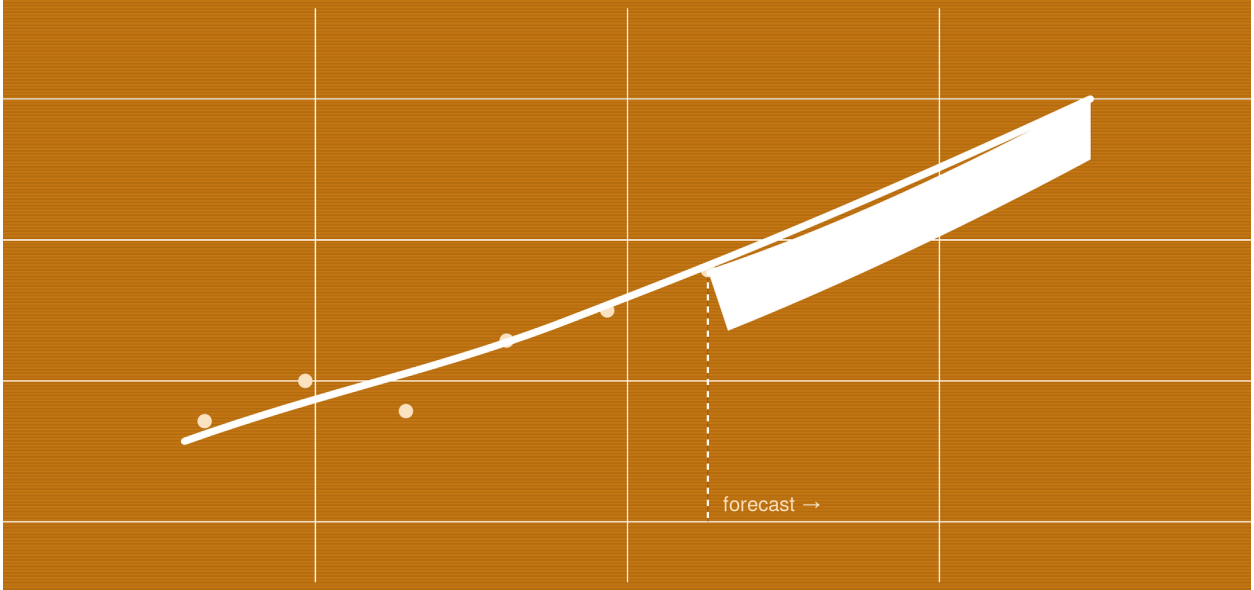
ccplanning.net
workforce planning, done well



A CCPLANNING.NET WHITE PAPER · PART

The Forecasting Masterclass

From the three building blocks to a defensible, accurate forecast — the complete craft.

October 2026 · ccplanning.net

Includes a free worked-methods spreadsheet + calculators

Contents

Right-click and choose "Update Field" to populate page numbers in Word.

Executive summary

Forecasting is the foundation of workforce planning, and most of what separates an accurate forecast from a poor one is not the sophistication of the method. It is the discipline applied to the basics: clean inputs, the right pattern captured, the appropriate method for the data, and honest measurement of how well it worked. The planners who forecast well are rarely the ones with the fanciest models. They are the ones who do the fundamentals properly and resist the temptation to skip them.

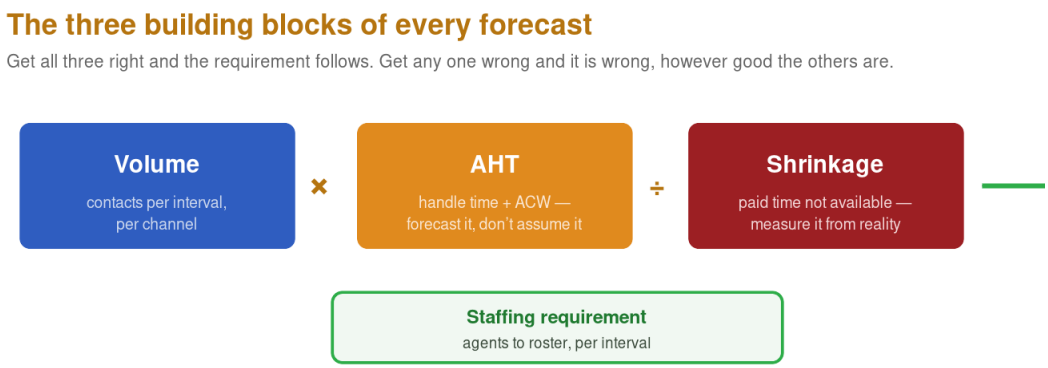
This masterclass walks through the whole craft in the order it should be done: the three building blocks every forecast rests on, getting the data right, choosing a method that fits, capturing the patterns that actually drive demand, translating a volume forecast into a staffing requirement, and measuring accuracy honestly enough to improve. It is written for the planner or analyst who does the forecasting and wants to do it better, and it is deliberately practical — every section is something you can apply to next week's forecast.

