Online Forecast Evaluation with High-Frequency Proxies

Jia Li[*], Yifan Li[†], Andrew J. Patton[‡], Tian Xie[§]

March 4, 2025

Abstract

We present methods to evaluate risk forecasts "online," enabling the analyst to determine in real time whether a forecasting model is producing "statistically acceptable" forecasts. We use the asymptotic distribution of the volatility proxy used for forecast evaluation to construct a confidence set for the true, latent, risk measure and we use the inclusion, or not, of the forecast in that set as a formal measure of acceptability. We consider the construction of the confidence sets in both large- and small-sample settings, allowing us to consider forecasts of quantities like integrated variance and spot volatility. We illustrate the method by evaluating spot volatility forecasts for a newly listed stock, including on the day of its IPO.

JEL classifications: C12, C53, C58, G14

Keywords: spot volatility, realized variance, volatility forecasting, initial public offerings

[*]School of Economics, Singapore Management University, Singapore.

[†]Accounting and Finance Division, University of Manchester, UK.

[‡]School of Banking & Finance, UNSW Sydney, Australia; Department of Economics, Duke University, USA; School of Economics, Singapore Management University, Singapore.

[§]College of Business, Shanghai University of Finance and Economics, Shanghai, China.

1 Introduction

Economic forecasts are being produced and acted upon at an ever-increasing pace: algorithmic traders place buy and sell orders at speeds below a millisecond; A/B tests run by online retailers can conclude within hours of starting; advances in natural language processing allow economists and traders to make financial decisions in just seconds after a lengthy statement is released by the Federal Reserve Board.[1] The need for methods to evaluate forecasts in an accurate and timely manner is more pressing than ever.

Economic forecasts are typically evaluated using the average predictive loss over a pseudo-out-of-sample, or test sample, of data, usually spanning a significant period of time (e.g., see Diebold and Mariano, 1995 and West, 1996). Complicating matters, in financial forecasting the targets are often measures of risk, such as volatility or correlation, which are not directly observable. This presents a significant challenge, as the accuracy of forecasts cannot be directly assessed. Over the past two decades, the standard practice has been to approximate these latent targets using "realized measures" derived from high-frequency asset price data. Researchers commonly rely on technical conditions to argue that the proxy error is "asymptotically negligible," enabling them to evaluate forecasts as if the realized measures were the true targets (Hansen and Lunde, 2006; Patton, 2011; Li and Patton, 2018).

We propose a new period-by-period, or "online," evaluation methodology for forecasts of latent risk measures, enabling researchers to determine whether a forecast is "statistically acceptable," in a sense defined precisely below, after a single realization of the target. This approach reveals two equivalent formal interpretations: as a hypothesis test for oracle forecasts, or a construction of loss-induced "evaluation confidence sets" for the latent risk measure. Unlike conventional approaches, our framework explicitly accounts for the sampling uncertainty in the realized proxy, making it especially useful for evaluating forecasts of quantities like spot (or instantaneous) volatility, where the proxy error can be substantial.

We explore two distinct asymptotic settings for implementing this methodology. The first assumes that the realized measure is constructed from a "large" number of high-frequency ob-

[1] See Lewis (2014), Kohavi et al. (2020) and Saret and Mitra (2016) for examples.

servations, resulting in standard large-sample asymptotics. This setting encompasses most high-frequency estimators, including various integrated functionals of volatility, jumps, and microstructure noise (Andersen et al. (2003); Barndorff-Nielsen et al. (2008); Jacod and Rosenbaum (2013); Li (2013); Wang and Mykland (2014); Jacod et al. (2019)). The second setting adopts a small-sample framework, where the realized measure is based on a "fixed" number of observations. As advocated by Bollerslev et al. (2021), the latter approach provides more reliable inference for spot estimators, making it well-suited for evaluating forecasts of spot volatility and spot beta. Extensions to multivariate and multi-period frameworks are also discussed.

The theoretical high-frequency econometrics literature has predominantly focused on deriving the asymptotic properties of the estimation error in various volatility proxies, while the forecasting literature, which is one of the most popular applications of high frequency volatility measures, has, for the most part, made little use of these results, instead treating the estimation error as "negligible." This paper represents a bridge between these two literatures, using the theoretical insights of the former to obtain novel and useful methods for the latter.

The rest of the paper is organized as follows. Section 2 introduces the conceptual framework for online forecast evaluation and details its implementation under both large-sample and small-sample asymptotic frameworks. Section 3 presents an empirical application of the method, evaluating spot volatility forecasts immediately following the initial public offering (IPO) of a firm. Section 4 concludes. The Supplemental Appendix contains all proofs and extensions of the methodology to multivariate settings and multi-period forecast horizons, as well as additional empirical results.

2 Online forecast evaluation

2.1 Motivation and framework

Let R_t denote the true latent forecast target, such as integrated volatility or spot volatility of an asset, and let F_t represent the corresponding forecast. Additionally, consider a realized measure $\hat{R}_t$, typically constructed as an estimator of R_t using high-frequency asset price data (Foster and Nelson, 1996; Andersen et al., 2003; Barndorff-Nielsen and Shephard, 2004). If R_t were observable,

forecast accuracy would ideally be evaluated using a loss function $\mathcal{L}(F_t, R_t)$, which quantifies the error between the forecast and the true latent target. However, since R_t is unobservable in practice, the proxy loss $\mathcal{L}(F_t, \hat{R}_t)$ is commonly employed. This approach assumes that the estimation error in $\hat{R}_t$ can be ignored under certain conditions on the loss function and technical assumptions about the proxy $\hat{R}_t$ (Hansen and Lunde, 2006; Patton, 2011; Li and Patton, 2018). These assumptions, however, can be difficult to validate in practice, particularly when $\hat{R}_t$ exhibits finite-sample bias or has a slow rate of convergence.

Forecast evaluation typically relies on the average loss over a "testing" sample, expressed as $T^{-1}\sum_{t=1}^{T}\mathcal{L}(F_t, \hat{R}_t)$. The conventional approach (Diebold and Mariano, 1995; West, 1996) employs a long-span asymptotic framework, assuming $T \to \infty$. Inference based on the average loss pertains to the expected loss $\mathbb{E}[\mathcal{L}(F_t, \hat{R}_t)]$, which reflects the "long-run" performance of the forecast relative to the proxy target. In academic work, this method is frequently applied to "pseudo out-of-sample" observations, comprised of a subsample of recent observations. In practice, however, the focus is on genuine out-of-sample performance, and new data typically arrives only slowly.

