SUPPLY CHAIN RESEARCH · EDITORIAL FEATURE

Forecast Accuracy Theatre

Why Everyone Reports MAPE and Nobody Improves

Forecast accuracy is measured with a metric that can be gamed, at an aggregation that flatters it, over the items that suit it. Meanwhile a large share of forecasting effort makes forecasts worse than doing nothing. The number improves and the business does not, because the metric was never the point.

Independent, vendor-neutral research for supply chain and technology leaders
Supply Chain Research | supplychainresearch.com | 2026

Contents

Why Everyone Reports MAPE and Nobody Improves

01 Executive summary

Nearly every supply chain organization measures forecast accuracy, reports it monthly, sets targets against it, and celebrates when the number improves. Nearly none of them can show that the improvement changed a decision or saved a dollar. This is forecast accuracy theatre: the ritual of measuring, reporting, and defending an accuracy percentage that has been detached from the business outcomes it is supposed to serve. The metric usually reported, mean absolute percentage error, is mathematically ill-suited to the demand it measures, gameable through the choice of aggregation level, and frequently improved by biasing forecasts in ways that have nothing to do with accuracy. Meanwhile, study after study finds that a large share of the forecasting process, including the expensive human judgment layered on top of statistical models, makes forecasts worse than a naive baseline that anyone could compute for free.

This article argues that the problem is not a lack of forecasting effort or sophistication but a misdirection of it. Organizations chase an accuracy metric as though the metric were the goal, when the goal is a better ordering, staffing, or production decision, and the two come apart routinely. A forecast that is more accurate by some percentage but changes no decision has created no value, and a forecasting process that cannot beat carrying last period forward is subtracting value while consuming salaries. We say honestly that measurement is necessary and that the critics can overstate the case. But the evidence, drawn from the academic literature on forecast metrics and the practitioner literature on forecast value added, is that most organizations are measuring the wrong thing, at the wrong level, and rewarding the wrong behavior, and that the remedy is not a better accuracy number but a different question: does each step of the forecasting process beat a naive baseline, and does the forecast change a decision for the better.

52%	**~half**	**4**
of forecasts found worse than a random walk in a study of eight supply chain companies	of judgmental planner overrides that made the forecast worse, upward ones most destructive	the specifications a forecasting accuracy claim needs, and almost never has: metric, baseline, level, population

The argument in brief

- **MAPE is the wrong ruler.** It is undefined at zero, explodes for low-volume items, and is asymmetric, so it can be improved by biasing forecasts down. The metric distorts exactly the intermittent demand that is hardest to forecast.
- **Aggregation determines the answer.** Errors cancel when aggregated, so a national total looks accurate while the item-location forecast that runs the warehouse is poor. The reported number flatters the process by hiding where decisions are made.
- **Much of the process subtracts value.** Studies find a large share of forecasts worse than naive, and about half of judgmental overrides make forecasts worse. The human effort layered on models frequently destroys value rather than adding it.

- **Accuracy is not value.** A more accurate forecast that changes no decision creates nothing. The percentage is a means; the inventory, service, and cost outcome is the end, and the two come apart constantly.
- **Reward the decision, not the metric.** Measure Forecast Value Added against a naive baseline, use metrics suited to the demand, measure at the decision level, and tie forecasting to outcomes rather than an abstract accuracy number.

02 The theatre, stated plainly

Forecast accuracy theatre is the performance of measurement in place of the substance of improvement. It has a recognizable choreography. Each month, a forecasting or demand-planning team computes an accuracy figure, usually a percentage derived from mean absolute percentage error, and reports it up the organization. The figure is compared to a target and to prior months. If it improved, the team is credited; if it declined, the team explains why, usually citing demand volatility or data issues. Targets are set for the coming period, initiatives are launched to improve the number, and software is bought on the promise of improving it further. The entire apparatus revolves around the accuracy percentage, which is treated as the measure of forecasting success and, implicitly, as the goal.

What is almost never present in this choreography is a link between the accuracy percentage and any business outcome. The team that improved its accuracy from one figure to another rarely demonstrates that the improvement reduced inventory, raised service levels, cut expediting costs, or changed any decision at all. The number improved; whether anything else did is not examined, because the number itself has become the object of the exercise. This is the defining feature of theatre: the performance of an activity, measured and rewarded on its own terms, disconnected from the outcome it was meant to produce. The forecasting organization is busy, measured, and accountable, and it may be creating no value whatsoever, because value was never what it was measuring.

The theatre persists for understandable reasons. An accuracy percentage is easy to compute, easy to report, and easy to compare, which makes it an attractive management metric regardless of whether it means anything. It gives the forecasting function something concrete to be accountable for and to improve, and it gives management something to track. And it feels rigorous: a number that goes up or down, with targets and trends, has the appearance of serious measurement. The appearance is the problem. The metric is rigorous in form and hollow in substance, because it measures a property of the forecast, its closeness to the actual, without measuring whether that closeness matters to any decision, and closeness that does not change a decision is worth nothing.

This article's purpose is to replace the theatre with something substantive. The substitution is not more sophisticated forecasting or a better accuracy metric, both of which leave the fundamental disconnect intact. It is a different question. Instead of asking how accurate the forecast is, the organization should ask two things: does each step of the forecasting process improve on a naive baseline that costs nothing, and does the resulting forecast change a decision for the better. These questions, developed through the rest of this article, connect forecasting to value in a way that the accuracy percentage does not, and they frequently reveal that a forecasting process consuming considerable resources is adding little or nothing, which is uncomfortable but is the beginning of actually improving rather than performing improvement.

The theatre has a characteristic emotional texture that helps explain its persistence. Producing an accuracy number and watching it improve is satisfying in a way that confronting the disconnect from value is not. The number is concrete, the trend is legible, and an improvement feels like progress, whereas the question of whether the improvement changed a decision is abstract, uncomfortable, and

frequently answered in the negative. Organizations gravitate toward the satisfying, legible activity and away from the uncomfortable, abstract one, which is a general feature of human institutions and not a failing peculiar to forecasting. The accuracy metric endures partly because it feels good to improve it, and the value question is avoided partly because it frequently feels bad to answer it.

There is also a self-protective function to the theatre that is worth naming. A forecasting function measured on an accuracy percentage can demonstrate its diligence and its improvement through that percentage, defending its budget and its headcount by pointing to a rising number, regardless of whether the number corresponds to value. A forecasting function measured on value would have to demonstrate that its work changed decisions and improved outcomes, which is harder to show and sometimes impossible to show favorably, because the honest answer is sometimes that much of the work adds little. The accuracy metric is therefore a safer basis for the function's self-justification than the value question, and the function has an institutional interest in being measured on accuracy rather than value, which is one more reason the theatre is sticky. Changing the measure threatens the function's ability to justify itself on comfortable terms, and institutions resist measures that expose them.

