

An Information-Theoretic Diagnostic Analytics Framework for Mapping Past-Future Dependence in Horizon-Specific Forecastability

Peter M. Catt

Abstract— In many social, business, economic, and physical systems, the true data-generating process is unknown, requiring forecasters to rely exclusively on observed time series. This study proposes a pre-modeling diagnostic analytics framework for horizon-specific forecastability assessment that evaluates forecastability before model selection begins, enabling informed decisions about whether additional modeling effort is likely to justify its cost. Forecastability is operationalized using auto-mutual information at lag h , which quantifies how much past observations reduce uncertainty about future values, and is estimated via a k -nearest-neighbor estimator computed strictly on training data to preserve out-of-sample validity. The diagnostic signal is validated against realized out-of-sample symmetric mean absolute percentage error across 42,355 time series spanning six temporal frequencies, using benchmark and higher-capacity probe models under a rolling-origin protocol. The results reveal a strong frequency-dependent relationship between measurable dependence and realized forecast error: for five of six frequencies, auto-mutual information exhibits a consistent negative rank association with realized error, supporting its use as a forecast triage signal for modeling investment decisions, whereas the daily series shows weaker discrimination despite measurable dependence. Across all frequencies, median forecast error declines monotonically from low to high forecastability terciles, demonstrating clear decision-relevant separation. Overall, the findings establish measurable past-future dependence as a practical screening tool for analytics-driven forecasting strategy, identifying when advanced models are likely to add value, when simple baselines suffice, and when attention should shift from accuracy improvement to robust decision design, thereby supporting a diagnostic-first approach to modeling effort and resource allocation in organizational forecasting contexts.

Index Terms—Diagnostic analytics, Time series forecasting, Forecastability assessment, Forecast triage, Information theory, Resource allocation.

I. INTRODUCTION

Organisations routinely allocate forecasting effort without first assessing whether sufficient predictive information exists at the decision horizon. Some series respond well to sophisticated methods. Others show no improvement over simple baselines, regardless of model complexity or computational

investment [1]. The practical question is therefore which series justify sophisticated modelling effort and which do not.

Current practices rely on trial and error. Practitioners build models, evaluate performance, then refine or abandon approaches based on results. This is operationally inefficient. If a series lacks exploitable structure at relevant horizons, model tuning is unlikely to improve forecasts substantially. Time and resources should be allocated elsewhere or focused on mitigating forecast error impacts rather than reducing errors that cannot be reduced [2].

Forecastability assessment should happen before model building begins because modelling effort is only justified when predictive information is present at the horizons that matter for decisions [3]. The central question is therefore whether a series contains such information. Forecastability is a property of the time series itself, conditional on a declared information set, rather than a property of any particular forecasting model. In this paper, the information set is restricted to the univariate history of the series, that is, past realisations.

$$\mathcal{S}_t = \{X_t, X_{t-1}, \dots, X_{t-k}\} \quad (1)$$

We therefore study univariate forecastability conditional on the series' own history and exclude exogenous variables. The measure therefore evaluates intrinsic past-future dependence rather than model-specific explanatory power. A series with strong organised patterns is forecastable even if current models perform poorly. Conversely, a series with weak temporal structure is difficult to forecast regardless of method sophistication.

We estimate horizon-specific forecastability and validate it against realised accuracy; we do not claim a universal 'forecast horizon' boundary. The goal is operational triage, not theoretical characterisation of forecastability limits.

Classical variability measures like standard deviation or coefficient of variation do not address this. A volatile series with large seasonal swings can be easier to forecast than a smooth series drifting randomly. What matters is not how much a series varies, but how much the past tells us about the future [4]. High variability with strong patterns is forecastable. Low variability with random drift is not.

The challenge is making forecastability operational. A measure is required that can be computed from historical data alone, estimated before forecasting begins, and validated against actual out-of-sample performance. It must handle heterogeneous series lengths and acknowledge that forecastabil-

ity varies by horizon. A single score ignoring horizon can be misleading [5].

This paper adopts an information-theoretic view of forecastability as measurable past–future dependence [6, 7]. We define forecastability at horizon h as mutual information $I(\text{Past}; \text{Future}_h)$, the reduction in uncertainty about the future obtained by observing the past. High mutual information indicates organised, exploitable past–future dependence. Low mutual information indicates weak or unexploitable dependence between past and future observations. We make no assumptions about the underlying data-generating process, which may be latent and time-varying, and instead focus on measurable past–future dependence in the realised data. AMI therefore quantifies an observable consequence of the data-generating mechanism—namely, how much predictive information persists at a given horizon without requiring parametric, structural, or stationarity assumptions about its form.

AMI is interpreted ordinally (rank-based), not metrically; our conclusions depend on whether AMI correctly orders series by forecastability, not on absolute mutual information values or estimator-specific units.

Contributions. This study makes four contributions:

(1) It formalises forecastability as horizon-conditioned past–future dependence, distinct from model-specific exploitability.

(2) It operationalises this construct using auto-mutual information estimated via a non-parametric k -nearest-neighbour framework, providing a practical method for classifying series into action categories (invest in modelling / model cautiously / manage uncertainty) based on ex-ante forecastability assessment. This diagnostic classification framework is the primary contribution for decision support.

(3) It evaluates the empirical relationship between AMI and out-of-sample forecast error (sMAPE) across 42,355 M4 series spanning six frequencies. Three probe models (Seasonal Naïve, ETS, N-BEATS) spanning different representational capacities confirm that the relationship holds across model classes, though strength and direction vary.

(4) It derives a diagnostic workflow linking measured dependence structure to forecasting investment decisions by series and horizon.

The methodology is tested on M4 competition series across six frequencies [8]. The goal is not competition performance but decision analytics. The question is not ‘Can we beat benchmark X?’ but ‘Which series justify investment in methods beyond benchmark X?’ Heterogeneity in dependence across series implies differential returns to forecasting investment. The diagnostic addresses not ‘which model is best?’ but ‘is this series worth modelling at all at the decision horizon?’

We test AMI as a pre-modelling diagnostic under a fixed-information protocol with rolling-origin evaluation (10 origins per series). Diagnostics are computed strictly pre-origin from training data only, while forecast errors are evaluated post-origin and then averaged across origins to reduce

single-origin variance. Observed associations are empirical and frequency-conditional rather than universal laws of predictability. Conceptually, we separate *forecastability* (available past–future dependence) from *exploitability* (model-specific ability to convert that dependence into lower loss). This study proposes a diagnostic measure of forecastability, not a causal model of forecasting performance nor a new forecasting algorithm.

The remainder of the paper is organised as follows: Section II reviews related work, Section III describes the experimental protocol, Section IV reports the empirical results, Section V discusses implications and limitations, and Section VI concludes.

II. CONCEPTUAL FOUNDATIONS

A. The Forecastability Problem

Forecasting competitions from M1 onwards show that simple methods remain competitive on average [9–11]. This is sometimes misinterpreted as ‘sophistication doesn’t help’. The correct interpretation is that average forecastability across competition portfolios is moderate. Sophisticated methods excel on high-structure series but struggle on low-signal series where they overfit noise. Simple methods perform adequately on low-signal series but underperform on high-signal series. The average difference narrows.

This supports rather than contradicts the forecastability thesis. Different series have materially different information content. Method selection should respect this. Applying sophisticated methods to all series wastes resources on those that cannot benefit.

B. Existing Approaches and Their Limitations

The forecastability problem has been addressed through statistical, information-theoretic, model-based, and practical approaches, each with distinct strengths and limitations.

Statistical measures. The coefficient of variation ($CV = \sigma/\mu$) is widely used in practice to assess forecast difficulty [12], but CV conflates signal and noise. A series with large seasonal swings (high CV) may be easier to forecast than one with small random drift (low CV). Classical measures like autocorrelation functions and spectral decomposition identify structure but do not aggregate information into horizon-specific forecastability scores.

Information-theoretic and spectral approaches. Entropy-based measures quantify regularity and pattern complexity [13]. Sample entropy [14] measures the likelihood that similar patterns persist, with lower entropy indicating higher predictability. Catt [15] demonstrated that sample entropy provides useful ex-ante indication of forecast accuracy on M3 data, though subject to limitations with intermittent demand and structural breaks. Goerg [16] introduced Forecastable component analysis (ForeCA), defining forecastability via spectral entropy, with lower spectral entropy indicating higher forecastability. These measures capture pattern com-

plexity that CV cannot, but they produce single global scores that do not vary by horizon.

