Classical Forecasting vs. Time-Series Foundation Models

Comparing established methods and pretrained models across 12 datasets

5 classical methods **3 foundation-model families** **12 datasets**

Prepared by
Amirhossein Bagheri
Politecnico di Milano Supervisor: Tommaso Bianchi

Why this comparison matters

Do time-series foundation models consistently outperform classical forecasting approaches across datasets with different statistical characteristics?

Zero-shot promise

Can a pretrained model transfer across domains, frequencies, and horizons without dataset-specific training?

Model families

Compare that promise against established statistical, algorithmic, and neural forecasting approaches.

Different metrics

Separate point accuracy from quantile and predictive-distribution quality; they answer different questions.

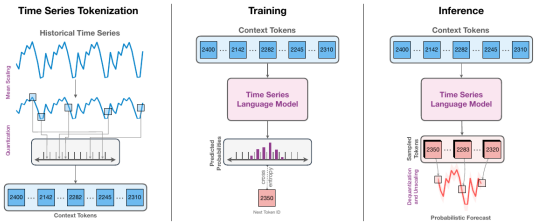
Statistical tests

Test whether observed performance differences are systematic or compatible with sampling variation.

Classical and conventional forecasting models

Model	Forecasting mechanism	Main strength	Main limitation
Seasonal Naive	Repeats the most recent seasonal cycle.	Transparent and robust reference for seasonal data.	Cannot adapt to trend or changing seasonal structure.
ARIMA $(2, 1, 2)$	Models differenced values through autoregressive and moving-average terms.	Captures short-range linear dynamics with a compact local model.	Fixed order; limited under nonlinear behavior or regime changes.
AutoARIMA	Searches candidate ARIMA orders and selects the best fit by AIC.	Automates model specification for each series.	Search can be costly or unstable on short and noisy histories.
Prophet	Decomposes a series into trend, changepoints, and calendar seasonalities.	Interpretable components and flexible trend changes.	A rigid additive decomposition may miss local dynamics.
DeepAR	Learns an autoregressive recurrent network under a probabilistic likelihood.	Represents nonlinear temporal relationships and uncertainty.	Requires training and hyperparameter choices for each task.

Chronos: forecasting as a language-model task



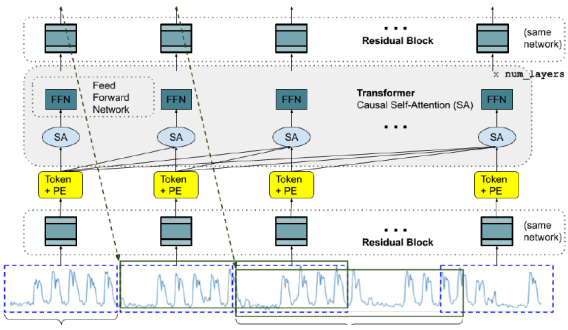
- **Representation:** mean-scale the series, quantize values into bins, and map them to a fixed token vocabulary.
- **Architecture:** a T5 encoder–decoder learns next-token probabilities using cross-entropy.
- **Forecast:** autoregressive token samples are dequantized into trajectories. Inference is **stochastic**; the sampled trajectories define a **probabilistic forecast**.

Chronos–T5 variants

Variant	Params	Enc./Dec.
Tiny	8M	4 / 4
Mini	20M	4 / 4
Small	46M	6 / 6
Base	200M	12 / 12
Large	710M	24 / 24

The variants use the same tokenization and forecasting mechanism. Capacity increases through greater depth and width.

TimesFM: patch the signal, decode the future



Patch tokens

Non-overlapping blocks of 32 observations are normalized and projected into token embeddings.

Decoder-only

Causal self-attention lets each patch representation use only its available history.

Multi-horizon

Each output head predicts a block of future time points; blocks are composed to reach the requested horizon.

Forecast behavior

TimesFM 2.5 directly outputs the mean and quantiles q10–q90. The values are **probabilistic forecasts**, but inference is **deterministic**: the same input returns the same distribution.

Moirai: one model for any frequency and variate count

Multiple patch sizes

A bank of input/output projections supports several patch lengths. The model can match its temporal resolution to series with different sampling frequencies and local dynamics.

Any-variate attention

Patches from every variable are flattened into one token sequence. Time and variable identifiers preserve where each patch came from, so attention can learn both within- and cross-variable dependence.

Distribution forecast

Masked future patches are decoded into an uncertainty representation. The original Moirai uses a flexible mixture density; Moirai 2.0 directly predicts quantiles. Inference is deterministic.

Twelve datasets I: competition and archive panels

Dataset	Domain	Series used	Main patterns
M1 Yearly	Business / economic	20	Trends and changes in level
M3 Quarterly	Business / economic	20	Trends and quarterly cycles
M4 Hourly	Mixed competition	20	Daily cycles and noise
Weather (rain)	Environment	20	Many zeros and seasonal rain patterns
Car Parts	Retail / inventory	20	Sparse demand with many zeros
NN5 Weekly	Banking / ATM	20	Yearly patterns, holidays, changing demand

Twelve datasets II: health, infrastructure, and macroeconomics

Dataset	Domain	Series used	Main patterns
National Illness	Public health	1	Yearly flu cycle and sudden peaks
Traffic	Transport	5	Strong daily and weekly traffic cycles
PSM	Server telemetry	5	Sudden anomalies and noisy sensor data
Elexon Demand	Energy	1	Weekly demand cycle and weather effects
USGS Streamflow	Hydrology	5	Yearly water cycle, floods, droughts, gaps
BLS Macro	Macroeconomics	3	Long trends and structural changes

Point and probabilistic forecast evaluation

MASE point

$$\mathrm{MASE} = \frac{H^{-1}\sum_{h=1}^{H}|y_h - \hat{y}_h|}{(T-m)^{-1}\sum_{t=m+1}^{T}|y_t - y_{t-m}|}$$

Error scaled by an in-sample seasonal-naive error. It is comparable across units and datasets.

