferentiated crowd (Hora 2004, Mannes et al. 2014, Goldstein et al. 2014, Budescu & Chen 2015). Accuracy-based trimming is consequently both an established approach and an essential benchmark for any more elaborate rule.

Diversity captures information that an accuracy ranking omits. Combining forecasts with aligned errors offers little protection against common mistakes, whereas forecasts with weakly aligned or offsetting errors can provide diversification benefits (Armstrong 2001). Theoretical and empirical studies relate lower error dependence to better combination performance and show that accuracy and diversity jointly determine the quality of an aggregate (Atiya 2020, Thomson et al. 2019). Diversity has informed meta-learning for forecast combination (Lemke & Gabrys 2010), and it can be extracted directly from generated forecasts without a large set of time-series features (Kang et al. 2022). Direct use of diversity for trimming is less common. Cang & Yu (2014) select subsets through mutual information, but evaluating candidate subsets becomes demanding as the pool expands; Lichtendahl Jr & Winkler (2020) consider accuracy and error correlation as separate screens rather than as terms in a single adaptive trimming criterion.

2.2 Research gap

These criteria can conflict. Accuracy-based deletion may remove a complementary forecast; alignment-only selection may retain a poor individual forecast; and an error-stability screen need not improve predictive value. The unresolved issue is how to combine these criteria and identify the pools in which an alignment adjustment is useful. We address it through a modular comparison, data-driven selection of the alignment weight, and the RelDiv diagnostic before rule-specific trimming. The analysis also distinguishes predictive gains from inventory gains, which depend on the replenishment policy and shortage penalty.

3 Forecast trimming rules

3.1 Trimming criteria: robustness, accuracy, and diversity

Trimming selects a subset before aggregation. We evaluate each candidate along three dimensions. Robustness describes the dispersion of its absolute errors across validation horizons; accuracy describes error magnitude; and diversity describes complementarity with the remaining pool.

Let D_{train}, D_{valid}, and D_{test} denote the training, validation, and test samples, respectively. Models fitted on D_{train} generate H-step forecasts for the validation sample. These validation forecasts, together with the corresponding observations, are used for trimming. The test sample is kept separate and has the same length H.

For candidate i, let $\boldsymbol{f}_i = (f_{i,1}, f_{i,2}, \dots, f_{i,H})^\top$ be its H-step validation-forecast vector, and let $\boldsymbol{F} = [\boldsymbol{f}_1^\top; \dots; \boldsymbol{f}_M^\top] \in \mathbb{R}^{M \times H}$ collect these vectors row-wise. Thus, the ith row of $\boldsymbol{F}$ is $\boldsymbol{f}_i^\top$, and M is the size of the initial pool. Let $\boldsymbol{y} = (y_1, y_2, \dots, y_H)^\top$ contain the corresponding observations,

with error $e_{i,h} = y_h - f_{i,h}$. We assume $H > 1$, $M > 1$, and $2 \leq m_{\min} \leq M$. A trimming rule returns an index set $\widehat{S} \subseteq \{1, \ldots, M\}$ for subsequent combination.

3.1.1 Robustness

The robustness criterion identifies candidates whose absolute errors vary unusually across validation horizons. For candidate i, let $s^2_{|e_i|}$ denote the sample variance of its absolute errors. Under the $H-1$ divisor used below,

$$s^2_{|e_i|} = \frac{H}{H-1}\left(\mathrm{MSE}_i - \mathrm{MAE}_i^2\right),$$

where $\mathrm{MSE}_i = \frac{1}{H}\sum_{h=1}^{H} e^2_{i,h}$ and $\mathrm{MAE}_i = \frac{1}{H}\sum_{h=1}^{H} |e_{i,h}|$. Absolute-error dispersion is therefore mechanically related to error magnitude and cannot, by itself, isolate instability.

To reduce this dependence on error magnitude, we measure robustness using the Regularised Coefficient of Variation (RCV) of absolute errors, which is defined as

$$\mathrm{RCV}_i = \frac{\left[\frac{1}{H-1}\sum_{h=1}^{H}\left(|e_{i,h}| - \overline{|e_i|}\right)^2\right]^{1/2}}{\overline{|e_i|} + \epsilon_R},$$

where $\overline{|e_i|} = \frac{1}{H}\sum_{h=1}^{H} |e_{i,h}|$ and $\epsilon_R > 0$, with the units of the observations, regularises the denominator when the mean absolute error is close to zero. The RCV expresses dispersion relative to mean absolute error and is therefore less sensitive to scale. As a population reference, if $e_{i,h} \sim N(0, \sigma_i^2)$, the coefficient of variation of $|e_{i,h}|$ equals $\sqrt{\pi/2 - 1}$, independently of σ_i^2. This motivates a relative, rather than absolute, measure of cross-horizon error instability.

We flag unusually large RCV values with Tukey's upper fence. Candidate i is removed when $\mathrm{RCV}_i > Q_3 + 1.5(Q_3 - Q_1)$, where Q_1 and Q_3 are the first and third quartiles across the initial screening pool. If this screen would violate the minimum pool size, we retain the candidates with the smallest RCV values until the minimum is met.

3.1.2 Accuracy

We measure the validation accuracy of candidate i by its MSE,

$$\mathrm{MSE}_i = \frac{1}{H}\sum_{h=1}^{H} e^2_{i,h}.$$

This loss penalises large errors more heavily than small ones. Under accuracy-based trimming, the candidate with the largest MSE is considered first for removal. This criterion ignores interactions among forecasts and therefore provides a transparent benchmark against which the value of diversity can be assessed.

3.1.3 Diversity

Combination performance depends on both individual accuracy and error co-movement. Strongly aligned errors limit diversification, even when individual forecasts perform reasonably well in isolation. We therefore define diversity from pairwise validation-error co-movement.

For candidates i and j, let $\sigma_{ij} = \frac{1}{H}\sum_{h=1}^{H} e_{i,h}e_{j,h}$ denote the uncentred cross-moment of their validation errors. Unlike a covariance, it reflects shared bias and error magnitude as well as synchronised deviations. We use it because it appears exactly in the MSE of the equal-weight combination over a subset S:

$$\mathrm{MSE}_{\mathrm{EW}} = \frac{1}{|S|^2}\left(\sum_{i\in S}\mathrm{MSE}_i + 2\sum_{i<j,\ i,j\in S}\sigma_{ij}\right).$$

Consequently, a large positive σ_{ij} makes a large positive cross-error contribution to combination MSE. A value near zero contributes little to the cross-error term, whereas a negative value reduces combination MSE through error cancellation.

This definition differs from measures of forecast disagreement such as the Mean Squared Error for Coherence (MSEC, Thomson et al. 2019),

$$\mathrm{MSEC}_{ij} = \frac{1}{H}\sum_{h=1}^{H}(f_{i,h} - f_{j,h})^2 = \mathrm{MSE}_i + \mathrm{MSE}_j - 2\sigma_{ij}.$$

MSEC combines individual error magnitudes and pairwise cross-errors. The cross-moment σ_{ij} isolates the interaction term in this decomposition, although it is not itself a pure, scale-free measure of diversity.

For a current subset S, define candidate i's average error-alignment penalty as

$$\mathrm{Align}_i(S) = \frac{1}{|S|-1}\sum_{j\in S,\, j\neq i}\sigma_{ij}.$$

Larger $\mathrm{Align}_i(S)$ values indicate a greater average cross-error contribution with the other candidates and hence less MSE-relevant complementarity. Because this score is unnormalised, it can reflect error scale as well as angular alignment; it should be interpreted as an MSE-based redundancy proxy rather than as pure diversity. Alignment-based trimming considers the candidate with the largest $\mathrm{Align}_i(S)$ first for removal.

3.1.4 Accuracy-diversity trade-off

To balance individual accuracy and error complementarity, we define the Accuracy-Diversity Trade-off (ADT) score for candidate i in a current subset S as

$$\mathrm{ADT}_i(S;\kappa) = \mathrm{MSE}_i + \kappa\ \mathrm{Align}_i(S),$$

where $\kappa \geq 0$ controls the weight assigned to the error-alignment term. Both terms have squared-error units: the first is the candidate's MSE, and the second is its average error cross-moment with the remaining pool. Thus, κ is dimensionless.

The ADT score is an alignment-adjusted accuracy criterion. At $\kappa = 0$, it is identical to the accuracy score. As κ increases, candidates with large positive cross-moments become less desirable, whereas negative cross-moments lower the score by rewarding potential error cancellation. ADT is a tunable surrogate for accuracy and redundancy, not the exact marginal contribution of a candidate to combination MSE. Validation performance selects κ separately for each forecasting task. This adapts the role of alignment without estimating combination weights.

3.2 A modular framework for forecast trimming

Starting from $S_0 = \{1, \ldots, M\}$, the framework returns a subset $\widehat{S}$ through optional RCV screening followed by score-based backward elimination. Different module and score choices retain the same underlying selection procedure.