Model-based and dynamical systems approaches. In econometrics, predictability is measured through variance decomposition or predictive R^2 [17]. Lyapunov exponents quantify sensitivity to initial conditions in dynamical systems, with positive exponents indicating chaos and limited predictability [18]. These methods assume deterministic dynamics and require long, high-quality series for reliable estimation.

Practical and empirical approaches. Forecast Value Added (FVA) evaluates forecast performance relative to a simple baseline [12]. If sophisticated methods cannot beat simple baselines, the series has low forecastability or the methods are poorly specified. FVA has proven valuable in operational contexts for identifying which forecasting activities add value, but it requires actual forecasting attempts and cannot be computed ex-ante from historical data alone. Selvam et al. (2024) demonstrate that algorithm choice substantially affects forecast errors across horizons for short univariate time series, illustrating the material consequences of method selection decisions [19].

Limitations of existing approaches. Two practical limitations constrain existing approaches in operational settings. First, horizon-dependence is ignored and explicit alignment with forecast horizons is absent. Nearly all existing metrics produce single-valued scores (CV, ApEn, SampEn, spectral entropy, Lyapunov exponents). A series highly predictable at 1-step ahead may be unpredictable at 12-steps ahead, or vice versa. The relationship between short-term and long-term forecastability is complex and horizon-specific assessment is essential [5]. Spectral entropy summarises frequency-domain concentration but does not map directly to h -step-ahead prediction difficulty. Sample entropy provides a global regularity measure but does not specify which horizons benefit from that regularity. Lyapunov exponents indicate long-term predictability horizons in chaotic systems but not specific h -step-ahead difficulty. Second, comprehensive empirical validation is limited. Sample entropy showed promise in initial M3 studies [15] but comprehensive validation across diverse frequencies, horizons, and methods was not performed. Lyapunov exponents require long, well-resolved series to be estimated reliably and tend to perform poorly on the short, sparse series common in business forecasting settings [20]. CV continues to be used despite weak theoretical justification and inconsistent empirical support.

Our approach. We address these limitations through horizon-specific auto-mutual information (AMI). Spectral entropy measures overall frequency-domain concentration, but AMI at lag τ directly quantifies dependence between present values and values τ -steps in the past. Unlike entropy or spectral measures, AMI is explicitly indexed to forecast horizon, which is the key object of decision relevance. Sample entropy requires parameter selection, whereas AMI has a well-established non-parametric estimation framework via

k -nearest neighbours. ACF captures only linear dependencies, but AMI captures both linear and nonlinear dependencies through its information-theoretic foundation. The key advantage of our framework is explicit lag-horizon mapping: $\text{AMI}(\tau)$ aligns to τ -step-ahead forecast difficulty. We compute AMI at lags corresponding to each forecast horizon h , making the connection between measurement and prediction operationally transparent. We then validate this framework comprehensively across M4 series, six frequencies, multiple horizons, and three forecasting methods, providing the empirical grounding that previous entropy-based approaches lacked.

C. Information-Theoretic Framework

Shannon [21] defined entropy to quantify uncertainty in random variables:

$$H(X) = -\sum p(x) \log p(x) \quad (2)$$

For time series prediction, the relevant quantity is not marginal entropy $H(\text{Future})$ but conditional entropy $H(\text{Future} | \text{Past})$. The difference is mutual information:

$$I(\text{Past}; \text{Future}) = H(\text{Future}) - H(\text{Future} | \text{Past}) \quad (3)$$

This quantity captures the reduction in uncertainty about the future attributable to the past.

For operational implementation, we use auto-mutual information (AMI), which quantifies shared information between a time series and its lagged self. At lag τ :

$$\text{AMI}(\tau) = I(X_t; X_{t+\tau}) = H(X_{t+\tau}) - H(X_{t+\tau} | X_t) \quad (4)$$

This is the mutual information between past and future. It measures how much the past tells you about the future. High predictive information means there's structure available to exploit. Entropy and mutual information are not redundant concepts. Two processes can have the same entropy rate but different 'organised structure'. AMI is the information-theoretic generalisation of the autocorrelation function, capturing both linear and nonlinear temporal dependencies [22, 23].

Bialek et al. (2001) formalised this as predictive information. Related constructs in computational mechanics include excess entropy and statistical complexity [24, 25]. The framework has deep precedent in physics, neuroscience, and computational theory where prediction and information processing are fundamental concerns [26].

1) Forecastability, Exploitability, and Validation

A key distinction must be made between forecastability as dependence and forecastability as exploitability. AMI measures statistical dependence between past and future, a property of the data-generating process that exists independently of any forecasting method. However, whether that dependence translates into reduced forecast error depends on whether the forecasting method can represent and exploit the relevant structure.

AMI is therefore a diagnostic for available dependence, not a guarantee of error reduction under any particular loss function or hypothesis class. We validate AMI against realised sMAPE using Seasonal Naïve, ETS, and N-BEATS as probes of exploitability across different model classes. Where AMI correlates negatively with sMAPE, the dependence is being exploited. Where correlation is weak or absent, the mismatch is itself informative: it may indicate representation limitations (the model cannot capture the dependence structure), nonstationarity (in-sample dependence does not persist out-of-sample), or insufficient sample size (the dependence cannot be reliably estimated or learned).

III. EXPERIMENTAL PROTOCOL

This study evaluates whether a pre-modelling diagnostic of temporal dependence can predict realised forecast accuracy under a realistic evaluation protocol. The diagnostic of interest is horizon-specific auto-mutual information (AMI), estimated strictly from historical data available before any forecast origin. Forecast accuracy is assessed using symmetric mean absolute percentage error (sMAPE) under an expanding-window rolling-origin evaluation with 10 origins per series.

The protocol enforces asymmetric temporal separation between diagnostics and outcomes. Diagnostics are computed once per series from a fixed base training window that excludes all evaluation data. Forecast errors are computed at each rolling origin and averaged across origins to reduce single-origin variance. This asymmetry is deliberate: AMI functions as a genuine pre-modelling screening tool, not an in-sample explanatory variable recomputed as information accumulates.

A. Data

The M4 competition dataset [8] comprises 100,000 time series across six sampling frequencies spanning diverse domains including finance, industry, demographics, and macroeconomic indicators. Spiliotis et al. [27] validated the representativeness of the M4 dataset by comparing its time series feature distributions with those of earlier competition datasets (M1, M3, Tourism) using instance space analysis and Kullback–Leibler divergence. Their analysis confirmed that the M4 dataset, randomly sampled from approximately 900,000 real-world business series, provides broad coverage of the feature space defined by trend, seasonality, spectral entropy, and distributional characteristics, making it a credible empirical testbed for evaluating forecasting diagnostics across diverse series types. Table 1 summarises the dataset structure.

The maximum forecast horizon H_{\max} and seasonal period m follow M4 conventions. These horizon sets define the evaluation window length at every rolling origin. The original M4 test splits are not used directly; instead, rolling-origin windows are constructed entirely within the available historical data for each series.

Table 1. M4 Dataset Structure by Frequency

Frequency	Series	H_{\max}	Seasonal period m
Yearly	23,000	6	1 (non-seasonal)
Quarterly	24,000	8	4
Monthly	48,000	18	12
Weekly	359	13	52
Daily	4,227	14	7
Hourly	414	48	24
Total	100,000		

B. Inclusion Principle: Exhaustive Over the Feasible Domain

This study does not construct fixed-size survivor panels and does not apply category balancing or stratified sampling. Instead, all M4 series are attempted, and a series is included if and only if it satisfies pre-specified feasibility conditions that ensure the diagnostic and evaluation are well-defined and numerically stable.

A series is included for a given frequency if it satisfies all of the following conditions:

(i) Rolling-origin feasibility. The training history must support an expanding-window rolling-origin design with exactly 10 origins ($\text{ROLLS} = 10$) and a full evaluation window of length H_{\max} at each origin. Concretely, the series must be long enough to provide a fixed base training segment, plus a rolling evaluation pool of length $\text{pool_len} = H_{\max} + (\text{ROLLS} - 1) \times \text{ROLL_STEP}$, where $\text{ROLL_STEP} = 1$. Series that cannot support 10 complete rolling-origin evaluations are excluded.