03 Why MAPE is the wrong ruler

The metric at the center of the theatre, mean absolute percentage error, is not merely imperfect; it is mathematically ill-suited to the demand most supply chains actually forecast, and its defects are well documented in the forecasting literature. Understanding them is the first step to seeing why the reported accuracy number frequently means less than it appears to. Figure 1 illustrates the core pathologies.

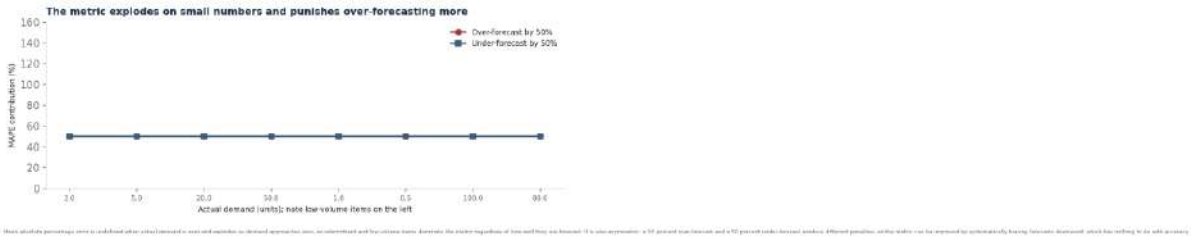

Figure 1. MAPE is undefined at zero and explodes as demand approaches zero, so intermittent items dominate it; and it is asymmetric, penalizing over-forecasting more, so it can be gamed by biasing forecasts down.

Mean absolute percentage error expresses the forecast error as a percentage of the actual demand, then averages those percentages across items and periods. The definition contains its first defect: when actual demand is zero, the percentage error is undefined, because it requires dividing by zero, and when actual demand is small, the percentage error is enormous even for a small absolute miss, because a small denominator inflates the ratio. For a product that sells one unit in a period and was forecast at three, the percentage error is two hundred percent; for a product forecast at two that sold zero, the error is undefined. Since most supply chains carry a long tail of intermittent, low-volume items alongside their high-volume ones, MAPE is dominated by exactly the items it handles worst, and the reported average is distorted by the low-volume tail in ways that have little to do with forecasting skill.

The second defect is asymmetry. MAPE penalizes over-forecasting and under-forecasting differently, because the percentage error is bounded below by zero when the forecast is too low but unbounded when the forecast is too high relative to a small actual. This asymmetry means that the metric can be improved, made to show a lower average percentage error, by systematically biasing forecasts downward, because low forecasts incur smaller percentage penalties when they miss. A forecaster optimizing for MAPE is therefore incentivized to under-forecast, which produces a better-looking accuracy number and worse inventory outcomes, since under-forecasting causes stockouts. The metric rewards a behavior that harms the business, which is the precise opposite of what a good metric should do, and a forecasting organization managed on MAPE may be quietly biasing its forecasts to hit the number at the cost of service.

These are not obscure technical quibbles; they are the reason the reported accuracy number is frequently uninformative or misleading. An organization reporting a MAPE figure is reporting an average dominated by its intermittent tail, computed with a metric that rewards downward bias, and treating that figure as a measure of forecasting quality. The forecasting literature has proposed better metrics, examined later in this article, that scale the error by the performance of a naive baseline and behave sensibly for intermittent demand. But the point here is diagnostic: the near-universal reliance on MAPE

means that much of the accuracy theatre is built on a metric whose own properties undermine its meaning, and a reader confronted with a MAPE figure should ask what it actually measures before crediting any improvement in it.

The pathologies of MAPE have a common root worth articulating, which is that it expresses error as a ratio to the actual value, and ratios behave badly when the denominator is small or zero. This is not a quirk specific to forecasting; it is a general property of percentage-based measures, and it appears wherever a percentage is computed against a small base. In forecasting it is especially damaging because supply chains are full of small bases: the intermittent items, the new products, the seasonal goods out of season, the long tail of slow movers. A metric built on ratios to the actual is a metric that will be dominated by, and distorted by, exactly these small-base items, and since they are numerous in most portfolios, the distortion is not a marginal effect but a central one.

A subtle consequence of the ratio construction is that MAPE is not comparable across items with different volumes in the way its users assume. A ten percent MAPE on a high-volume item and a ten percent MAPE on a low-volume item represent substantially different absolute errors and substantially different business consequences, but the metric presents them as equivalent, and averaging them treats them as equivalent. An organization that manages by MAPE is therefore implicitly weighting its attention by a quantity, the percentage error, that does not correspond to business impact, spending as much concern on a ten percent error on a trivial item as on a ten percent error on a critical one. Volume-weighted metrics correct this by weighting errors by the volume that gives them their business significance, which is why they give a truer picture, but plain MAPE, by construction, misallocates attention across the portfolio in proportion to a number that does not track what matters.

04 The aggregation game

Even setting aside the choice of metric, the reported accuracy number is determined as much by the level at which it is measured as by the quality of the forecast, and this creates a second avenue for the number to flatter the process while concealing what matters. The mechanism is aggregation, and its effect is shown in Figure 2.

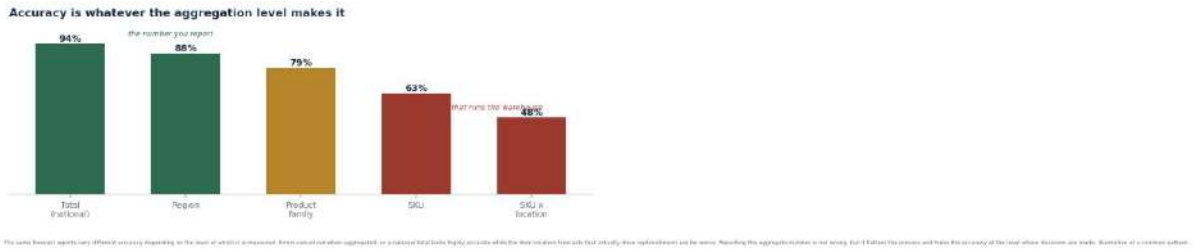

Figure 2. The same forecast reports substantially different accuracy at different aggregation levels. Errors cancel upward, so the national total looks accurate while the item-location forecast that drives replenishment is poor.

Forecast errors partially cancel when they are aggregated. If a forecast is too high for one store and too low for another, the errors offset when the two are summed, so the forecast for the combined total is more accurate than either individual forecast. Aggregate far enough, to a regional or national total, and the offsetting errors produce an impressively accurate number, because the individual errors that would show up at the item-location level have washed out against each other. The national total forecast for a product might be accurate to within a few percent while the store-level forecasts that actually determine what gets shipped to each location are wrong by large margins in both directions, their errors canceling in the aggregate.