For a current subset S, let $q_i(S; \kappa)$ denote the trimming score assigned to candidate $i \in S$. The three score choices used in this paper are

$$q_i(S;\kappa) = \begin{cases} \mathrm{MSE}_i, & \text{accuracy-based trimming,} \\ \mathrm{Align}_i(S), & \text{alignment-based trimming,} \\ \mathrm{ADT}_i(S;\kappa), & \text{accuracy-diversity trimming.} \end{cases}$$

For any score definition, let the pool-level average score be

$$B_q(S;\kappa) = \frac{1}{|S|} \sum_{i \in S} q_i(S;\kappa).$$

Backward elimination considers the candidate with the largest score, $r = \arg\max_{i \in S} q_i(S;\kappa)$, for deletion. With $S^- = S \setminus \{r\}$, all subset-dependent scores are recomputed on S^-. The deletion is accepted only when it reduces the pool-level average score by at least the relative threshold δ:

$$\frac{B_q(S;\kappa) - B_q(S^-;\kappa)}{\max\{|B_q(S;\kappa)|, \epsilon_B\}} \geq \delta,$$

where $\epsilon_B > 0$, with squared-error units, protects against a near-zero denominator. The threshold $\delta \geq 0$ stops elimination when the reduction in average trimming score is too small; for any fixed κ, larger values therefore produce more conservative trimming and weakly larger retained pools. Elimination also stops at the prespecified minimum pool size $m_{\min}$. Ties in the largest candidate score follow the fixed input order.

For an ADT rule, κ is selected rather than fixed. Each value on a finite grid K generates a backward-elimination path and a retained subset $S^{(\kappa)}$. We choose the value that minimises the validation MSE of the corresponding equal-weight combination:

$$\hat{\kappa} = \arg\min_{\kappa \in K} \frac{1}{H} \sum_{h=1}^{H} \left(y_h - \frac{1}{|S^{(\kappa)}|} \sum_{i \in S^{(\kappa)}} f_{i,h} \right)^2 .$$

Ties are resolved in favour of the smallest κ. Because zero belongs to K, autoAD reduces to A when $\hat{\kappa} = 0$, whereas autoRAD reduces to robustness-screened accuracy trimming. The rules thus adapt the role of error alignment across pools while preserving a common aggregation method.

Table 1 summarises the six configurations. R, A, and D apply robustness, accuracy, and the alignment-based diversity proxy separately. autoAD selects κ and trims with ADT; autoRAD applies the robustness screen before the same adaptive procedure. Algorithm 1 gives the complete framework.

Table 1: *Forecast trimming rules in the modular framework.*

Rule	Robustness screen	Score criterion	κ selected	Purpose
None	✗	–	✗	No-trimming benchmark
R	✓	–	✗	Robustness only
A	✗	MSE_i	✗	Accuracy only
D	✗	$\text{Align}_i(S)$	✗	Alignment only (diversity proxy)
autoAD	✗	$\text{ADT}_i(S;\kappa)$	✓	Accuracy-diversity trade-off
autoRAD	✓	$\text{ADT}_i(S;\kappa)$	✓	Accuracy-diversity trade-off with robustness screening

Algorithm 1 Modular forecast trimming framework

Require: Initial pool $S_0 = \{1, \dots, M\}$, validation forecasts $\{\boldsymbol{f}_i\}_{i \in S_0}$, observations $\{y_h\}_{h=1}^{H}$, trimming rule g, stopping threshold δ, minimum pool size $2 \le m_{\min} \le M$, numerical safeguards ϵ_R and ϵ_B, and ADT grid K.

Ensure: Retained pool $\widehat{S}$ and, when applicable, selected value $\hat{\kappa}$.

Compute validation errors $e_{i,h} = y_h - f_{i,h}$ for all $i \in S_0$.
if $g \in \{\mathsf{R}, \mathsf{autoRAD}\}$ **then**
    Apply the RCV Tukey-fence screen to S_0 and denote the retained pool by S_R.
else
    Set $S_R \leftarrow S_0$.
end if
if $g \in \{\mathsf{None}, \mathsf{R}\}$ **then**
    return $\widehat{S} = S_R$.
end if
Compute MSE_i and σ_{ij} for all $i, j \in S_R$.
Define the candidate grid K_g as

$$K_g = \begin{cases} K, & \text{if } g \in \{\mathsf{autoAD}, \mathsf{autoRAD}\}, \\ \{0\}, & \text{otherwise.} \end{cases}$$

for $\kappa \in K_g$ **do**
    Set $S \leftarrow S_R$.
    while $|S| > m_{\min}$ **do**
        Compute $q_i(S;\kappa)$ for all $i \in S$ and $B_q(S;\kappa)$ according to rule g.
        Set $r \leftarrow \arg\max_{i \in S} q_i(S;\kappa)$ and $S^- \leftarrow S \setminus \{r\}$.
        Recompute the subset-dependent scores and $B_q(S^-;\kappa)$ on S^-.
        if $\frac{B_q(S;\kappa) - B_q(S^-;\kappa)}{\max\{|B_q(S;\kappa)|, \epsilon_B\}} \ge \delta$ **then**
            Set $S \leftarrow S^-$.
        else
            break
        end if
    end while
    Record $S^{(\kappa)} \leftarrow S$.
end for
if $g \in \{\mathsf{autoAD}, \mathsf{autoRAD}\}$ **then**
    Select $\hat{\kappa}$ by the validation MSE of the equal-weight combination over $S^{(\kappa)}$.
    return $\widehat{S} = S^{(\hat{\kappa})}$.
else
    return $\widehat{S} = S^{(0)}$.
end if

3.3 Relative diversity

ADT scores individual candidates conditional on the current subset. To describe the initial pool as a whole, we define Relative Diversity (RelDiv) as the average cosine dissimilarity between nonzero validation-error vectors. Let

$$\mathcal{P}_0 = \{(i,j) : i < j,\ i, j \in S_0,\ \mathrm{MSE}_i \mathrm{MSE}_j > 0\}$$

denote the admissible pairs in the initial pool $S_0 = \{1, \ldots, M\}$. When $\mathcal{P}_0$ is nonempty, define

$$\mathrm{RelDiv}(S_0) = \frac{1}{|\mathcal{P}_0|} \sum_{(i,j)\in\mathcal{P}_0} \left(1 - \frac{\sigma_{ij}}{\sqrt{\mathrm{MSE}_i \mathrm{MSE}_j}}\right). \tag{1}$$

Here, $\sigma_{ij}/\sqrt{\mathrm{MSE}_i \mathrm{MSE}_j}$ is the cosine similarity between two error vectors. Similarity near one indicates aligned errors and little diversity; smaller, including negative, values indicate weaker or offsetting co-movement. Each pairwise cosine dissimilarity, and hence RelDiv, lies between zero and two. RelDiv is invariant to positive rescaling of an individual nonzero error vector, so it summarises angular alignment rather than error magnitude. Pairs involving a zero-norm error vector are excluded through $\mathcal{P}_0$; RelDiv is recorded as missing if $\mathcal{P}_0$ is empty.

RelDiv describes the initial pool and does not enter the selection rule. We use it in Section 4.5.2 to examine the incremental performance of ADT and the selected alignment weight.

4 Large-scale evaluation on the M competitions

4.1 Data

We first evaluate the trimming rules on the yearly, quarterly, monthly, weekly, daily, and hourly series from the M1, M3, and M4 forecasting competitions (Makridakis et al. 1982, Makridakis & Hibon 2000, Makridakis et al. 2020). The three competitions provide a demanding testbed with heterogeneous frequencies, sample lengths, seasonal patterns, and application domains. Each series contains a historical sample and a held-out test period. We split the historical sample into training and validation periods, setting the validation length equal to the competition forecast horizon.

The processed data come from the `Mcomp` (Hyndman et al. 2018) and `M4comp2018` (Montero-Manso et al. 2019) packages. We exclude a series if its training period has fewer than two observations or is constant. The resulting sample contains 103,826 series. Because M4 supplies nearly all of these series, the overall summaries pool series without reweighting competitions and are numerically dominated by M4. For yearly, quarterly, monthly, weekly, daily, and hourly data, the forecast horizons are 6, 8, 18, 13, 14, and 48, and the seasonal periods are 1, 4, 12, 52, 7, and 168, respectively.

4.2 Design

The candidate pool comprises exponential-smoothing state-space (ETS) specifications (Hyndman et al. 2002), a standard family in studies of forecast selection, trimming, and combination (Kourentzes et al. 2019, Montero-Manso et al. 2020, Wang et al. 2022, Petropoulos & Siemsen 2023). We exclude multiplicative trends and specifications that combine additive errors with

multiplicative seasonality. The resulting pool contains six specifications for nonseasonal yearly series and 15 for seasonal series. We estimate yearly, quarterly, monthly, and daily models with `ets()` from the `forecast` package (Hyndman et al. 2022), and weekly and hourly models with `es()` from `smooth` (Svetunkov 2022).

