

Forecastability as an Information-Theoretic Limit on Prediction

Peter Maurice Catt

Independent Researcher
Auckland, New Zealand

Abstract—Forecasting is usually framed as a problem of model choice. This paper starts earlier, asking how much predictive information is available at each horizon. Under logarithmic loss, the answer is exact: the mutual information between the future observation and the declared information set equals the maximum achievable reduction in expected loss [6]. This paper develops the consequences of that identity. Forecastability, defined as this mutual information evaluated across horizons, forms a profile whose shape reflects the dependence structure of the process and need not be monotone. Three structural properties are derived: compression of the information set can only reduce forecastability; the gap between the profile under a finite lag window and the full history gives an exact truncation error budget; and for processes with periodic dependence, the profile inherits the periodicity. Predictive loss decomposes into an irreducible component fixed by the information structure and an approximation component attributable to the method; their ratio defines the exploitation ratio, a normalised diagnostic for method adequacy. The exact equality is specific to log loss, but when forecastability is near zero, classical inequalities imply that no method under any loss can materially improve on the unconditional baseline. The framework provides a theoretical foundation for assessing, prior to any modelling, whether the declared information set contains sufficient predictive information at the horizon of interest.

Index Terms—Forecastability profile, Mutual information, Conditional entropy, Log loss, Prediction limits, Data processing inequality, Exploitation ratio, Forecasting methodology.

I. INTRODUCTION

Forecasting is usually framed as a problem of model choice. This paper starts earlier. It asks how much predictive information the available data contain at each horizon, and what that implies for the best accuracy any method can achieve. The question has antecedents in Wiener’s theory of prediction [24] and Granger’s characterisation of predictive relations [25], but neither framework provides an exact, model-free, horizon-specific bound under a proper scoring rule.

The central result is that, under logarithmic loss, the mutual information $I(Y_{t+h}; \mathcal{I}_t)$ between the future ob-

servation and the declared information set is exactly the maximum achievable reduction in expected loss relative to the unconditional baseline. This identity is established in the predictability literature by DelSole [6]; the present paper develops its consequences for forecasting practice. It is not an approximation or an asymptotic bound; it is an exact equality, achieved when the forecasting method coincides with the true conditional distribution. The quantity $I(Y_{t+h}; \mathcal{I}_t)$, evaluated across horizons, defines the forecastability profile: a horizon-resolved function that maps each forecast horizon to the maximum predictive improvement available from the declared information set.

Forecastability is not a scalar quantity but a structure over horizons. The shape of the profile, not any single number, determines where modelling effort is justified and where it is not.

Prior work has studied predictability using information-theoretic quantities, but in different forms and for different purposes. In climate science, DelSole [6] and DelSole and Tippett [7] characterise predictability within a given dynamical model, building on the signal-to-noise framework of Leith [26]—their framework conditions on a specified dynamical model, whereas the present paper conditions on a declared information set without assuming a model. Bialek et al. [18] study predictive information as a complexity quantity over observation length, extending ideas from computational mechanics [27, 28]; Goerg [19] ranks components by spectral entropy; Pennekamp et al. [22], Ponce-Flores et al. [23], and Papacharalampous et al. [20] use entropy-based measures descriptively to relate time series structure to realised forecast performance. Related model-free approaches include Garland et al. [29] on permutation entropy-based quantification and Sugihara and May [30] on nonlinear forecasting as a diagnostic for chaos. None develops the identity into a systematic, model-free, horizon-resolved diagnostic framework conditional on a declared information set. The underlying information-theoretic results are standard [4, 14]; the contribution is their assembly into a coherent framework for forecasting, together with the structural properties of the forecastability profile (Corollary 5, Propositions 6 and 7) and the conceptual vocabulary (forecastability profile, exploitability, exploitation ratio) needed to make them operational.

The theoretical foundations rest on the entropy and

mutual information framework of classical information theory [4, 8, 11, 16, 17], the divergence-based approach to statistical discrimination [5, 14], the theory of proper scoring rules and their connection to probabilistic forecasting [9], and prior work establishing forecastability as an information-theoretic diagnostic for time series [1–3].

The paper is organised as follows. Section II defines the setting. Section III establishes the information bound. Section IV introduces the forecastability profile and derives its structural properties, including the data processing inequality corollary, the finite-window information loss, and a periodicity result, illustrated by two Gaussian examples. Section V gives the loss decomposition. Section VI characterises behaviour at horizons of low forecastability. Section VII discusses estimation. Section VIII develops implications for practice. Section IX concludes.

II. SETTING

Let $\{Y_t\}_{t \in \mathbb{Z}}$ be a stochastic process with values in a measurable space $(\mathcal{Y}, \mathcal{B})$. At time t , the forecaster uses a declared information set \mathcal{I}_t , represented as a sub- σ -algebra. All results are conditional on this choice: different information sets imply different limits on achievable performance.

The true predictive object is the conditional distribution $p_h^*(y \mid \mathcal{I}_t) = P(Y_{t+h} \in dy \mid \mathcal{I}_t)$ at horizon $h \geq 1$. A forecasting method supplies an approximation $q_h(\cdot \mid \mathcal{I}_t)$. Performance is evaluated under logarithmic loss, $\ell(q, y) = -\log q(y)$. Logarithmic loss is used here not because it exhausts all forecasting objectives, but because it is the setting in which predictive information admits an exact and transparent decision-theoretic interpretation: it is strictly proper, equals cross-entropy in expectation, and has a direct coding interpretation [4, 9, 10, 17]. The connection between proper scoring rules and Bayesian decision theory is developed in Dawid [31] and the relationship between calibration and refinement in DeGroot and Fienberg [32].

The mutual information between the future and the information set is

$$I(Y_{t+h}; \mathcal{I}_t) = H(Y_{t+h}) - H(Y_{t+h} \mid \mathcal{I}_t), \quad (1)$$

which measures the reduction in uncertainty about Y_{t+h} obtainable from \mathcal{I}_t [4, 8]. The variation of this quantity across horizons constitutes the forecastability profile of the process. The information set may include lagged values, exogenous variables, covariates, or external signals; additional variables improve forecasting only insofar as they increase $I(Y_{t+h}; \mathcal{I}_t)$.