< 1 beats the scale; $= 1$ matches it; > 1 is worse.

Central forecast
How accurate is the median path?

WQL quantiles

$$\mathrm{WQL} = \frac{1}{|Q|}\sum_{q\in Q}\frac{2\sum_h \rho_q(y_h - \hat{y}_{q,h})}{\sum_h |y_h|}$$

Weighted quantile (pinball) loss on $Q = \{0.1, \ldots, 0.9\}$, normalized by target magnitude.

Multiple quantiles reveal asymmetric risk and interval quality.

Quantile accuracy
Where are under- and over-predictions?

CRPS distribution

$$\mathrm{CRPS}(F, y) = \int_{-\infty}^{\infty}\left(F(z) - \mathbf{1}\{y \le z\}\right)^2 dz$$

Scores the entire predictive CDF against the observation; rewards **calibration** and **sharpness** jointly.

Implemented as a finite-quantile approximation from q10–q90.

Distribution quality
How well does the predictive CDF match reality?

Together they separate scale-free point error from quantile and distributional quality; lower is better.

Empirical winners across the three core losses

Dataset	MASE	WQL	CRPS
M1 Yearly	**TimesFM**	**ARIMA**	**ARIMA**
M4 Hourly	**Chronos Base**	**Chronos Base**	**Chronos Base**
Weather	**Chronos Base**	**Moirai**	**Moirai**
M3 Quarterly	**TimesFM**	**TimesFM**	**TimesFM**
Car Parts	**Moirai**	**Moirai**	**Moirai**
NN5 Weekly	**TimesFM**	**TimesFM**	**TimesFM**
National Illness	**TimesFM**	**TimesFM**	**TimesFM**
Traffic	**Moirai**	**Moirai**	**Moirai**
PSM	**Chronos Small**	**Chronos Small**	**Chronos Small**
Elexon Demand	**TimesFM**	**TimesFM**	**TimesFM**
USGS Streamflow	**Moirai**	**TimesFM**	**TimesFM**
BLS Macro	**Auto-ARIMA**	**TimesFM**	**TimesFM**

foundation model **statistical model**

Each cell is the lowest aggregate loss among 11 models; this is a **descriptive ranking**, not a significance claim.

How often each model records the lowest loss

Model	MASE	WQL	CRPS	Total
TimesFM 2.5 200M	5	6	6	**17**
Moirai 2.0 Small	3	3	3	**9**
Chronos Base	2	1	1	**4**
Chronos Small	1	1	1	**3**
ARIMA(2,1,2)	0	1	1	**2**
Auto-ARIMA	1	0	0	**1**
Chronos Mini	0	0	0	**0**
Chronos Tiny	0	0	0	**0**
DeepAR	0	0	0	**0**
Prophet	0	0	0	**0**
Seasonal Naive	0	0	0	**0**
Column total	**12**	**12**	**12**	**36**

36 comparisons

- Each model is ranked on **MASE, WQL, and CRPS** across 12 datasets.
- A win means recording the lowest aggregate loss for one dataset under one metric.
- TimesFM records the lowest loss **17 times**; Moirai does so **9 times**.
- Across all comparisons, a foundation model ranks first **33 times**.

When every foundation model beats every comparator

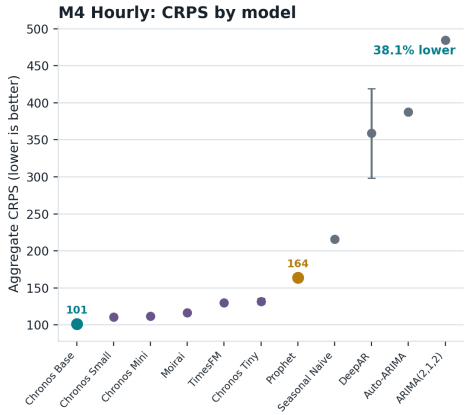
Strict family dominance: $\max_{m \in \mathrm{FM}} L(m) < \min_{m \in \mathrm{C}} L(m)$ — even the weakest foundation model has lower aggregate loss than the strongest classical/conventional comparator.

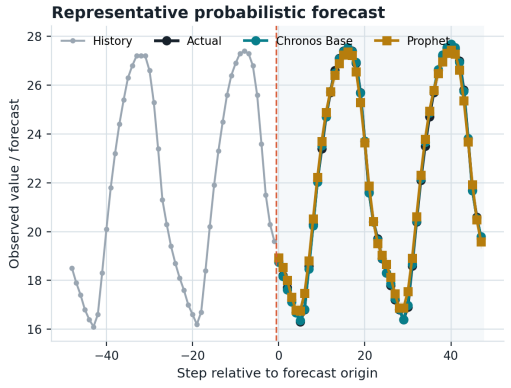
Dataset	MASE	WQL	CRPS
M4 Hourly	✓ Chronos Base	✓ Chronos Base	✓ Chronos Base
PSM	✓ Chronos Small	✓ Chronos Small	✓ Chronos Small
Elexon Demand	✓ TimesFM	✓ TimesFM	✓ TimesFM
USGS Streamflow	✓ Moirai	✓ TimesFM	✓ TimesFM
Car Parts	*not strict*	✓ Moirai	✓ Moirai
Traffic	*not strict*	✓ Moirai	✓ Moirai

16 strict comparisons

Four datasets satisfy the condition for all three losses; Car Parts and Traffic satisfy it for WQL and CRPS.