This matters because decisions are made at the disaggregated level, not the aggregate. A warehouse does not replenish a national total; it replenishes a specific item at a specific location, and the accuracy that matters for that decision is the accuracy of the item-location forecast, which is the worst level, not the aggregate, which is the best. An organization that reports its forecast accuracy at an aggregate level is reporting the level at which the forecast looks best and is least relevant to decisions, while the level that actually drives the ordering and stocking decisions, and where the forecast is worst, goes unreported. The reported number is not false, but it is measured where it flatters and is disconnected from where it would inform, which is a form of the theatre.

The aggregation game can be played across several dimensions, not only geography. Aggregating across products, from an individual item to a product family, cancels errors the same way; aggregating across time, from a weekly to a monthly or quarterly forecast, cancels errors the same way; and aggregating across customers cancels errors the same way. In each case, the more aggregated the measurement, the better the accuracy looks and the further it is from any specific decision. An organization truly interested in whether its forecasts help decisions would measure accuracy at the level and horizon at which decisions are actually made, which is usually granular and short-horizon, and would accept the worse-looking numbers that result. An organization performing accuracy theatre measures at whatever level produces the best number, which is usually aggregate and long-horizon, and reports that. The

choice of aggregation level is therefore a tell: measuring where the number looks good rather than where decisions are made is the signature of the theatre.

The aggregation game interacts with organizational structure in a way that entrenches it. Forecasting is frequently organized so that the accuracy reported to senior management is an aggregate figure, because senior management thinks in aggregate terms, about total demand, regional performance, and category trends, and an aggregate accuracy number matches that frame. But the aggregate number that satisfies senior management is precisely the number that conceals the item-location accuracy that determines operational outcomes, so the reporting structure that suits management attention is the structure that hides the accuracy that matters. The mismatch is not anyone's fault, but its effect is that the number rising to the top of the organization is systematically the flattering aggregate, and the granular truth stays buried in the operational layer where it is felt but not reported upward.

There is a defensible use of aggregate forecasting that should be acknowledged so the critique is not overstated. Some decisions truly are made at an aggregate level, capacity planning, procurement of long-lead materials, financial planning, and for those decisions the aggregate forecast is the relevant one and its higher accuracy is materially useful. The error is not measuring aggregate accuracy, which is appropriate for aggregate decisions, but reporting aggregate accuracy as though it characterized the forecasting overall, including the disaggregated forecasts that drive operational decisions. An organization should measure accuracy at each level for the decisions made at that level, reporting aggregate accuracy for aggregate decisions and item-location accuracy for item-location decisions, rather than reporting the flattering aggregate as a summary of forecasting quality and allowing it to stand in for the granular accuracy it does not represent.

05 The metric menagerie, and choosing one well

If MAPE is the wrong ruler, what is the right one? The forecasting literature offers a menagerie of alternatives, each with its own assumptions and its own suitable domain, and the honest answer is that no single metric is correct for all situations. The pathology is not that MAPE exists but that one metric is applied everywhere, including where it misleads, and chosen for its flattery rather than its fit. Figure 3 surveys the main alternatives.

Figure 3. There is no single correct accuracy metric; each suits some demand patterns and distorts others. The error is applying one metric everywhere and choosing it because it flatters the current process.

Weighted MAPE, sometimes called WAPE, addresses part of MAPE's problem by weighting the percentage errors by volume, so that the low-volume tail no longer dominates. Because it effectively divides total absolute error by total actual demand, it does not explode for small items and gives a more representative picture of overall accuracy. It is more robust than plain MAPE for the mixed-volume portfolios most supply chains carry, and it is a reasonable default where a percentage-style metric is wanted.

The mean absolute scaled error, MASE, takes a more principled approach that connects directly to this article's central theme. It scales the forecast error by the error of a naive baseline forecast, so a MASE below one means the forecast beats the naive baseline and a MASE above one means it does worse. This is exactly the right question, because it measures not the raw accuracy but whether the forecasting effort improves on doing nothing, and it behaves sensibly for intermittent demand where MAPE breaks down. The root mean squared scaled error, RMSSE, used in a well-known forecasting competition, applies the same scaling idea while penalizing large misses more heavily, which suits situations where big errors are especially costly. And a bias metric, measuring the direction rather than the magnitude of error, is essential alongside any spread metric, because it catches the systematic over- or under-forecasting that a magnitude metric can miss and that MAPE actively encourages.

The practical guidance that follows from the menagerie is not to adopt one alternative metric universally, which would repeat the original error in a new form, but to choose the metric that fits the demand being forecast and the decision being made. For stable, high-volume items, plain MAPE is adequate and its defects rarely bite. For intermittent demand, MASE or RMSSE are far more appropriate because they do not explode at low volumes and they measure improvement over naive. Where volume weighting matters, WMAPE gives a truer overall picture. And bias should be measured always, alongside whatever spread metric is chosen, because systematic bias is both common and directly harmful to inventory. The discipline is to select the metric deliberately, for its fit, and to be suspicious of any metric

chosen because it makes the current process look good, since choosing the metric for its flattery is the beginning of the theatre.

A practical note on transitioning metrics is worth adding, because organizations that recognize MAPE's limitations frequently stumble on the change. Switching the reported metric is disruptive: targets set in MAPE terms do not translate directly into MASE or WMAPE terms, historical comparisons break, and people accustomed to interpreting one metric must learn to interpret another. This friction leads some organizations to keep MAPE despite understanding its flaws, simply because changing is costly, which is a version of the sunk-cost reasoning that keeps many bad practices in place. The friction is real, but it is a one-time transition cost against a permanent improvement in the meaningfulness of the metric, and an organization that lets transition friction lock it into a misleading metric is paying a recurring cost to avoid a one-time one, which is the wrong trade.

The recommended path through the transition is to run the old and new metrics in parallel for a period, so the organization can build intuition for the new metric while retaining the familiar one for continuity, and can see how the two relate on its own data. During the parallel period, the organization learns what a given MASE or WMAPE value means in its context, recalibrates its targets to the new metric, and builds the interpretive fluency that the old metric had accumulated over years. Once the new metric is understood and its targets are set, the old metric can be retired, having served as a bridge. This measured transition avoids both the disruption of an abrupt switch and the permanent cost of never switching, and it reflects the general principle that the difficulty of changing a measure is a transition cost to be managed, not a reason to retain a measure known to mislead.

06 Forecast Value Added: the right question

The single most clarifying idea in the practitioner literature, and the antidote to accuracy theatre, is Forecast Value Added analysis. It reframes the question from how accurate the forecast is to whether each step of the forecasting process improves on doing nothing, and that reframing exposes how much of the typical forecasting process adds no value or subtracts it. Figure 4 illustrates the analysis.