A. Regularity conditions

The following conditions are assumed throughout.

Assumption 1. (i) All distributions admit densities with respect to a common dominating measure μ on

$(\mathcal{Y}, \mathcal{B})$, and q_h is absolutely continuous with respect to p_h^* wherever KL divergence is invoked. (ii) The entropies $H(Y_{t+h})$, $H(Y_{t+h} \mid \mathcal{I}_t)$, and the expected KL divergence $\mathbb{E}[D_{\text{KL}}(p_h^* \parallel q_h)]$ are finite. (iii) When the forecastability profile is treated as a function of horizon alone, the process $\{Y_t\}$ is stationary [35], so that $I(Y_{t+h}; \mathcal{I}_t)$ does not depend on t .

Condition (i) ensures that the log score and KL divergence are well-defined; the absolute continuity requirement follows standard conventions in information theory [33, 34]. In the continuous case, all entropies are differential entropies defined relative to the dominating measure μ ; the mutual information $I(Y_{t+h}; \mathcal{I}_t)$ is invariant to the choice of μ and is therefore well-defined without ambiguity [4]. Condition (ii) excludes pathological cases where entropy is infinite. Condition (iii) is sufficient for Proposition 4 and for the interpretation of the forecastability profile as a function of horizon alone. It is not necessary for the remaining results: all formal results other than Proposition 4 hold pointwise at each (t, h) without stationarity, and require only that $I(Y_{t+h}; \mathcal{I}_t)$ is well-defined under conditions (i) and (ii). When the process is not stationary, the forecastability profile should be interpreted as local in time and indexed by both h and t .

III. THE INFORMATION BOUND ON PREDICTION

Theorem 1 (Conditional-distribution optimality). *For any predictive distribution $q_h(\cdot \mid \mathcal{I}_t)$ satisfying Assumption 1,*

$$\mathbb{E}[-\log q_h(Y_{t+h} \mid \mathcal{I}_t)] = H(Y_{t+h} \mid \mathcal{I}_t) + \mathbb{E}[D_{\text{KL}}(p_h^* \parallel q_h)]. \quad (2)$$

Consequently,

$$\mathbb{E}[-\log q_h(Y_{t+h} \mid \mathcal{I}_t)] \geq H(Y_{t+h} \mid \mathcal{I}_t), \quad (3)$$

with equality if and only if $q_h = p_h^*$ almost surely.

Proof.

$$\begin{aligned} & \mathbb{E}[-\log q_h(Y_{t+h} \mid \mathcal{I}_t)] \\ &= \mathbb{E}[-\log p_h^*(Y_{t+h} \mid \mathcal{I}_t)] + \mathbb{E}\left[\log \frac{p_h^*(Y_{t+h} \mid \mathcal{I}_t)}{q_h(Y_{t+h} \mid \mathcal{I}_t)}\right] \\ &= H(Y_{t+h} \mid \mathcal{I}_t) + \mathbb{E}[D_{\text{KL}}(p_h^* \parallel q_h)]. \end{aligned} \quad (4)$$

Since $D_{\text{KL}}(p_h^* \parallel q_h) \geq 0$ with equality iff $p_h^* = q_h$ a.s., the bound follows. \square

Theorem 1 identifies the benchmark: the best achievable loss is the conditional entropy, and every other method pays an additional KL penalty. The non-negativity of KL divergence, central to the proof, is a consequence of Jensen’s inequality; its role as the

unique divergence consistent with the log score was established by Bernardo [36] and is reviewed in Grünwald and Dawid [37]. The marginal predictor is the unconditional distribution $p(Y_{t+h})$, which ignores \mathcal{I}_t and achieves expected log loss $H(Y_{t+h})$.

Theorem 2 (Information bound on prediction). *The maximum achievable reduction in expected log loss from exploiting \mathcal{I}_t , relative to the marginal predictor, is exactly*

$$H(Y_{t+h}) - H(Y_{t+h} | \mathcal{I}_t) = I(Y_{t+h}; \mathcal{I}_t). \quad (5)$$

This bound is achieved if and only if $q_h = p_h^$ almost surely.*

Proof. The marginal predictor achieves expected log loss $H(Y_{t+h})$. The minimum achievable loss using \mathcal{I}_t is $H(Y_{t+h} | \mathcal{I}_t)$ by Theorem 1. The difference is $I(Y_{t+h}; \mathcal{I}_t)$ by (1). \square

The bound is tight, but generally unattainable in practice because p_h^* is unknown. The gap between the bound and achieved performance is the approximation error $\mathbb{E}[D_{\text{KL}}(p_h^* \| q_h)]$ [13]. Theorem 2 recovers, in the present notation, the identity established by DelSole [6] in the context of climate predictability; the contribution of this paper lies not in the identity itself but in the structural consequences developed in Sections IV–VI. Convergence rates for the approximation term under various model classes are studied in Barron [38] and Yang and Barron [39]. A direct consequence is the horizon-specific bound: for any method q_h ,

$$H(Y_{t+h}) - \mathbb{E}[-\log q_h(Y_{t+h} | \mathcal{I}_t)] \leq F(h; \mathcal{I}_t), \quad (6)$$

since the left-hand side equals $F(h; \mathcal{I}_t) - \mathbb{E}[D_{\text{KL}}(p_h^* \| q_h)] \leq F(h; \mathcal{I}_t)$ by non-negativity of KL divergence.

IV. FORECASTABILITY PROFILES AND THEIR PROPERTIES

A. Forecastability

Definition 3 (Forecastability). *The forecastability of the process $\{Y_t\}$ at horizon $h \geq 1$ relative to information set \mathcal{I}_t is*

$$F(h; \mathcal{I}_t) = I(Y_{t+h}; \mathcal{I}_t). \quad (7)$$

By Theorem 2, forecastability is the maximum number of nats by which any probabilistic method can improve on the marginal predictor at horizon h , under logarithmic loss. It is defined relative to the declared information set and is therefore not an intrinsic property of the time series alone. Changes in the composition of \mathcal{I}_t will, in general, change the forecastability profile. Hereafter, $F(h; \mathcal{I}_t)$ is used throughout; $I(Y_{t+h}; \mathcal{I}_t)$ appears only when the chain rule or data processing inequality is applied directly to the mutual information.