(ii) Scale feasibility. A scale proxy, scale_0 , must be defined on the base training window using mean absolute seasonal (or first) differences. Series with undefined scale (e.g., too short after differencing, degenerate values) are excluded.

(iii) Scale floor filter. For each frequency, a scale floor is computed as the 5th percentile ($q = 0.05$) of scale_0 across all candidate series in that frequency. Series whose base-training scale_0 falls below this floor are excluded. This prevents pathological near-constant series from dominating error behaviour or causing numerical instability.

(iv) AMI feasibility at the worst horizon. AMI must be computable at the maximum horizon H_{\max} on the base training window under a frequency-specific minimum effective sample size requirement. The effective sample size at horizon h is $n_{\text{eff}} = T_{\text{base}} - h$. If n_{eff} falls below the frequency-specific threshold, AMI is treated as undefined and the series is excluded. The minimum effective sample size thresholds are: Yearly 30, Quarterly 80, Monthly 100, Weekly 120, Daily 250, Hourly 400.

These feasibility conditions jointly define the domain of applicability for the diagnostic. Exclusions are not treated as sampling bias but as explicit acknowledgement that forecastability diagnostics are meaningful only where they can be

Single-Series Forecastability Intuition

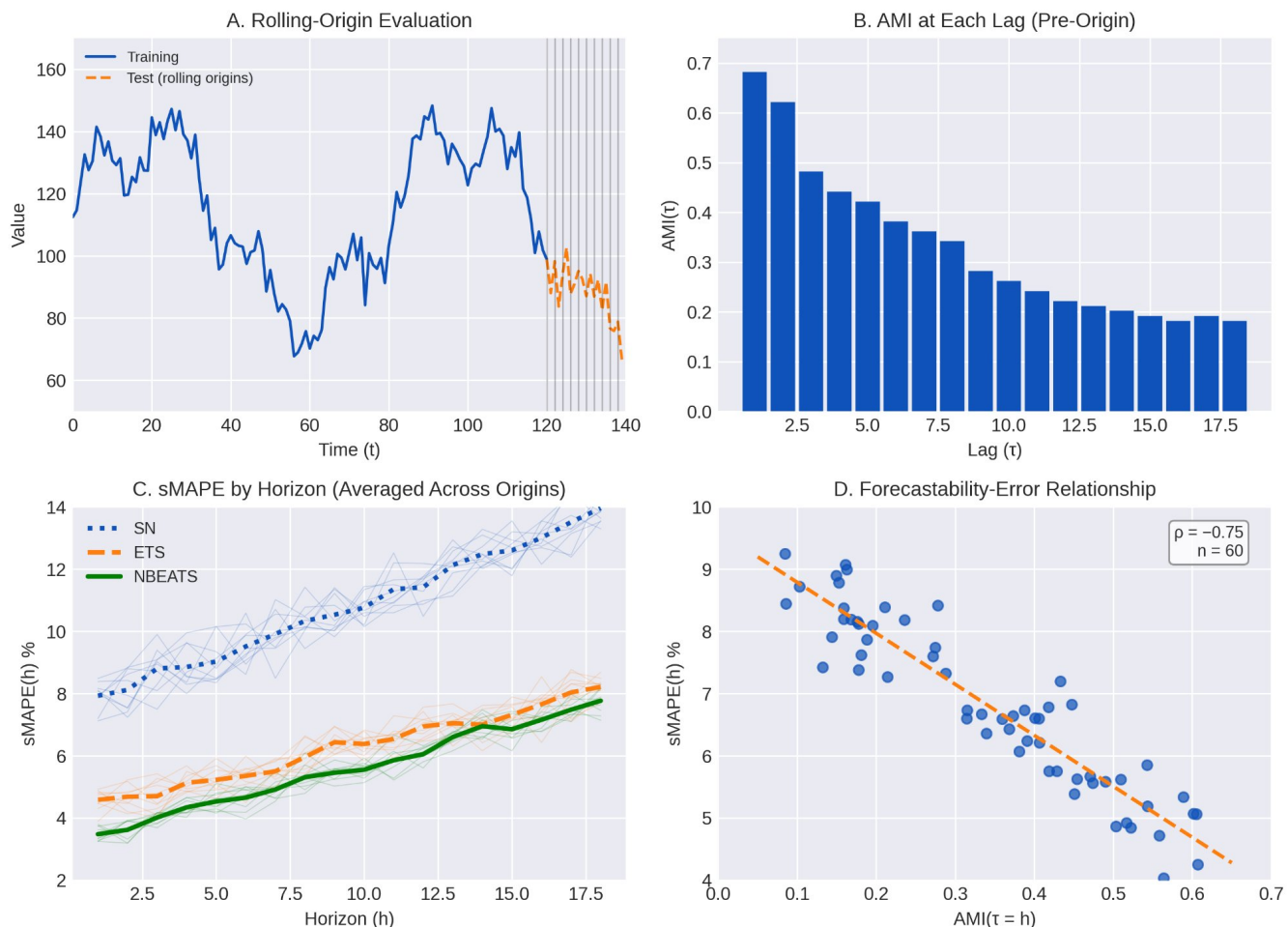


Figure 1. Conceptual forecastability intuition. (A) Expanding-window rolling-origin evaluation with 10 origins. (B) Horizon-specific $AMI(h)$ computed once per series using a fixed base training window that excludes all evaluation windows. (C) sMAPE evaluated post-origin for each horizon and then averaged across origins. (D) Relationship tested empirically: within frequency and horizon, higher $AMI(h)$ is associated with lower realised mean sMAPE(h) across series.

reliably computed and where the evaluation protocol can be executed as specified. A diagnostic that cannot be estimated on a series cannot support pre-modelling decisions for that series.

C. Rolling-Origin Evaluation

For each included series, forecasts are generated using an expanding-window rolling-origin design with exactly 10 origins. The origins are positioned in the final portion of available history such that each origin supports a full H_{\max} -step evaluation window. At origin o : (1) models are estimated (or, for the global model, applied) using only data available up to o ; (2) point forecasts are generated for all horizons $h = 1, \dots, H_{\max}$; (3) sMAPE is computed for each horizon.

For each series and horizon, the reported error metric is the mean sMAPE across the 10 origins. Results are retained only

when all 10 origin-level errors exist for that series–horizon pair; no partial averaging is performed. This reduces single-origin variance while maintaining strict pre-/post-origin separation. sMAPE is bounded between 0 and 2 (or 0% to 200%), scale-independent, and symmetric in over- and under-prediction [28]. We report sMAPE as a percentage in all tables and figures.

D. Probe Model 1: Seasonal Naïve (SN)

Seasonal Naïve repeats the observation from the same seasonal position one cycle earlier. Formally:

$$\hat{y}_{t+h|t} = y_{t+h-km}, \quad k = \lceil h/m \rceil \quad (5)$$

where m is the seasonal period (12 Monthly, 4 Quarterly, 52 Weekly, 7 Daily, 24 Hourly). For Yearly data ($m = 1$), the