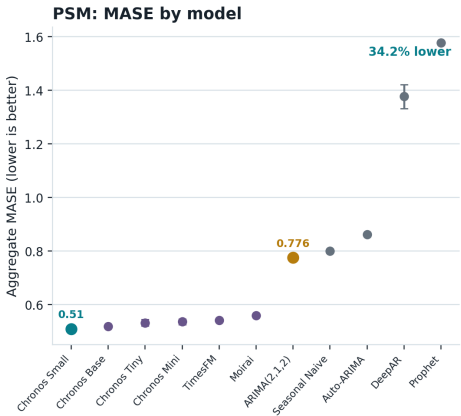
Clear empirical separation: M4 Hourly CRPS

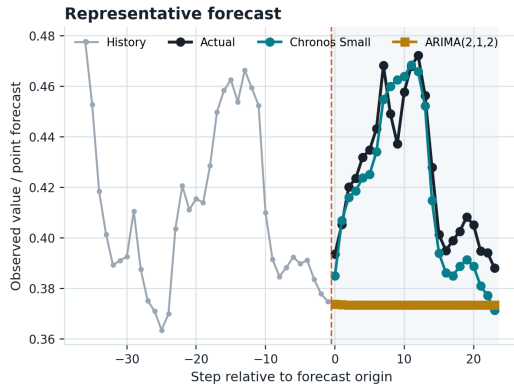




38.1% lower aggregate CRPS Similar point forecasts can still have differently scored distributions.

Clear empirical separation: PSM MASE





34.2% lower aggregate MASE

Chronos Small follows the rise and decline; ARIMA remains nearly flat.

Where a foundation model ranks first in a mixed ordering

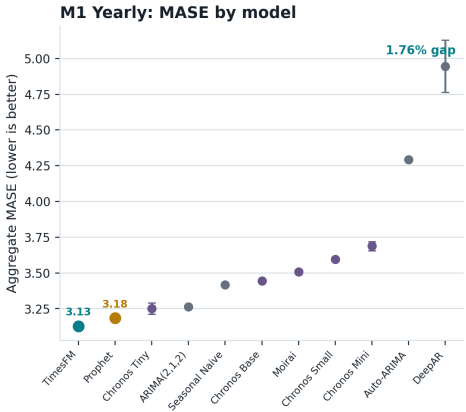
Mixed ordering: a foundation model has the lowest loss, but at least one classical/conventional model still outranks another foundation model.

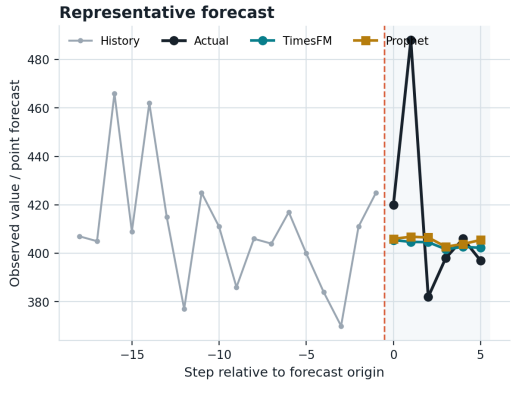
Dataset	MASE	WQL	CRPS
Car Parts	**Moirai**	*strict FM*	*strict FM*
M1 Yearly	**TimesFM**	*classical first*	*classical first*
M3 Quarterly	**TimesFM**	**TimesFM**	**TimesFM**
NN5 Weekly	**TimesFM**	**TimesFM**	**TimesFM**
Weather	**Chronos Base**	**Moirai**	**Moirai**
National Illness	**TimesFM**	**TimesFM**	**TimesFM**
Traffic	**Moirai**	*strict FM*	*strict FM*
BLS Macro	*classical first*	**TimesFM**	**TimesFM**

17 mixed comparisons strict FM dominance classical model ranks first

Purple cells name the empirical foundation-model leader.

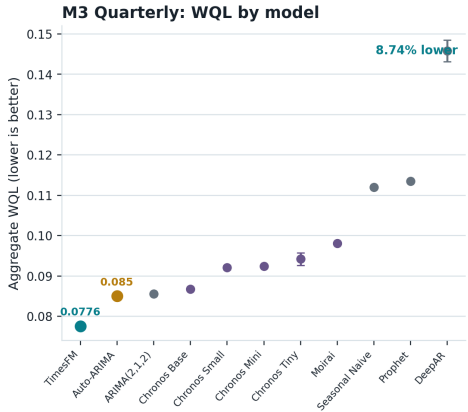
Close empirical contest: M1 Yearly MASE

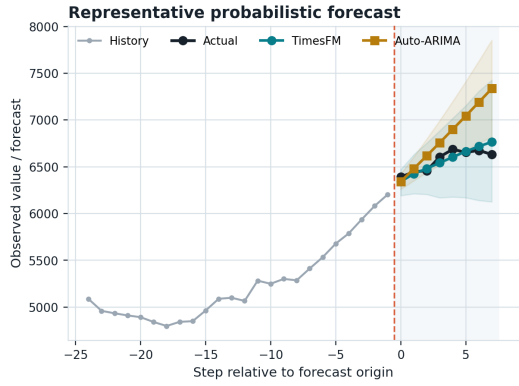




1.76% aggregate gap TimesFM ranks first, while Prophet is close and the forecasts nearly overlap.