While the standard "long-run" approach offers valuable insights into average forecast performance, it falls short in scenarios requiring immediate responses to new or changing conditions. To address this, we propose a complementary "online" evaluation framework that evaluates forecasts on a period-by-period basis. This method provides timely and actionable feedback by statistically assessing whether the observed proxy loss in a single period is "statistically acceptable" relative to a theoretical threshold. Such granularity is crucial for identifying abrupt shifts in data patterns, allowing practitioners to adapt their strategies in real time.

To formalize this framework, we test the null hypothesis $H_0 : R_t = F_t$, which represents the "oracle" scenario where the latent target R_t perfectly matches the forecast F_t. Since R_t is unobserved, we measure the gap between R_t and F_t using the proxy loss $\mathcal{L}(F_t, \hat{R}_t)$, which serves as a test statistic. At a significance level α, we aim to derive an asymptotically valid critical value function $\bar{L}_{1-\alpha}(\cdot)$ such that:

$$\mathbb{P}(\mathcal{L}(F_t, \hat{R}_t) > \bar{L}_{1-\alpha}(F_t)) \to \alpha \quad \text{under} \quad H_0 : R_t = F_t. \tag{1}$$

A forecast F_t is considered "acceptable" if the proxy loss $\mathcal{L}(F_t, \hat{R}_t)$ does not exceed the tolerance level $\bar{L}_{1-\alpha}(F_t)$. Allowing the critical value to depend on F_t simplifies the construction of certain confidence sets in subsequent analysis.

By the duality between hypothesis tests and confidence sets, the methodology can also be interpreted through a confidence set for R_t. Under the null restriction $R_t = F_t$, condition (1) can be reformulated as:

$$\mathbb{P}(\mathcal{L}(R_t, \hat{R}_t) \leq \bar{L}_{1-\alpha}(R_t)) \to 1 - \alpha, \tag{2}$$

implying that:

$$\mathrm{CS}_{1-\alpha} \equiv \{r : \mathcal{L}(r, \hat{R}_t) \leq \bar{L}_{1-\alpha}(r)\} \tag{3}$$

is an asymptotic $1-\alpha$ confidence set for the latent risk measure R_t, satisfying $\mathbb{P}(R_t \in \mathrm{CS}_{1-\alpha}) \to 1-\alpha$. The forecast F_t is deemed "acceptable" if and only if it lies within the confidence set $\mathrm{CS}_{1-\alpha}$.

The two interpretations—hypothesis testing and confidence sets—provide complementary perspectives, offering flexibility based on the context. For practitioners, the confidence-set approach is particularly intuitive, as it shifts the focus from achieving an exact point target to ensuring forecasts fall within an acceptable range that accounts for the estimation noise in the proxy $\hat{R}_t$. This represents a fundamental departure from conventional methods, which often disregard the error introduced by the proxy.

The confidence set for R_t defined in (3) notably depends on the choice of the loss function, reflecting the decision maker's economic preferences. This also contrasts sharply with confidence intervals for high-frequency estimators commonly discussed in the literature, which are primarily designed to minimize length.

It is useful to note, however, that the confidence set is invariant under strictly increasing transformations of the loss function.[2] As a direct consequence, both the absolute deviation loss, $|F_t - \hat{R}_t|$, and the quadratic loss, $(F_t - \hat{R}_t)^2$, produce the same confidence set. This invariance highlights the ordinal nature of the proposed evaluation framework, in contrast to conventional evaluation pro-

[2]If $\mathcal{L}'(\cdot,\cdot)$ is another loss function such that $\mathcal{L}'(\cdot,\cdot) = g(\mathcal{L}(\cdot,\cdot))$ for some strictly increasing function $g(\cdot)$, then $g(\bar{L}_{t,1-\alpha}(F_t))$ serves as the critical value for $\mathcal{L}'(F_t, \hat{R}_t)$, and the associated confidence set $\{r : \mathcal{L}'(r, \hat{R}_t) \leq g(\bar{L}_{t,1-\alpha}(r))\}$ is identical to $\{r : \mathcal{L}(r, \hat{R}_t) \leq \bar{L}_{t,1-\alpha}(r)\}$.

cedures based on average loss, which adopt a cardinal perspective akin to expected utility theory. By focusing on ordinal properties, our framework emphasizes the "statistical acceptability" of forecasts as an alternative interpretation of forecast rationality (cf. Granger (1969)) rather than the expected predictive loss, offering a fundamentally different and complementary approach to forecast evaluation.

In Sections 2.2 and 2.3, we provide a precise characterization of the confidence set under more specific assumptions about the asymptotic properties of the high-frequency estimator $\hat{R}_t$. Section 2.2 examines a classic setting (Foster and Nelson, 1996; Andersen et al., 2003; Barndorff-Nielsen and Shephard, 2004; Jacod and Protter, 2012) where $\hat{R}_t$ is constructed using a "large" number of observations, resulting in its concentration within an asymptotically shrinking neighborhood of the true value. In this context, only the local shape of the loss function is relevant. When the loss function is symmetric, the resulting confidence interval aligns with conventional length-minimizing confidence intervals. By contrast, Section 2.3 explores an alternative setting where $\hat{R}_t$ is constructed using a "small" (i.e., fixed) number of high-frequency observations. This approach is particularly relevant for inference on instantaneous, or "spot," quantities estimated over short time windows (Bollerslev et al., 2021). In this setting, the global, rather than local, behavior of the loss function becomes important, leading to loss-driven confidence sets that generally differ from conventional confidence intervals. We now turn to the details.

2.2 The case with a large number of high-frequency observations

In this subsection, we implement the general inference procedure in a setting where $\hat{R}_t$ is constructed from a "large" number, $n \to \infty$, of high-frequency observations. A prominent example is the realized variance estimator for integrated variance, defined as the sum of squared intraday high-frequency returns (Andersen et al., 2003; Barndorff-Nielsen and Shephard, 2004) over a day, which exhibits asymptotic mixed normality with a $\sqrt{n}$ convergence rate.

Other estimators explicitly addressing microstructure noise have also been proposed, achieving asymptotic mixed normality but with slower convergence rates, such as $n^{1/4}$ or lower (Zhang et al., 2005; Barndorff-Nielsen et al., 2008; Jacod et al., 2009; Da and Xiu, 2021). Kernel-based nonpara-

metric estimators for spot volatility exhibit asymptotic mixed normality but converge at a slower, nonparametric rate (Foster and Nelson, 1996; Kristensen, 2010).