AMI Illustration Across Canonical Processes

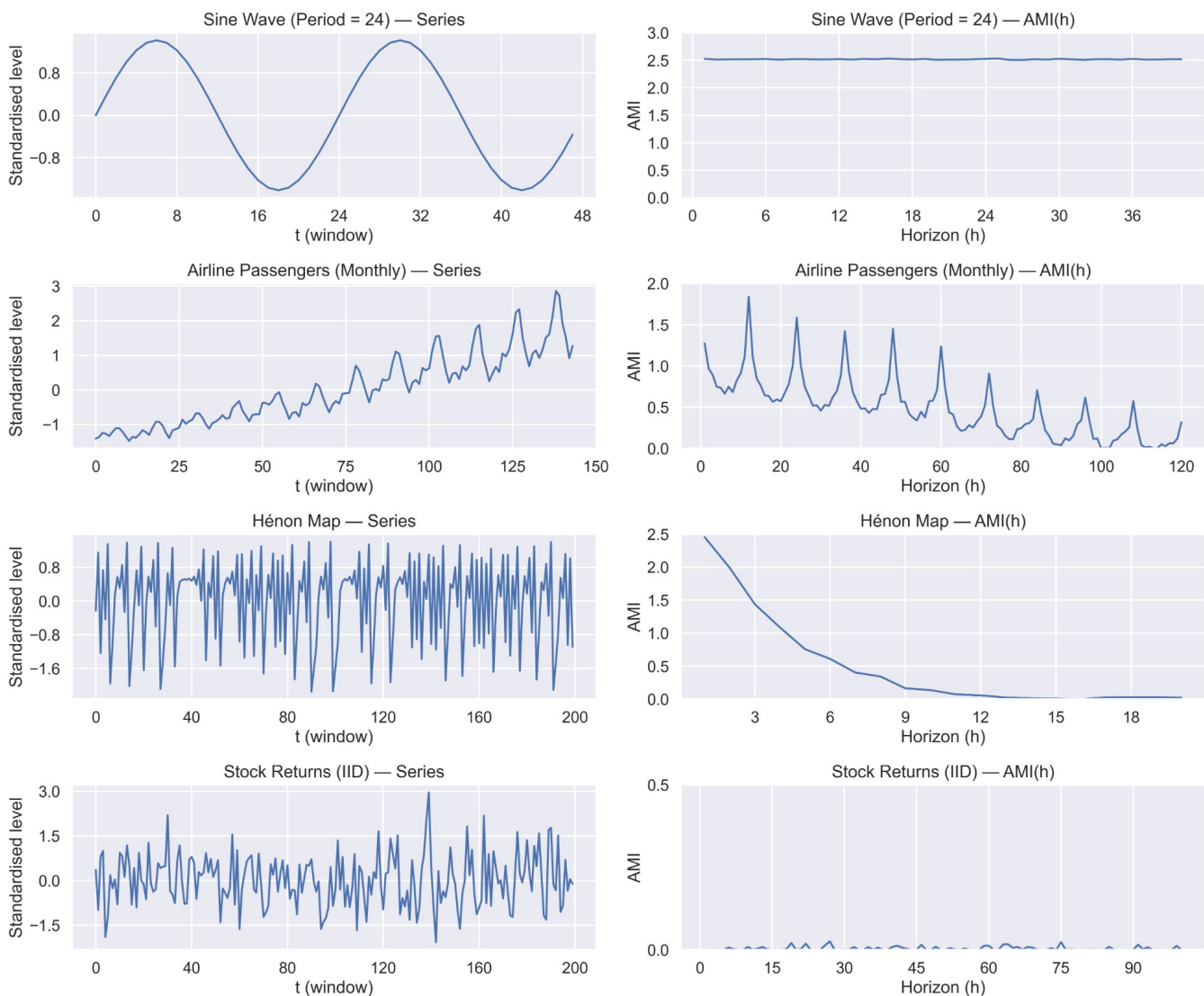


Figure 2. Schematic illustration of the AMI diagnostic framework. AMI is computed once per series from training data alone. The resulting forecastability profile informs pre-modelling triage decisions across frequencies and horizons.

method reduces to a simple naïve forecast repeating the most recent observation.

Seasonal Naïve represents minimal modelling effort and provides a low-capacity baseline.

E. Probe Model 2: ETS Specification

ETS forecasts are generated using the exponential smoothing state-space framework with automatic selection over: trend (none, additive, additive damped) and seasonality (none, additive, and multiplicative when admissible).

Model selection is performed by minimising AIC among the candidate configurations using `statsmodels.tsa.holtwinters.ExponentialSmoothing` [29]. Multiplicative seasonality is permitted only when all observations in the

fitted window are strictly positive. Box-Cox transformation is disabled.

The seasonal period is fixed by frequency (as above) for comparability across series. Point forecasts are produced for all horizons $h = 1, \dots, H_{\max}$ at each origin.

F. Probe Model 3: N-BEATS Specification (Global, Per Frequency)

N-BEATS [30] is implemented using the `neuralforecast` library [31]. A single global model is trained per frequency on a pooled set of base training windows (i.e., the same base-window definition used for AMI, excluding the rolling evaluation tail for each pooled series). The pool size is capped for computational feasibility (default 5,000 series per frequency;

15,000 for Yearly), with pooling series randomly selected using a fixed seed to ensure reproducibility. The training loss is sMAPE, aligning the optimisation objective with the evaluation metric.

Key implementation detail (time axis). For all frequencies, a synthetic daily timestamp index is used internally for NeuralForecast to avoid datetime overflow and to keep a consistent interface across frequencies:

$$ds_i = ds_{\text{base}} + i \text{ days} \quad (6)$$

The frequency passed to the library is therefore “D” for all frequencies. This is an implementation convenience and does not alter the sequence order or information content. Seasonal periods and forecast horizons are defined in observation steps, not calendar time; the synthetic timestamp exists solely to satisfy library interface requirements and does not affect model semantics.

Rolling-origin application. The trained per-frequency global model is applied under the same rolling-origin protocol. At each origin, N-BEATS is asked to forecast using only the history available up to that origin (no future leakage). Forecasts are generated for all horizons $h = 1, \dots, H_{\text{max}}$.

G. AMI Computation

AMI at horizon h quantifies shared information between present observations and observations h steps ahead. This is the mutual information between past and future at lag h —the reduction in uncertainty about X_{t+h} obtained by observing X_t . AMI generalises the autocorrelation function to capture both linear and nonlinear temporal dependencies [22, 23].

For each series, AMI is computed on standardised values (zero mean, unit variance) of the base training window. If standardisation is undefined (e.g., constant series), AMI is treated as undefined and the series is excluded by the feasibility gates.

kNN estimator: AMI is estimated using the Kraskov–Stögbauer–Grassberger (KSG) k -nearest-neighbour mutual information estimator [32] with $k = 8$. The kNN estimator avoids discretisation and explicit density estimation, providing consistent estimates for continuous real-valued time series without binning artifacts that affect histogram-based Shannon entropy approaches. The choice of $k = 8$ reflects a bias–variance tradeoff explicitly acknowledged in the estimation literature [32, 33]. Increasing k reduces variance but increases bias through oversmoothing; decreasing k reduces smoothing but increases variance and sensitivity to local noise. Since validation in this study relies on rank association (Spearman ρ) rather than absolute MI magnitudes, a moderate k suffices for stable cross-series ordering under heterogeneous series characteristics. Mild monotone bias in AMI does not invalidate the validation objective, provided the diagnostic preserves ordinal information.

Critical implementation detail: AMI is not recomputed per roll. AMI(h) is computed once per series and horizon using the base training window that excludes all rolling evaluation windows. It is then held fixed and paired with the

mean rolling-origin sMAPE(h). This ensures the diagnostic is strictly ex-ante and avoids “moving target” behaviour where the diagnostic changes with each origin.

H. Validation

Validation proceeds within each frequency and model as follows. First, for each horizon h , compute Spearman’s rank correlation ρ_h across series between AMI(h) (fixed, computed pre-origin) and mean sMAPE(h) across the 10 rolling origins. Second, aggregate the horizon-level correlations by taking the mean Spearman ρ across horizons within each (frequency, model). Critically, AMI at horizon h is correlated exclusively with forecast error at the same horizon; no cross-horizon pooling is performed in the primary analysis.

A negative ρ indicates that higher AMI (stronger past–future dependence) associates with lower forecast error, consistent with the forecastability hypothesis. This two-stage aggregation—per-horizon Spearman across series, then average across horizons—preserves horizon-specificity (avoiding conflation of fundamentally different forecast distances) while providing interpretable summary statistics for cross-frequency and cross-model comparison.

We additionally report tercile-based analysis: series are partitioned into Low, Medium, and High AMI terciles within each frequency across all valid (series, horizon) pairs, and median sMAPE is compared across terciles to assess decision-relevant separation.

I. Scope and Interpretation

The study is diagnostic rather than explanatory. It does not aim to infer data-generating processes, estimate causal effects, or optimise forecasting performance. The tested proposition is narrow and operationally motivated: whether a fixed, pre-modelling diagnostic of past–future dependence provides decision-relevant ranking of series by forecast difficulty under a realistic rolling-origin protocol.