> **The thesis in one paragraph**
>
> A good forecast is built, not modelled. Volume, AHT, and shrinkage are the three building blocks; get those right on clean, granular data and almost any sensible method produces a usable forecast. Get them wrong and the most advanced model in the world produces a confident, precise, wrong answer. The craft is in the fundamentals — clean inputs, the right pattern, an appropriate method, and honest accuracy measurement — and the planners who master those out-forecast the ones chasing sophistication every time.

This is the fourth ccplanning white paper. The earlier three covered the strategy of planning — AI, the business case, and building the function. This one goes to the craft itself: the actual work of producing a forecast you can defend.

1. The three building blocks

Every contact centre forecast, however it is produced, rests on three quantities. Forecast each one well and the staffing requirement follows; forecast any one badly and the requirement is wrong no matter how carefully the others were done.



Volume is how many contacts will arrive, by interval, by channel. It is the most visible building block and the one most forecasting effort goes into. **Average handle time (AHT)** is how long each contact takes to handle, including after-call work — and it is the building block most often treated as a static assumption when it should be forecast in its own right, because AHT drifts with channel mix, complexity, product changes, and the deflection of easy contacts to self-service. **Shrinkage** is the proportion of paid time that is not available for handling contacts — holidays, training, breaks, meetings, sickness, and the rest. It is the building block most often modelled optimistically, and because it sits at the end of the calculation, an error in shrinkage flows straight through to the requirement.

> *Volume gets the attention; AHT and shrinkage get the staffing wrong. The forecast that nails volume and fudges the other two will still under- or over-staff the floor.*

The discipline is to treat all three as things to be forecast, not just volume. A planner who forecasts volume to two decimal places and applies a single inherited shrinkage figure across the whole year has done the easy third of the job well and the hard two-thirds badly.

Forecasting AHT, not assuming it

AHT deserves its own forecast because it moves for reasons a single assumption cannot capture. It drifts up as self-service deflects the simple, short contacts and leaves agents the complex residue. It shifts with product launches and policy changes that lengthen calls. It varies by time of day, by day of week, and seasonally — calls in a December peak are often shorter and more transactional than the considered enquiries of a quiet February. A forecast that applies one AHT to the whole year will be systematically wrong in exactly the periods where being wrong is most expensive. The practical approach is to forecast AHT with the same pattern-aware methods used for volume — capturing its trend and its day-of-week and intraday shape — rather than freezing it at a historical average.

Measuring shrinkage from reality

Shrinkage is the building block most corrupted by wishful thinking. The figure inherited from a previous plan, or chosen because it makes the headcount look affordable, is almost always lower than reality. Real shrinkage is built from the bottom up — holiday, sickness, training, coaching, meetings, breaks, system downtime, and the dozens of small off-phone activities — and measured from what actually happened, not what was

budgeted. It also varies through the year: holiday clusters in summer and December, training lands in waves, sickness rises in winter. A single annual shrinkage number applied flat across the year under-staffs the high-shrinkage periods and over-staffs the low ones. Forecasting shrinkage by period, from measured history, is as much a part of the craft as forecasting volume.

2. Getting the data right

The most common cause of a bad forecast is not the method. It is the data feeding it. A forecast is a statement about the future built from the past, and if the record of the past is dirty, incomplete, or mis-aggregated, no method recovers from it. This is the least glamorous part of the craft and the highest-leverage.