The following high-level condition is designed to accommodate these classic examples and related cases, with a_n representing the estimator's rate of convergence and $\xrightarrow{\mathcal{L}\text{-}s}$ denoting stable convergence in law.[3]

Assumption 1 *For some real sequence $a_n \to \infty$ and estimator $\hat{\Sigma}_t$: (i) $a_n(\hat{R}_t - R_t) \xrightarrow{\mathcal{L}\text{-}s} \mathcal{MN}(0, \Sigma_t)$ for some strictly positive random variable Σ_t; (ii) $\hat{\Sigma}_t \xrightarrow{\mathbb{P}} \Sigma_t$.*

Part (a) of Assumption 1 states that $\hat{R}_t$ consistently estimates R_t at rate a_n, with the scaled error following an asymptotically mixed Gaussian distribution with conditional variance Σ_t. Besides the aforementioned examples, this condition is satisfied by many other estimators for general integrated volatility or jump functionals (Jacod and Protter, 2012; Jacod and Rosenbaum, 2013) and estimators for "deep" parameters such as integrated leverage effect and volatility-of-volatility (Wang and Mykland, 2014; Kalnina and Xiu, 2017; Li et al., 2022).[4] Part (b) ensures that $\hat{\Sigma}_t$ consistently estimates Σ_t, enabling feasible inference.

Together, these conditions imply the "standard" $1-\alpha$ confidence interval for R_t that is routinely employed in practice:

$$\mathrm{CI}_{1-\alpha} = [\hat{R}_t - q_{1-\alpha}\hat{\Sigma}_t^{1/2}/a_n, \hat{R}_t + q_{1-\alpha}\hat{\Sigma}_t^{1/2}/a_n], \tag{4}$$

where $q_{1-\alpha}$ denotes the $1-\alpha$ quantile of the absolute value of a standard normal variable. This standard benchmark will be contrasted with the loss-driven confidence set proposed in (3).

To analyze the properties of the confidence set, we focus on loss functions that can be expressed in a structured form, consisting of a leading term and a "residual" term:

$$\mathcal{L}(u, \hat{u}) = \frac{L(\hat{u} - u)}{S(u, \hat{u})} + G(u, \hat{u}), \tag{5}$$

[3]Stable convergence in law is a stronger notion than the usual weak convergence, often employed in high-frequency econometric analysis to account for the randomness of the asymptotic variance in limit theorems. See Jacod and Protter (2012) for theoretical details.

[4]Important exceptions include realized semivariance and covariance measures (Barndorff-Nielsen et al., 2010; Bollerslev et al., 2020), which do not admit a central limit theorem due to second-order biases.

where the components of the loss function are required to satisfy the following assumption.

Assumption 2 *The loss function has the form (5) such that (i) $L(\cdot)$ is pth order homogeneous for some constant $p > 0$; (ii) $S(\cdot,\cdot)$ is continuous and $S(R_t, R_t) > 0$; (iii) $G(R_t, \hat{R}_t) = o_p(a_n^{-p})$ with a_n defined in Assumption 1.*

Assumption 2 accommodates most commonly used loss functions. For example, when $S(u,\hat{u}) \equiv 1$, $G(u,\hat{u}) \equiv 0$, and $L(x) = |x|^p$, the resulting loss function $\mathcal{L}(u,\hat{u}) = |\hat{u} - u|^p$ includes both the absolute deviation loss ($p = 1$) and the quadratic loss ($p = 2$). Asymmetric loss functions, such as the lin-lin loss, are covered by setting $L(x) = |x|^p 1_{\{x \geq 0\}} + \lambda |x|^p 1_{\{x<0\}}$. The scaling factor $S(u,\hat{u})$ can be introduced to ensure scale invariance; for instance, when $S(u,\hat{u}) \equiv |u|^p$ and $L(x) = |x|^p$, we have $\mathcal{L}(u,\hat{u}) = |\hat{u}/u - 1|^p$, which is useful for analyzing "scale quantities" such as volatility. The residual term $G(u,\hat{u})$ further broadens the class of loss functions. For example, the q-like loss

$$\mathcal{L}(u,\hat{u}) = (\hat{u}/u) - \log(\hat{u}/u) - 1 \tag{6}$$

satisfies this assumption, with a leading term proportional to $|\hat{u}/u - 1|^2$, the scale-invariant quadratic loss.

Proposition 1, below, characterizes the confidence set defined in (3) under Assumptions 1 and 2, where $Q_{Z,1-\alpha}$ denotes the $1-\alpha$ quantile of a generic random variable Z.

Proposition 1 *Let ξ denote a standard normal random variable. Under Assumptions 1 and 2, the following statements hold true:*

(a) $\mathbb{P}\left(\mathcal{L}(R_t, \hat{R}_t) \leq \bar{L}_{1-\alpha}(R_t)\right) \to 1-\alpha$ holds for

$$\bar{L}_{1-\alpha}(R_t) = \frac{a_n^{-p} Q_{L(\xi),1-\alpha} \hat{\Sigma}_t^{p/2}}{S(R_t, \hat{R}_t)} + G(R_t, \hat{R}_t). \tag{7}$$

(b) The confidence set defined in (3) can be expressed as

$$\mathrm{CS}_{1-\alpha} \equiv \left\{ r : L\left(\frac{a_n\left(\hat{R}_t - r\right)}{\hat{\Sigma}_t^{1/2}} \right) \leq Q_{L(\xi),1-\alpha} \right\}.$$

(c) If $x \mapsto L(x)$ is symmetric, the confidence set coincides with the "standard" confidence interval defined in (4).

COMMENT. Part (a) establishes a valid critical value function $\bar{L}_{1-\alpha}(\cdot)$, as defined in (7). Under Assumption 2(iii), the adjustment term $G(R_t, \hat{R}_t) = o_p(a_n^{-p})$ is asymptotically negligible compared to the leading term, which is of order $O_p(a_n^{-p})$. While this adjustment term is not strictly required for the derivation in part (a), its inclusion simplifies the confidence set expressions in parts (b) and (c) and reveals the underlying common structure across seemingly distinct problems.

The proposition demonstrates that many seemingly distinct loss functions belong to the same "equivalence class" in terms of the resulting confidence set. Specifically, loss functions that share the same $L(\cdot)$ up to a strictly increasing transformation yield identical confidence sets. For instance, widely used loss functions such as absolute deviation and quadratic losses (with or without scale-invariance normalization), and q-like loss are equivalent in this sense. This equivalence arises because, in the large-sample setting, the asymptotic analysis is governed solely by the local behavior of the loss function near $\hat{R}_t \approx R_t$, where many globally distinct loss functions exhibit similar local properties.

By contrast, the next subsection examines a setting where the large-sample assumption no longer applies, leading to distinct inference results and highlighting the importance of the global properties of the loss function.