The profile should be interpreted as a horizon-resolved structure rather than a scalar quantity. Its shape reflects the temporal dependence in the data and may be non-monotone. Fraser and Swinney [40] introduced lagged mutual information for embedding dimension selection in dynamical systems; Li [41] established the relationship between mutual information functions and correlation functions. In the present framework, the profile serves a different purpose: in many practical series, forecastability is concentrated at specific lags, particularly seasonal periods, rather than distributed smoothly across horizons.

B. The entropy rate and maximum forecastability

For a stationary process with the full history as information set, $\mathcal{I}_t = \sigma(Y_t, Y_{t-1}, \dots)$, the forecastability at the shortest horizon has a direct interpretation in terms of the entropy rate.

Proposition 4 (Maximum forecastability and entropy rate). *Let $h_\infty = \lim_{n \rightarrow \infty} H(Y_n | Y_{n-1}, \dots, Y_1)$ denote the entropy rate of the stationary process $\{Y_t\}$ (assumed finite under Assumption 1), and let $\mathcal{I}_t = \sigma(Y_t, Y_{t-1}, \dots)$. Then*

$$F(1; \mathcal{I}_t) = H(Y_{t+1}) - h_\infty. \quad (8)$$

Proof. For a stationary process, $H(Y_{t+1} | \mathcal{I}_t) = H(Y_{t+1} | Y_t, Y_{t-1}, \dots) = h_\infty$ by definition of the entropy rate. Therefore $F(1; \mathcal{I}_t) = H(Y_{t+1}) - H(Y_{t+1} | \mathcal{I}_t) = H(Y_{t+1}) - h_\infty$. \square

The entropy rate h_∞ is the irreducible uncertainty per observation that remains when the full past is known [4, 11]. For stationary ergodic processes, the Shannon–McMillan–Breiman theorem [42, 43] ensures that sample averages of log-likelihood converge to the entropy rate almost surely. Hence $H(Y_{t+1}) - h_\infty$ gives the maximum one-step forecastability and anchors the profile at its shortest horizon.

C. Information-set compression and the data processing inequality

Any transformation or compression of the information set cannot increase forecastability. This is a direct consequence of the data processing inequality [4].

Corollary 5 (Forecastability under information-set compression). *Let $T(\mathcal{I}_t)$ be any measurable function of the information set. Then*

$$F(h; T(\mathcal{I}_t)) \leq F(h; \mathcal{I}_t) \quad (9)$$

for all $h \geq 1$. Equality holds if and only if T is a sufficient statistic for Y_{t+h} given \mathcal{I}_t in the sense of Fisher [44] and Halmos and Savage [45].

Proof. By the data processing inequality, $F(h; T(\mathcal{I}_t)) \leq F(h; \mathcal{I}_t)$ for any measurable T . Equality holds iff the Markov chain $Y_{t+h} - \mathcal{I}_t - T(\mathcal{I}_t)$ satisfies the sufficiency condition [4]. \square

This result has immediate consequences: feature extraction, dimensionality reduction, lag truncation, and any other summarisation of the raw information set can only reduce the forecastability profile. The profile under the raw information set is therefore an upper bound on what any reduced representation can achieve.

D. Finite-window information loss

Even when the full history $\mathcal{I}_t = \sigma(Y_t, Y_{t-1}, \dots)$ is available, forecasters typically condition on a finite lag window $\mathcal{I}_t^{(p)} = \sigma(Y_t, Y_{t-1}, \dots, Y_{t-p+1})$ for some $p \geq 1$. The information lost by this truncation is itself a mutual information quantity.

Proposition 6 (Finite-window information loss). *For a stationary process satisfying Assumption 1, the forecastability lost by truncating the information set from the full history to a p -lag window is*

$$\begin{aligned} F(h; \mathcal{I}_t) - F(h; \mathcal{I}_t^{(p)}) \\ = I(Y_{t+h}; Y_{t-p}, Y_{t-p-1}, \dots | Y_t, \dots, Y_{t-p+1}). \end{aligned} \quad (10)$$

This quantity is non-negative and equals zero if and only if Y_{t+h} is conditionally independent of the remote past given the p most recent observations.

Proof. By the chain rule for mutual information,

$$\begin{aligned} I(Y_{t+h}; \mathcal{I}_t) &= I(Y_{t+h}; Y_t, \dots, Y_{t-p+1}) \\ &\quad + I(Y_{t+h}; Y_{t-p}, Y_{t-p-1}, \dots | Y_t, \dots, Y_{t-p+1}). \end{aligned} \quad (11)$$

The first term is $F(h; \mathcal{I}_t^{(p)})$. The second is non-negative by definition of conditional mutual information and equals zero iff the stated conditional independence holds. \square

This gives the empirical programme an exact error budget: the gap between the estimated forecastability profile under a finite window and the true profile under the full history is a well-defined information-theoretic quantity. For processes with short memory, this gap vanishes for moderate p . For processes with long-range dependence [46], it may remain substantial. The choice of lag order p connects to the embedding problem in nonlinear dynamics [47, 48].

E. Periodicity and non-monotonicity

Proposition 7 (Periodic forecastability). *Let $\{Y_t\}$ be a stationary process such that the conditional distribution $P(Y_{t+h} | \mathcal{I}_t)$ has period s in h , in the sense that $P(Y_{t+h+s} | \mathcal{I}_t) = P(Y_{t+h} | \mathcal{I}_t)$ for all $h \geq 1$. Then $F(h+s; \mathcal{I}_t) = F(h; \mathcal{I}_t)$ for all $h \geq 1$.*

Proof. Under the stated periodicity, $H(Y_{t+h+s} | \mathcal{I}_t) = H(Y_{t+h} | \mathcal{I}_t)$ and $H(Y_{t+h+s}) = H(Y_{t+h})$ by stationarity. Therefore $F(h+s; \mathcal{I}_t) = H(Y_{t+h+s}) - H(Y_{t+h+s} | \mathcal{I}_t) = H(Y_{t+h}) - H(Y_{t+h} | \mathcal{I}_t) = F(h; \mathcal{I}_t)$. \square

The forecastability profile therefore need not be monotone: a process with seasonal period s will show recurring peaks at multiples of s . The proof follows directly from the definitions; the substantive content is the implication that scalar diagnostics, which compress the dependence structure into a single number, cannot detect this horizon-resolved structure. More generally, even without exact periodicity, strong seasonal or cyclical dependence can produce local peaks in forecastability at seasonal lags superimposed on a long-run decay. Under standard mixing conditions [49, 50] with sufficient regularity, forecastability converges to zero as $h \rightarrow \infty$ [11], but the path need not be monotone at any finite range of horizons. This non-monotonicity is precisely what distinguishes the horizon-resolved profile from scalar diagnostics such as sample entropy or spectral entropy, which cannot identify horizons where predictive structure re-emerges after an intervening gap.