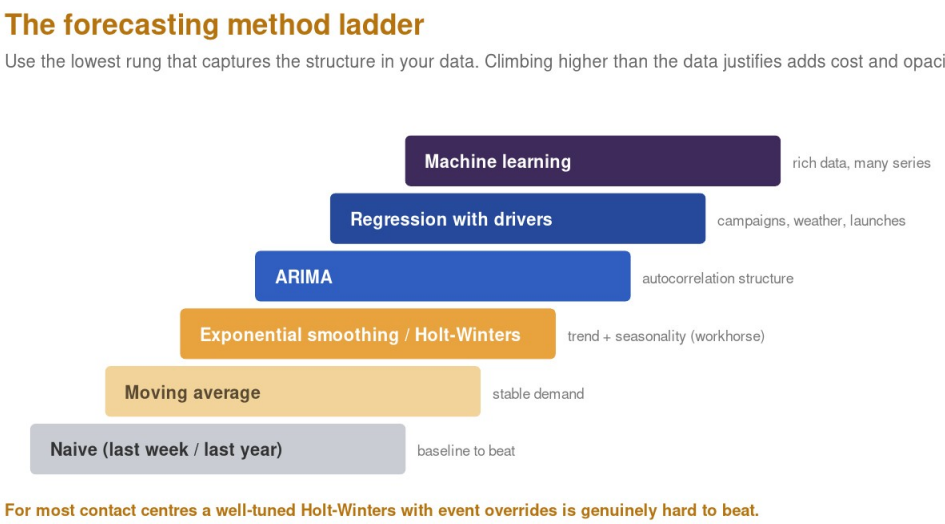
Several data problems recur. **Granularity**: a forecast needs interval-level history — typically 15 or 30 minutes — because the staffing decision is made at interval level; daily totals hide the intraday shape the schedule has to cover. **Channel separation**: voice, chat, email, and back-office work behave differently and must be forecast separately, not lumped into a single contact count. **Clean actuals**: ACD reports that double-count abandoned contacts, wrap codes that drift, IVR-deflection numbers nobody can reproduce — each quietly corrupts the history. **Event flags**: the marketing campaign, the outage, the system migration that distorted a week of history needs to be marked, so the model learns the underlying pattern rather than the anomaly.

> **Clean the history before you model it**
> Outliers and known anomalies should be identified and handled before any forecast is fitted — not because they are inconvenient, but because an unmarked anomaly teaches the model the wrong lesson. The week a system outage tripled abandons is not a demand signal; left in the history, it inflates every future forecast for that week. Marking and adjusting known events is tedious, manual, and the single most reliable way to lift forecast accuracy without touching the method at all.

The practical rule: spend the time on the data before the model. A planner who invests an afternoon cleaning and structuring history will out-forecast one who invests the same afternoon tuning an algorithm on dirty inputs, every time.

3. Choosing a method that fits

Forecasting methods form a ladder from simple to sophisticated, and the right rung is the lowest one that captures the structure in your data. Climbing higher than the data justifies adds complexity, cost, and opacity without adding accuracy — and is one of the most common mistakes in the craft.



At the bottom sits the **naive** forecast — last week, or the same week last year — useful only as a baseline to beat. **Moving averages** smooth recent history and suit stable demand with little trend or seasonality. **Exponential smoothing**, and Holt-Winters in particular, handles trend and seasonality together and is the workhorse of contact centre forecasting: for most operations, a well-tuned Holt-Winters with sensible event overrides is genuinely hard to beat. **ARIMA** models the autocorrelation structure and can edge ahead on certain series. **Regression with drivers** brings in the external variables — marketing, weather, holidays, product launches — that a time-series-only model cannot see, and is where real accuracy gains live for driver-sensitive queues. **Machine learning** sits at the top, valuable in specific places (covered in the first paper of this series) and wasted on small or clean-but-simple data.