Close empirical contest: M3 Quarterly WQL





8.74% lower aggregate WQL TimesFM follows the realized level; Auto ARIMA drifts upward.

Where a classical model beats every foundation model

3 of 36 cells

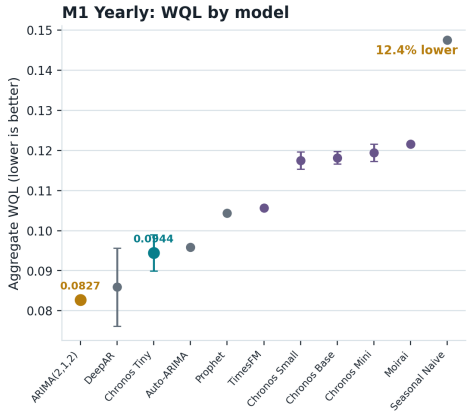
Dataset	Loss	Classical winner	Best foundation model	Lower loss
M1 Yearly	WQL	ARIMA 0.0827	Chronos Tiny 0.0944	**12.4%**
M1 Yearly	CRPS	ARIMA 202,882	Chronos Tiny 231,712	**12.4%**
BLS Macro	MASE	Auto-ARIMA 0.141	TimesFM 0.174	**18.8%**

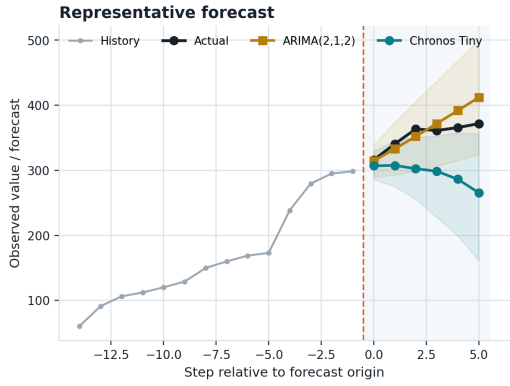
M1 is the substantive exception
ARIMA leads both probabilistic losses across 60 paired tasks.

BLS remains tentative
Its MASE result is based on only seven paired tasks.

Values are aggregate losses; percentages compare the classical winner with the best-performing foundation model. Lower is better.

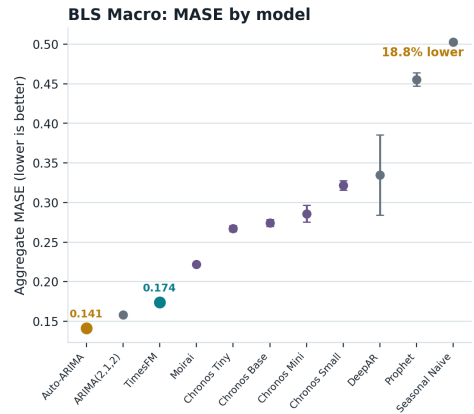
Classical exception: M1 Yearly WQL

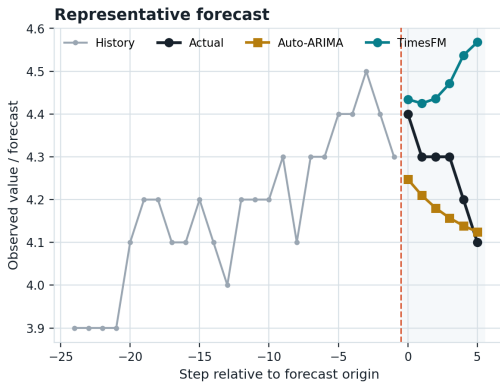




12.4% lower aggregate WQL ARIMA follows the rise; its 10–90% interval covers the observations.

Classical exception: BLS Macro MASE





18.8% lower aggregate MASE Auto-ARIMA follows the decline; TimesFM extrapolates upward.

How to interpret the repetition test

Why are 22 p-values adjusted?

36 comparisons (12 datasets × 3 losses) − **14** SI cases (no numerical p-value) = **22** numerical p-values adjusted together

Holm correction treats the 22 values as one family, limiting false positives caused by testing many comparisons.

Supported
adjusted $p < .05$
The lead remains distinguishable from seed variation after correction.

Seed-invariant
SI: no p-value
The same positive gap occurs in all three runs; its standard deviation is zero.

Unresolved
adjusted $p \geq .05$
Three runs do not distinguish the observed lead from seed variation.

Scope: repeatability across three random seeds.

Do the rankings persist across random seeds?

Winner versus the best model from the opposite family: $d_r = L_{\text{opposite},r} - L_{\text{winner},r}$, $H_0 : \mathbb{E}[d] \leq 0$, one-sided paired t test, $n = 3$.

Dataset	MASE	WQL	CRPS
Car Parts	**0.025**	**SI**	**SI**
M1 Yearly	**0.004**	0.068	0.068
M3 Quarterly	**SI**	**SI**	**SI**
M4 Hourly	< 0.001	**0.010**	**0.010**
NN5 Weekly	**0.047**	**0.025**	**0.025**
Weather	0.091	**0.025**	**0.025**
National Illness	< 0.001	< 0.001	< 0.001
PSM	< 0.001	< 0.001	< 0.001
Traffic	**SI**	**SI**	**SI**
BLS Macro	**SI**	**SI**	**SI**
Elexon Demand	**SI**	**SI**	< 0.001
USGS Streamflow	**SI**	0.052	0.052

17 significant **14 seed invariant** **5 unresolved**

Holm-adjusted p values cover the 22 finite tests. SI means the same positive gap occurred in all three runs, so finite t and p values do not exist. The table measures repeatability across the three random seeds only.