We distinguish *forecastability* (available past–future dependence, a property of the series) from *exploitability* (the degree to which a given model class converts that dependence into reduced forecast error). AMI measures the former; the probe models test the latter across different representational capacities. Where AMI correlates negatively with sMAPE, dependence is being exploited. Where correlation is weak or absent, the mismatch may reflect representation limitations, nonstationarity, or estimation noise, but each informative for diagnostic deployment.

IV. RESULTS

This section reports empirical findings from the experimental protocol described in Section III. We present sample coverage, the relationship between forecastability measures (AMI) and realised forecast error (sMAPE), robustness analysis by training length, and decision utility of AMI-based triage.

Horizon Profile of Training-Only AMI by Frequency

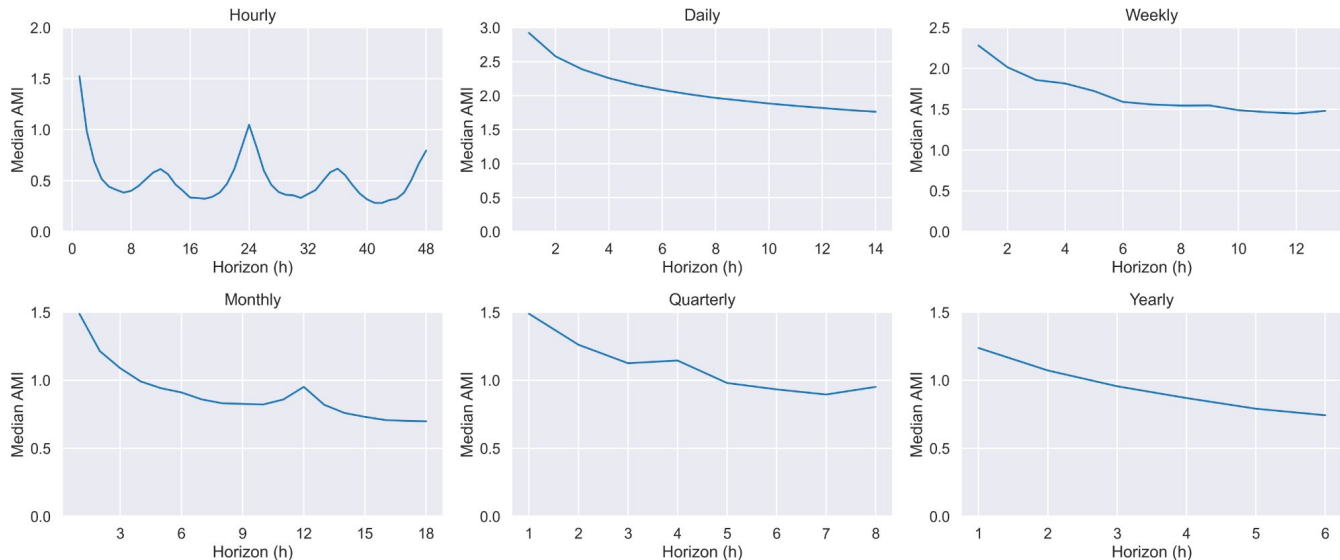


Figure 3. Realised forecast difficulty by horizon and frequency. Each panel shows median sMAPE across survivor series, where each series’ sMAPE(h) is first averaged across the 10 rolling origins. Median error generally increases with horizon, with occasional local non-monotonicity. Horizon ranges are frequency-specific (Hourly 1–48, Daily 1–14, Weekly 1–13, Monthly 1–18, Quarterly 1–8, Yearly 1–6).

A. Sample Coverage

Applying the feasibility gates exhaustively across all 100,000 M4 series yielded the following survivor counts: Hourly ($n = 393$), Daily ($n = 3,245$), Weekly ($n = 275$), Monthly ($n = 29,240$), Quarterly ($n = 8,219$), and Yearly ($n = 983$), totalling 42,355 series. All three models (Seasonal Naïve, ETS, and N-BEATS) produced forecasts at all horizons for the survivor series included in the reported analyses.

B. Probe Model Context

Before examining the core AMI-sMAPE relationship, we briefly characterise the probe models to establish that they span different representational capacities and that subsequent results are not artefacts of a particular forecasting architecture. Probe models are used to test whether forecastability rankings persist across model capacity, not to compete for best accuracy; relative probe performance is incidental to the diagnostic question. Expanding the model set would increase computational cost without changing the diagnostic question.

Probe models differ in absolute accuracy by frequency; these differences are incidental to the diagnostic objective. N-BEATS and ETS are the stronger discriminators across all frequencies, with ETS strongest at Weekly ($\rho = -0.72$) and N-BEATS most consistent overall. Seasonal Naïve serves as a minimal-complexity baseline throughout. Daily exhibits weak cross-series discrimination by AMI across all three probes ($\rho \approx -0.09$ to -0.10), despite non-trivial absolute dependence in the series.

Across all frequencies, forecast error increases monotonically with horizon (Fig. 3), confirming that horizon itself

drives forecast difficulty.

C. Forecastability Measures and Realised Error

A clarification is essential before presenting results. We do not estimate which frequency is harder overall (absolute difficulty); we estimate whether AMI ranks series within a frequency by forecast difficulty at that frequency’s horizons (cross-series discrimination). The contribution is within-frequency triage, not between-frequency difficulty rankings.

Fig. 4 illustrates the decay of forecastability (AMI) with horizon. Auto-mutual information profiles computed on training data exhibit systematic decay with lag across all frequencies, consistent with diminishing past–future dependence as forecast distance increases. Lag axes reflect frequency-specific horizon ranges. This decay confirms that forecastability is inherently horizon-specific, reinforcing the need for horizon-aligned diagnostics.

The AMI-sMAPE relationship varies systematically by frequency, with strong negative associations at some frequencies and materially weaker associations at others. We assess whether AMI behaves as an operational measure of forecastability by examining Spearman rank correlations between AMI and sMAPE. Consistent with the interpretation of AMI as an ordinal diagnostic, we report rank association rather than linear fit: the objective is stable ordering of series by expected difficulty within each frequency, not prediction of absolute error levels.

Table 2 reports the primary validation: Spearman rank correlations computed per horizon across survivor series, then averaged across horizons within each (frequency, model). Under this protocol, Hourly, Weekly, Quarterly,

Realised Forecast Difficulty by Horizon and Frequency

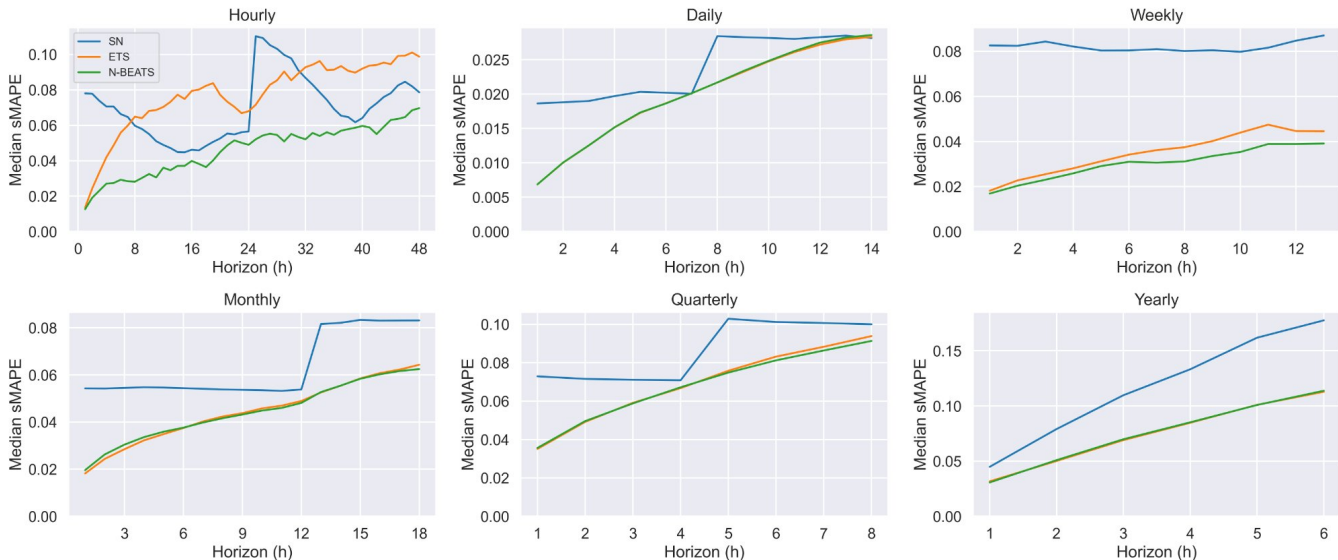


Figure 4. AMI decay with horizon. Profiles show the median AMI(h) across survivor series. AMI(h) is computed once per series on the base training window. Auto-mutual information profiles computed on training data exhibit systematic decay with lag across all frequencies, consistent with diminishing past–future dependence as forecast distance increases. Lag axes reflect frequency-specific horizon ranges.