2.3 The case with a small number of high-frequency observations

We now turn to an alternative asymptotic framework where the estimator $\hat{R}_t$ is constructed using a fixed number, k, of high-frequency observations. This small-sample setting, introduced by Bollerslev et al. (2021) for spot volatility estimation, has since been extended to various contexts (Li et al., 2024; Bollerslev et al., 2024a). By avoiding the assumption of a "large" sample size, the fixed-k approach minimizes nonparametric estimation biases and addresses distortions that arise from the inadequacy of conventional asymptotic Gaussian approximations in small samples. This framework is particularly well-suited for "local" estimation problems, such as those involving spot volatility and similar quantities.

Unlike in the large-sample setting, where $\hat{R}_t$ closely approximates R_t and only the local behavior of the loss function near $\hat{R}_t \approx R_t$ matters, the fixed-k framework depends on the global behavior of the loss function. This distinction arises because $\hat{R}_t$, based on a fixed number of observations, can no longer be claimed as a consistent estimator for R_t. Furthermore, as illustrated in the examples below, the fixed-k limit distribution of $\hat{R}_t$ is generally nonstandard and not mixed Gaussian, further distinguishing this framework from the large-sample asymptotics discussed in Section 2.2.

Following the existing literature, we analyze two estimation settings. The first focuses on cases where R_t represents a positive "scale parameter," such as volatility or idiosyncratic variance. To facilitate this analysis, we impose the following regularity conditions:

Assumption 3 *The following conditions hold: (i)* $\hat{R}_t/R_t \xrightarrow{d} \xi$ *for some continuous random variable* ξ *with a known distribution; (ii) the loss function takes the form* $\mathcal{L}(u, \hat{u}) = L(\hat{u}/u)$ *for a non-constant convex loss function* $L : \mathbb{R}_+ \to \mathbb{R}_+$.

Condition (i) describes the asymptotic property of the $\hat{R}_t$ estimator through the "self-normalized" quantity $\hat{R}_t/R_t$. This formulation differs from the central limit theorems typically encountered in large-sample analysis as seen in Assumption 1, reflecting the distinctive features of small-sample frameworks. The resulting limit distribution in these settings is often nonstandard, as illustrated in the examples below. The self-normalized limit theorem naturally aligns with the scale-invariant loss function specified in condition (ii). Moreover, the additional convexity requirement on the loss function, satisfied by many commonly used examples, enables the simplification of confidence sets into confidence intervals.

Example 1 (Return-based Spot Variance Estimator). Let $(r_i)_{1\le i\le k}$ be high-frequency asset returns sampled at frequency $\Delta_n \to 0$ near time t. The standard spot variance estimator is given by $\hat{R}_t = (k\Delta_n)^{-1}\sum_{i=1}^{k} r_i^2$. Let R_t denote the asset's spot variance at time t. With the sample size k fixed, Bollerslev et al. (2021) exploited the local Gaussianity property of the standard Itô semimartingale model and employed a coupling argument to establish that $\hat{R}_t/R_t \xrightarrow{d} \xi \sim \chi_k^2/k$, where χ_k^2/k represents a scaled chi-squared distribution with k degrees of freedom. □

Example 2 (Candlestick-based Spot Volatility Estimator). Li et al. (2024) proposed

the Best Linear Unbiased Estimator (BLUE) for spot volatility using high-frequency candlestick observations. By leveraging the richer information in candlesticks, this estimator is substantially more accurate than those based only on returns. Let w_i denote the price range associated with the return r_i for the ith time interval. The BLUE takes the form $\hat{R}_t = k^{-1}\Delta_n^{-1/2}\sum_{i=1}^{k}(0.811w_i - 0.369|r_i|)$, where the weights are chosen to minimize asymptotic variance. Let R_t be the asset's spot volatility. These authors show that $\hat{R}_t/R_t \xrightarrow{d} \xi$, where the limit variable is represented as

$$\xi = \frac{1}{k}\sum_{i=1}^{k}\left\{0.811 \cdot \sup_{s,u\in[0,1]}|B_s^{(i)} - B_u^{(i)}| - 0.369 \cdot |B_1^{(i)}|\right\},$$

for independent copies of standard Brownian motions, $(B^{(i)})_{1\leq i\leq k}$. Bollerslev et al. (2024a) extend this framework to study optimal estimators without shape restrictions and established similar convergence results for risk-minimizing estimators. We omit the discussion on their more complicated estimators for brevity. □

Proposition 2, below, characterizes the confidence set under Assumption 3. As noted earlier, the convexity of $L(\cdot)$ enables the simplification of confidence sets into confidence intervals. To formalize this, we introduce the following notation: for a threshold level $Q > 0$, the lower level set $\{x : L(x) \leq Q\}$, if nonempty, forms a closed interval, denoted as $[\underline{c}(L, Q), \overline{c}(L, Q)]$.

Proposition 2 *Under Assumption 3,* $\mathbb{P}(\mathcal{L}(R_t, \hat{R}_t) \leq \bar{L}_{1-\alpha}) \to 1-\alpha$ *where* $\bar{L}_{1-\alpha} = Q_{L(\xi),1-\alpha}$. *Moreover, the corresponding confidence set defined in (3) can be expressed as:*

$$\mathrm{CS}_{1-\alpha} = \left[\frac{\hat{R}_t}{\overline{c}\left(L, Q_{L(\xi),1-\alpha}\right)}, \frac{\hat{R}_t}{\underline{c}\left(L, Q_{L(\xi),1-\alpha}\right)}\right].$$

The analysis thus far has focused on cases where the forecast target R_t is a "scale parameter," such as spot volatility. While Assumption 3 is tailored for this class of problems, it is not suitable for other types of forecast targets. A notable example is spot beta, as examined in Bollerslev et al. (2024b). To extend the framework to accommodate a broader range of applications, we introduce the following assumption.

Assumption 4 *The following conditions hold: (i)* $(\hat{R}_t - R_t)/\hat{\Sigma}_t^{1/2} \xrightarrow{d} \xi$ *for some continuous random variable* ξ *with a known distribution and some estimator* $\hat{\Sigma}_t$*; (ii) the loss function takes the form* $\mathcal{L}(u, \hat{u}) = L(\hat{u} - u)$ *for a pth-order homogeneous loss function* $L : \mathbb{R} \to \mathbb{R}_+$ *for some constant* $p > 0$*.*

Condition (i) specifies the asymptotic behavior required for the $\hat{R}_t$ estimator, representing a significant departure from Assumption 1. Unlike the large-sample framework, no scaling factor (e.g., a_n) is introduced, and $\hat{R}_t$ is not assumed to converge to R_t. Additionally, $\hat{\Sigma}_t$ is not required to consistently estimate any "asymptotic variance," which is irrelevant in the small-sample context, particularly because the limit distribution is not (mixed) Gaussian.