F. Interpretation of the forecastability profile

The forecastability profile partitions the horizon space into regions where modelling effort has different expected returns. At horizons where forecastability is large, improvements in model specification can translate into measurable gains. At horizons where forecastability is small, the expected reduction in log loss achievable by any method is bounded by the forecastability itself, and improvements in model complexity yield diminishing returns. Rather than necessarily indicating model inadequacy, weak performance relative to the unconditional baseline may reflect the absence of exploitable structure in the declared information set. The profile therefore separates model-limited regimes from information-limited regimes.

G. Illustrative examples

Two Gaussian AR processes illustrate the forecastability profile and connect it to the structural properties derived above.

Example 1 (AR(1): monotone decay). Let $Y_t = \phi Y_{t-1} + \varepsilon_t$ with $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$ and $|\phi| < 1$. The process is Markov, so $\mathcal{I}_t = \sigma(Y_t)$ is sufficient for the full history. The conditional distribution of Y_{t+h} given $Y_t = y$ is $\mathcal{N}(\phi^h y, \sigma^2(1 - \phi^{2h})/(1 - \phi^2))$, and the marginal is $\mathcal{N}(0, \sigma^2/(1 - \phi^2))$. The forecastability profile is

$$F(h) = -\frac{1}{2} \log(1 - \phi^{2h}). \quad (12)$$

Since the population coefficient of determination is $R_h^2 = \phi^{2h}$, this is equivalently $F(h) = -\frac{1}{2} \log(1 - R_h^2)$, an instance of the Gaussian relationship developed in Section VIII.

Fig. 1(a) plots the profile for $\phi = 0.95$ and $\phi = 0.3$. For the weakly dependent process ($\phi = 0.3$), forecastability is 0.047 nats at $h = 1$ and falls below 0.005 nats by $h = 2$: no method can materially improve on

the marginal predictor beyond the first step. For the strongly dependent process ($\phi = 0.95$), forecastability exceeds 0.5 nats through $h = 10$ and remains above 0.1 nats past $h = 30$, so modelling effort is justified across a wide horizon range. The contrast between the two profiles is the diagnostic in action: prior to fitting any model, the forecastability profile identifies where predictive improvement is available.

Since the AR(1) is Markov of order 1, the finite-window information loss (Proposition 6) satisfies $\Delta_p(h) = 0$ for all $p \geq 1$: a single lag captures all predictive information, and the truncation error budget is identically zero.

Example 2 (Seasonal AR: non-monotone profile). Consider the multiplicative seasonal process

$$Y_t = \phi Y_{t-1} + \Phi Y_{t-s} - \phi\Phi Y_{t-s-1} + \varepsilon_t \quad (13)$$

with $\phi = 0.5$, $\Phi = 0.8$, seasonal period $s = 12$, and $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$. This is the stationary Gaussian AR(1, 0) \times (1, 0)₁₂ model. Consider first the restricted information set $\mathcal{I}_t = \sigma(Y_t)$, so that $F(h; \sigma(Y_t)) = -\frac{1}{2} \log(1 - \rho_h^2)$ where ρ_h is the autocorrelation at lag h .

Fig. 1(b) plots the resulting profile. It is non-monotone. Forecastability falls from 0.14 nats at $h = 1$ to a minimum of approximately 0.0004 nats near $h = 6$, where the AR(1) component has largely dissipated and the seasonal component has not yet engaged. It then rises sharply to 0.49 nats at $h = 12$, where the annual dependence dominates. Further peaks appear at $h = 24$ (0.25 nats) and $h = 36$ (0.13 nats), attenuated by successive powers of Φ .

The profile partitions the horizon axis into qualitatively distinct regimes. At short horizons ($h = 1$ –3), forecastability is dominated by the AR(1) component. At intermediate horizons ($h \approx 5$ –9), the marginal predictor is nearly optimal and increased model complexity yields negligible returns. At seasonal horizons ($h \approx 12, 24, 36$), forecastability re-emerges; a method that fails to exploit the seasonal structure would show a low exploitation ratio χ_q at those horizons. This non-monotonicity is the phenomenon formalised in Proposition 7: scalar diagnostics that compress the dependence structure into a single number cannot detect this re-emergence.

The finite-window information loss (Proposition 6) is instructive here. The profile under $\mathcal{I}_t = \sigma(Y_t)$ misses all dependence beyond lag 1. Extending the information set to include the seasonal lag, $\mathcal{I}_t^{(13)} = \sigma(Y_t, \dots, Y_{t-12})$, captures the seasonal structure and raises the profile at all horizons. The gap between the two profiles is the truncation budget $\Delta_p(h)$: it is negligible at short horizons (where the AR(1) component dominates) and large at seasonal horizons (where remote lags carry substantial conditional information). The error budget thus provides a principled criterion for lag-order selection.