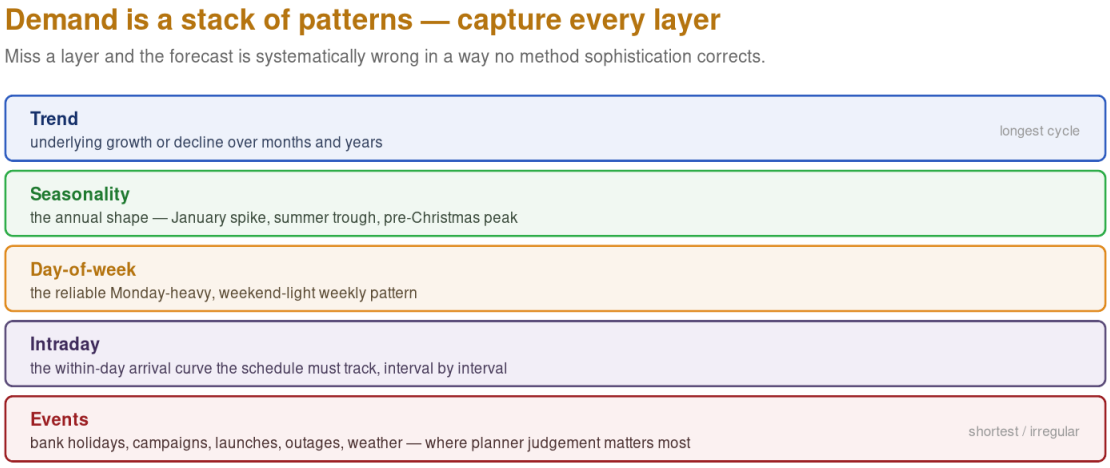
> *The right method is the lowest rung that captures the structure in your data. A well-tuned Holt-Winters beats a badly-used neural network in most contact centres — and is far easier to defend.*

The judgement is to match the rung to the series. A stable, low-volume queue is well served by exponential smoothing; a campaign-driven queue needs driver-based regression; a queue with rich history and strong external drivers may justify machine learning. Use the simplest method that captures the pattern, and always benchmark it honestly against the rung below.

A worked intuition helps. Take a queue with a clear weekly cycle and gentle upward trend. A moving average lags the trend and smears the weekly shape, so it is consistently a little behind. Holt-Winters, which models level, trend, and seasonality as three separate components and updates each as new data arrives, tracks both the climb and the weekly shape — and on this kind of series it will typically cut the error of a moving average by a third or more, with parameters a planner can actually understand and explain. Reach past it to ARIMA or ML and, on a series this simple, you will usually spend far more effort for a fraction of a point of accuracy, while losing the explainability that lets you defend the number. That trade-off — effort and opacity up, accuracy barely moving — is the signature of climbing the ladder too far, and it is worth learning to recognise.

4. The patterns you must capture

Contact demand is not random; it is a stack of overlapping patterns, and a good forecast captures each layer. Miss a layer and the forecast is systematically wrong in a way no amount of method sophistication corrects.



The layers, from longest to shortest cycle: **trend**, the underlying growth or decline in demand over months and years; **seasonality**, the annual shape — the January spike, the summer trough, the pre-Christmas peak; **day-of-week**, the reliable Monday-heavy, weekend-light pattern most operations show; **intraday**, the within-day curve of arrivals that the schedule has to track interval by interval; and **events**, the one-offs and irregular recurrences — bank holidays, campaigns, launches, outages, weather — that sit on top of the regular structure.

Each layer is forecast and then combined. The reliable patterns — day-of-week and intraday — are stable enough to model from history with confidence. Seasonality needs enough history (ideally two to three years) to estimate cleanly. Events are where human judgement matters most: the model handles the regular structure, and the planner overlays knowledge of what is coming that the history cannot contain. A forecast that captures trend, seasonality, day-of-week, and intraday, then has a planner overlay the known events, is most of the way to accurate before any sophistication is added.

5. Forecasting beyond voice

Most forecasting technique is taught on voice, but the modern contact centre is multi-channel, and the other channels do not behave like calls. Applying voice assumptions to them is a common and costly error. Each channel needs its own forecast and, in some cases, its own staffing logic.