Before rule-specific trimming, every configuration receives the same prediction-interval screen following Meira et al. (2021) and Petropoulos & Siemsen (2023). At the longest horizon, Tukey's rule is applied separately to the lower and upper bounds across specifications; extreme specifications and any model whose lower bound exceeds its upper bound are removed. Consequently, None denotes the complete *post-screen* pool, and all reported trimming effects are incremental to this common preprocessing step. RelDiv is likewise calculated on the post-screen pool before any rule-specific trimming.

For each series, we estimate the eligible ETS specifications on the training period and generate validation forecasts. The trimming rule uses only these forecasts and their realised errors to select a subset. We then re-estimate the same specifications on the complete historical sample available before the test period and equally weight the test forecasts from the selected subset. Holding aggregation fixed in this way attributes performance differences to pool construction rather than to estimated combination weights.

The comparison covers the six configurations in Table 1: None, R, A, D, autoAD, and autoRAD. We set $\delta = 0.05$ and $m_{\min} = 3$. The adaptive configurations search the 21-point grid $K = \{0, 0.1, 0.2, \ldots, 2.0\}$. Only R and autoRAD apply the RCV screen.

4.3 Evaluation metrics

We evaluate point forecasts with three scale-independent measures: Mean Absolute Scaled Error (MASE, Hyndman & Koehler 2006), symmetric Mean Absolute Percentage Error (sMAPE, Makridakis & Hibon 2000), and forecast bias (Petropoulos & Siemsen 2023). Bias is the mean signed test error divided by the in-sample mean. Lower MASE and sMAPE indicate greater accuracy, whereas bias is judged by its absolute distance from zero. Scale independence permits aggregation across series.

Average errors can mask heterogeneous method performance, so we compare within-series MASE ranks using Nemenyi Multiple-Comparisons-with-the-Best (MCB) analysis at the 5% significance level (Koning et al. 2005). Lower mean ranks indicate better performance, and methods not significantly different from the best are highlighted.

4.4 Results

Figure 1 illustrates selection for M1 yearly series YAM14. R and A retain all six post-screen specifications: no candidate breaches the RCV fence, and accuracy-based deletion fails the stopping threshold. D, autoAD and autoRAD retain three specifications, although the alignment-only subset differs from the adaptive subsets. Selection thus changes both pool size and composition.

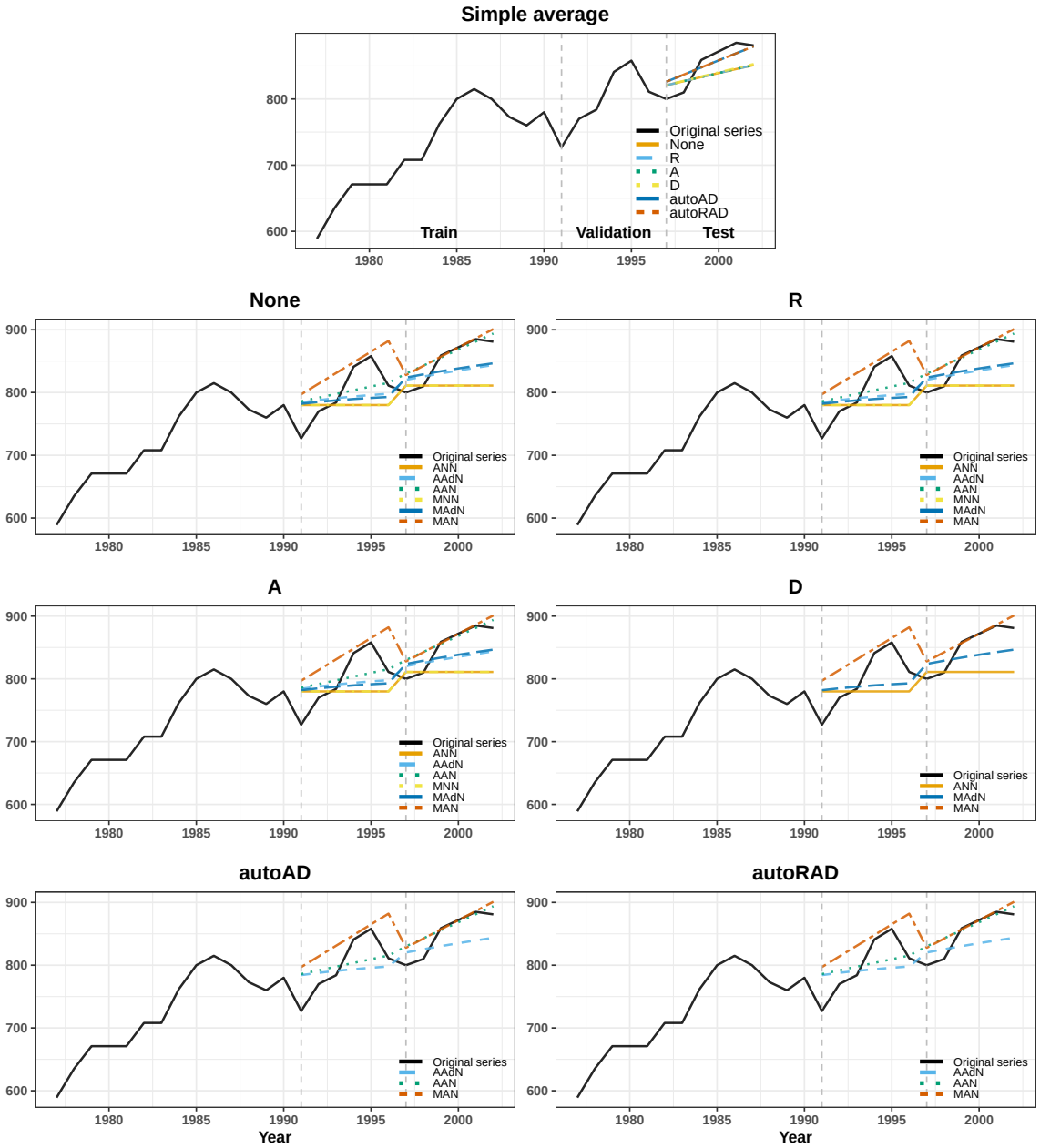


Figure 1: *Forecast subsets for M1 yearly series YAM14. The upper panel compares the equal-weight point forecasts; the lower panels show the ETS specifications retained by each configuration and the observed series. Vertical dashed lines separate the training, validation, and test periods.*

Table 2 reports mean test performance by competition. Overall, A and autoAD both reduce MASE from 1.570 to 1.513, a 3.6% improvement over None; autoRAD is nearly identical at 1.514.

A records the lowest overall sMAPE. Improvements are more apparent in M1 and M4 than in M3, while R changes average accuracy little.

The alignment-only rule D produces the bias closest to zero but the highest overall MASE. A favourable cross-error score alone therefore does not guarantee an accurate combination. The contrast between D and the two ADT configurations shows the importance of retaining an explicit individual-MSE term when error complementarity is used for selection.

Table 2: *Mean point-forecast performance by competition data set. The final block pools all usable series without reweighting competitions. Bold denotes the smallest MASE or sMAPE and the bias closest to zero.*

Data set	Measure	None	R	A	D	autoAD	autoRAD
M1	MASE	1.693	1.692	**1.600**	1.634	1.609	1.605
	sMAPE	16.157	16.166	**15.284**	15.872	15.318	15.286
	Bias	0.071	0.071	**0.058**	**0.058**	0.059	0.059
M3	MASE	1.387	1.386	**1.383**	1.411	1.385	**1.383**
	sMAPE	13.399	13.385	13.333	13.534	13.330	**13.296**
	Bias	0.014	0.014	**0.004**	**0.004**	**0.004**	**0.004**
M4	MASE	1.574	1.571	**1.516**	1.630	**1.516**	1.517
	sMAPE	12.284	12.291	**12.132**	12.238	12.142	12.146
	Bias	0.027	0.029	0.022	**0.010**	0.022	0.023
Overall	MASE	1.570	1.567	**1.513**	1.624	**1.513**	1.514
	sMAPE	12.352	12.358	**12.195**	12.308	12.205	12.207
	Bias	0.027	0.029	0.022	**0.011**	0.022	0.023

Figure 2 shows the accompanying pool reduction. R usually leaves the post-screen pool almost intact, whereas score-based rules remove a substantial share of eligible specifications.