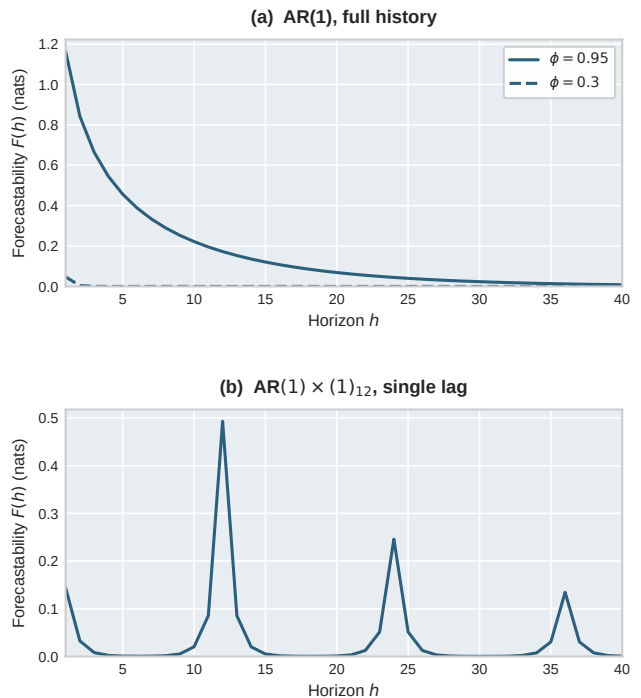


Figure 1. Forecastability profiles for Gaussian processes. (a) AR(1) with $\phi = 0.95$ (solid) and $\phi = 0.3$ (dashed), using the full history. (b) Multiplicative seasonal AR(1) \times (1)₁₂ with $\phi = 0.5$, $\Phi = 0.8$, using $\mathcal{I}_t = \sigma(Y_t)$ to illustrate non-monotonicity.

For a non-Gaussian process with the same autocovariance structure, the forecastability profile would in general differ because mutual information captures dependence beyond second moments. By the maximum-entropy property of the Gaussian, the profile computed here is a lower bound on the forecastability of any process with identical autocovariances.

H. Complexity futility

Corollary 8 (Complexity futility). *At any horizon, the maximum achievable reduction in expected log loss relative to the unconditional baseline is $F(h; \mathcal{I}_t)$, regardless of the complexity of the forecasting method or the computational resources deployed.*

Increasing model complexity can reduce only the approximation component $\mathbb{E}[D_{\text{KL}}(p_h^* || q_h)]$; it cannot alter the irreducible component $H(Y_{t+h} | \mathcal{I}_t)$, which is a property of the data and the information set. At horizons where forecastability is small, even a perfect method offers only a negligible gain over the unconditional baseline under logarithmic loss. The corollary does not imply that forecasts cannot be improved; it implies that improvement requires enriching the information set, not the model class.

V. DECOMPOSITION OF PREDICTIVE LOSS

Proposition 9 (Loss decomposition). *For any forecasting procedure $q_h(\cdot | \mathcal{I}_t)$ satisfying Assumption 1, the expected log loss separates as*

$$\mathbb{E}[-\log q_h(Y_{t+h} | \mathcal{I}_t)] = \underbrace{H(Y_{t+h} | \mathcal{I}_t)}_{\text{irreducible}} + \underbrace{\mathbb{E}[D_{\text{KL}}(p_h^* \| q_h)]}_{\text{approximation}}. \quad (14)$$

Proof. Direct consequence of Theorem 1: the expected log loss of any predictive distribution equals the conditional entropy plus its expected KL divergence from the true conditional distribution. \square

The first term is irreducible and fixed by the joint distribution of the process and information set. The second is approximation error and can, in principle, be reduced by better modelling. This decomposition parallels the calibration–refinement factorisation in probability forecasting [51], but is exact under log loss rather than approximate under other scoring rules.

Equivalently, the expected log loss can be written as

$$\mathbb{E}[-\log q_h] = H(Y_{t+h}) - F(h; \mathcal{I}_t) + \mathbb{E}[D_{\text{KL}}(p_h^* \| q_h)]. \quad (15)$$

This form identifies two distinct levers for improvement: enriching the information set to increase $F(h; \mathcal{I}_t)$, and improving the forecasting method to reduce the KL term. Even a perfect method can eliminate only the approximation term; the irreducible component is fixed by the information structure of the data.

For a specified method q_h , the realised reduction relative to the unconditional baseline, $X_q(h; \mathcal{I}_t) = H(Y_{t+h}) - \mathbb{E}[-\log q_h(Y_{t+h} | \mathcal{I}_t)]$, may be termed the exploitability. The ratio $\chi_q = X_q/F$ gives the fraction of available forecastability that the method achieves. Forecastability is a property of the process and information set; exploitability is method-dependent and can only be evaluated post-modelling. The distinction echoes Granger’s [52] separation of predictability from achieved forecast performance.

VI. HORIZONS OF LOW FORECASTABILITY

At horizons where forecastability is small, the irreducible floor is high. When $F(h; \mathcal{I}_t)$ is close to zero,

$$H(Y_{t+h} | \mathcal{I}_t) \approx H(Y_{t+h}). \quad (16)$$

At such horizons, the best possible probabilistic forecast is already close to the unconditional baseline in expected log loss. Low forecastability does not imply that prediction is impossible, only that the maximum achievable improvement over the unconditional baseline is small under logarithmic loss.

Two classical inequalities make the consequences precise. Fano’s inequality [8] gives a lower bound on classification error: for any predictor \hat{Y}_{t+h} of a discrete Y_{t+h}

taking M values,

$$\begin{aligned} P(\hat{Y}_{t+h} \neq Y_{t+h}) &\geq \frac{H(Y_{t+h} | \mathcal{I}_t) - 1}{\log M} \\ &= \frac{H(Y_{t+h}) - F(h; \mathcal{I}_t) - 1}{\log M}, \end{aligned} \quad (17)$$

so when forecastability is small, the minimum achievable error probability is high regardless of method. Pinsker’s inequality [33] bounds the total variation distance between the conditional and marginal predictive distributions:

$$\|p_h^*(\cdot | \mathcal{I}_t) - p(\cdot)\|_{\text{TV}} \leq \sqrt{\frac{1}{2} F(h; \mathcal{I}_t)}, \quad (18)$$

so when forecastability is small, the two distributions are close: conditioning on the information set barely changes the predictive distribution.

This is not a failure of the forecasting method; it is a consequence of the information structure. The standard response to poor forecasting accuracy, namely increasing model complexity, is effective only if the approximation component is the binding constraint. When it is the irreducible component that dominates, no amount of additional complexity can help: when forecastability is near zero, all methods converge toward the same floor. The minimum description length principle [53] provides a complementary perspective: model complexity is justified only when it captures genuine regularity. At low-forecastability horizons, the data do not support complex descriptions, and the optimal code length approaches the marginal entropy [54].