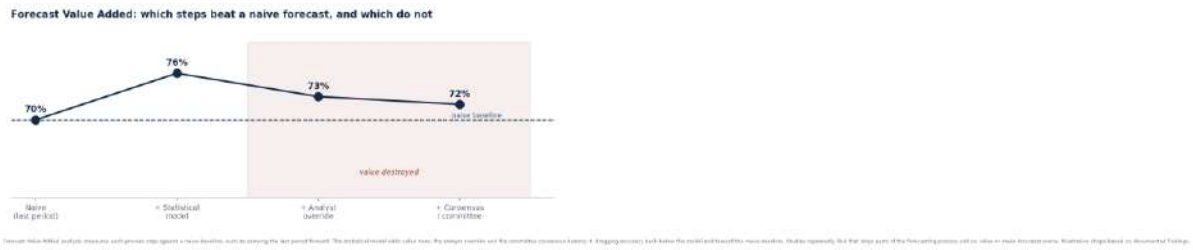

Figure 4. Forecast Value Added measures each process step against a naive baseline. The statistical model adds value; the analyst override and committee consensus destroy it, dragging accuracy back toward naive.

The core idea, associated with Michael Gilliland and developed at a forecasting-software company, is disarmingly simple. Take a naive forecast, the cheapest possible baseline, such as carrying the last period's actual forward or using a seasonal average. Then measure whether each subsequent step of the forecasting process, the statistical model, the planner's manual adjustment, the consensus forecast agreed in a demand-planning meeting, improves accuracy relative to that baseline. The value added by each step is the improvement it produces over the step before, and over the naive baseline overall. Gilliland likens the naive forecast to the placebo in a drug trial: a treatment that does not beat the placebo is not working, however elaborate it is, and a forecasting step that does not beat naive is not adding value, however much effort it consumes.

The results of applying this analysis are frequently sobering and are the empirical heart of the case against accuracy theatre. In many organizations, the statistical model adds value over the naive baseline, as it should. But the manual adjustments that planners make to the model's output, and the consensus process that adjusts it further in meetings, frequently subtract value, producing a final forecast that is worse than the model alone and sometimes worse than the naive baseline. The elaborate, expensive, human-intensive part of the forecasting process, the part that consumes the most salaries and the most meeting time, is precisely the part that most often destroys value, and the organization would produce better forecasts, at lower cost, by using the statistical model's output directly or even the naive baseline, without the human overlay that it believes is adding sophistication.

This is a profound and uncomfortable finding, because it inverts the intuition that more effort and more judgment produce better forecasts. Forecast Value Added analysis repeatedly shows that much of the judgment layered onto forecasts makes them worse, and that a large fraction of the forecasting process could be eliminated with no loss, or a gain, in accuracy, at considerable savings in cost. An organization that performs this analysis rigorously frequently discovers that its forecasting process is a value-destroying apparatus wrapped around a value-adding statistical core, and that the theatre of accuracy measurement has concealed this by focusing on the level of the final accuracy number rather than on whether each step earned its place. The question Forecast Value Added asks, does this step beat doing

nothing, is the question accuracy theatre never asks, and it is the question that actually connects forecasting effort to value.

The placebo analogy that Gilliland uses repays a moment of reflection, because it captures precisely why the naive baseline is the right comparison. In a drug trial, the question is not whether patients improved after taking the drug, since they might have improved anyway, but whether they improved more than patients taking a placebo. The placebo controls for everything that would have happened without the drug, isolating the drug's actual contribution. The naive forecast plays the same role: it captures what accuracy an organization would achieve without any real forecasting effort, so that comparing the actual forecast to the naive baseline isolates the contribution of the forecasting effort itself. A forecast that beats naive has added something; a forecast that does not has added nothing, however much effort it consumed, just as a drug that does not beat placebo has done nothing however expensive it was.

This framing exposes why absolute accuracy metrics are so misleading: they report the equivalent of whether patients improved, without the placebo comparison that would reveal whether the treatment caused the improvement. An organization reporting an absolute accuracy figure is reporting that its patients improved, which sounds like success, without checking whether they would have improved anyway under the naive baseline. Frequently they would have, and the forecasting effort added nothing or subtracted value, but the absolute metric cannot reveal this because it lacks the control that the naive comparison provides. Forecast Value Added is, at its core, the insistence on running the placebo comparison, and the reason it so often produces uncomfortable results is that many forecasting processes, subjected to the comparison for the first time, turn out not to beat their placebo.

07 Worse than doing nothing

The finding that much of the forecasting process subtracts value is not a single study or a theoretical claim; it is a repeated empirical result, and its most striking form deserves to be stated directly. Figure 5 presents the headline numbers.

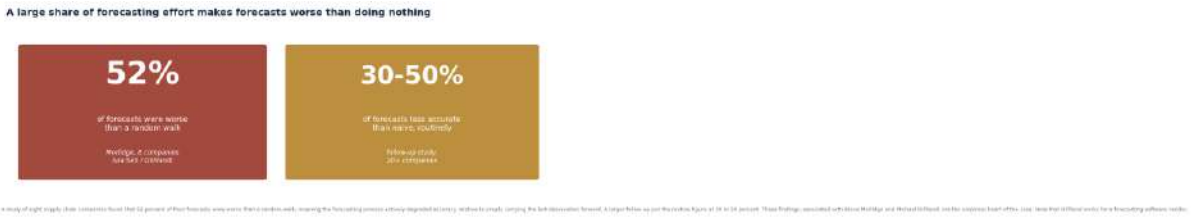

Figure 5. A study of eight supply chain companies found 52 percent of forecasts worse than a random walk; a larger follow-up put the routine figure at 30 to 50 percent. Note the source works for a forecasting-software vendor.

In research presented at forecasting conferences and published through a forecasting-software company, Steve Morlidge examined the forecasts of eight supply chain companies and found that fifty-two percent of them were worse than a naive forecast, specifically worse than a random walk that simply carries the last observation forward. More than half of the forecasts these companies produced, with all their models, planners, systems, and meetings, were less accurate than the cheapest possible baseline that requires no forecasting at all. A larger follow-up study of more than twenty companies found that thirty to fifty percent of forecasts were routinely worse than naive. These are not isolated failures; they describe a systematic condition in which a large fraction of professional forecasting effort produces results inferior to doing nothing.

The implication is stark. A forecast worse than naive is not merely a weak forecast; it is a value-destroying one, because the organization would have been better off, in accuracy terms, using the free naive baseline instead of the forecast it spent money to produce. When half of an organization's forecasts fall into this category, half of its forecasting effort is not merely wasted but counterproductive, actively degrading the accuracy it was meant to improve. And because the organization measures accuracy in absolute terms rather than relative to naive, it does not see this: a forecast worse than naive can still show a respectable-looking MAPE, and the organization reports that MAPE with satisfaction, unaware that a free alternative would have done better. The absolute accuracy metric conceals the value destruction that a comparison to naive would reveal.