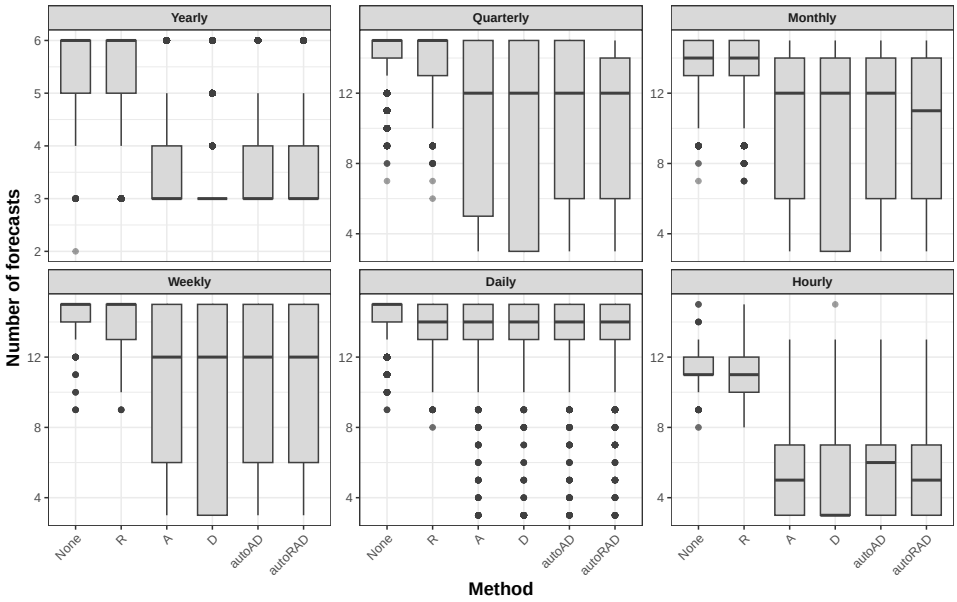


Figure 2: *Distribution of retained ETS forecasts by data frequency.* None *is the complete post-screen candidate pool; the other configurations show the additional reduction produced by each rule-specific criterion.*

Figure 3 complements the mean errors with within-series MASE ranks. Overall, autoAD has the lowest mean rank, followed closely by A and autoRAD; the difference between autoAD and A is not statistically significant. Thus, autoAD matches the strong overall performance of accuracy-based trimming rather than uniformly dominating it. None and R have the least favourable overall ranks.

The frequency panels show where the rankings separate. The clearest evidence for diversity-aware trimming occurs in monthly data: autoAD ranks first and significantly outperforms A. D also ranks ahead of A, although not significantly, suggesting that an alignment adjustment can be useful in this frequency but works most reliably when paired with an explicit accuracy term. The differences are small for weekly and daily series, where the MCB procedure cannot discriminate among the configurations.

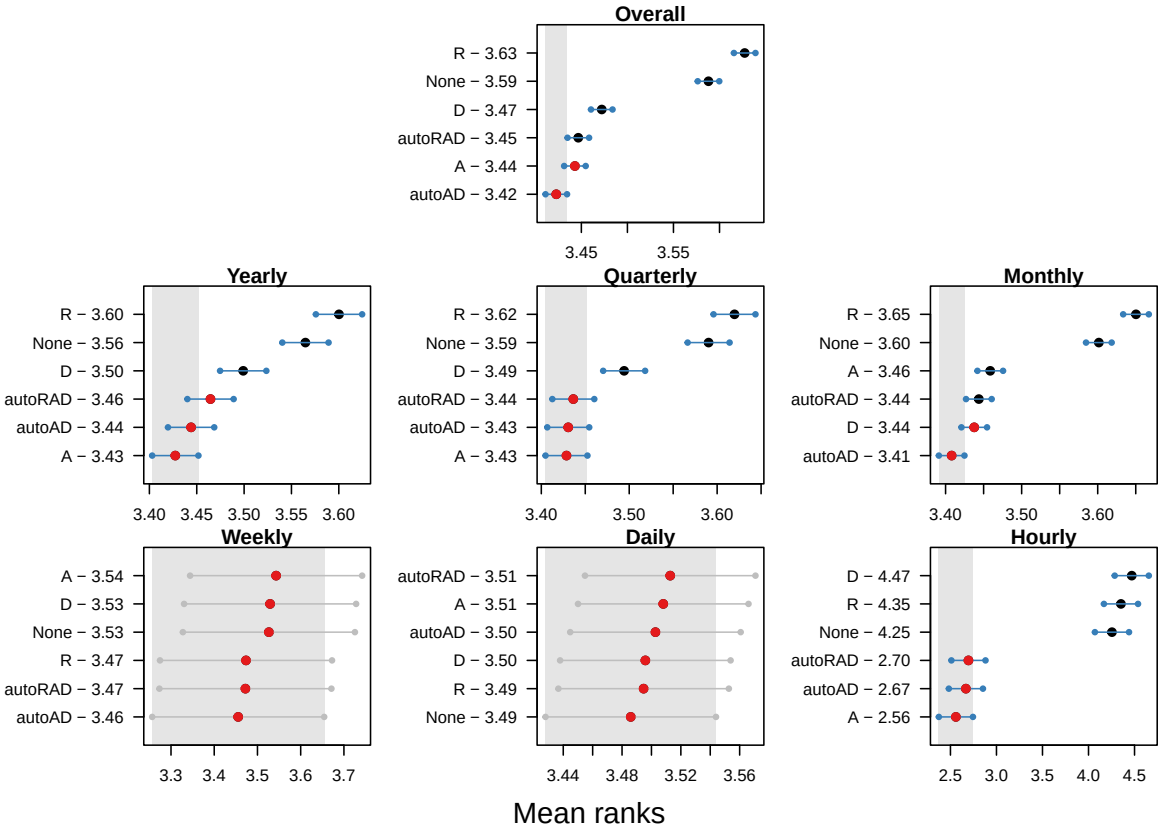


Figure 3: *Nemenyi MCB analysis of within-series MASE ranks, overall and by frequency.*

4.5 Analysis

4.5.1 The effect of the stopping threshold

The threshold δ is the minimum relative reduction in average trimming score required for another deletion. For a fixed score and alignment weight, larger values stop backward elimination no later and retain at least as many candidates; reselecting the alignment weight can alter this

relationship. Figure 4 reports MASE over $\delta \in [0, 0.10]$ overall and by frequency; None and R are invariant because they do not use score-based elimination.

Sensitivity is greatest near $\delta = 0$. Overall MASE for the accuracy-based and ADT configurations declines until approximately 0.04–0.06 and then changes little. The pattern is frequency dependent: aggressive trimming at very small δ is detrimental for the yearly and weekly series, whereas the quarterly and monthly curves have shallow interior minima. The chosen value, $\delta = 0.05$, lies in a low-error region for most panels, although no single threshold dominates at every frequency. The smaller yearly pools are especially sensitive because removing even one specification changes a relatively large share of the combination.