VII. ESTIMATION

The results above are stated in terms of the true forecastability $F(h; \mathcal{I}_t)$. In practice, this must be estimated from data. Several approaches exist. Histogram-based estimators partition the observation space but suffer from binning artefacts [55]. The standard non-parametric approach is the k -nearest-neighbour estimator of Kraskov, Stögbauer and Grassberger [12], which avoids explicit density estimation and is consistent under regularity conditions [15]. Gao et al. [56] extend the KSG framework to improve performance under strong dependence, and Kozachenko and Leonenko [57] provide the foundational entropy estimator on which KSG builds. The main practical constraint is the dimensionality of \mathcal{I}_t : in low-dimensional lag-based settings, the estimator is reliable for moderate sample sizes, but in higher-dimensional settings, finite-sample bias and variance grow and may make the diagnostic as demanding as the modelling it is intended to precede. In principle, richer information sets can only increase true forecastability, but in finite samples they may make estimated forecastability less reliable. In practice, the diagnostic is most tractable in the univariate lag-based

setting where $\mathcal{I}_t = \sigma(Y_t, \dots, Y_{t-p+1})$ with $p \leq 3$, series lengths $n \geq 200$, and horizons where the effective sample size $n - h$ remains sufficient for stable estimation. The companion empirical work [3] validates the diagnostic in this regime across more than 42,000 series. Permutation-based significance testing [58, 59], in which the dependence between past and future is destroyed by random shuffling, provides a natural null for determining whether estimated forecastability at a given horizon differs from zero [3]. Fuller estimation methodology is developed in the companion empirical work [3]. Formal estimation theory—including finite-sample convergence rates, minimax bounds, and the interaction between lag dimension and sample size—is not developed here and remains an open direction.

VIII. IMPLICATIONS FOR FORECASTING PRACTICE

Under logarithmic loss and for the declared information set, forecasting is constrained by information before it is constrained by modelling. The forecastability profile provides a theoretical basis for identifying where predictive structure exists prior to selecting and fitting a forecasting model class. The framework is diagnostic rather than prescriptive: it characterises the information-limited structure of the forecasting problem, but does not by itself determine a forecasting model or decision rule.

The data processing inequality (Corollary 5) has direct implications for model design: any feature extraction, dimensionality reduction, or data compression applied to the information set can only reduce the forecastability ceiling. The finite-window result (Proposition 6) gives an exact error budget for lag truncation, connecting the theoretical framework to the practical reality that forecasters always work with finite histories.

In many modern forecasting workflows [60, 61], model development already involves competing specifications, repeated validation, and sometimes ensembling; the practical role of forecastability is to constrain that wider search, not to replace modelling altogether [62]. When forecastability is high, model improvement can reduce the approximation component and is therefore productive. When forecastability is low, the achievable improvement is tightly bounded, and increasing complexity risks overfitting without meaningful gain. The forecastability profile provides a principled basis for making this distinction.

Regime shifts alter the underlying dependence structure [63, 64] and can render previously estimated forecastability profiles unrepresentative. Reliable re-estimation may require substantial post-shift data, so forecastability should be interpreted locally in such settings.

A. Illustrative example: seasonal information structure

To illustrate how the profile changes a structural decision that a single diagnostic would miss, consider a monthly series with strong annual seasonality [65]. Suppose that $F(h)$ is substantial for $h \in \{1, 2, 3, 4, 11, 12, 13\}$ and negligible for $h \in \{5, 6, 7, 8, 9, 10\}$. A single first-crossing diagnostic would report $h = 5$ as the point at which prediction becomes uninformative. The forecastability profile reveals that this conclusion is wrong: horizons 11–13 contain substantial predictive information because the seasonal dependence re-emerges at the annual lag, as formalised in Proposition 7.

Now suppose an external leading indicator is added that is correlated with the target at horizons 6–8 but uncorrelated elsewhere. The updated profile may show substantial forecastability at $h \in \{1, 2, 3, 4, 6, 7, 8, 11, 12, 13\}$, expanding the productive horizons from seven to ten. This demonstrates that the forecastability profile is not a fixed property of the target series but changes structurally when the declared information set is enriched.

B. Extension to alternative losses

For Gaussian processes, the forecastability profile has a direct connection to the MSE-based predictability familiar from the Wiener–Granger tradition. If (Y_{t+h}, \mathcal{I}_t) is jointly Gaussian, the forecastability at horizon h reduces to

$$F(h; \mathcal{I}_t) = -\frac{1}{2} \log(1 - R_h^2), \quad (19)$$

where R_h^2 is the population R^2 from regressing Y_{t+h} on \mathcal{I}_t . Granger-causal predictability under squared-error loss is captured by R_h^2 ; forecastability under log loss is a monotone transformation of it. In the Gaussian case, the two order series identically: a series that is more forecastable under log loss is also more predictable under MSE. In the non-Gaussian case, the two can diverge because F captures nonlinear dependence that R_h^2 , being a second-moment quantity, misses entirely.

More generally, logarithmic loss occupies a distinguished position among proper scoring rules: in binary and finite-alphabet settings, its associated divergence upper-bounds the divergences induced by broad classes of smooth, proper, convex losses up to normalisation [9, 21, 66]. Low forecastability under log loss therefore implies low forecastability under these losses as well, but the implication is one-directional and may be loose at moderate forecastability levels. For practitioners who evaluate forecasts under squared-error loss, CRPS, or quantile loss, the exact equality proved here does not transfer: forecastability under log loss bounds but does not equal the maximum achievable improvement under those losses. The exact equality is specific to log loss; for other losses, the relationship is an inequality rather than an identity. For continuous-outcome forecasting

under squared-error loss, related bounds may be developed through entropy-based inequalities, but these are weaker and more model-dependent. A full loss-specific treatment belongs more naturally to a rate-distortion framework [4,67] and lies beyond the scope of the present paper.