A notable example within this framework is spot beta, the small-sample behavior of which has been studied by Bollerslev et al. (2024b). In this case, R_t is the spot beta of a stock relative to a market portfolio and $\hat{R}_t$ is obtained by regressing the asset's high-frequency returns on the market portfolio's returns over a sample of size k. The variance estimator $\hat{\Sigma}_t$ is defined as proportional to the ratio of the asset's spot idiosyncratic variance estimator to the market's spot variance estimator, ensuring that $(\hat{R}_t - R_t)/\hat{\Sigma}_t^{1/2}$ corresponds to the t-statistic of the regression coefficient. By coupling the nonparametric spot beta estimation problem with a finite-sample Gaussian linear regression limit experiment, Bollerslev et al. demonstrated that this t-statistic converges in distribution to a random variable ξ following a t-distribution with $k - 1$ degrees of freedom.

The loss function in condition (ii) of Assumption 4 supports a broad class of location-invariant possibly asymmetric polynomial losses. In contrast to Assumption 3, scale-invariant losses are not considered here, as such restriction is less relevant for quantities like spot beta. Like convex losses, this type of scale-homogeneous loss functions also permits the characterization of confidence sets as confidence intervals, as described in the following proposition.

Proposition 3 *Under Assumption 4,* $\mathbb{P}(\mathcal{L}(R_t, \hat{R}_t) \leq \bar{L}_{1-\alpha}) \to 1 - \alpha$ *where* $\bar{L}_{1-\alpha} = \hat{\Sigma}_t^{p/2} Q_{L(\xi),1-\alpha}$*. Moreover, the accompanying confidence set defined in (3) can be expressed as*

$$\mathrm{CS}_{1-\alpha} = \left[\hat{R}_t - \hat{\Sigma}_t^{1/2} \overline{c}\left(L, Q_{L(\xi),1-\alpha}\right), \hat{R}_t - \hat{\Sigma}_t^{1/2} \underline{c}\left(L, Q_{L(\xi),1-\alpha}\right)\right].$$

If $x \mapsto L(x)$ *is symmetric, then the confidence set simplifies to* $\mathrm{CS}_{1-\alpha} = [\hat{R}_t - \hat{\Sigma}_t^{1/2} Q_{|\xi|,1-\alpha}, \hat{R}_t -$

$\hat{\Sigma}_t^{1/2} Q_{|\xi|,1-\alpha}]$.

COMMENT. In comparison to the "standard" confidence interval (4), the confidence intervals described in Proposition 3 may be asymmetric unless the loss function $L(\cdot)$ is symmetric. Even under symmetry, these intervals differ because the limit variable ξ may follow a non-Gaussian distribution (e.g., t-distribution), resulting in a different critical value than in the "standard" case.

2.4 Extensions

The previous subsections have focused on single-period and univariate settings. The same conceptual framework for forecast evaluation can also be applied in more general settings. In Supplemental Appendix A, we present extensions in two key directions: multi-period evaluation and multivariate forecasting targets.

The multi-period extension enables the joint evaluation of forecast performance across a sequence of time periods, effectively addressing a multiple testing problem. This extension introduces confidence bands to assess the overall accuracy of forecast paths and provides a formal framework for hypothesis testing over extended time horizons. The methodology incorporates a "sup" statistic, capturing the worst-case proxy loss across multiple periods, ensuring robust and uniform coverage of confidence bands.

The multivariate extension generalizes the framework to vector-valued forecast targets, such as integrated covariance matrices and spot betas. This extension highlights the subtleties of trade-offs across forecast components and introduces a broader class of loss functions to accommodate the added complexity. The interplay between components and their weighting, as governed by the loss function, leads to distinct inferential outcomes compared to the univariate setting.

3 Spot volatility forecasting for newly IPO'd stocks

3.1 The empirical setting

To highlight the novelty of the proposed evaluation framework in an empirical setting, we consider forecasting the spot volatility of a stock immediately after its initial public offering (IPO), when

it first starts trading on the stock market. This is an extreme scenario where historical data is inherently limited for both model training and forecast evaluation, but it is an economically important one given that there are an average of over 200 IPOs in the U.S. per year.[5] Since the new method is designed for real-time forecast evaluation with short samples, this setting naturally highlights its key advantages. We emphasize that the method proposed in this paper is also valuable when the available sample is "effectively short," such as after major corporate events like mergers, spin-offs, index inclusions or exclusions, and other structural shifts that alter a firm's characteristics.

We analyze forecasts for Advanced RISC Machines (ARM), a British semiconductor and software design company, as our leading example. ARM's IPO on September 14, 2023, was one of the most significant market events of the year, drawing widespread investor attention. Indeed, this firm was included in the NASDAQ 100 index on June 24, 2024, less than a year after its IPO.[6] We consider the problem faced by an analyst tasked with forecasting the spot volatility of ARM's stock price immediately after it begins trading, and evaluating the quality of those forecasts in real time. As high-frequency transaction data accumulate, we compute spot volatility estimates (i.e., $\hat{R}_t$) progressively on a granular 5-minute time grid and assess whether the corresponding forecasts are "statistically acceptable" at a given confidence level using the proposed inference method.

It is worth noting that constructing spot volatility forecasts in this empirical scenario is a non-standard and challenging task, due to the lack of historical data for model training. In particular, when forecasting for the first trading day, the analyst in this scenario has no prior data for the newly listed stock. To address this issue we adopt a simple transfer learning framework with a dynamic scale-adjustment mechanism. Specifically, historical data from the QQQ ETF, which tracks the NASDAQ 100 index, serve as "source data" for training the forecasting models.[7] As new stock data accumulate, its own "target data" gradually replace the "source data" within the rolling window for parameter estimation. Recognizing the scale differences between source and target data, we

[5] Source: https://site.warrington.ufl.edu/ritter/files/IPO-Statistics.pdf.

[6] In the supplemental appendix, we conduct similar analyses for five other recently IPO'd stocks that were later included in the S&P 500 index.

[7] To assess the robustness of our approach, we also consider an alternative specification where historical data from the SPY ETF, which tracks the S&P 500 index, are used as the source data. The results, reported in Supplemental Appendix C, show that this alternative choice leads to minimal differences, suggesting that the proposed method is not sensitive to the specific selection of source data.

dynamically adjust the source data scale to align with the target data as the latter becomes progressively available. We emphasize, however, that our primary contribution is the new evaluation framework, not the transfer learning approach we adopt. The latter is employed primarily as a practical tool for constructing forecasts in a short-sample setting, and its further refinement is left for future research.