95% Model Confidence Set: method and results

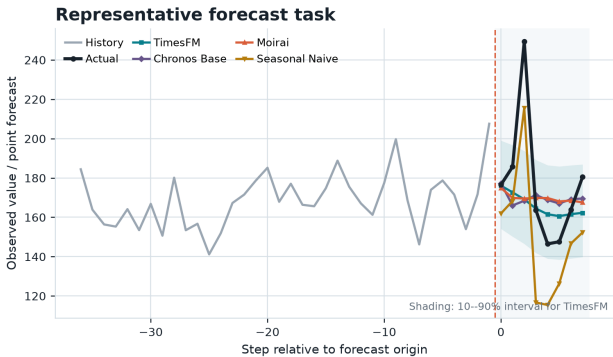
11 candidate models ⟶ **test equality, remove the weakest, repeat** ⟶ **95% MCS**

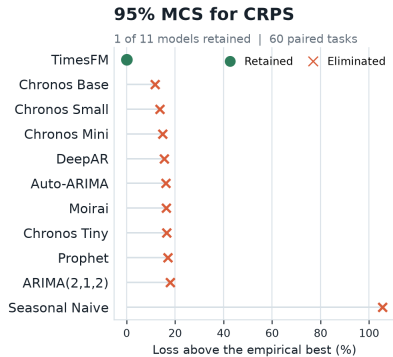
MCS starts with all 11 models, tests equal predictive ability on paired losses, and removes the weakest after a rejection. The final 95% set contains models not distinguished from the best. Each cell is **MCS size / opposite-family flag**: 1 means the best opposite-family model is excluded; 0 means it remains.

Dataset	MASE	WQL	CRPS
M1 Yearly	9/0	11/0	11/0
M4 Hourly	6/1	11/0	11/0
Weather	8/0	2/1	1/1
M3 Quarterly	4/0	2/1	5/0
Car Parts	7/0	7/1	7/1
NN5 Weekly	1/1	1/1	1/1
National Illness	7/0	5/1	5/1
Traffic	6/0	5/1	5/1
PSM	7/1	7/1	7/1
Elexon Demand	3/1	2/1	2/1
USGS Streamflow	8/0	11/0	11/0
BLS Macro	5/0	11/0	11/0

1: unique | **2–3: narrow** | **4–7: competitive** | **8–11: broad**

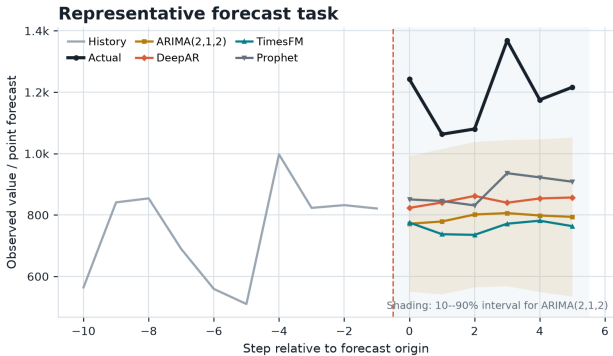
NN5 Weekly: CRPS leaves one model in the MCS

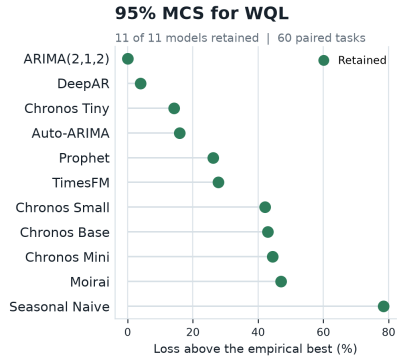




TimesFM has the lowest CRPS and is the only model retained. In this task, its predictive interval covers most of the realized path, although the central forecast misses the early spike.

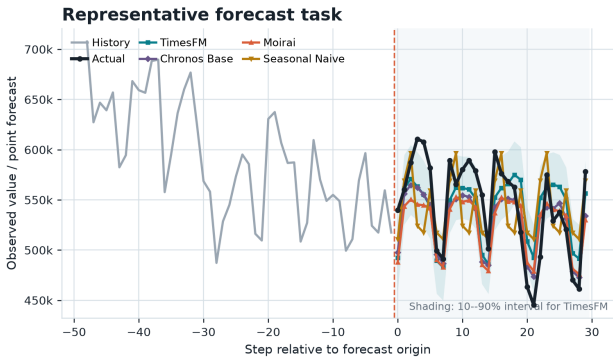
M1 Yearly: WQL leaves all models in the MCS

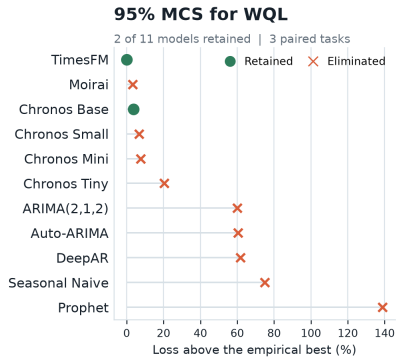




ARIMA records the lowest aggregate WQL, but the MCS cannot eliminate any alternative. A first-place empirical rank is therefore not a uniquely supported winner on this dataset.