Chat

Chat looks like voice — real-time, interactive — but differs in one decisive way: concurrency. An agent handles several chats at once, so the staffing translation is not a

simple Erlang calculation on contact count. Volume must be forecast as for voice, but the requirement depends on the concurrency rate, which itself varies with complexity and with how hard the operation pushes agents to multitask. Forecast chat volume with the same pattern-aware methods, but translate to staffing through a concurrency-adjusted model, and watch that AHT-per-chat rises as concurrency rises — the two are not independent.

Email and back-office

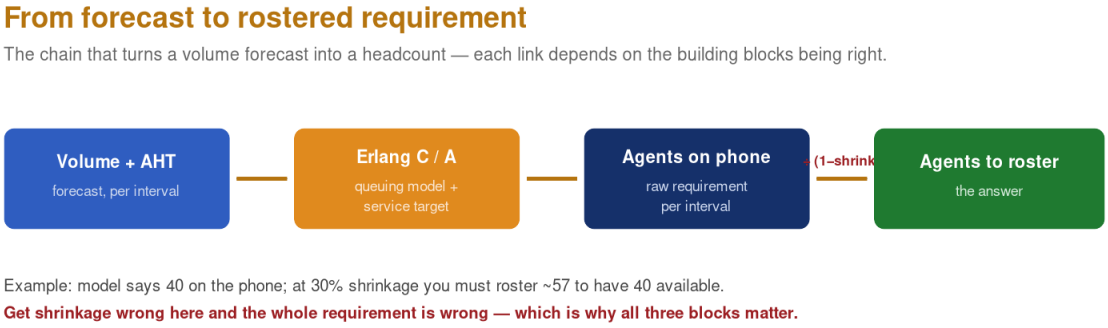
Email, cases, and back-office work are deferrable, and deferrability changes everything. The work does not have to be handled the instant it arrives; it has to be handled within a service window — same day, 24 hours, 48 hours. This means the forecast drives a workload-and-backlog model rather than an interval-by-interval Erlang requirement: what matters is whether enough capacity exists across the window to clear arrivals plus any backlog, not whether each interval is individually covered. Forecasting these channels means forecasting arrivals and modelling the backlog dynamics — a genuinely different discipline from real-time channels, and one many operations get wrong by forcing email into a voice-shaped model.

New operations and thin history

Sometimes there is little or no history to forecast from — a new operation, a new product line, a new queue. Here the craft shifts from extrapolation to construction: build the forecast bottom-up from drivers (customer base, contact propensity, expected contacts per customer), borrow the shape from a comparable existing queue, and widen the uncertainty range honestly because the forecast is genuinely less certain. As history accumulates, blend the driver-built forecast into a data-driven one. The mistake to avoid is false precision — quoting a confident single number for a queue with no track record invites exactly the credibility damage the ranges discipline exists to prevent.

6. From volume to staffing requirement

A volume forecast is not the end of the job. The point of forecasting is to produce a staffing requirement — how many people are needed in each interval to hit the service target — and the translation from one to the other is where the building blocks come back together.



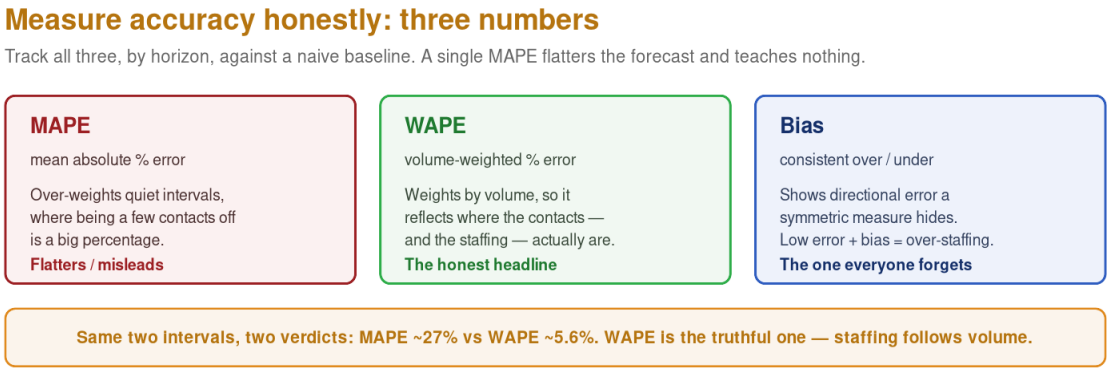
For voice and other real-time channels, the translation runs through queuing theory — the Erlang models. **Erlang C** takes forecast volume, AHT, and the service target and returns the number of agents required to hit it in a given interval; **Erlang A** extends this to account for customers abandoning the queue, which makes it more realistic for operations with meaningful abandonment. The forecast volume and AHT feed the model; the result is a raw agent requirement per interval.