Fairness requires noting that the researcher most associated with these findings works for a company that sells forecasting software, which has an interest in demonstrating that current forecasting is broken and improvable. This does not invalidate the findings, which are methodologically sound and have been presented in peer-reviewed forecasting venues, but it is a source characteristic a careful reader should weigh, and it is noted here in keeping with this publication's practice of flagging interested parties. The underlying result, that a substantial share of forecasts underperform a naive baseline, is corroborated across multiple studies and is consistent with the Forecast Value Added findings from independent practitioners, so the conclusion does not rest on a single interested source. But the specific figures should be read as coming from someone with a commercial interest in the problem they describe, and

the appropriate response is to run the analysis on one's own forecasts rather than to accept the headline number on faith.

It is worth dwelling on why a forecast worse than naive is so common, because the phenomenon seems paradoxical: how can a sophisticated process with models, data, and expert planners produce results inferior to carrying last period forward? Part of the answer is that each layer of the process introduces opportunities for error as well as improvement, and when a layer introduces more error than it removes, it makes things worse. A statistical model fit to noisy history can overfit and project noise forward; a planner adjustment can inject bias; a consensus process can average toward a politically comfortable number rather than an accurate one. Each layer is intended to improve the forecast, and each can degrade it, and when the degradations outweigh the improvements, the elaborate forecast underperforms the naive baseline that has no layers to introduce error. The sophistication that is supposed to add accuracy is also a source of error, and the net effect is frequently negative.

08 The override that subtracts value

If a large share of forecasts is worse than naive, where does the value destruction come from? The academic literature points to a specific and well-studied culprit: the judgmental adjustments that human planners make to statistical forecasts. Figure 6 shows the pattern.

Figure 6. Across tens of thousands of forecasts, roughly half of judgmental overrides made forecasts worse, and upward adjustments were systematically the most destructive, while larger downward corrections more often helped.

A body of research examining tens of thousands of forecasts across supply chain companies has studied what happens when planners override the output of a statistical forecasting model with their own judgment. The finding is consistent and striking: roughly half of these judgmental overrides make the forecast worse, not better. The planners believe they are adding information the model lacks, knowledge of a promotion, a market shift, a customer conversation, and sometimes they are. But as often as not, they are adding noise, bias, or wishful thinking, and the adjusted forecast is less accurate than the model's output would have been. The human overlay that organizations believe improves their forecasts degrades them about half the time.

The research refines this in a way that is directly actionable. Not all overrides are equally harmful: the direction and size of the adjustment matter. Upward adjustments, where the planner raises the forecast above the model's output, are systematically the most destructive, because they are frequently driven by optimism, sales pressure, or a desire to avoid stockouts, and they push the forecast above what demand will actually be, creating excess inventory. Downward adjustments, and especially larger downward corrections, are more likely to improve the forecast, because they often reflect genuine knowledge that the model is over-forecasting. The asymmetry is informative: the overrides most likely to help are the cautious downward ones, and the overrides most likely to hurt are the optimistic upward ones, which are unfortunately also the ones organizational pressure most encourages.

This points to a specific reform that Forecast Value Added analysis enables. If overrides subtract value about half the time, and upward overrides subtract value most reliably, then an organization can improve its forecasts by constraining overrides, requiring them to be justified, tracking their value added, and being especially skeptical of upward adjustments. Some organizations that have measured the value added by their planners' overrides have concluded that many planners should override less, or not at all, and that the model's output should stand unless there is documented, specific reason to change it. This is a difficult message organizationally, because it tells experienced planners that their judgment frequently makes things worse, but it is what the evidence supports, and an organization serious about forecast quality rather than forecast theatre will measure the value its overrides add and

act on the finding, even when the finding is that less human intervention would produce better forecasts.

The finding that overrides frequently subtract value collides with a deeply held professional identity, which is part of why organizations resist it. Demand planners understand their role as adding judgment and market knowledge to the mechanical output of a model, and the proposition that this judgment frequently makes forecasts worse is not merely a technical finding but a challenge to the value of their work and their expertise. The resistance is understandable and human, and it means that acting on Forecast Value Added findings about overrides requires more than presenting the data; it requires managing the professional and emotional response of people being told that a core part of their job frequently subtracts value. Organizations that have navigated this successfully have done so by reframing the planner's role rather than eliminating it, directing planners to the overrides that truly add value and relieving them of the ones that do not.

The reframing is important because the finding is not that human judgment is worthless but that it is valuable in specific circumstances and harmful in others, and the skill is knowing which is which. Judgmental overrides add value when the planner has genuine, specific information the model cannot have, a known future promotion, a confirmed large order, a discontinuation, and the override incorporates that specific information. They subtract value when they are driven by generic optimism, by discomfort with the model's output, by a desire to be seen adding value, or by organizational pressure to forecast in a particular direction. The productive response to the override findings is therefore to channel judgment toward the circumstances where it helps, requiring overrides to cite specific information and tracking their value added, so that planners learn which of their interventions improve forecasts and which do not. This preserves the genuine value of human judgment while curbing the reflexive adjustment that the evidence shows is frequently harmful, and it treats planners as professionals capable of learning from feedback rather than as a source of noise to be eliminated.

09 Accuracy is not value

Underlying every specific critique in this article is a single conceptual error that sustains the theatre: the conflation of forecast accuracy with forecast value. These are different things, and the gap between them is where the theatre lives. Figure 7 states the distinction.

Figure 7. A more accurate forecast that changes no decision creates no value. Accuracy is a means; the inventory, service, and cost outcome is the end, and the two come apart routinely.

Forecast accuracy is a property of the forecast: how close it came to the actual demand. Forecast value is a property of the decisions the forecast informed: whether it led to better ordering, staffing, production, or inventory outcomes than would have occurred without it. The two are related, because a more accurate forecast can support better decisions, but they are not the same, and the relationship is far weaker than the theatre assumes. A forecast can become more accurate without changing any decision, in which case its improved accuracy created no value. And a forecast can be quite inaccurate yet still support a good decision, if the decision is robust to the forecast error or if the forecast is accurate enough on the dimension that matters for the specific decision.

Consider a concrete case. Suppose a forecasting team improves its accuracy on a product from one figure to a better one, a genuine improvement in the metric. If the product is ordered in fixed quantities from a supplier with a long lead time, and the improved accuracy does not change the order quantity or timing, because the ordering policy is not sensitive to that range of forecast variation, then the accuracy improvement changed no decision and created no value, however real the metric improvement. The team will report the improved accuracy as a success, and in accuracy terms it is one, but in value terms nothing happened, because the decision the forecast was supposed to inform was unaffected. Multiply this across a portfolio and much of the accuracy improvement an organization celebrates may be creating no value, because it does not change decisions.

The converse is equally important and equally neglected. A forecast that is only modestly accurate can create substantial value if it meaningfully improves a decision that is sensitive to it, and a small improvement in forecast quality at a decision-critical point can be worth more than a large improvement where no decision hangs on it. This means that forecasting effort should be directed not to where accuracy is lowest or where improvement is easiest, but to where improved forecasts would most change decisions for the better, which is a completely different targeting principle from the one accuracy theatre uses. An organization that allocates its forecasting effort by decision sensitivity rather than by accuracy metric will create more value from less effort, because it concentrates on the forecasts that matter to decisions and stops polishing the ones that do not. The reorientation from accuracy to value is not a refinement of the theatre; it is a replacement of it, and it changes what the organization measures, rewards, and works on.