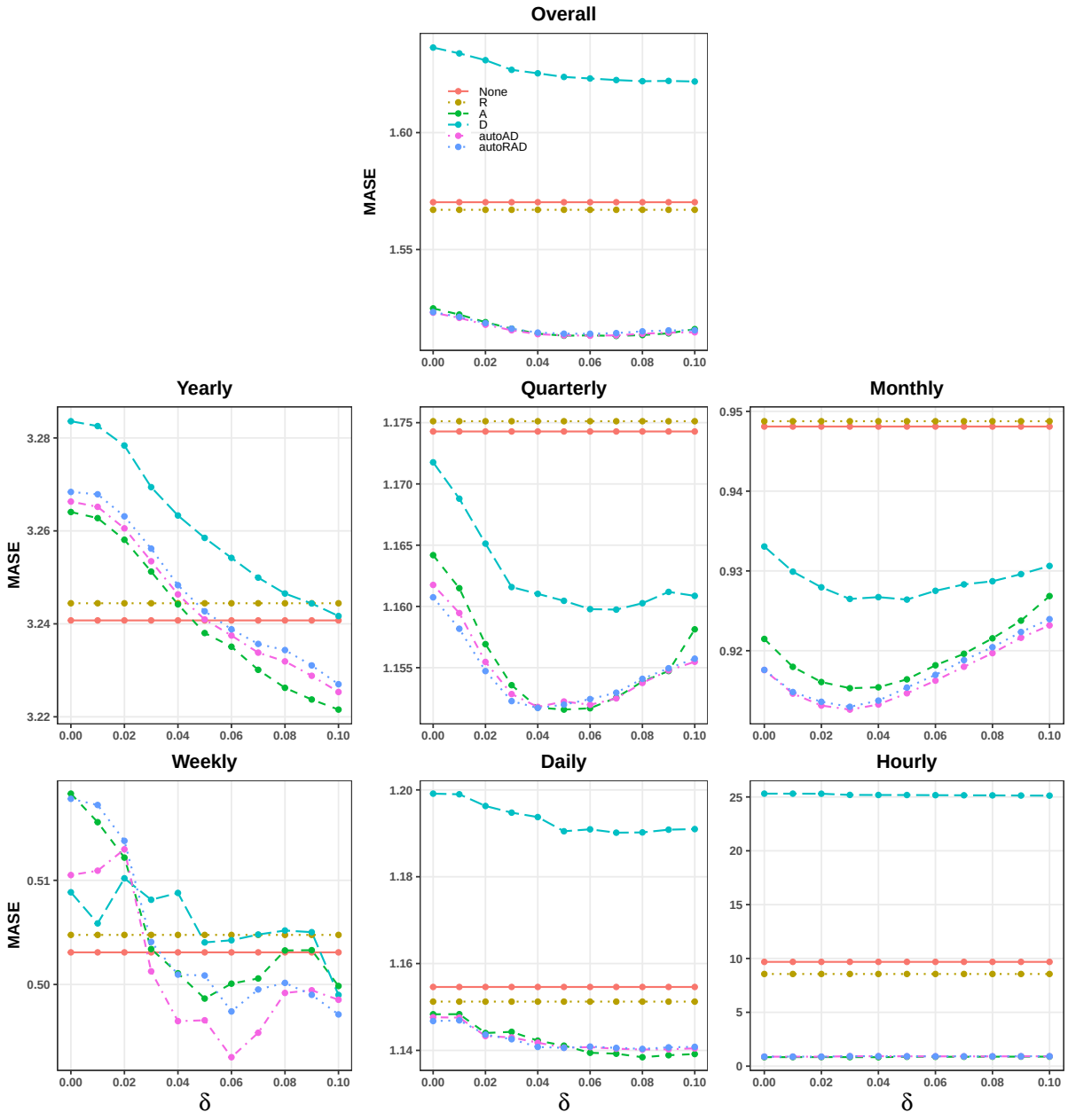


Figure 4: *Sensitivity of mean MASE to the stopping threshold δ, overall and by frequency.*

Among the score-based rules, autoAD performs best for monthly series over most of the grid and attains the lowest weekly MASE near $\delta = 0.06$. D is much more frequency dependent and is particularly uncompetitive for hourly data. These results reinforce the distinction between exploiting complementarity selectively and optimising error alignment alone.

4.5.2 When does diversity help?

We next use RelDiv as an exploratory descriptor of the conditions under which diversity-aware trimming adds value. Figure 5 plots bin-level mean performance and the selected ADT weight across 30 approximately equal-frequency RelDiv bins. Gains of autoAD over None are small in the lowest bins and positive across much of the middle of the distribution, but the pattern is not monotonic.

The comparison with A isolates the incremental contribution of error complementarity. Differences are negligible and noisy at low RelDiv, become more favourable to autoAD across much of the middle range, and reverse in the extreme upper tail. This descriptive relationship suggests that the clearest opportunity for diversity-aware trimming occurs at moderate, not necessarily maximal, RelDiv.

The selected κ describes the selection response. In the very lowest bins, autoAD selects $\kappa = 0$ for nearly every series. Positive weights remain uncommon, although not absent, across the broader low-RelDiv range; their share then increases and exceeds one half in several higher bins. This pattern is consistent with pool-specific adaptation: the procedure can revert to A when the validation data do not favour an alignment adjustment.

The fixed groups in Figure 6 provide a more interpretable comparison. In the low-RelDiv group, D, A, and autoAD have similar ranks, offering little evidence of an incremental benefit from ADT. In the moderate group, autoAD has the best mean rank and significantly outperforms A. In the high group, autoAD remains best ranked, but its differences from A and autoRAD are not significant; D performs poorly despite the high pool-level diversity.

RelDiv is therefore associated with the incremental value of error complementarity, but the relationship is neither monotonic nor prescriptive. It provides a descriptive indicator of potential complementarity, particularly in the moderate range, rather than a validated threshold for choosing a trimming rule.

5 Application to the RAF data

5.1 Data and design

We next evaluate the trimming rules on the Royal Air Force (RAF) aerospace spare-parts data, available in the `fable.intermittent` package (Damato et al. 2026). The data comprise 5,000

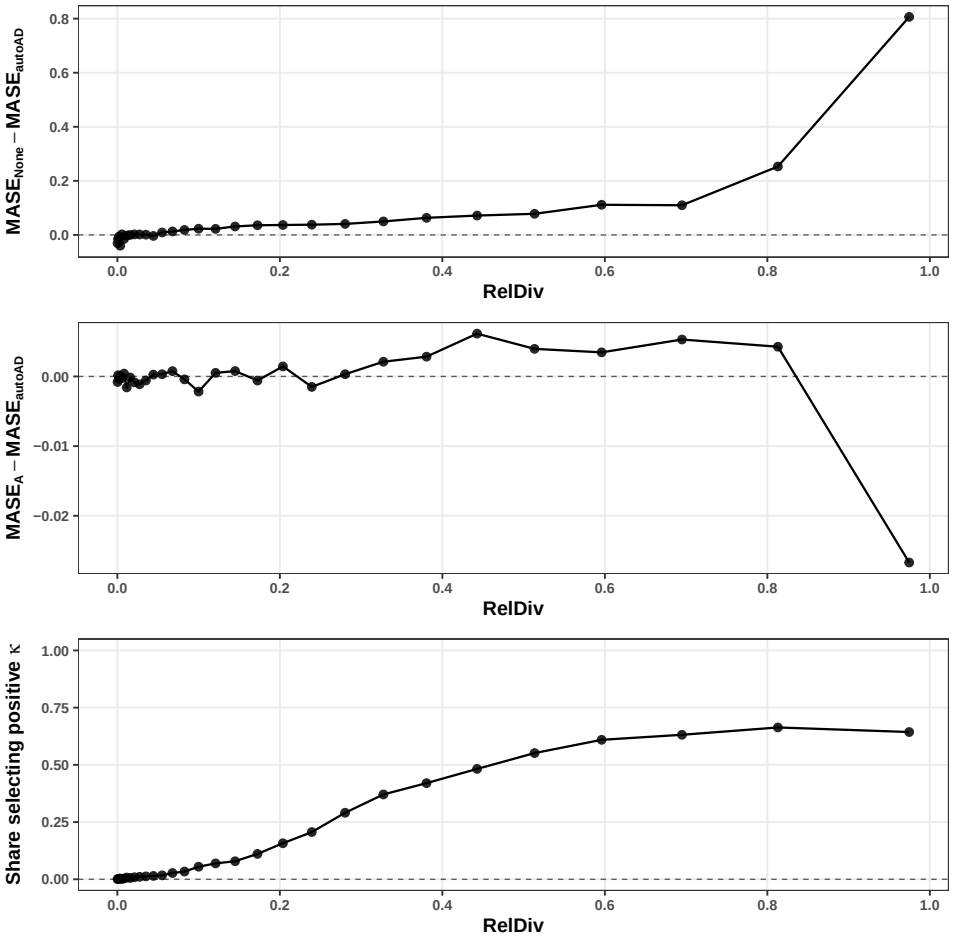


Figure 5: *RelDiv, autoAD performance, and the selected diversity weight across approximately equal-frequency bins. Points are bin means. The upper and middle panels report MASE differences relative to None and A; positive values favour autoAD. The lower panel reports the share selecting* $\kappa > 0$.

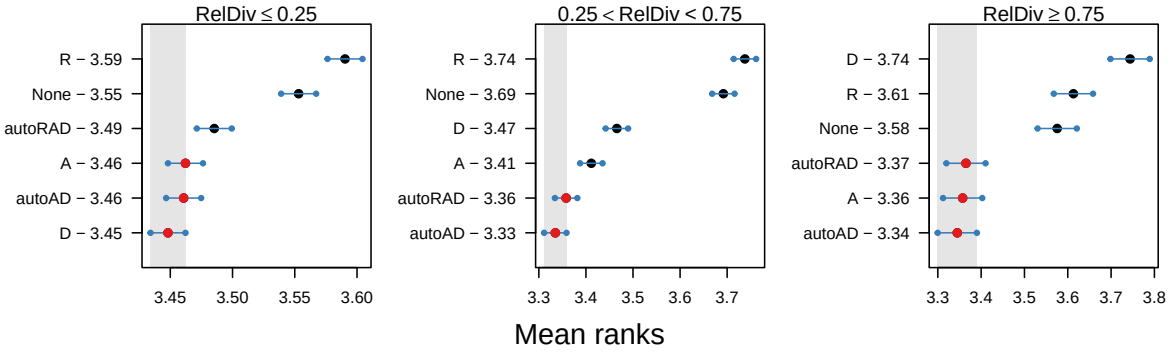


Figure 6: *Nemenyi MCB analysis of within-series MASE ranks for low (RelDiv* ≤ 0.25*), moderate (*$0.25 < \text{RelDiv} < 0.75$*), and high (RelDiv* ≥ 0.75*) groups.*

monthly demand series with 84 observations each and are established in the inventory and intermittent-demand literature (Petropoulos & Kourentzes 2015, Kourentzes & Athanasopoulos 2021, Li et al. 2023). Spare-parts forecasting is a persistent production challenge because intermittency complicates both model choice and inventory control (Jiang et al. 2021, Sanguri

et al. 2024). This application connects the benchmark analysis to an environment in which forecasts support spares provisioning and maintenance readiness.