Then shrinkage converts the raw requirement into a rostered requirement. If the model says you need 40 agents handling contacts in an interval and shrinkage is 30%, you need to roster roughly 57 to have 40 available — and if your shrinkage assumption is wrong, this is exactly where the error bites. The full chain is: forecast volume and AHT, apply the queuing model to get agents-on-the-phone, then gross up by realised shrinkage to get agents-to-roster. Each link depends on the building blocks being right, which is why the craft begins and ends with them.

> **Use the calculators to run the chain**
>
> The free Erlang C and Erlang A calculators on ccplanning.net take your forecast volume, AHT, and service target and return the agent requirement directly — and the shrinkage calculator grosses it up to the rostered figure. Running your own numbers through them is the fastest way to see how sensitive the requirement is to each building block: nudge AHT or shrinkage and watch the headcount move. That sensitivity is the best argument for forecasting all three blocks properly rather than just volume.

7. Measuring accuracy honestly

A forecast you do not measure is a forecast you cannot improve. Accuracy measurement is the feedback loop of the whole craft, and it has to be honest — measured the right way, at the right level, against the right benchmark — or it flatters the forecast and teaches nothing.



Three measures matter. **MAPE** (mean absolute percentage error) is the most quoted, but it has a flaw: it over-weights errors on low-volume intervals, where being a few

contacts off is a large percentage. **WAPE** (weighted absolute percentage error) weights by volume and is the more honest headline for an operation, because it reflects where the contacts — and the staffing decisions — actually are. **Bias** is the most overlooked: it shows whether the forecast is consistently over or under, which a symmetric error measure hides. A forecast with low WAPE but persistent positive bias is quietly over-staffing all year.

Two disciplines make the measurement useful. Measure **by horizon** — accuracy at the two-week lead time that drives scheduling matters more than accuracy at the day-before, and the two behave differently. And benchmark against a **naive baseline** — a method that cannot beat last-year-same-week is not earning its complexity. Track WAPE and bias by horizon against a baseline, and the forecast tells you honestly whether it is improving. Track a single headline MAPE and it mostly tells you what you want to hear.

A small worked example shows why the choice of measure matters. Suppose two intervals: a busy one forecast at 1,000 against an actual of 1,050, and a quiet one forecast at 10 against an actual of 20. The quiet interval is off by 50% and the busy one by under 5%. Averaged naively, MAPE reports about 27% — dragged up almost entirely by ten contacts in a quiet interval that barely touches staffing. WAPE, weighting by volume, divides the total error (60 contacts) by the total actual (1,070) for about 5.6% — which honestly reflects that the forecast was excellent where the contacts actually were. The same two numbers, two very different verdicts. Report MAPE and the forecast looks shaky; report WAPE and it looks strong. For an operation making staffing decisions, WAPE is the truthful one, because staffing follows volume.

> *MAPE flatters; WAPE is honest; bias is what everyone forgets. A forecast with great error scores and persistent bias is over- or under-staffing the floor every single week.*

8. The forecasting cadence

Forecasting is not a one-off act but a cycle, and the operations that forecast well run it to a rhythm. Different horizons serve different decisions and want different cadences.

The **long-range** forecast — twelve months and beyond — drives capacity planning, hiring, and budget, and is refreshed monthly or quarterly as the picture clarifies. The **operational** forecast — the next several weeks — drives scheduling and is refreshed weekly, because that is the lead time at which schedules are built and changed. The **short-term and intraday** forecast — this week and today — drives real-time decisions and is refreshed continuously as actuals arrive. Each horizon has an owner, a cadence, and a decision it serves; conflating them, or running them all at one frequency, is a common source of both wasted effort and missed signals.