The accuracy-value distinction has a practical corollary for how forecasting improvement projects should be justified and evaluated, which most organizations get wrong. A project to improve forecast accuracy is typically justified by the accuracy improvement it promises and evaluated by the accuracy improvement it delivers, with the implicit assumption that accuracy improvement equals benefit. But the correct justification and evaluation is in terms of the decisions the improved accuracy will change and the value those changed decisions will create, which requires tracing the path from accuracy to decision to outcome, and that path is frequently broken or weak. A forecasting project that improves accuracy but changes no decision has delivered its promised accuracy and created no value, and it should be evaluated as a failure despite hitting its accuracy target, because the accuracy target was a proxy for value and the proxy did not hold.

This reframing changes which forecasting investments an organization should make. Rather than investing wherever accuracy can be improved most, or most cheaply, an organization should invest wherever improved accuracy would most change decisions for the better, which requires understanding the decision sensitivity of each forecast, how much a given improvement in accuracy would change the associated ordering, stocking, or production decision, and how much value that change would create. Some forecasts are highly decision-sensitive, where a modest accuracy improvement meaningfully changes decisions and creates substantial value; others are decision-insensitive, where even large accuracy improvements change nothing. Directing investment by decision sensitivity rather than by accuracy improvability concentrates forecasting effort where it creates value and withdraws it from where it does not, which is a fundamentally different and more valuable allocation than the accuracy-driven one, and it follows directly from taking seriously the distinction between accuracy and value.

10 The vendor claim, and its four missing numbers

The accuracy theatre is sustained and amplified by a marketing ecosystem, because the vendors of forecasting and demand-planning software sell their products on the promise of improved accuracy, and their claims are a case study in how an accuracy figure can be made to sound impressive while meaning nothing. Figure 8 sets out the interrogation such claims require.

Figure 8. The ubiquitous claim of a large accuracy improvement is close to meaningless without four specifications: the metric, the baseline, the aggregation level, and the population of items. Almost all such claims omit them.

The claims are everywhere and they follow a pattern. A vendor states that its software improves forecast accuracy by some impressive-sounding range, twenty to fifty percent, or delivers a specified accuracy uplift with associated financial benefits. The figure is presented as a headline benefit, unattributed to any specific study or defined methodology, and it is designed to convey that the software will make forecasts substantially better. A buyer under pressure to improve forecasting hears the number, finds it compelling, and moves toward purchase, without noticing that the claim as stated cannot be evaluated because it omits everything needed to make sense of it.

Four specifications are needed to evaluate any forecast-accuracy claim, and their absence is the tell. First, which metric: a twenty percent improvement in MAPE, in WMAPE, in MASE, or in bias are entirely different things, and the claim rarely says which. Second, against what baseline: an improvement measured against a naive forecast, against a spreadsheet, against a prior system, or against nothing at all means completely different things, and the claim rarely specifies the comparison. Third, at what aggregation level: an improvement at the national-total level, where accuracy is easy, is far less valuable than one at the item-location level where decisions are made, and the claim rarely states the level. Fourth, over which population of items: an improvement measured on stable high-volume items, ignoring the difficult intermittent tail, is not the improvement a buyer needs, and the claim rarely defines the population. A claim that omits all four, as almost all do, is not a measurement; it is a marketing number, and it cannot be reproduced or verified.

The discipline for a buyer is to demand the four specifications and to treat their absence as disqualifying. When a vendor claims an accuracy improvement, the buyer should ask: measured by which metric, against which baseline, at which aggregation level, over which items, and in which documented study with what population. A vendor who can answer has made a claim that can be evaluated; a vendor who cannot has made a marketing assertion dressed as a measurement, and the assertion should carry no weight in the decision. Traced to their source, most vendor accuracy claims dissolve into a customer testimonial without methodology, or an unattributed study, or a figure from sponsored research whose provenance the vendor cannot or will not specify. The four missing numbers are missing because

supplying them would make the claim evaluable, and an evaluable claim is a weaker marketing instrument than an impressive but unverifiable one. A buyer who insists on the four numbers cuts through most of the accuracy marketing at a stroke.

11 The fairness case: measurement is not the enemy

This article has argued hard against the way forecast accuracy is measured and used, and fairness requires stating the opposing case, which has real merit. A reader who concludes that forecast accuracy measurement is worthless and should be abandoned would be overcorrecting as badly as the theatre this article criticizes.

Begin with the necessity of measurement. An organization cannot manage what it does not measure, and forecast accuracy measurement, for all its flaws, gives forecasting a discipline and an accountability it would otherwise lack. Without any accuracy measurement, forecasting would drift into pure guesswork with no feedback, and the situation would be worse, not better. The problem this article identifies is not that accuracy is measured but that it is measured badly, with the wrong metric, at the wrong level, disconnected from value, and the remedy is better measurement, not the absence of measurement. A reader should not take the critique of MAPE and aggregation as license to stop measuring; measurement is essential, and the point is to measure the right thing well.

MAPE itself deserves a fairer hearing than the critique so far allows. For stable, high-volume items with no zeros and modest variability, MAPE is intuitive, easy to communicate, and perfectly adequate, and its pathologies simply do not bite in that regime. Its defects arise specifically with intermittent and low-volume demand, and an organization whose portfolio is dominated by stable high-volume items may reasonably use MAPE without much distortion. The error is not using MAPE ever; it is using MAPE everywhere, including where it misleads, and treating it as the only metric. Used in its appropriate domain and supplemented where it fails, MAPE is a reasonable tool, and the menagerie of alternatives is not a set of universally superior replacements but a set of tools for the situations MAPE handles badly.

The critics can also overstate the worse-than-naive findings. A naive forecast is not always operationally available or appropriate; some demand truly requires forecasting that a random walk cannot provide, and the fact that a naive baseline sometimes wins does not mean forecasting is generally pointless. Forecast Value Added analysis over short windows is noisy, and a single month in which a process step underperforms naive does not condemn it, as the practitioners who developed the method are careful to note. And judgmental overrides, while frequently harmful in aggregate, are sometimes exactly right, incorporating genuine information the model lacks, and the goal should be better overrides, not the mechanical elimination of human judgment. The honest synthesis is that the theatre is real and the critique is sound, but the response is disciplined measurement directed at value, not the abandonment of forecasting or measurement, and a reader should take from this article a better way to measure and reward forecasting, not a nihilism about whether forecasting is worth doing.