Following Li et al. (2023), the pool contains 12 methods. Six are conventional: Naive, seasonal Naive (sNaive), Simple Exponential Smoothing (SES), Moving Average (MA), ARIMA, and ETS. Six target intermittent demand: Croston's method (CRO), optimised Croston (optCro), the Syntetos–Boylan Approximation (SBA, Syntetos & Boylan 2005), the Teunter–Syntetos–Babai method (TSB, Teunter et al. 2011), the Aggregate–Disaggregate Intermittent-Demand Approach (ADIDA, Nikolopoulos et al. 2011), and the Intermittent Multiple-Aggregation Prediction Algorithm (IMAPA, Petropoulos & Kourentzes 2015).

We remove leading zeros so that each series starts with its first nonzero demand. Forecast horizons are 3, 6, and 12 months. For each $H \in \{3, 6, 12\}$, the final H observations form the test period and the preceding H form the validation period. Each method is estimated before the validation period to generate validation forecasts, then re-estimated through the end of validation to generate test forecasts. Negative base forecasts are truncated at zero, consistent with the nonnegative demand domain. We retain the trimming parameters from Section 4.2.

5.2 Point forecast performance

Table 3 reports mean MASE for the individual methods and trimming configurations. At every horizon, at least one trimming rule lowers mean MASE relative to None. autoRAD is best at $H = 3$ and $H = 6$, lowering mean MASE by 11.1% and 6.8%, while D is best at $H = 12$ with a 3.4% reduction and is followed closely by autoAD. At $H = 3$, autoRAD (0.918) also has lower mean MASE than the best individual method, ADIDA (0.963). At $H = 6$ and $H = 12$, however, the best trimmed combinations (0.917 and 0.921) do not surpass the Naive method (0.852 and 0.849). The horizon-level means therefore support improvement over the full combination, but not guaranteed dominance over the strongest individual method.

The RAF pools have relatively low RelDiv. Across horizons, median RelDiv ranges from approximately zero to 0.1, and 90.0–97.6% of pools satisfy RelDiv ≤ 0.25. Correspondingly, autoAD selects $\kappa > 0$ for 22.3%, 20.7%, and 13.5% of series at $H = 3$, 6, and 12. The rule therefore behaves primarily like accuracy-based trimming in this application while retaining the option to penalise error alignment for a minority of series.

The within-series ranks in Figure 7 add an important qualification. autoAD significantly outperforms A at all three horizons, although its gains in mean MASE are modest, approximately 1.3–1.6%. These rank improvements are achieved with smaller forecasting crowds: autoAD retains an average of 5.7, 7.2 and 9.8 of the 12 candidate methods at $H = 3$, 6 and 12, respectively. Adding the RCV screen yields no systematic point-rank advantage despite autoRAD's

Table 3: *Mean MASE for individual methods and trimming configurations on the RAF data. Bold denotes the best trimming configuration at each horizon.*

Method	$H = 3$	$H = 6$	$H = 12$
Naive	0.997	0.852	0.849
sNaive	1.109	1.032	0.971
SES	1.053	1.013	0.986
MA	0.966	0.963	0.986
ARIMA	1.032	0.991	0.947
ETS	1.019	0.975	0.949
CRO	1.085	1.043	0.997
optCro	1.083	1.042	0.996
SBA	1.080	1.038	0.991
TSB	1.042	1.002	0.965
ADIDA	0.963	0.928	0.909
IMAPA	1.008	0.967	0.924
None	1.032	0.984	0.953
R	1.022	0.981	0.967
A	0.941	0.940	0.937
D	0.937	0.926	**0.921**
autoAD	0.929	0.925	0.925
autoRAD	**0.918**	**0.917**	0.939

lower mean MASE at the two shorter horizons. Its contribution to interval and inventory performance is examined separately below.

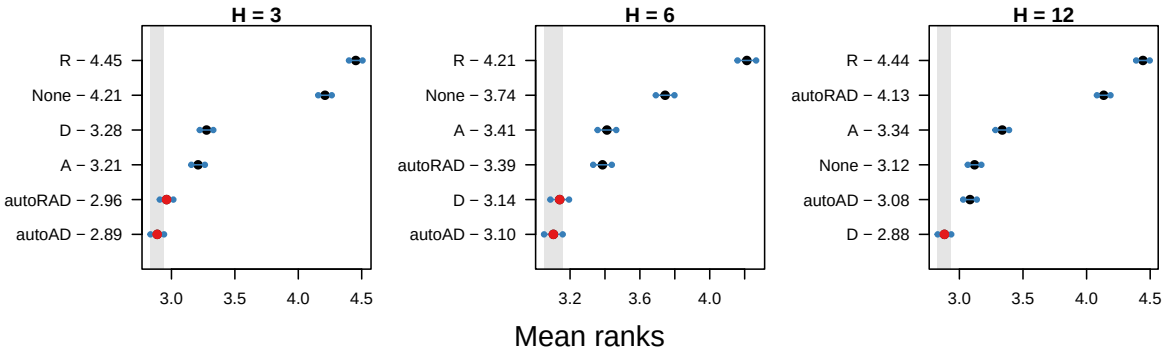


Figure 7: *Nemenyi MCB analysis of within-series MASE ranks at RAF horizons of 3, 6, and 12 months.*

5.3 Prediction interval performance

To examine whether a crowd selected for point accuracy also provides useful uncertainty estimates, we assess its 95% prediction intervals without further selection or tuning. Following the implementation accompanying Li et al. (2023), we use native intervals for Naive, sNaive, SES, ARIMA and ETS, rescale aggregate-model intervals for ADIDA, and add kernel-smoothed fitted-residual quantiles to the remaining point forecasts. Bounds are truncated at zero. Lower and upper endpoints are combined separately using their simple average, median or interior-trimmed mean, as suggested by Grushka-Cockayne & Jose (2020). The last rule removes the largest $\lfloor 0.2|\widehat{S}| \rfloor$ lower bounds and the same number of smallest upper bounds before averaging, yielding intervals at least as wide as simple endpoint averaging.

Table 4 reports coverage and the Mean Scaled Interval Score (MSIS, Gneiting & Raftery 2007) at $H = 12$, which penalises interval width and noncoverage. Lower MSIS is better, while coverage is assessed by proximity to 95%. autoRAD has the lowest MSIS under all three combination rules, whereas autoAD has scores close to those of None and A. Among the single-criterion rules, R has the lowest MSIS under all three combination rules. Coverage differs more across combination rules than across trimming configurations: simple averaging undercovers, while median and interior trimming approach 95%. Lower interval scores do not imply better calibration in every case, and the interval ranking differs from the point-forecast ranking at this horizon. Additional interval results for $H = 3$ and $H = 6$ are omitted for brevity and are available from the authors upon request.

Table 4: *MSIS and coverage (%) of 95% prediction intervals on RAF at $H = 12$. Bold denotes the lowest MSIS within each combination rule.*

	Simple average		Median		Interior trimmed	
Method	MSIS	Coverage	MSIS	Coverage	MSIS	Coverage
None	14.798	83.42	13.931	96.11	14.108	94.69
R	14.659	84.09	13.910	95.92	13.997	94.53
A	14.707	84.21	13.917	96.20	14.110	94.76
D	14.752	83.75	13.913	96.22	14.114	94.74
autoAD	14.701	84.26	13.915	96.21	14.116	94.78
autoRAD	**14.531**	84.65	**13.883**	95.95	**13.971**	94.44

5.4 Inventory performance

We evaluate the selected forecast pools under a monthly order-up-to policy adapted from Wang & Petropoulos (2016). Forecasts are fixed at the test origin, orders arrive one month after placement, and unmet demand is fully backordered. Order quantities are continuous and nonnegative; capacity constraints and fixed ordering costs are excluded. For item j under configuration g, net stock $x_{jg,t}$ and the order $o_{jg,t}$ placed after demand in month t satisfy

$$\begin{aligned} x_{jg,t} &= x_{jg,t-1} + o_{jg,t-1} - y_{j,t}, \\ o_{jg,t} &= \max\{0, f_{jg,t+1|0} + s_{jg,t} - x_{jg,t}\}. \end{aligned} \tag{2}$$

Here, $f_{jg,t+1|0}$ denotes the fixed-origin forecast, and negative $x_{jg,t}$ represents backorders. The safety allowance is updated sequentially as

$$s_{jg,t} = \Phi^{-1}(\tau)\,\mathrm{sd}\left(e^{\mathrm{valid}}_{jg,1:12}, e^{\mathrm{test}}_{jg,1:t}\right),$$

where Φ is the standard normal distribution function and sd is the sample standard deviation of errors observed by month t. Thus, the selected subsets and forecasts remain fixed throughout the test period, while the safety allowance adapts as new forecast errors become available. We set the unit holding cost to $c_h = 1$, consider backorder costs $c_b \in \{9, 19, 99\}$, and set $\tau = c_b/(c_h + c_b)$. This common safety-stock rule is used to compare the operational consequences of the competing

forecast pools rather than to optimise the replenishment policy itself; it does not guarantee the service level implied by τ.