The cycle also closes on itself. Every forecast becomes, in time, an actual to measure against — and the accuracy review feeds back into the next forecast. A planning function that forecasts, measures, learns, and adjusts on a steady cadence improves quarter on quarter; one that forecasts and never looks back repeats the same errors indefinitely. The cadence is what turns forecasting from a task into a discipline that compounds.

9. Common forecasting mistakes

The same mistakes recur across operations and sectors. Knowing them is the cheapest way to avoid them.

- **Forecasting volume only.** Treating AHT and shrinkage as fixed assumptions while lavishing effort on volume — the staffing requirement is only as good as the weakest of the three blocks.
- **Modelling on dirty data.** Fitting a method to history full of unmarked anomalies, so the model learns the noise. Clean first.
- **Climbing the method ladder too far.** Reaching for ML or ARIMA when a tuned Holt-Winters would do, adding opacity without accuracy.
- **Forecasting the wrong granularity.** Daily totals when the staffing decision lives at interval level; the intraday shape is where coverage is won or lost.
- **Single-number forecasts.** Quoting one figure to finance instead of a range, then being judged wrong when the actual inevitably differs.
- **Measuring with MAPE alone.** A flattering headline that hides bias and over-weights low-volume noise. Track WAPE and bias by horizon.
- **Never benchmarking.** Not knowing whether the sophisticated forecast actually beats a naive baseline — and often it does not.

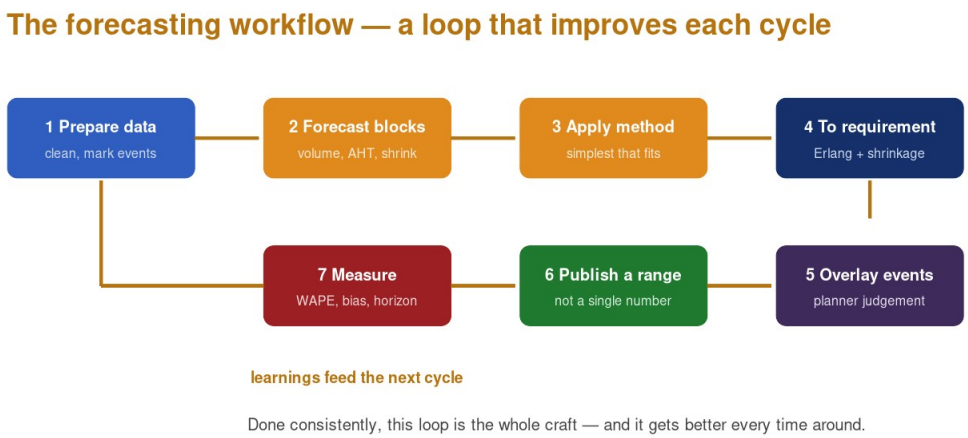
10. Forecasting in the AI era

No forecasting paper written now can ignore AI, but the honest place for it is in proportion. Machine learning and the new time-series foundation models add real, measurable accuracy in specific situations — driver-rich queues with ample clean history, large numbers of related series, and thin-history queues where pre-trained models forecast competently out of the box. The gains are real and they are smaller than the marketing suggests.

The decisive point for this paper is that AI sits at the top of the method ladder, not beside it — it is the most sophisticated rung, and it only pays off on foundations that are already solid. An ML forecast on dirty data automates the existing problem faster and wraps it in opacity. The order of operations is unchanged by AI: clean the data, get the building blocks right, beat a naive baseline with a tuned classical model, layer in drivers where they matter, and only then consider machine learning. The first paper in this series covers the AI question in full; for the working forecaster, the takeaway is that AI is an addition to mastery of the fundamentals, never a substitute for it.

11. A practical forecasting workflow

Pulling the craft together, a reliable forecasting workflow runs in a repeatable loop — and following it is most of what produces an accurate, defensible forecast.



Done consistently, this loop is the whole craft — and it gets better every time around.