Table 2. Forecastability–Accuracy Relationship (AMI Validation)

Frequency	Seasonal Naïve	ETS	N-BEATS
Hourly	−0.51	−0.41	−0.55
Daily	−0.09	−0.10	−0.10
Weekly	−0.29	−0.72	−0.70
Monthly	−0.32	−0.45	−0.47
Quarterly	−0.47	−0.58	−0.62
Yearly	−0.22	−0.50	−0.51

Notes: Spearman rank correlation computed per horizon across survivor series, then averaged across horizons within each (frequency, model).

and Yearly series show strong negative rank association between training-only AMI and realised out-of-sample sMAPE across all probe models. Weekly exhibits the strongest relationships for the higher-capacity probes (ETS $\rho = -0.72$, N-BEATS $\rho = -0.70$), consistent with AMI identifying dependence that is exploitable by both classical state-space structure and global nonlinear architectures. Hourly shows substantial association, with Seasonal Naïve exhibiting the strongest rank association ($\rho = -0.51$) and N-BEATS now at $\rho = -0.55$ (strengthened relative to ETS at $\rho = -0.41$ following alignment of training loss with the evaluation metric), and the relationship remains strong at Quarterly (N-BEATS $\rho = -0.62$) and Yearly (N-BEATS $\rho = -0.51$). Monthly exhibits substantial negative association for ETS ($\rho = -0.45$) and N-BEATS ($\rho = -0.47$), with Seasonal Naïve weaker ($\rho = -0.32$).

Daily is an empirical weak-discrimination case under this

protocol. Although AMI indicates non-trivial past–future dependence within daily series, its rank association with realised sMAPE is materially weaker than for other frequencies (approximately $\rho = -0.09$ to -0.10 across probes). This indicates reduced cross-series discriminative power for triage: variation in AMI corresponds to comparatively smaller variation in realised error within the daily survivor panel. This does not imply absence of dependence, but it does limit the usefulness of AMI as a within-frequency screening structure for Daily in the present setting.

As a robustness check, we confirmed that the qualitative conclusions in Table 2 are unchanged under alternative aggregation choices, including pooling horizons within frequency to compute a single Spearman correlation and reporting the median rather than the mean of horizon-specific correlations. These alternatives change magnitudes (as expected) but do not induce sign reversals or frequency-level reordering, indicating that the results reflect stable within-frequency rank association rather than artefacts of a specific summarisation.

Fig. 5 presents a heatmap showing the Spearman correlation between AMI and sMAPE by frequency and horizon. Blue cells indicate negative correlation (higher AMI associated with lower error), confirming that AMI serves as a consistent forecastability indicator for most frequencies, with Daily exhibiting weak and non-monotone association. The contribution lies in ordinal separation for triage, not linear predictability.

Role of categories relative to frequency. Series categories are used solely for post-hoc robustness checks to verify that results are not driven by category composition. To assess

Forecastability–Accuracy Association by Horizon (N-BEATS Probe)

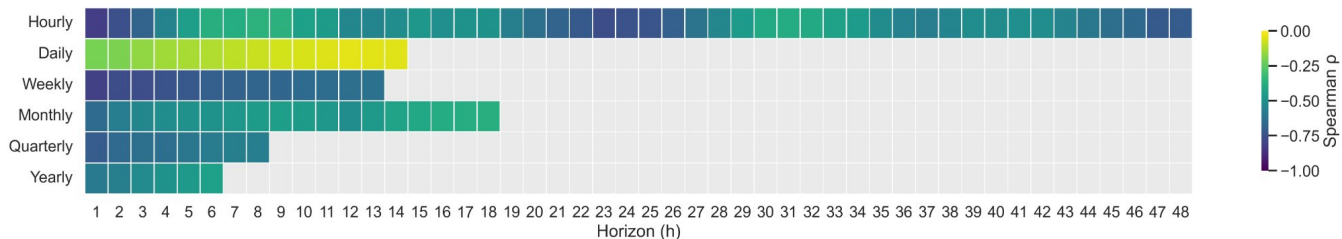


Figure 5. Forecastability–Accuracy Association by Frequency and Horizon (N-BEATS probe). Spearman correlation (ρ) between AMI and sMAPE by frequency and horizon. Each cell shows series-level Spearman ρ computed across all survivor series within each (frequency, horizon) combination; blue indicates negative correlation (higher AMI associated with lower error). Cells are shown only for horizons defined for each frequency (Hourly 1–48, Daily 1–14, Weekly 1–13, Monthly 1–18, Quarterly 1–8, Yearly 1–6); blank regions are not missing data.

Table 3. AMI–sMAPE Correlations by Training Length Tercile (N-BEATS)

Frequency	Short	Medium	Long
Hourly	+0.01	−0.56	−0.56
Daily	−0.04	−0.10	−0.15
Weekly	−0.47	−0.63	−0.85
Monthly	−0.47	−0.37	−0.51
Quarterly	−0.49	−0.61	−0.67
Yearly	−0.38	−0.56	−0.55

Notes: Series are

stratified into terciles by training length within each frequency. Values are Spearman ρ between AMI and sMAPE within each tercile. N-BEATS reported as an illustrative global probe; ETS shows qualitatively similar patterns.

whether semantic domain materially alters the relationship between dependence and forecast difficulty, we examined the direction of the AMI–sMAPE association within categories for each frequency. Across frequencies, we do not observe systematic category-level sign reversals relative to the corresponding frequency-level pattern. While association strength varies across categories and statistical power is limited in some cells (notably for Weekly series), the direction of association is broadly consistent. Accordingly, category is treated as a secondary conditioning dimension: forecastability is shaped primarily by temporal resolution and forecast horizon, with semantic domain playing a supporting rather than dominant role.

D. Robustness: Training Length Effects

A natural concern is whether AMI–sMAPE correlations are confounded by series length. Stratified analysis within each frequency shows that negative correlations persist across nearly all training-length terciles, with the exception of short Hourly series where the association is near zero (Table 3), indicating that the forecastability structure is not merely a proxy for series length. We report tercile stratification for the global probe (N-BEATS) as an illustrative check; results are qualitatively similar for ETS.

E. Decision Utility of Forecastability Assessment

The practical value of forecastability assessment lies in its ability to inform resource allocation decisions before forecasting begins [34]. Table 4 demonstrates this by showing median sMAPE conditional on AMI terciles.

Across all six frequencies, median sMAPE decreases monotonically from Low to High AMI terciles for each probe model (Table 4), confirming that AMI supports decision-relevant ordinal separation even when rank correlations are weaker. The discrimination is steepest for Hourly and Weekly: for ETS, median sMAPE falls by approximately 95% from Low to High AMI terciles at Hourly (14.05 to 0.64) and by approximately 87% at Weekly (8.15 to 1.08); N-BEATS shows strong separation (Hourly 9.76 to 0.85, 91%; Weekly 6.65 to 0.95, 86%). These gradients indicate that AMI-based triage can function as an operational screening layer: high-AMI series justify investment in model development and monitoring, while low-AMI series are better served by baselines and by shifting effort towards consequence mitigation and robust decision design. Tercile boundaries are equal-frequency and serve as a presentation device; the underlying ranking pattern is continuous and is captured more precisely by the Spearman correlations reported in Tables 2 and 3.

V. DISCUSSION

These results support a simple conceptual point: forecastability is not a single scalar property of a series, but a horizon-specific relationship between past and future under a declared information set. Forecast models operate within these informational constraints; they do not create predictive information where the series contains little usable past–future dependence at the decision horizon.