We obtain high, low, open, and close observations (also known as "candlestick" observations) on ARM at a 5-minute frequency ($\Delta_n = 5$) from PiTrading Inc., covering the period from September 14, 2023, to February 7, 2024. Each candlestick is used to estimate spot volatility at 5-minute granularity using the candlestick-based BLUE estimator proposed in (Li et al., 2024). The rolling window size is set to 50 trading days, comprising 3,900 5-minute observations. We consider a range of forecast horizons, τ, from five minutes to one trading day.

We consider three forecasting models. The first is the multiplicative component GARCH (MC-GARCH) model, proposed by Engle and Sokalska (2012). This model decomposes spot volatility into daily, diurnal, and stochastic intraday components, providing a flexible framework for modeling volatility dynamics both across days and within intraday periods. Specifically, the stock return over the ith high-frequency observation interval on day t is modeled as follows for $1 \leq i \leq n$ and $1 \leq t \leq T$:

$$r_{t,i} = \sqrt{h_t s_i q_{t,i}}\epsilon_{t,i}, \quad \text{where} \quad \epsilon_{t,i} \sim \mathcal{N}(0,1),$$

where h_t represents the daily variance component, s_i captures the diurnal (intraday seasonal) pattern, $q_{t,i}$ is the high-frequency volatility component, and $\epsilon_{t,i}$ is the error term. For identification we impose $\mathbb{E}(q_{t,i}) = 1$ and $\mathbb{E}(s_i) = 1/n$.

Following Engle and Sokalska (2012), we estimate the daily conditional variance, h_t, using a realized variance estimator, $\hat{h}_{t,i-1}$, defined as the sum of the 78 squared high-frequency returns ending in the previous intradaily interval.[8] The diurnal variance component, s_i, is then estimated

[8] In the first interval of the day, this estimator corresponds to the familiar open-to-close realized variance for the previous trade day. For other intradaily intervals, it is an realized variance measure over the same number of observations, but ending at the start of the interval of interest.

as

$$\hat{s}_i = \frac{1}{T}\sum_{t=1}^{T}\frac{r_{t,i}^2}{\hat{h}_t},$$

for each intraday timestamp i. Finally, the normalized returns, $z_{t,i} = r_{t,i}/\sqrt{\hat{h}_t\hat{s}_i}$, are modeled as having the conditional distribution $\mathcal{N}(0, q_{t,i})$, where the conditional variance $q_{t,i}$ follows a GARCH(1,1) model:

$$q_{t,i} = (1-\alpha-\beta) + \alpha z_{t,i-1}^2 + \beta q_{t,i-1}.$$

The estimated GARCH model is then used to compute a τ-period-ahead forecast, $\hat{q}_{T,i+\tau}$, based on the training sample. The corresponding spot volatility forecast is given by $\sqrt{\hat{h}_T\hat{s}_{i+\tau}\hat{q}_{T,i+\tau}/\bar{q}}$, where the normalizing factor $\bar{q}$ is the sample average of $q_{t,i}$ over the training sample.

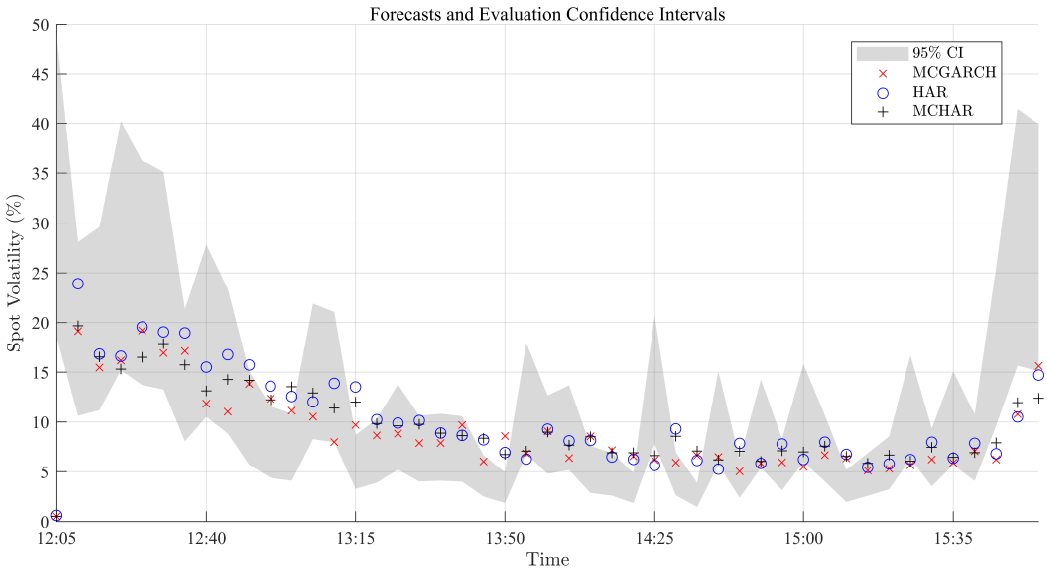
The second forecasting method is adapted from the heterogeneous autoregressive (HAR) model (Corsi, 2009), originally developed to forecast *daily* integrated volatility using lagged moving averages over daily, weekly, and monthly intervals. We modify the HAR model to predict spot volatility as follows. Let $\hat{R}_{t,i}$ denote the realized spot volatility at day t and interval i, computed using the candlestick-based BLUE estimator from Li et al. (2024). The τ-period-ahead HAR model is specified as

$$\hat{R}_{t,i+\tau} = \beta_0 + \beta_s\hat{R}_{t,i}^{(s)} + \beta_m\hat{R}_{t,i}^{(m)} + \beta_l\hat{R}_{t,i}^{(l)} + \epsilon_{t,i}, \tag{8}$$

where $\hat{R}_{t,i}^{(s)}$, $\hat{R}_{t,i}^{(m)}$, and $\hat{R}_{t,i}^{(l)}$ are backward-looking moving-averages of short-term (5 minutes), medium-term (1 hour), and long-term (1 day) volatility, respectively.

Finally, we note that compared to MCGARCH, a key limitation of the HAR model is its failure to explicitly account for intradaily seasonality in volatility dynamics. While this shortcoming is inconsequential in applications where HAR is used for daily volatility forecasting, it becomes relevant when forecasting intraday spot volatility. This motivates us to propose a new variant, the multiplicative component HAR (MCHAR). We estimate the daily volatility component and the diurnal component as in the MCGARCH model described above, and we model the normalized spot volatility component, $\hat{Z}_{t,i} = \hat{R}_{t,i}/(\hat{h}_{t,i-1}\hat{s}_i)$, using a HAR model analogous to equation (8).