All configurations start from the same inventory state based on the full-pool combination (None), with $o_{j,0} = \bar{y}_j^{\text{valid}}$ and $x_{j,0} = s_{j,0}^{\text{None}} + \bar{f}_j^{\text{None,valid}} - \bar{y}_j^{\text{valid}}$. This common initialisation holds starting conditions fixed across configurations, although initial stock can affect subsequent costs. Orders are placed for $t = 1, \ldots, 11$, with no terminal salvage value, and performance is evaluated over the 12-month test horizon.

For each item, inventory cost is measured as the average monthly holding and backorder cost over the test period. To facilitate aggregation across items with different demand scales, the cost is divided by the item's demand standard deviation over the available history and test period and then averaged across items. Test demand enters this retrospective scale only, not decisions. Service is the proportion of item-months ending without backorders, not the demand-weighted fill rate.

Table 5: *Pool size and inventory performance on RAF. Relative cost changes are computed against the mean standardised cost of None, with negative values indicating savings. Bold denotes the smallest mean pool size, lowest cost change, or highest no-backorder rate.*

	Mean	Relative cost change (%)			No-backorder rate (%)		
Method	pool size	$c_b = 9$	$c_b = 19$	$c_b = 99$	$c_b = 9$	$c_b = 19$	$c_b = 99$
None	12.00	0.000	0.000	0.000	**94.120**	**94.642**	95.527
R	10.72	-0.195	-0.098	**-0.037**	94.108	94.637	**95.537**
A	10.26	-0.368	-0.072	+0.374	94.072	94.597	95.473
D	9.81	-0.245	-0.028	+0.268	94.085	94.597	95.475
autoAD	9.81	-0.444	-0.072	+0.511	94.065	94.588	95.467
autoRAD	**9.16**	**-0.588**	**-0.175**	+0.379	94.077	94.600	95.503

Table 5 reports the resulting pool sizes, inventory costs, and service levels. Both adaptive methods substantially reduce the forecast pool while maintaining similar service performance. In particular, autoRAD retains 23.7% fewer models than None and achieves the lowest inventory cost when $c_b = 9$ and $c_b = 19$, with reductions of 0.588% and 0.175%, respectively. autoAD also lowers costs in these two settings while retaining 18.2% fewer models. The no-backorder rate at $c_b = 19$ varies by less than 0.06 percentage points across all configurations, suggesting that these cost differences are not accompanied by material changes in observed service.

The gains are nevertheless sensitive to the relative cost of backorders. When $c_b = 99$, R yields the lowest cost, whereas both adaptive methods are more costly than None. At $c_b = 19$, autoAD has a 0.0006% higher mean standardised cost than A. A paired percentile bootstrap with 5,000 item-level resamples gives a marginal 95% confidence interval of $[-0.023\%, 0.026\%]$ for this relative difference. The interval assumes independent items and does not establish cost equivalence.

The direction of the comparison between autoAD and None persists under the evaluated alternatives. Excluding the common first month yields a 0.079% cost reduction for autoAD relative to None at $c_b = 19$, while scaling costs by the pretest demand standard deviation gives a 0.031% reduction over the full test period. Across the evaluated cost ratios, these alternative specifications preserve the direction of the comparison with None.

Overall, the inventory experiment shows that substantial forecast-pool reduction need not translate into poorer operational performance. The adaptive trimming methods retain service levels close to those of the full pool and can reduce inventory costs under moderate backorder penalties, although the gains are modest and depend on the cost setting.

6 Managerial implications and practical guidance

The empirical results provide a clear starting point for point forecast combinations. When the primary objective is point accuracy, A remains a transparent and competitive benchmark, while autoAD provides a natural adaptive extension when validation information is sufficiently rich. Because the candidate grid includes $\kappa = 0$, autoAD can revert to accuracy-only trimming when additional error alignment is not supported by validation performance. In practice, this makes accuracy the baseline criterion while allowing forecast complementarity to contribute when it is empirically useful. RelDiv can be used as a diagnostic indicator of when such complementarity may matter, but it should not be interpreted as a mechanical selection threshold. The less stable performance of D across datasets, horizons, and diversity levels further suggests that error alignment is better used as a complement to accuracy than as a standalone trimming objective. The RCV screen should likewise be viewed as an application-specific safeguard rather than a universally beneficial preprocessing step.

When prediction intervals are also required, the RAF results show that the adaptive trimming methods remain competitive despite being developed primarily for point forecasting. In particular, autoRAD achieves the lowest MSIS across the interval-combination rules considered, while the median and interior-trimmed mean yield coverage closer to the nominal level than simple averaging. These results support evaluating interval quality separately: autoAD largely preserves the benchmark scores, whereas autoRAD improves them at $H = 12$.

For downstream inventory decisions, the relevant operational loss should be specified before the final configuration is chosen. The adaptive methods substantially reduce the forecast pool while maintaining no-backorder rates close to those of the full pool. autoRAD also achieves the lowest inventory costs at $c_b = 9$ and $c_b = 19$, although the ranking changes when the backorder penalty becomes large. Forecast-pool trimming can therefore preserve, and in some settings improve, downstream operational performance, but the preferred configuration should

be evaluated against the shortage penalty, service requirement, and loss function relevant to the application.

7 Conclusions and discussion

This paper asks when a forecasting crowd benefits from being selective. We propose a modular trimming framework that combines robustness screening with backward selection based on forecast accuracy, error alignment, or both. The ADT criterion allows the role of alignment to adapt to each forecast pool, while RelDiv provides a descriptive measure of the pool's initial error diversity. The framework therefore offers an interpretable way to construct a smaller forecast pool rather than leaving pool composition as an implicit outcome of forecast combination.

The empirical results support accuracy as the primary criterion and error alignment as a useful complement when supported by validation evidence. Across the M-competition data, autoAD and A perform similarly in mean MASE, while autoAD achieves the lowest mean rank and shows clearer gains for monthly series and pools with moderate diversity. The gains are not monotonic in RelDiv, suggesting that diversity is better used as a diagnostic of potential complementarity than as a mechanical rule for selecting a trimming method. The RAF results reinforce this conclusion in an intermittent-demand setting, where autoAD consistently improves on A in point-forecast rankings across horizons.

The interval and inventory experiments assess whether smaller pools retain value beyond point forecasting. At $H = 12$, both autoAD and autoRAD remain competitive for interval prediction, with autoRAD achieving the lowest MSIS across the interval-combination rules considered. In the inventory experiment, both adaptive methods substantially reduce the pool while maintaining service levels close to those of the full pool and reducing costs under moderate backorder penalties, with autoRAD achieving the largest savings. These gains do not persist uniformly when the backorder penalty becomes large, indicating that the preferred configuration can depend on the downstream objective. The results support smaller forecasting crowds with comparable observed performance in several settings, but no configuration dominates across all evaluation criteria.

This study has several limitations that also suggest directions for future research. First, the proposed framework focuses on subset construction under equal-weight combinations, leaving open how adaptive trimming interacts with estimated combination weights and other aggregation schemes. Second, the analysis evaluates forecast-pool quality primarily through predictive performance, while operational applications may benefit from selecting pools directly against downstream decision losses and service objectives. Third, the empirical evidence is drawn from benchmark forecasting data and one intermittent-demand application, so broader validation

across industries, forecast architectures, and probabilistic settings would help establish the generality of the approach. Future research could therefore extend adaptive accuracy-diversity trimming to weighted and probabilistic forecast combinations and integrate pool selection more directly with decision-oriented objectives.

Acknowledgments

The authors gratefully acknowledge financial support from the National Natural Science Foundation of China under numbers 72601399, 72542010, 72571014, 72671008 and 72495123. Xiaoqian Wang acknowledges the Presidential Foundation of the Academy of Mathematics and Systems Science, Chinese Academy of Sciences, China (No. E555930101). Feng Li acknowledges the Fundamental Research Funds for the Central Universities, Peking University.

Disclosure statement

The authors report there are no competing interests to declare.