IX. CONCLUSION

Under logarithmic loss, the mutual information between the future and the declared information set is exactly the maximum achievable reduction in expected loss. This paper has developed the consequences of that identity for forecasting: the forecastability profile as a horizon-resolved diagnostic, its behaviour under information-set compression (Corollary 5), the exact error budget for finite-window truncation (Proposition 6), the inheritance of periodicity (Proposition 7), and the decomposition of predictive loss into irreducible and approximation components, with the forecastability exploitation ratio χ_q providing a normalised measure of how much predictive structure a given method captures. The framework provides a theoretical foundation for assessing whether the declared information set contains sufficient predictive information at the horizon of interest, prior to committing to a model class. Companion empirical work [3] validates the diagnostic against realised forecast accuracy across more than 42,000 time series: auto-mutual information computed strictly from training data exhibits Spearman rank correlations of -0.41 to -0.72 with out-of-sample sMAPE for five of six temporal frequencies, and median forecast error declines monotonically from low to high forecastability terciles across all frequencies and probe models. Open extensions include loss-specific bounds via rate-distortion theory, the behaviour of the profile under non-stationarity and regime change, computable bounds on the information loss for specific compression schemes (e.g. principal component projections or feature maps), and integration with established forecast evaluation protocols [68,69]. The framework extends directly to multivariate information sets: all results are stated in terms of sub- σ -algebras and hold without modification when \mathcal{I}_t includes vector-valued histories, exogenous variables, or mixed-frequency data. The central implication is that forecastability is a measurable property of the data, not an emergent property of the model. It can be quantified before any model is built, and it bounds what every model can achieve.

REFERENCES

- [1] P.M. Catt, “Forecastability: Insights from physics, graphical decomposition, and information theory,” *Foresight: Int. J. Appl. Forecasting*, no. 13, pp. 24–33, 2009.
- [2] P.M. Catt, “Entropy as an a priori indicator of forecastability,” SSRN, 2014. doi:10.2139/ssrn.6235738.
- [3] P.M. Catt, “An information-theoretic diagnostic analytics framework for mapping past–future dependence in horizon-specific forecastability,” manuscript under review, 2025. Available: <https://doi.org/10.2139/ssrn.6416626>.
- [4] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Wiley, 2006.
- [5] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, 2nd ed. Cambridge Univ. Press, 2011.
- [6] T. DelSole, “Predictability and information theory. Part I: Measures of predictability,” *J. Atmos. Sci.*, vol. 61, no. 20, pp. 2425–2440, 2004.
- [7] T. DelSole and M. K. Tippett, “Predictability: Recent insights from information theory,” *Rev. Geophys.*, vol. 45, no. 4, p. RG4002, 2007.
- [8] R. M. Fano, *Transmission of Information: A Statistical Theory of Communications*. MIT Press, 1961.
- [9] T. Gneiting and A. E. Raftery, “Strictly proper scoring rules, prediction, and estimation,” *J. Amer. Statist. Assoc.*, vol. 102, no. 477, pp. 359–378, 2007.
- [10] T. Gneiting and M. Katzfuss, “Probabilistic forecasting,” *Annu. Rev. Statist. Appl.*, vol. 1, pp. 125–151, 2014.
- [11] R. M. Gray, *Entropy and Information Theory*, 2nd ed. Springer, 2011.
- [12] A. Kraskov, H. Stögbauer, and P. Grassberger, “Estimating mutual information,” *Phys. Rev. E*, vol. 69, no. 6, p. 066138, 2004.
- [13] S. Kullback and R. A. Leibler, “On information and sufficiency,” *Ann. Math. Statist.*, vol. 22, no. 1, pp. 79–86, 1951.
- [14] S. Kullback, *Information Theory and Statistics*. Wiley, 1959.
- [15] L. Paninski, “Estimation of entropy and mutual information,” *Neural Comput.*, vol. 15, no. 6, pp. 1191–1253, 2003.
- [16] A. Rényi, “On measures of entropy and information,” in *Proc. 4th Berkeley Symp. Math. Statist. Probab.*, vol. 1, Univ. California Press, 1961, pp. 547–561.
- [17] C. E. Shannon, “A mathematical theory of communication,” *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, 1948.