Figure 1: Spot Volatility Forecasts on the IPO Date



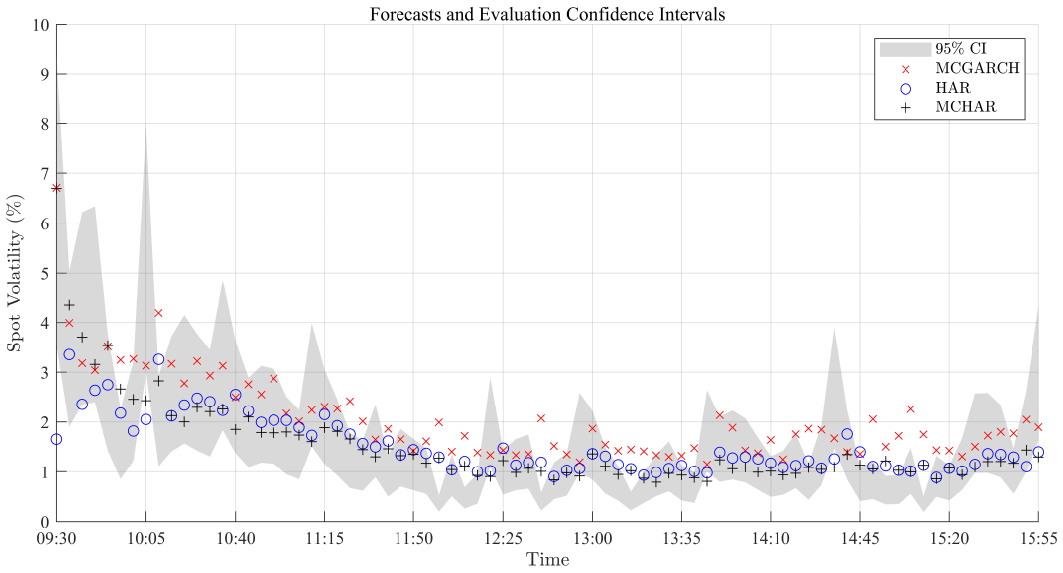
The figure presents 5-minute-ahead spot volatility forecasts of ARM based on the MCGARCH, HAR, and MCHAR models, along with 95% evaluation confidence intervals derived from the q-like loss, for ARM's IPO date, September 14, 2023. Like most IPOs, trading commenced at noon and not the usual opening time of 9:30am. Volatility is expressed in percentage terms per day.

3.2 Results

As an initial illustration, Figure 1 presents the 5-minute-ahead spot volatility forecasts from the three models described above, along with the 95% evaluation confidence intervals[9] derived in Proposition 2, throughout the day of ARM's IPO, September 24, 2023. This is obviously an atypical day, due to the very limited information available about ARM's stock price dynamics, the late start of trading (noon rather than 9:30am), and the elevated level of volatility generated by investors seeking to determine the equilibrium price for this asset. Nevertheless, this day represents a fascinating example of the usefulness of the proposed method: we see that aside from the very first forecast, when no historical data on this stock was available at all, for about the first 45 minutes all of the three methods produce "statistically acceptable" forecasts, perhaps predominantly due to the very wide confidence intervals for spot volatility during those first few periods of this stock's trading life. As the trading day progresses the level of volatility falls (from around 20% at the open to around

[9]For all confidence intervals we use the q-like loss function from equation (6); results based on quadratic loss are very similar and are omitted for brevity.

Figure 2: Spot Volatility Forecasts on a Representative Day



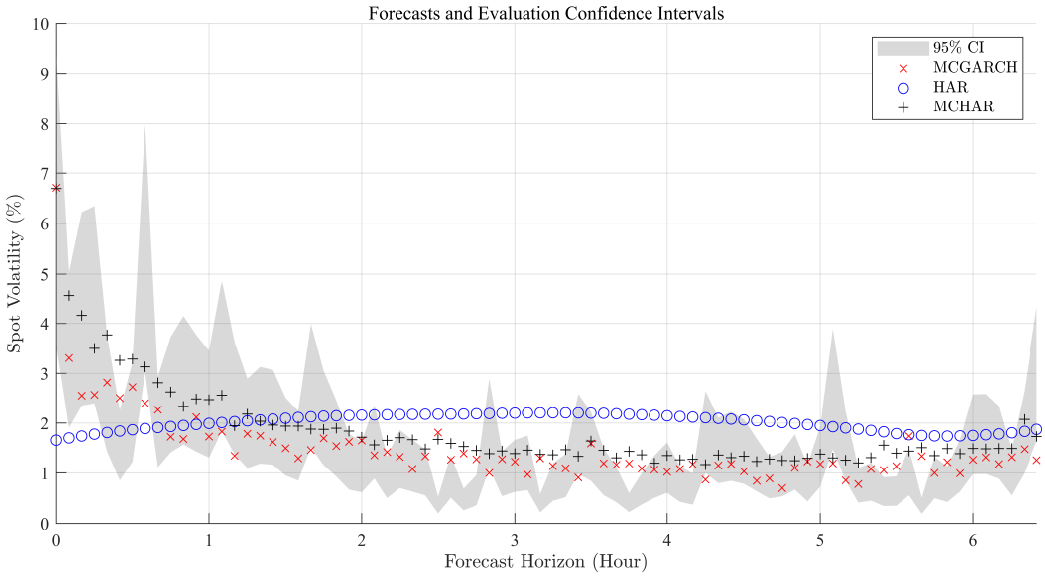
The figure presents 5-minute-ahead spot volatility forecasts of ARM based on the MCGARCH, HAR, and MCHAR models, along with 95% evaluation confidence intervals derived from the q-like loss, for November 29, 2023. Volatility is expressed in percentage terms per day.

5% an hour before the close) and the confidence intervals tighten, and we observe some forecasts falling outside of the confidence interval, representing a forecast failure, and one that is detected in (almost) real time.

In Figure 2 we next consider a more representative day for this stock, November 29, 2023, about two months after its IPO.[10] Although each confidence interval is constructed from a single high-frequency candlestick observation, they remain reasonably tight, consistent with the theoretical result that optimal candlestick-based spot volatility estimators are substantially more accurate than those based solely on returns at the same frequency (Li et al., 2024; Bollerslev et al., 2024a). Furthermore, the confidence intervals appear informative enough to distinguish among the alternative forecasts. Specifically, forecasts from the two HAR-type models generally fall within the evaluation confidence intervals, indicating their statistical acceptability. In contrast, MCGARCH forecasts more frequently fall outside the confidence intervals, suggesting the model's relative inferiority for short-horizon ($\tau = 5$ minutes) forecasting.

[10]The selected day is representative because the average acceptance rates of the forecasts on this day closely align with the overall sample averages, making it a suitable depiction of the forecasts' typical performance.