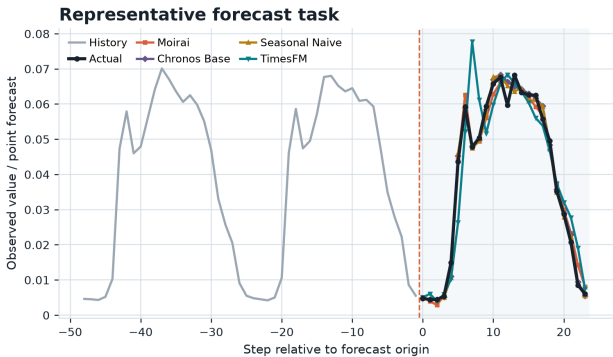
Elexon Demand: WQL narrows the MCS to two models

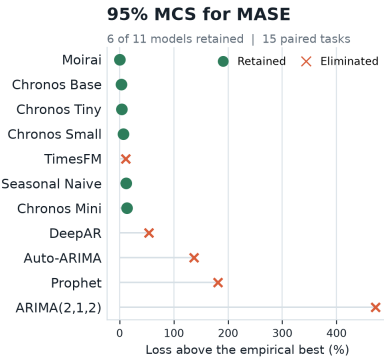




TimesFM and Chronos Base remain in the set. Because Elexon contributes only three paired tasks, this narrow MCS should be treated as exploratory rather than as broad evidence of separation.

Traffic: MASE leaves a competitive set of six





Moirai has the lowest aggregate MASE, while six models remain. MCS membership is determined by paired loss differences, so it need not follow the aggregate ranking exactly.

Conclusion: empirical results

11 of 12
foundation winner for MASE

11 of 12
foundation winner for WQL

11 of 12
foundation winner for CRPS

Main pattern

- A foundation model has the lowest loss in 33 of the 36 dataset and metric comparisons.
- Foundation models win all three metrics on 10 of the 12 datasets.
- They lead on seasonal and higher-frequency data such as M4 Hourly, NN5 Weekly, Traffic, and Elexon Demand.
- TimesFM wins 17 comparisons and Moirai wins 9.

The three classical wins

Dataset	Metric	Winner
M1 Yearly	WQL	ARIMA
M1 Yearly	CRPS	ARIMA
BLS Macro	MASE	Auto-ARIMA

Classical methods are most competitive on the economic datasets with long trends and changes in level.

Conclusion: Model Confidence Set results

4 of 12
meaningful comparison for MASE

8 of 12
meaningful comparison for WQL

7 of 12
meaningful comparison for CRPS

Main pattern

- The best model from the other family is excluded in 19 of the 36 comparisons.
- WQL and CRPS account for 15 of these 19 results. MASE accounts for only 4.
- The strongest pattern appears on recurring demand and on sparse or noisy series.
- Only four comparisons leave one model: all three NN5 losses and Weather CRPS.

Meaningful on all three losses

Dataset	MASE	WQL	CRPS
NN5 Weekly	1	1	1
PSM	7	7	7
Elexon Demand	3	2	2

The table reports the full MCS size. M1 Yearly, BLS Macro, and USGS Streamflow do not exclude the other-family winner for any loss.

MCS evidence is stronger for WQL and CRPS than for MASE, especially on recurring demand and sparse or noisy series.

Future work

Use more datasets

Add more datasets for each frequency and domain. This would let us test whether model families behave differently on seasonal, sparse, high-frequency, or trend-dominated data.

Build stronger classical and hybrid models

Remove trend or seasonality first, forecast the residual with another model, and then add the components back. Compare STL with ARIMA or ETS, model ensembles, and classical plus foundation-model hybrids.

Measure the data characteristics

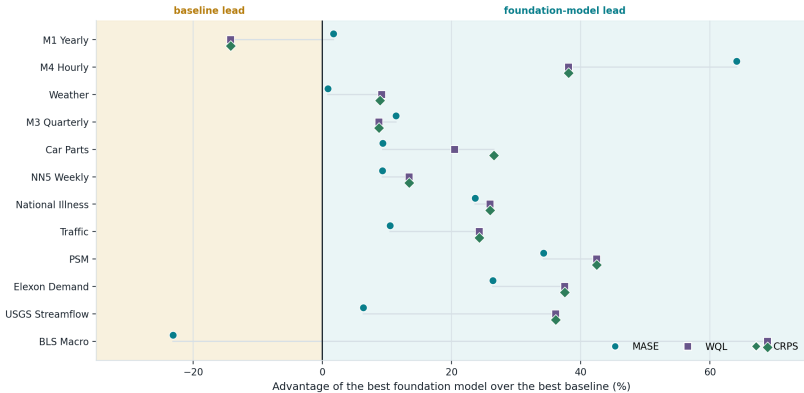
Use the Pettitt test for level changes, trend tests, seasonality strength, autocorrelation, and the share of zero values. Relate these measurements to the loss differences between model families.

Strengthen the statistical analysis

Add more rolling forecast origins. Use hierarchical models across datasets and series, and check how the MCS changes with different bootstrap blocks and significance levels.

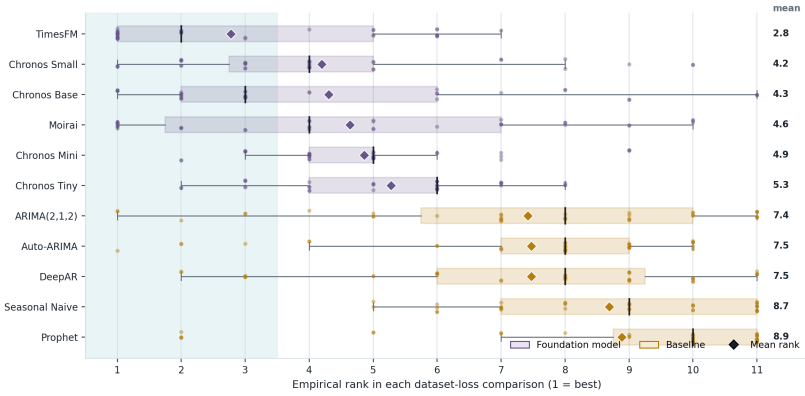
Goal: choose methods from measurable properties of the series, not from one overall ranking.