A further point in fairness concerns the relationship between this critique and the interests of those who make it, including software vendors and consultants who benefit from persuading organizations that their forecasting is broken. Some of the loudest voices arguing that current forecasting adds no value are voices with something to sell, whether software that promises to fix it or consulting that promises to diagnose it, and a reader should apply to the critics the same skepticism this article applies to vendor accuracy claims. The worse-than-naive findings, the override research, and the Forecast Value Added

framework are sound and independently corroborated, but they are also commercially useful to parties who profit from the perception of a problem, and the appropriate response is neither to dismiss the findings because of who benefits from them nor to accept them uncritically, but to verify them on one's own forecasts.

This is, in the end, the deepest fairness point and the one that should govern how a reader uses this article. The claims here, including the critical ones, should be tested rather than believed, and the tools to test them, computing Forecast Value Added against a naive baseline on one's own forecasts, measuring the value added by one's own overrides, checking whether accuracy improvements changed one's own decisions, are available to any organization willing to do the work. An organization should not take this article's word that its forecasting is theatre any more than it should take a vendor's word that its software will fix it. It should run the analysis, and let its own results determine the conclusion. If the analysis shows that the forecasting process beats naive and adds value, then the critique does not apply and the process is sound. If it shows the opposite, the organization has learned something important that its accuracy metrics were concealing. Either way, the verification, not the assertion, is what should drive the decision, and this article's role is to identify the right questions and the right tools, not to pronounce a verdict on any particular organization's forecasting.

12 Measure value, not accuracy

The constructive core of this article is a reorientation from measuring accuracy to measuring value, and it can be stated as a governing principle: reward the decision the forecast informs, not the accuracy percentage the forecast achieves. Figure 9 sets out the resulting discipline.

Figure 9. From accuracy theatre to forecast value: measure against naive, publish what destroys value, match the metric to the demand, measure at the decision level, and reward the decision not the metric.

The reorientation has several components, each of which replaces a piece of the theatre with something substantive. First, measure Forecast Value Added against a naive baseline at every step of the process, so the organization knows which steps add value and which subtract it, rather than only knowing the level of the final accuracy number. Second, publish and act on the finding, naming the steps that make forecasts worse and eliminating or constraining them, rather than preserving the whole process because it feels sophisticated. This is the step most organizations flinch from, because it means telling planners their overrides frequently hurt and dismantling parts of a process people are invested in, but it is where the value is.

Third, match the metric to the demand, using MASE or RMSSE for intermittent items and WMAPE where volume weighting matters, rather than applying MAPE universally, and measure bias always. Fourth, measure accuracy at the level and horizon where decisions are actually made, the item-location and the lead-time-relevant horizon, rather than at the aggregate level where the number looks best, accepting the worse-looking figures as the price of relevance. Fifth, and most fundamentally, tie forecasting to the business outcomes it is supposed to serve, measuring whether forecast improvements actually change inventory, service, and cost outcomes, and directing forecasting effort to where it most changes decisions rather than to where the accuracy metric is easiest to move.

The cumulative effect of these changes is to connect forecasting to value in a way the accuracy theatre never does. An organization that adopts them stops asking how accurate its forecasts are and starts asking whether its forecasting process beats doing nothing and whether it changes decisions for the better. These are harder questions, and they frequently yield uncomfortable answers, that much of the process adds no value, that many overrides hurt, that some accuracy improvements change no decision. But they are the questions that matter, because they measure what forecasting is for, and an organization that answers them candidly will produce better decisions at lower cost than one that continues to perform accuracy theatre, polishing a percentage that was never the point. The

reorientation is not a new metric to add to the dashboard; it is a different conception of what forecasting is measured and rewarded for, and it replaces the theatre rather than refining it.

A practical obstacle to the reorientation deserves acknowledgment, because organizations attempting it routinely encounter it. Measuring value rather than accuracy is harder than measuring accuracy, and the difficulty is not incidental but fundamental. Accuracy is a property of the forecast alone, computable from the forecast and the actual, whereas value is a property of the decisions the forecast informed and the outcomes those decisions produced, which requires connecting the forecast to the decision to the outcome through a chain that is frequently untracked and sometimes untrackable. An organization that wants to measure whether a forecast improvement changed inventory outcomes must be able to trace the forecast to the inventory decision to the inventory result, isolating the forecast's contribution from everything else that affected inventory, which is truly difficult and is why organizations default to the easier accuracy measurement.

The difficulty is real but it is not a reason to retreat to accuracy theatre, because the difficulty of measuring value does not make accuracy a valid substitute for it. An organization that cannot fully measure value can still do better than measuring accuracy alone: it can measure Forecast Value Added against naive, which at least establishes whether the forecasting effort beats doing nothing; it can measure bias, which directly affects inventory; it can measure accuracy at the decision-relevant level rather than the aggregate; and it can make qualitative judgments about decision sensitivity to direct its effort. These are partial measures of value, imperfect but far better aligned with what matters than an absolute accuracy percentage, and they are available even where full value measurement is not. The counsel is not to achieve perfect value measurement, which may be impossible, but to move as far from pure accuracy measurement toward value measurement as the organization's data and effort allow, because every step in that direction improves the alignment between what is measured and what matters, and the theatre is precisely the refusal to take any such step.

13 A forecasting protocol, and a scoring rubric

The principles above combine into a forecasting protocol that an organization can adopt in place of accuracy theatre, and a rubric a leader can use to assess whether a forecasting function is measuring value or performing measurement.

The protocol runs as follows. Establish a naive baseline for every forecast, the cheapest sensible one, such as a random walk or seasonal naive. Measure the value added by each process step against that baseline, and report the value added, not only the final accuracy. Constrain and track judgmental overrides, requiring justification and being especially skeptical of upward adjustments. Choose accuracy metrics that fit the demand, using scaled metrics for intermittent items and measuring bias always. Measure at the decision-relevant level and horizon, not the flattering aggregate. And connect forecast quality to business outcomes, directing effort to where improved forecasts most change decisions and tracking whether improvements actually change inventory, service, and cost.

A scoring rubric

The dimensions below distinguish a value-oriented forecasting function from one performing accuracy theatre.

Dimension	Value orientation	Accuracy theatre
Baseline	Every forecast measured against naive	Accuracy measured in absolute terms only
Process steps	Value added measured and acted on per step	Whole process preserved, unexamined
Overrides	Justified, tracked, upward ones scrutinized	Unconstrained, assumed to help
Metric	Matched to demand; bias measured	MAPE applied everywhere
Level	Measured where decisions are made	Measured where the number looks best
Value link	Tied to inventory, service, cost outcomes	Accuracy is the goal in itself
Vendor claims	Four specifications demanded	Headline percentage accepted

A function scoring in the left column measures whether its forecasting creates value and directs effort accordingly. A function scoring in the right column performs the ritual of accuracy measurement while remaining disconnected from the outcomes forecasting is supposed to serve. The rubric does not make forecasting easy, which it is not. It ensures the organization is measuring and rewarding the thing that matters, the value the forecast creates through the decisions it changes, rather than a percentage that can improve while the business does not.