Declaration of generative AI use

During the preparation of this work, the authors used ChatGPT (GPT-5.6, OpenAI) to assist with language editing, including improving spelling, grammar, and clarity. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Data availability statement

The M1 and M3 data are publicly available through version 2.8 of the `Mcomp` R package (Hyndman et al. 2018), the M4 data through version 0.2.0 of `M4comp2018` (Montero-Manso et al. 2019), and the RAF data through version 0.2.0 of `fable.intermittent` (Damato et al. 2026).

References

Armstrong, J. S. (2001), Combining forecasts, *in* 'Principles of Forecasting: A Handbook for Researchers and Practitioners', Springer, pp. 417–439.

Atiya, A. F. (2020), 'Why does forecast combination work so well?', *International Journal of Forecasting* **36**(1), 197–200.

Babai, M. Z., Boylan, J. E. & Rostami-Tabar, B. (2022), 'Demand forecasting in supply chains: a review of aggregation and hierarchical approaches', *International Journal of Production Research* **60**(1), 324–348.

Batchelor, R. & Dua, P. (1995), 'Forecaster diversity and the benefits of combining forecasts', *Management Science* **41**(1), 68–75.

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

Budescu, D. V. & Chen, E. (2015), 'Identifying expertise to extract the wisdom of crowds', *Management Science* **61**(2), 267–280.

Cang, S. & Yu, H. (2014), 'A combination selection algorithm on forecasting', *European Journal of Operational Research* **234**(1), 127–139.

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

Damato, S., Zambon, L. & Azzimonti, D. (2026), *fable.intermittent: Forecasting Models for Intermittent Time Series*. R package version 0.2.0.
URL: *https://CRAN.R-project.org/package=fable.intermittent*

Garred, W., Oger, R. & Lauras, M. (2026), 'Automatic demand forecast model selection in supply chains: a forecast value-added analysis of selection strategies, machine learning, and hyperparameter optimisation', *International Journal of Production Research* pp. 1–20.

Geweke, J. & Amisano, G. (2011), 'Optimal prediction pools', *Journal of Econometrics* **164**(1), 130–141.

Gneiting, T. & Raftery, A. E. (2007), 'Strictly proper scoring rules, prediction, and estimation', *Journal of the American Statistical Association* **102**(477), 359–378.

Goldstein, D. G., McAfee, R. P. & Suri, S. (2014), The wisdom of smaller, smarter crowds, *in* 'Proceedings of the fifteenth ACM conference on Economics and computation', New York, NY, USA, pp. 471–488.

Granger, C. W. J. & Jeon, Y. (2004), 'Thick modeling', *Economic Modelling* **21**(2), 323–343.

Grushka-Cockayne, Y. & Jose, V. R. R. (2020), 'Combining prediction intervals in the M4 competition', *International Journal of Forecasting* **36**(1), 178–185.

Hibon, M. & Evgeniou, T. (2005), 'To combine or not to combine: selecting among forecasts and their combinations', *International Journal of Forecasting* **21**(1), 15–24.

Hora, S. C. (2004), 'Probability judgments for continuous quantities: Linear combinations and calibration', *Management Science* **50**(5), 597–604.

Hyndman, R. J., Akram, M., Bergmeir, C. & O'Hara-Wild, M. (2018), *Mcomp: Data from the M-Competitions*. R package version 2.8.

URL: *https://CRAN.R-project.org/package=Mcomp*

Hyndman, R. J., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., Petropoulos, F., Razbash, S., Wang, E. & Yasmeen, F. (2022), *forecast: Forecasting functions for time series and linear models*. R package version 8.16.

URL: *https://pkg.robjhyndman.com/forecast/*

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

Hyndman, R. J., Koehler, A. B., Snyder, R. D. & Grose, S. (2002), 'A state space framework for automatic forecasting using exponential smoothing methods', *International Journal of Forecasting* **18**(3), 439–454.

Jiang, P., Huang, Y. & Liu, X. (2021), 'Intermittent demand forecasting for spare parts in the heavy-duty vehicle industry: a support vector machine model', *International Journal of Production Research* **59**(24), 7423–7440.

Kang, Y., Cao, W., Petropoulos, F. & Li, F. (2022), 'Forecast with forecasts: Diversity matters', *European Journal of Operational Research* **301**(1), 180–190.

Koning, A. J., Franses, P. H., Hibon, M. & Stekler, H. O. (2005), 'The M3 competition: Statistical tests of the results', *International Journal of Forecasting* **21**(3), 397–409.

Kourentzes, N. & Athanasopoulos, G. (2021), 'Elucidate structure in intermittent demand series', *European Journal of Operational Research* **288**(1), 141–152.

Kourentzes, N., Barrow, D. & Petropoulos, F. (2019), 'Another look at forecast selection and combination: Evidence from forecast pooling', *International Journal of Production Economics* **209**, 226–235.

Lemke, C. & Gabrys, B. (2010), 'Meta-learning for time series forecasting and forecast combination', *Neurocomputing* **73**(10), 2006–2016.

Li, L., Kang, Y., Petropoulos, F. & Li, F. (2023), 'Feature-based intermittent demand forecast combinations: accuracy and inventory implications', *International Journal of Production Research* **61**(22), 7557–7572.

Lichtendahl Jr, K. C. & Winkler, R. L. (2020), 'Why do some combinations perform better than others?', *International Journal of Forecasting* **36**(1), 142–149.

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

Makridakis, S. & Hibon, M. (2000), 'The M3-Competition: results, conclusions and implications', *International Journal of Forecasting* **16**(4), 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**(1), 54–74.

Mannes, A. E., Soll, J. B. & Larrick, R. P. (2014), 'The wisdom of select crowds', *Journal of Personality and Social Psychology* **107**(2), 276–299.

Meira, E., Oliveira, F. L. C. & Jeon, J. (2021), 'Treating and pruning: New approaches to forecasting model selection and combination using prediction intervals', *International Journal of Forecasting* **37**(2), 547–568.

Montero-Manso, P., Athanasopoulos, G., Hyndman, R. J. & Talagala, T. S. (2020), 'FFORMA: Feature-based forecast model averaging', *International Journal of Forecasting* **36**(1), 86–92.

Montero-Manso, P., Netto, C. & Talagala, T. S. (2019), *M4comp2018: Data from the M4-Competitions*. R package version 0.2.0.
URL: *https://github.com/carlanetto/M4comp2018*

Nikolopoulos, K., Syntetos, A. A., Boylan, J. E., Petropoulos, F. & Assimakopoulos, V. (2011), 'An aggregate–disaggregate intermittent demand approach (ADIDA) to forecasting: an empirical proposition and analysis', *Journal of the Operational Research Society* **62**(3), 544–554.

Petropoulos, F., Hyndman, R. J. & Bergmeir, C. (2018), 'Exploring the sources of uncertainty: Why does bagging for time series forecasting work?', *European Journal of Operational Research* **268**(2), 545–554.

Petropoulos, F. & Kourentzes, N. (2015), 'Forecast combinations for intermittent demand', *Journal of the Operational Research Society* **66**(6), 914–924.

Petropoulos, F. & Siemsen, E. (2023), 'Forecast selection and representativeness', *Management Science* **69**(5), 2672–2690.

Petropoulos, F. et al. (2022), 'Forecasting: theory and practice', *International Journal of Forecasting* **38**(3), 705–871.

Sanguri, K., Patra, S., Nikolopoulos, K. & Punia, S. (2024), 'Intermittent demand, inventory obsolescence, and temporal aggregation forecasts', *International Journal of Production Research* **62**(5), 1663–1685.

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

Stock, J. H. & Watson, M. W. (2004), 'Combination forecasts of output growth in a seven-country data set', *Journal of Forecasting* **23**(6), 405–430.

Svetunkov, I. (2022), *smooth: Forecasting Using State Space Models*. R package version 3.1.5.
URL: *https://CRAN.R-project.org/package=smooth*

Syntetos, A. A. & Boylan, J. E. (2005), 'The accuracy of intermittent demand estimates', *International Journal of Forecasting* **21**(2), 303–314.

Teunter, R. H., Syntetos, A. A. & Babai, M. Z. (2011), 'Intermittent demand: Linking forecasting to inventory obsolescence', *European Journal of Operational Research* **214**(3), 606–615.

Thomson, M. E., Pollock, A. C., Önkal, D. & Gönül, M. S. (2019), 'Combining forecasts: Performance and coherence', *International Journal of Forecasting* **35**(2), 474–484.

Wang, X., Hyndman, R. J., Li, F. & Kang, Y. (2023), 'Forecast combinations: An over 50-year review', *International Journal of Forecasting* **39**(4), 1518–1547.

Wang, X., Kang, Y., Petropoulos, F. & Li, F. (2022), 'The uncertainty estimation of feature-based forecast combinations', *Journal of the Operational Research Society* **73**(5), 979–993.

Wang, X. & Petropoulos, F. (2016), 'To select or to combine? the inventory performance of model and expert forecasts', *International Journal of Production Research* **54**(17), 5271–5282.