Start by **preparing the data**: pull interval-level history, separate channels, clean the actuals, and mark known events. Then **forecast the building blocks**: model volume capturing each pattern layer, forecast AHT rather than assuming it, and measure shrinkage from reality. **Apply the method** that fits the series, benchmarked against the rung below. **Translate to a requirement** through the queuing model and gross up by shrinkage. **Overlay judgement**: add what you know about coming events that the history cannot. **Publish a range**, not a single number. Then **measure** when the actuals arrive — WAPE and bias by horizon against a baseline — and feed what you learn into the next cycle. Done consistently, this loop is the whole craft, and it improves every time around.

Conclusion: the craft is in the fundamentals

Forecasting well is not about owning the most sophisticated model. It is about doing the fundamentals properly and in order: clean, granular data; all three building blocks forecast rather than assumed; the simplest method that fits; every pattern layer captured; the requirement translated honestly through the queuing model and shrinkage; accuracy measured by WAPE and bias against a baseline; and a cadence that closes the loop. The planners who do these things out-forecast the ones chasing sophistication, reliably and indefinitely.

> *A good forecast is built, not modelled. Master the fundamentals — clean inputs, the three blocks, the right method, honest measurement — and almost any method produces a forecast you can defend.*

The methods will keep advancing, and the AI layer will keep improving in its place. None of it changes the foundation. Get the building blocks right on clean data, capture the patterns, measure honestly, and run the loop — and you will forecast better than operations spending ten times as much on tools they have not yet earned. That is the craft, and it is learnable by anyone willing to do the unglamorous parts well.

Appendix: using the calculators

This paper has three free companion tools on ccplanning.net, and together they run the whole volume-to-staffing chain from section 5. None requires a sign-up and nothing you enter leaves your browser.

The volume forecaster takes your history and projects forward, capturing trend and seasonality — a fast way to produce or sanity-check a baseline volume forecast before you refine it. **The Erlang C calculator** takes that forecast volume, your AHT, and your service target and returns the agents required per interval; **the Erlang A calculator** does the same while accounting for abandonment, which is the more realistic choice where customers hang up. **The shrinkage calculator** then grosses the agent requirement up to a rostered figure using your measured shrinkage.

> **Run the sensitivity test**
>
> The most instructive thing you can do with the calculators is change one building block at a time and watch the requirement move. Add a few seconds to AHT, or correct shrinkage from an optimistic 30% to a realistic 36%, and the rostered headcount jumps visibly. That sensitivity is the whole argument of this paper made concrete: the requirement is only as good as the weakest of the three blocks, so all three deserve real forecasting effort — not just the volume.

The forecasting methods workbook

There is also a free companion spreadsheet — the forecasting methods workbook, in the templates section at ccplanning.net. It applies four of the methods from section 3 — a naive last-week baseline, a 7-day moving average, simple exponential smoothing, and a day-of-week seasonal forecast — to the same sample of daily contact data, and scores each one by WAPE, MAPE, and bias on the Accuracy tab. Everything is live formulas: drop your own actuals into the blue column, or change the smoothing factor, and every forecast and score updates.

The workbook makes section 3's argument tangible. On the sample data, the two methods that ignore the weekly pattern — the moving average and exponential smoothing — miss badly, while the two that respect the weekday come out far ahead, and the naive baseline proves genuinely hard to beat. That is the lesson in miniature: match the method to the structure in the data, and always check whether your clever method actually beats the simple one. Watch the bias column too — the seasonal method carries a visible negative bias because it averages older, lower weeks and lags the trend, exactly the kind of systematic error a single error score would hide.

About ccplanning.net

ccplanning.net is an opinionated, practitioner-focused resource for contact centre workforce planning — forecasting, scheduling, real-time management, capacity planning, MI, and the leadership of the planning function. It publishes free articles, browser-based planning calculators, and a fortnightly newsletter for working planners.

This is the fourth paper in a series. **Paper one** covered AI in planning; **paper two**, the business case for planning, with a value calculator; **paper three**, building a planning function, with a maturity assessment; and **this paper**, the forecasting craft, with the forecaster and Erlang calculators as its companions. All four are free at ccplanning.net.

Put it into practice

Run your next forecast through the free calculators at ccplanning.net/calculators — volume forecaster, Erlang C, Erlang A, and shrinkage. Pair this paper with the articles on the three building blocks, forecast accuracy metrics, and the forecasting methods series for the deeper detail behind each section.