14 The politics of the forecast number

A dimension that the technical treatment of forecasting metrics tends to omit, and that sustains the theatre as much as any mathematical defect, is the politics of the forecast number: who is accountable for it, how it is used, and what incentives that use creates. A forecast is not only a prediction; it is a number that people are measured against, that drives commitments, and that different functions want to move in different directions, and these political forces shape the forecast as much as any model.

Consider the incentives around the number. If the forecast is used to set sales targets, the sales organization has an incentive to forecast low, so the target is easy to beat, and the forecast is biased downward by the pressure of the target it feeds. If the forecast is used to justify capacity or inventory, the operations organization may have an incentive to forecast high, to secure the resources it wants, and the forecast is biased upward. If the forecast is used to evaluate the forecasting team, that team has an incentive to choose the metric, level, and baseline that make its accuracy look best, which is the aggregation game and the metric-selection game described earlier, driven by the political need to show a good number. The forecast that emerges from these crosscurrents is shaped by who is accountable for what and how the number is used, and its biases reflect the incentives as much as the demand.

This political dimension explains why purely technical fixes to forecasting frequently fail. An organization can adopt better metrics and better methods, but if the forecast number is still used in ways that create incentives to bias it, the biases will persist, expressed through whatever the new methods allow. A forecasting process is embedded in an organizational context of targets, evaluations, and resource allocations, and that context exerts pressure on the number regardless of the sophistication of the methods producing it. Improving forecasting therefore requires attending to the politics, to how the number is used and what incentives that use creates, and not only to the methods, because the methods operate within the incentive structure and cannot override it.

The practical response is to separate, as far as possible, the forecast used for planning from the numbers used for targets and evaluations, so that the planning forecast can be an honest best estimate rather than a negotiated or gamed number. Some organizations maintain a distinction between an unbiased demand forecast, produced for operational planning and measured on its value added, and the separate commitment and target numbers used for performance management, precisely so that the forecast is not corrupted by the incentives attached to the targets. This separation is difficult to maintain, because the numbers tend to collapse into one another under organizational pressure, but it is the structural response to the politics of the forecast, and an organization serious about forecast quality must address the incentive structure around the number, not only the methods that produce it. Otherwise it will find that its improved methods produce the same biased numbers, because the politics that biased the old numbers is still in place, and the politics, not the methods, was the binding constraint.

15 Conclusion: reward the decision, not the metric

Forecast accuracy theatre endures because it is comfortable. It gives the forecasting function a clear number to improve, gives management a clear number to track, and gives everyone the feeling of rigorous measurement, all without the discomfort of asking whether any of it creates value. The number goes up, the function is credited, the software is bought, and the ritual continues, self-contained and self-justifying, disconnected from the decisions and outcomes it was supposed to serve. The theatre is not a failure of effort or sophistication; forecasting organizations are frequently highly capable and hardworking. It is a misdirection of that effort toward a metric that was never the point.

The evidence assembled in this article is that the metric is not merely beside the point but actively misleading. MAPE distorts the intermittent demand that is hardest to forecast and rewards downward bias that harms service. Aggregation flatters the process by measuring where the number looks best rather than where decisions are made. A large share of forecasts are worse than a naive baseline that costs nothing, and about half of judgmental overrides make forecasts worse, with optimistic upward adjustments the most destructive. And the entire apparatus is conflated with value, when a more accurate forecast that changes no decision creates nothing. These are not marginal criticisms; together they indicate that much of what forecasting organizations measure, report, and reward is disconnected from the value forecasting is supposed to create.

The remedy is a different question, asked consistently: does each step of the forecasting process beat doing nothing, and does the forecast change a decision for the better. Forecast Value Added analysis operationalizes the first question, revealing which steps add value and which subtract it, and frequently showing that much of the process could be eliminated with no loss. The reorientation to decision value operationalizes the second, directing forecasting effort to where it most changes decisions and measuring forecasting by the outcomes it improves rather than the accuracy it achieves. Neither is easy, and both yield uncomfortable answers about processes and judgments that organizations are invested in. But they are the questions that connect forecasting to value, and an organization that asks them will stop performing accuracy theatre and start doing the harder, more valuable work of forecasting to change decisions. The number on the dashboard may not improve, and may even look worse once measured rigorously at the decision level. The business will be better off, because the forecast will finally be measured and rewarded for the thing it was always supposed to do, which is not to be accurate but to help someone decide.

16 Methodology, caveats, and sources

Methodology

- This article synthesizes the peer-reviewed forecasting literature on accuracy metrics, the practitioner literature on Forecast Value Added, and published research on judgmental adjustment, current to mid-2026. Supply Chain Research is independent and accepts no payment from the software vendors discussed.
- Where a finding originates with a party that sells forecasting software, that interest is identified, and the finding is corroborated against independent sources where possible.

Caveats

- The worse-than-naive figures are drawn from research presented at forecasting conferences and published through a forecasting-software company. They are methodologically sound and corroborated across studies, but the primary source has a commercial interest in the problem described, and readers should run the analysis on their own forecasts rather than rely on the headline figures.
- Figures 2, 4, and 6 are illustrative of documented patterns and directions rather than exact reproductions of specific datasets, and are labelled as such. The specific percentages shown are representative, not measured from a single named study.
- The judgmental-override findings are drawn from research across tens of thousands of forecasts in a limited number of supply chain companies. The direction of the finding is robust; the exact proportions vary by study, industry, and context.
- MAPE's suitability for stable high-volume demand is genuine. The critique applies to its universal application, especially to intermittent demand, not to its use within its appropriate domain.
- Forecast Value Added over short windows is noisy, and a single period's underperformance relative to naive does not condemn a process step. Conclusions require a sufficient window, as the method's developers emphasize.

Sources

[1] Hyndman, R.J. and Koehler, A.B. Another Look at Measures of Forecast Accuracy (International Journal of Forecasting, 2006).

[2] Gilliland, M. (SAS). Forecast Value Added Analysis: Step by Step.

[3] Gilliland, M. Forecast accuracy, FVA, and worse-than-naive findings (International Symposium on Forecasting 2019 materials).

[4] Institute of Business Forecasting. Forecast Value Added (FVA) blog series.

[5] Fildes, R., Goodwin, P., Lawrence, M. and Nikolopoulos, K. Effective forecasting and judgmental adjustments (research on overrides across supply chain companies).

[6] Kolassa, S. On the evaluation of point forecasts for intermittent demand.

[7] Makridakis, S. Accuracy measures: theoretical and practical concerns (International Journal of Forecasting, 1993).

[8] Makridakis, Spiliotis, Assimakopoulos. The M5 competition and RMSSE (International Journal of Forecasting).

Additional context drawn from the peer-reviewed forecasting literature and from practitioner materials on Forecast Value Added. Findings originating with parties that sell forecasting software are identified as such. This article is analysis, not statistical or operational advice, and its conclusions should be validated against your own forecasts and circumstances before any decision.