A. The AMI–Error Relationship is Frequency-Conditional

The dominant finding of this study is that the AMI–sMAPE relationship is strongly frequency-conditional. At Hourly, Weekly, Monthly, Quarterly, and Yearly frequencies, AMI computed from training data provides a consistent negative

Table 4. Decision Utility — Median sMAPE (%) by AMI Tercile

Frequency	SN Low	SN Mid	SN High	ETS Low	ETS Mid	ETS High	NB Low	NB Mid	NB High
Hourly	11.80	9.19	1.46	14.05	12.23	0.64	9.76	7.31	0.85
Daily	2.59	2.41	1.98	2.29	2.04	1.30	2.31	2.03	1.29
Weekly	11.14	10.17	5.94	8.15	3.83	1.08	6.65	3.37	0.95
Monthly	10.44	6.15	4.12	8.82	4.58	2.14	9.14	4.52	2.08
Quarterly	13.98	9.09	5.16	12.63	7.50	2.83	13.13	7.39	2.68
Yearly	12.91	12.97	6.02	13.04	8.12	3.09	12.91	7.99	3.18

Notes: AMI terciles computed within each frequency across all valid (series, horizon) pairs. Values are median sMAPE (%).

rank association with out-of-sample forecast difficulty across all probe models. For Daily series, AMI indicates non-trivial past–future dependence, but its rank association with realised sMAPE is materially weaker than for other frequencies under the present protocol, implying reduced discriminative power for within-frequency triage.

This finding establishes an operational constraint for AMI-based triage: the diagnostic should be validated within each application context rather than assumed universal. These observed associations are empirical and protocol-specific rather than universal laws of predictability.

B. Daily Frequency: Dependence Present, Weaker Discriminative Power

Daily series show substantial past–future dependence (AMI) in absolute terms, yet forecast error is low in absolute terms across the probe models. Consistent with this compression of error, the AMI–sMAPE association is weaker for Daily than for most other frequencies: variation in AMI provides limited discrimination among already accurate series, even though dependence itself is clearly present. This should be read as a frequency-conditional limitation on the triage value of AMI for Daily under the present rolling-origin evaluation protocol, not as evidence that temporal dependence is absent. In this sense, the method self-diagnoses its limited applicability: where error variation is already compressed, ex-ante triage provides little additional value.

C. Forecastability versus Exploitability

A conceptual distinction is necessary. *Forecastability* as we define it, namely AMI between past and future, is a property of the series itself, independent of any forecasting model. It measures statistical dependence in the data-generating process. *Exploitability* is the degree to which a particular model class can convert that dependence into accurate forecasts.

AMI measures total dependence; the fraction that translates to error reduction depends on the model’s hypothesis class. At Weekly frequency, all probes exhibit strong negative association between AMI and sMAPE, with ETS and N-BEATS strongest (ETS $\rho = -0.72$, N-BEATS $\rho = -0.70$) and Seasonal Naïve weaker ($\rho = -0.29$). This pattern, where different probe models exhibit varying association strengths against the same underlying dependence, supports the distinction between forecastability (available dependence) and

exploitability (model capacity to convert that dependence into accuracy under the chosen loss and protocol).

D. Moderating Factors

Beyond frequency and horizon, several factors moderate the forecastability-accuracy relationship. Series length affects both AMI estimation variance and the training signal available to statistical and neural models; shorter series yield noisier diagnostics and less reliable forecasts regardless of method. Seasonal strength matters because stable, repeating patterns benefit structured models like ETS and even Seasonal Naïve, whereas weak or evolving seasonality favours more flexible approaches. Variance structure and scale effects may also influence the AMI-sMAPE relationship, though this study did not directly test these dimensions.

Structural breaks and regime instability present a more fundamental challenge, as they can sever the link between in-sample dependence and out-of-sample difficulty. Since we treat the data-generating process as latent and measure only its observable consequence (past-future dependence in the realised training segment), forecastability diagnostics are inherently conditional on the assumption that the future segment is governed by a similar process. When regimes shift, historical dependence patterns may not persist. Finally, cross-series heterogeneity affects global models: pooled learning benefits portfolios with shared structure but offers diminishing returns when series are fundamentally dissimilar.

For short Hourly series, the AMI–sMAPE association is near zero and non-significant ($\rho = +0.01$, $p = 0.21$), indicating that AMI-based discrimination requires sufficient training length at this frequency. This boundary case reinforces the importance of validating diagnostic power within specific frequency and length contexts.

This study explicitly tested frequency, horizon, and series length effects. Seasonal strength, variance structure, and regime stability remain avenues for future investigation.

E. Implications for Multivariate Forecasting

Although the empirical analysis focuses on univariate series, the findings carry direct implications for multivariate forecasting contexts. The use of global models introduces a form of implicit multivariate learning by exploiting shared structure across a large panel of series, without requiring explicit

specification of cross-series dependencies. Our results suggest that forecastability is primarily governed by temporal resolution and effective information content within individual series, which implies that univariate forecastability is a necessary, though not sufficient, condition for value creation in multivariate models.

This leads to a practical gating rule: multivariate enrichment should be attempted only when univariate forecastability is low but plausible external drivers exist at the decision horizon. If univariate AMI is already high, covariates are unlikely to add value because the target series already contains sufficient predictive structure. Conversely, if univariate AMI is low and no leading indicators exist, covariates will not rescue an inherently unforecastable target. A related consideration is measurement cadence: low-frequency series are unlikely to benefit from covariates unless those covariates are measured at higher cadence and genuinely lead the target. Annual series paired with annual covariates simply compound information scarcity.

F. Implications for Decision Workflows

These findings support a concrete pre-modelling diagnostic workflow. Before committing forecasting resources, practitioners should first identify the temporal resolution and decision-relevant horizon, since frequency determines baseline expectations for forecastability.

Based on frequency, AMI level, and series characteristics, each series can then be assigned to an appropriate modelling regime. Series with high AMI at Weekly or Hourly frequency warrant investment in more sophisticated models such as ETS or global neural architectures, where the strong negative AMI-sMAPE correlations (Table 2) indicate meaningful accuracy gains over simple baselines. Series in the middle AMI tercile, or at Monthly and Quarterly frequencies, merit standard models with managed horizon expectations; ensemble or robustified approaches may be appropriate. Series with low AMI or unstable AMI estimation are better served by simple baselines. For these series, resources should shift from forecast refinement toward consequence mitigation: safety stock sizing, scenario planning, cadence adjustment, or decision architectures that are robust to forecast error.

This workflow operationalises the core finding: forecastability assessment should precede model selection, and the appropriate action depends on both frequency and measured dependence strength.

G. Limitations and Future Research

Several limitations warrant acknowledgement. Survivorship filtering was necessary to compute AMI reliably, which means that forecastability assessment is only meaningful where it is definable. The method is best framed as a ‘feasible where definable’ triage tool rather than a universal measure applicable to all series. Because AMI is interpreted ordinally, estimator bias that is approximately monotone across series does not affect the validity of triage conclusions. AMI is used as an associative diagnostic of available past–future

dependence and does not imply a causal mechanism linking past values to future outcomes.

Validation used three specific forecasting methods on M4 data. While the conceptual framework is general, empirical relationships are context-dependent, and domain-specific validation remains essential before operational deployment. This study also focuses exclusively on point forecast accuracy measured via sMAPE; probabilistic forecasting, density forecast evaluation, and decision-theoretic loss functions may exhibit different relationships with AMI.

Finally, horizon-specific forecastability is conditioned on the realised DGP segment observed in the training window. If the underlying data-generating process changes between estimation and forecast periods, in-sample dependence will not predict out-of-sample difficulty.

VI. CONCLUSIONS AND RECOMMENDATIONS

This study developed and validated an information-theoretic framework for assessing time series forecastability as horizon-specific past–future dependence. Using an expanding-window rolling-origin protocol with 10 origins per series and computing AMI strictly from training data, we evaluated whether AMI ranks series by realised out-of-sample error (sMAPE) within each frequency. The central finding is that the AMI–sMAPE relationship is strongly frequency-conditional under this protocol. For Hourly, Weekly, Monthly, Quarterly, and Yearly frequencies, AMI exhibits consistently negative rank association with sMAPE across probe models, with particularly strong effects at Weekly and substantial effects across the remaining four frequencies. Daily is the sole weak-discrimination case under this protocol, exhibiting materially weaker rank association despite measurable dependence.