Appendix: how large is the family-level advantage?



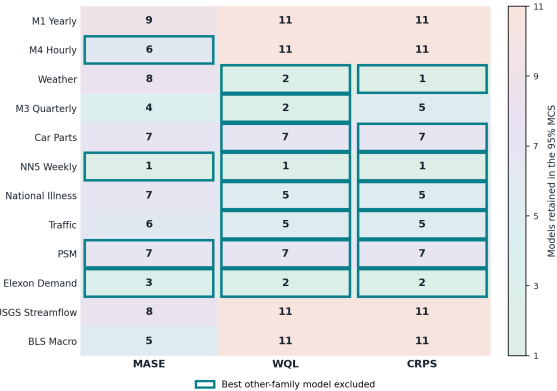
Foundation models lead in 33 of 36 comparisons, but the margin varies sharply. The three baseline wins are M1 Yearly WQL and CRPS, and BLS Macro MASE.

Appendix: rank consistency across datasets and losses



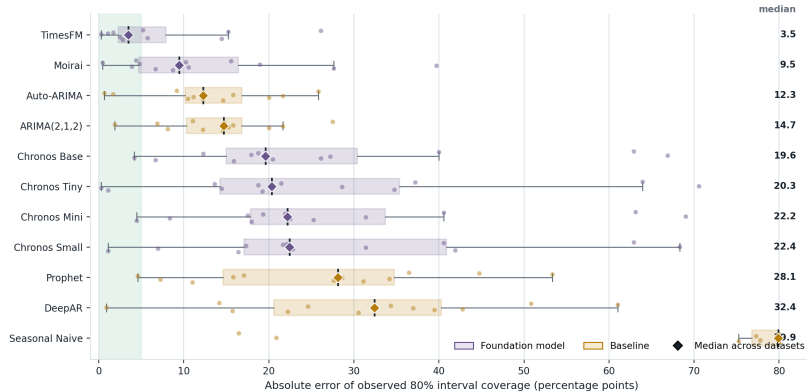
TimesFM has the best mean rank and the most consistently strong placement. The other foundation-model variants move more across datasets.

Appendix: where is the 95% MCS selective?



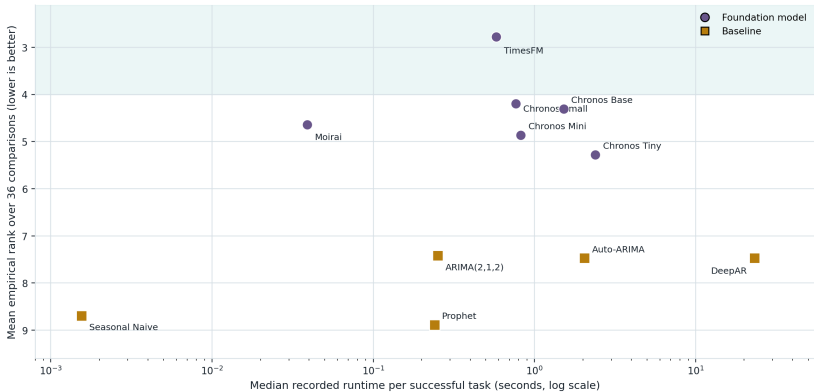
NN5 produces a one-model set for every loss. Several datasets retain many models even when their empirical winner is a foundation model.

Appendix: calibration of the 80% prediction interval



TimesFM has the smallest median coverage error across datasets. Coverage alone is not enough: WQL and CRPS also reward narrower, sharper forecasts.

Appendix: recorded runtime versus empirical rank



TimesFM has the best mean rank; Moirai combines a top-five mean rank with low recorded task time. Runtime values are specific to this benchmark setup.

Appendix: paired forecast tasks used by the MCS

Dataset	Main pattern	Series	MASE	WQL	CRPS	Evidence
Car Parts	Sparse demand, many zeros	20	39	31	40	adequate
M1 Yearly	Annual trends and level changes	20	60	60	60	adequate
M3 Quarterly	Trend and quarterly cycles	20	60	60	60	adequate
M4 Hourly	Daily cycles and noise	20	60	60	60	adequate
NN5 Weekly	Annual pattern, holidays, demand shifts	20	60	60	60	adequate
Weather	Zeros and seasonal rainfall	20	60	46	60	adequate
National Illness	Annual cycle and sudden peaks	1	3	3	3	very low
PSM	Anomalies and noisy sensors	5	15	15	15	low
Traffic	Daily and weekly cycles	5	15	15	15	low
BLS Macro	Long trends and structural changes	3	7	7	7	very low
Elexon Demand	Weekly cycle and weather effects	1	3	3	3	very low
USGS Streamflow	Seasonality, extremes, and gaps	5	13	13	13	low

A paired task is one series at one forecast origin after averaging the three model repetitions. Missing or undefined losses reduce some task counts.