- [18] W. Bialek, I. Nemenman, and N. Tishby, “Predictability, complexity, and learning,” *Neural Comput.*, vol. 13, no. 11, pp. 2409–2463, 2001.
- [19] G. M. Goerg, “Forecastable component analysis,” in *Proc. 30th Int. Conf. Machine Learning*, JMLR W&CP 28, 2013, pp. 64–72.
- [20] G. Papacharalampous, H. Tyrallis, I. G. Pechlivanidis, S. Grimaldi, and E. Volpi, “Massive feature extraction for explaining and foretelling hydroclimatic time series forecastability at the global scale,” *Geosci. Front.*, vol. 13, no. 3, p. 101349, 2022.
- [21] A. Painsky and G. W. Wornell, “On the universality of the logistic loss function,” in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, 2020, pp. 392–397.
- [22] F. Pennekamp *et al.*, “The intrinsic predictability of ecological time series and its potential to guide forecasting,” *Ecol. Monogr.*, vol. 89, no. 2, p. e01359, 2019.
- [23] M. Ponce-Flores, J. Frausto-Solís, G. Santamaría-Bonfil, J. Pérez-Ortega, and J. J. González-Barbosa, “Time series complexities and their relationship to forecasting performance,” *Entropy*, vol. 22, no. 1, p. 89, 2020.
- [24] N. Wiener, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*. MIT Press, 1949.
- [25] C. W. J. Granger, “Investigating causal relations by econometric models and cross-spectral methods,” *Econometrica*, vol. 37, no. 3, pp. 424–438, 1969.
- [26] C. E. Leith, “Predictability of climate,” *Nature*, vol. 276, no. 5686, pp. 352–355, 1978.
- [27] J. P. Crutchfield and K. Young, “Inferring statistical complexity,” *Phys. Rev. Lett.*, vol. 63, no. 2, pp. 105–108, 1989.
- [28] C. R. Shalizi and J. P. Crutchfield, “Computational mechanics: Pattern and prediction, structure and simplicity,” *J. Statist. Phys.*, vol. 104, no. 3–4, pp. 817–879, 2001.
- [29] J. Garland, R. G. James, and E. Bradley, “Model-free quantification of time-series predictability,” *Phys. Rev. E*, vol. 90, no. 5, p. 052910, 2014.
- [30] G. Sugihara and R. M. May, “Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series,” *Nature*, vol. 344, no. 6268, pp. 734–741, 1990.
- [31] A. P. Dawid, “Probability forecasting,” in *Encyclopedia of Statistical Sciences*, vol. 7, Wiley, 1986, pp. 210–218.
- [32] M. H. DeGroot and S. E. Fienberg, “The comparison and evaluation of forecasters,” *The Statistician*, vol. 32, no. 1–2, pp. 12–22, 1983.
- [33] M. S. Pinsker, *Information and Information Stability of Random Variables and Processes*. Holden-Day, 1964.
- [34] P. Billingsley, *Probability and Measure*, 3rd ed. Wiley, 1995.
- [35] J. L. Doob, *Stochastic Processes*. Wiley, 1953.
- [36] J. M. Bernardo, “Expected information as expected utility,” *Ann. Statist.*, vol. 7, no. 3, pp. 686–690, 1979.
- [37] P. D. Grünwald and A. P. Dawid, “Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory,” *Ann. Statist.*, vol. 32, no. 4, pp. 1367–1433, 2004.
- [38] A. R. Barron and T. M. Cover, “Minimum complexity density estimation,” *IEEE Trans. Inf. Theory*, vol. 37, no. 4, pp. 1034–1054, 1991.
- [39] Y. Yang and A. R. Barron, “Information-theoretic determination of minimax rates of convergence,” *Ann. Statist.*, vol. 27, no. 5, pp. 1564–1599, 1999.
- [40] 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.
- [41] W. Li, “Mutual information functions versus correlation functions,” *J. Statist. Phys.*, vol. 60, no. 5–6, pp. 823–837, 1990.
- [42] L. Breiman, “The individual ergodic theorem of information theory,” *Ann. Math. Statist.*, vol. 28, no. 3, pp. 809–811, 1957.
- [43] A. R. Barron, “The strong ergodic theorem for densities: Generalized Shannon–McMillan–Breiman theorem,” *Ann. Probab.*, vol. 13, no. 4, pp. 1292–1303, 1985.
- [44] R. A. Fisher, “On the mathematical foundations of theoretical statistics,” *Phil. Trans. R. Soc. A*, vol. 222, no. 594–604, pp. 309–368, 1922.
- [45] P. R. Halmos and L. J. Savage, “Application of the Radon–Nikodym theorem to the theory of sufficient statistics,” *Ann. Math. Statist.*, vol. 20, no. 2, pp. 225–241, 1949.
- [46] J. Beran, *Statistics for Long-Memory Processes*. Chapman & Hall, 1994.

- [47] H. Tong, *Non-Linear Time Series: A Dynamical System Approach*. Oxford University Press, 1990.
- [48] H. Kantz and T. Schreiber, *Nonlinear Time Series Analysis*, 2nd ed. Cambridge University Press, 2004.
- [49] R. C. Bradley, “Basic properties of strong mixing conditions. A survey and some open questions,” *Probab. Surv.*, vol. 2, pp. 107–144, 2005.
- [50] I. A. Ibragimov and Y. A. Rozanov, *Gaussian Random Processes*. Springer, 1978.
- [51] A. H. Murphy, “A new vector partition of the probability score,” *J. Appl. Meteorol.*, vol. 12, no. 4, pp. 595–600, 1973.
- [52] C. W. J. Granger, *Empirical Modeling in Economics: Specification and Evaluation*. Cambridge University Press, 1999.
- [53] J. Rissanen, “Modeling by shortest data description,” *Automatica*, vol. 14, no. 5, pp. 465–471, 1978.
- [54] A. R. Barron, J. Rissanen, and B. Yu, “The minimum description length principle in coding and modeling,” *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2743–2760, 1998.
- [55] G. A. Darbellay and I. Vajda, “Estimation of the information by an adaptive partitioning of the observation space,” *IEEE Trans. Inf. Theory*, vol. 45, no. 4, pp. 1315–1321, 1999.
- [56] 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.
- [57] L. F. Kozachenko and N. N. Leonenko, “Sample estimate of the entropy of a random vector,” *Probl. Peredachi Inf.*, vol. 23, no. 2, pp. 9–16, 1987.
- [58] P. Good, *Permutation, Parametric and Bootstrap Tests of Hypotheses*, 3rd ed. Springer, 2005.
- [59] J. Theiler, S. Eubank, A. Longtin, B. Galdrikian, and J. D. Farmer, “Testing for nonlinearity in time series: The method of surrogate data,” *Physica D*, vol. 58, no. 1–4, pp. 77–94, 1992.
- [60] 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.
- [61] F. Petropoulos, D. Apiletti, V. Assimakopoulos, et al., “Forecasting: Theory and practice,” *Int. J. Forecast.*, vol. 38, no. 3, pp. 845–1154, 2022.
- [62] R. Fildes, P. Goodwin, M. Lawrence, and K. Nikolopoulos, “Effective forecasting and judgmental adjustments: An empirical evaluation and strategies for improvement in supply-chain planning,” *Int. J. Forecast.*, vol. 25, no. 1, pp. 3–23, 2009.
- [63] M. P. Clements and D. F. Hendry, *Forecasting Economic Time Series*. Cambridge University Press, 1998.
- [64] J. D. Hamilton, “A new approach to the economic analysis of nonstationary time series and the business cycle,” *Econometrica*, vol. 57, no. 2, pp. 357–384, 1989.
- [65] R. J. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*, 3rd ed. OTexts, 2021.
- [66] W. Ehm, T. Gneiting, A. Jordan, and F. Krüger, “Of quantiles and expectiles: Consistent scoring functions, Choquet representations and forecast rankings,” *J. R. Statist. Soc. B*, vol. 78, no. 3, pp. 505–562, 2016.
- [67] T. Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression*. Prentice-Hall, 1971.
- [68] F. X. Diebold and R. S. Mariano, “Comparing predictive accuracy,” *J. Bus. Econ. Statist.*, vol. 13, no. 3, pp. 253–263, 1995.
- [69] L. J. Tashman, “Out-of-sample tests of forecasting accuracy: An analysis and review,” *Int. J. Forecast.*, vol. 16, no. 4, pp. 437–450, 2000.