A conceptual distinction between forecastability and exploitability is useful for interpretation. AMI measures horizon-aligned dependence in the realised training segment, whereas exploitability reflects the extent to which a given model class can convert that dependence into reduced error under a specific evaluation protocol. Consistent with this, association magnitudes differ across probes, with ETS and N-BEATS generally showing stronger rank associations than Seasonal Naïve.

For practitioners, the findings support a decision workflow that places forecastability assessment before model selection. First stratify by frequency and decision-relevant horizon. Then compute horizon-specific AMI as an ordinal screening signal to rank series by expected difficulty within frequency. Where AMI is high, investment in more expressive modelling approaches is warranted; where AMI is low, returns to modelling effort diminish and attention should shift towards robust decision design (for example, safety stock, scenario planning, or policies that reduce sensitivity to forecast error). The contribution is not a new forecasting algorithm but a diagnostic layer that addresses a practitioner-critical question: not “which model is best?”, but “is this se-

ries worth modelling at all?” under the declared information set and decision horizon.

DATA AVAILABILITY

Data used in this study are derived from the publicly available M4 competition dataset [8]. Replication code (`ami_forecastability_m4.py`) and full documentation are available on Zenodo at <https://doi.org/10.5281/zenodo.18828113> [software]. The repository includes the main analysis script, a README with configuration instructions and data requirements, and run outputs comprising per-series AMI profiles, sMAPE results, and the summary tables reported in this paper. The M4 dataset itself is available separately from the M4 competition organisers and the M4comp2018 R package.

REFERENCES

- [1] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, “The M4 competition: Results, findings, conclusion and way forward,” *Int. J. Forecast.*, vol. 34, no. 4, pp. 802–808, 2018.
- [2] S. Kolassa, “Can we obtain valid benchmarks from published surveys of forecast accuracy?” *Foresight: Int. J. Appl. Forecast.*, vol. 14, pp. 6–12, 2009.
- [3] D. Thomakos and P. Xidonas, “The origins of forward-looking decision making: Cybernetics, operational research, and the foundations of forecasting,” *Decis. Anal. J.*, vol. 8, p. 100284, 2023.
- [4] P. M. Catt, “Forecastability: Insights from physics, graphical decomposition, and information theory,” *Foresight: Int. J. Appl. Forecast.*, vol. 13, pp. 24–33, 2009.
- [5] G. C. Tiao and R. S. Tsay, “Some advances in non-linear and adaptive modelling in time-series,” *J. Forecast.*, vol. 13, no. 2, pp. 109–131, 1994.
- [6] W. Bialek, I. Nemenman, and N. Tishby, “Predictability, complexity, and learning,” *Neural Comput.*, vol. 13, no. 11, pp. 2409–2463, 2001.
- [7] E. T. Jaynes, “Information theory and statistical mechanics,” *Phys. Rev.*, vol. 106, no. 4, pp. 620–630, 1957.
- [8] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, “The M4 competition: 100,000 time series and 61 forecasting methods,” *Int. J. Forecast.*, vol. 36, no. 1, pp. 54–74, 2020.
- [9] S. Makridakis *et al.*, “The accuracy of extrapolation (time series) methods: Results of a forecasting competition,” *J. Forecast.*, vol. 1, no. 2, pp. 111–153, 1982.
- [10] S. Makridakis and M. Hibon, “The M3-competition: Results, conclusions and implications,” *Int. J. Forecast.*, vol. 16, no. 4, pp. 451–476, 2000.
- [11] F. Petropoulos, D. Apiletti, V. Assimakopoulos, *et al.*, “Forecasting: Theory and practice,” *Int. J. Forecast.*, vol. 38, no. 3, pp. 845–1154, 2022.
- [12] M. Gilliland, *The Business Forecasting Deal*. Wiley, 2010.
- [13] S. M. Pincus, “Approximate entropy as a measure of system complexity,” *Proc. Natl. Acad. Sci.*, vol. 88, no. 6, pp. 2297–2301, 1991.
- [14] J. S. Richman and J. R. Moorman, “Physiological time-series analysis using approximate entropy and sample entropy,” *Am. J. Physiol. Heart Circ. Physiol.*, vol. 278, no. 6, pp. H2039–H2049, 2000.
- [15] P. M. Catt, “Entropy as an a priori indicator of forecastability,” Working paper, ResearchGate, 2014.
- [16] G. M. Goerg, “Forecastable component analysis,” in *Proc. 30th Int. Conf. Mach. Learn.*, PMLR, 2013, pp. 64–72.
- [17] J. H. Stock and M. W. Watson, “Forecasting output and inflation: The role of asset prices,” *J. Econ. Lit.*, vol. 41, no. 3, pp. 788–829, 2003.
- [18] J.-P. Eckmann and D. Ruelle, “Ergodic theory of chaos and strange attractors,” *Rev. Mod. Phys.*, vol. 57, no. 3, pp. 617–656, 1985.
- [19] S. K. Selvam, C. Rajendran, and G. Sankaralingam, “A linear programming-based bi-objective optimization for forecasting short univariate time series,” *Decis. Anal. J.*, vol. 10, p. 100365, 2024.
- [20] R. Wang, S. Klee, and A. Roos, “Time series forecastability measures,” in *Proc. 1st Workshop AI Supply Chain, KDD '25*, ACM, 2025.
- [21] C. E. Shannon, “A mathematical theory of communication,” *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, 1948.
- [22] A. M. Fraser and H. L. Swinney, “Independent coordinates for strange attractors from mutual information,” *Phys. Rev. A*, vol. 33, no. 2, pp. 1134–1140, 1986.
- [23] H. Kantz and T. Schreiber, *Nonlinear Time Series Analysis*, 2nd ed. Cambridge University Press, 2004.
- [24] P. Grassberger, “Toward a quantitative theory of self-generated complexity,” *Int. J. Theor. Phys.*, vol. 25, no. 9, pp. 907–938, 1986.
- [25] J. P. Crutchfield and D. P. Feldman, “Regularities unseen, randomness observed: Levels of entropy convergence,” *Chaos*, vol. 13, no. 1, pp. 25–54, 2003.
- [26] S. E. Palmer, *Vision Science: Photons to Phenomenology*. MIT Press, 1999.

- [27] E. Spiliotis, A. Kouloumos, V. Assimakopoulos, and S. Makridakis, “Are forecasting competitions data representative of the reality?” *Int. J. Forecast.*, vol. 36, no. 1, pp. 37–53, 2020.
- [28] R. J. Hyndman and A. B. Koehler, “Another look at measures of forecast accuracy,” *Int. J. Forecast.*, vol. 22, no. 4, pp. 679–688, 2006.
- [29] S. Seabold and J. Perktold, “Statsmodels: Econometric and statistical modeling with Python,” in *Proc. 9th Python Sci. Conf.*, 2010, pp. 57–61.
- [30] B. N. Oreshkin, D. Carpov, N. Chapados, and Y. Bengio, “N-BEATS: Neural basis expansion analysis for interpretable time series forecasting,” in *Int. Conf. Learn. Represent.*, 2020.
- [31] K. G. Olivares, C. Challu, G. Marcjasz, R. Weron, and A. Dubrawski, “NeuralForecast: User-friendly state-of-the-art neural forecasting models,” in *Proc. 21st Python Sci. Conf.*, 2022, pp. 1–9.
- [32] A. Kraskov, H. Stögbauer, and P. Grassberger, “Estimating mutual information,” *Phys. Rev. E*, vol. 69, no. 6, p. 066138, 2004.
- [33] S. Gao, G. Ver Steeg, and A. Galstyan, “Efficient estimation of mutual information for strongly dependent variables,” in *Proc. 18th Int. Conf. Artif. Intell. Stat.*, 2015, pp. 277–286.
- [34] R. Gardas and S. Narwane, “An analysis of critical factors for adopting machine learning in manufacturing supply chains,” *Decis. Anal. J.*, vol. 10, p. 100377, 2024.