Appendix: lowest loss and MCS size

Dataset	MASE	WQL	CRPS
Car Parts	**Moirai (7)**	**Moirai (7)**	**Moirai (7)**
M1 Yearly	**TimesFM (9)**	**ARIMA (11)**	**ARIMA (11)**
M3 Quarterly	**TimesFM (4)**	**TimesFM (2)**	**TimesFM (5)**
M4 Hourly	**Chronos Base (6)**	**Chronos Base (11)**	**Chronos Base (11)**
NN5 Weekly	**TimesFM (1)**	**TimesFM (1)**	**TimesFM (1)**
Weather	**Chronos Base (8)**	**Moirai (2)**	**Moirai (1)**
National Illness	**TimesFM (7)**	**TimesFM (5)**	**TimesFM (5)**
PSM	**Chronos Small (7)**	**Chronos Small (7)**	**Chronos Small (7)**
Traffic	**Moirai (6)**	**Moirai (5)**	**Moirai (5)**
BLS Macro	**Auto-ARIMA (5)**	**TimesFM (11)**	**TimesFM (11)**
Elexon Demand	**TimesFM (3)**	**TimesFM (2)**	**TimesFM (2)**
USGS Streamflow	**Moirai (8)**	**TimesFM (11)**	**TimesFM (11)**

Each entry is **empirical winner (number of models in the 95% MCS)**. foundation winner nonfoundation winner

Appendix: who remains in each confidence set?

Dataset	MASE	WQL	CRPS
Car Parts	6 F + 1 B	6 F + 1 B	6 F + 1 B
M1 Yearly	4 F + 5 B	6 F + 5 B	6 F + 5 B
M3 Quarterly	2 F + 2 B	**2 F + 0 B**	3 F + 2 B
M4 Hourly	5 F + 1 B	6 F + 5 B	6 F + 5 B
NN5 Weekly	**1 F + 0 B**	**1 F + 0 B**	**1 F + 0 B**
Weather	6 F + 2 B	**2 F + 0 B**	**1 F + 0 B**
National Illness	5 F + 2 B	4 F + 1 B	4 F + 1 B
PSM	6 F + 1 B	6 F + 1 B	6 F + 1 B
Traffic	5 F + 1 B	**5 F + 0 B**	**5 F + 0 B**
BLS Macro	2 F + 3 B	6 F + 5 B	6 F + 5 B
Elexon Demand	**3 F + 0 B**	**2 F + 0 B**	**2 F + 0 B**
USGS Streamflow	5 F + 3 B	6 F + 5 B	6 F + 5 B

F = foundation-model variant; B = nonfoundation baseline. Green cells exclude every baseline. Coral cells retain all 11 candidates.

Appendix: empirical wins by model and loss

Model	MASE	WQL	CRPS	Total
TimesFM	5	6	6	**17**
Moirai	3	3	3	**9**
Chronos Base	2	1	1	**4**
Chronos Small	1	1	1	**3**
ARIMA(2,1,2)	0	1	1	**2**
Auto-ARIMA	1	0	0	**1**
Other five models	0	0	0	**0**

33 foundation wins

3 baseline wins

TimesFM accounts for just over half of all cells. Moirai is the only model to win three or more cells under every loss.

Models with no empirical wins can still remain in broad confidence sets.

Appendix: family dominance is not MCS separation

Dataset	Strict foundation dominance	MCS sizes: MASE / WQL / CRPS	Paired tasks
NN5 Weekly	0 of 3	**1 / 1 / 1**	60 / 60 / 60
M4 Hourly	**3 of 3**	6 / 11 / 11	60 / 60 / 60
PSM	**3 of 3**	7 / 7 / 7	15 / 15 / 15
Traffic	2 of 3	6 / 5 / 5	15 / 15 / 15
Elexon Demand	**3 of 3**	**3 / 2 / 2**	3 / 3 / 3
USGS Streamflow	**3 of 3**	8 / 11 / 11	13 / 13 / 13
Car Parts	2 of 3	7 / 7 / 7	39 / 31 / 40

Strict dominance means the worst foundation model has lower aggregate loss than the best nonfoundation baseline.

MCS separation depends on paired task-level loss differences. It can be narrow without family dominance, as NN5 shows.

Appendix: how the 95% MCS was constructed

Sequential procedure

1. Form the task-by-model loss matrix $\mathbf{L} = (L_{i,m})$ for one dataset and loss.
2. Start with all 11 candidates, $\mathcal{M}_0$.
3. Test equal predictive ability within the current set using the range statistic and a stationary bootstrap.
4. After a rejection, remove the weakest model and repeat. Stop when equality is no longer rejected.

Implementation choices

- Confidence level: 95%.
- Bootstrap replications: 5,000.
- Block length: $\lceil \sqrt{n_{\text{tasks}}} \rceil$.
- Three repetitions are averaged inside each series-origin task before comparison.

Output: a set of models whose predictive ability is not separated from the best at the 5% level.

Traffic MASE example: 1 Moirai | 2 Chronos Base | 3 Chronos Tiny | 4 Chronos Small | 5 TimesFM | 6 Seasonal Naive

TimesFM is eliminated while the sixth-ranked model remains: membership depends on paired differences, not aggregate rank alone.

Appendix: limits of the evidence

What the MCS supports

Within one dataset and loss, it identifies the candidates that cannot be separated from the best under the chosen bootstrap design.

What it does not support

It does not establish a universal model hierarchy or prove that a retained model is equally good on every individual series.

Small panels

National Illness and Elexon contribute three paired tasks; BLS Macro contributes seven. Their sets are descriptive and low powered.

Dependence and scope

Series and origins are not fully independent. The stationary bootstrap mitigates this, but conclusions remain specific to these 12 datasets, 11 model variants, and three losses.

WQL and CRPS are computed from the saved q10–q90 forecasts; CRPS is a finite-quantile approximation rather than a score from a continuous CDF.