Figure 3: Spot Volatility Forecasts on a Representative Day over Multiple Horizons



The figure presents the τ-period-ahead spot volatility forecasts of ARM based on the MCGARCH, HAR, and MCHAR models, formed at the start of a representative trading day, November 29, 2023, with τ ranging from 5 minutes to 6.5 hours throughout the day. The shaded area represents 95% evaluation confidence intervals derived from the q-like loss. Volatility is expressed in percentage terms per day.

For the same day, Figure 3 presents forecasts over different horizons, all formed at 9:30 a.m. These forecasts illustrate the expected spot volatility paths throughout the day based solely on information available at the market open. This task is inherently more challenging than that in Figure 2, as the latter involves only short-horizon forecasts and benefits from continuously updated intraday information. Nevertheless, forecasts from the MCGARCH and MCHAR models, which incorporate built-in mechanisms for capturing intraday seasonality in spot volatility dynamics, perform well, as they generally fall within the evaluation confidence intervals. In contrast, the "plain" HAR model, which does not explicitly account for diurnal patterns, produces distinct and noticeably poor volatility forecasts at longer horizons. In contrast with standard forecast evaluation methods that use long out-of-sample periods and focus on long-run average prediction errors, this forecast failure is detectable in a formal statistical manner using the proposed method in real time, on the very day of the failure.

To evaluate the performance of the alternative forecasting methods over a longer sample Table

Table 1: Acceptance Rates for the Alternative Forecasts

Model	Forecast Horizon			
	5 min	1 hour	2 hour	4 hour
Panel A: Q-like Loss				
MCGARCH	0.6906	0.6418	0.6216	0.6077
HAR	0.7478	0.5760	0.5556	0.5385
MCHAR	0.7825	0.7298	0.7071	0.6862
Panel B: Quadratic Loss				
MCGARCH	0.7456	0.6926	0.6748	0.6624
HAR	0.7907	0.6377	0.6114	0.5963
MCHAR	0.8113	0.7640	0.7453	0.7185

Note: The table reports the acceptance rates of spot volatility forecasts for ARM stock using the MCGARCH, HAR, and MCHAR models. The acceptance rate is defined as the proportion of forecasts falling within the 95% evaluation confidence intervals, based on the q-like loss (Panel A) or the quadratic loss (Panel B), over a prediction sample spanning 100 trading days after the company's IPO on September 14, 2023. All models are trained using a 50-day rolling window scheme with 3,900 high-frequency observations. A transfer learning scheme is employed, using QQQ ETF data as source data to augment the training sample prior to the IPO.

1 reports the "acceptance rates" for each model, defined as the proportion of forecasts falling within the evaluation confidence intervals. A higher acceptance rate indicates better performance. The analysis is conducted under both the scale-invariant q-like loss and the quadratic loss, presented in Panels A and B respectively. It is important to note that a perfect forecast would result in an acceptance rate of 95%, aligning with the confidence level. Consistent with the patterns observed in Figures 2 and 3, HAR outperforms MCGARCH at the short 5-minute horizon, but underperforms MCGARCH at longer intraday horizons. Meanwhile, MCHAR consistently outperforms both HAR and MCGARCH, achieving an acceptance rate close to 80% at the short horizon and remaining mostly above 70% across various intraday horizons.[11]

The analysis above reveals that an intraday version of the HAR model of Corsi (2009) underperforms at longer horizons due to its lack of diurnal adjustment, while the MCGARCH of Engle and Sokalska (2012) underperforms at short horizons despite capturing intraday seasonality. In contrast, a new model that combines elements of both, dubbed the "MCHAR" model, is the best-performing model across all forecast horizons. Supplemental Appendix C presents corresponding analyses for another five stocks that IPO'd in 2023-24 and were later included in the S&P 500

[11]We also used the test of Diebold and Mariano (1995) to determine whether, for each pair of models at a given forecast horizon, the acceptance rates are significantly different. In all cases the p-values for the tests were less than 0.001, revealing that the differences reported in Table 1 are statistically significant at all conventional levels.

index, and reports similar empirical findings to those for ARM presented above.

4 Conclusion

This paper proposes a new method to evaluate risk forecasts period-by-period, or "online," enabling the analyst to determine in real time whether or not a given forecasting model is producing "statistically acceptable" forecasts. We do this by acknowledging and exploiting the sampling error in the volatility proxy that is used for forecast evaluation. We show how to construct a confidence set for the true, latent, risk measure and we use the inclusion, or exclusion, of the forecast in that set as a formal measure of "acceptability." We construct evaluation confidence sets in both large- and small-sample settings, allowing us to consider forecasts of quantities like integrated variance (large samples) and spot volatility (small samples). The theory reveals how these confidence sets depend on the local or global shape of loss functions and possibly nonstandard limit distributions of the realized measures. We also consider extensions to multivariate forecast evaluation, allowing us to consider forecasts of covariance matrices for example, and multi-period forecast evaluation.

We demonstrate the utility of our method by applying it to forecasting the volatility of a newly listed stock. This application shows that our framework is effective even in environments with very limited historical data, where traditional evaluation methods fall short. Our results indicate that forecast evaluation can begin on the first day data become available, that forecast breakdowns can be detected almost in real time, and that long-term performance can be compared by examining the average acceptance rate of the forecasting model.

References

Andersen, T. G., Bollerslev, T., Diebold, F. X., and Labys, P. (2003). Modeling and forecasting realized volatility. *Econometrica*, 71(2):pp. 579–625.

Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A., and Shephard, N. (2008). Designing realized kernels to measure the ex post variation of equity prices in the presence of noise. *Econometrica*, 76(6):1481–1536.

Barndorff-Nielsen, O. E., Kinnebrouck, S., and Shephard, N. (2010). Measuring downside risk: Realised semivariance. In Bollerslev, T., Russell, J., and Watson, M., editors, *Volatility and Time*

Series Econometrics: Essays in Honor of Robert F. Engle, pages 117–136. Oxford University Press.

Barndorff-Nielsen, O. E. and Shephard, N. (2004). Econometric analysis of realized covariation: High frequency based covariance, regression, and correlation in financial economics. *Econometrica*, 72(3):pp. 885–925.

Bollerslev, T., Li, J., and Li, Q. (2024a). Optimal nonparametric range-based volatility estimation. *Journal of Econometrics*, 238(1):105548.

Bollerslev, T., Li, J., and Liao, Z. (2021). Fixed-k inference for volatility. *Quantitative Economics*, 12(4):1053–1084.

Bollerslev, T., Li, J., Patton, A. J., and Quaedvlieg, R. (2020). Realized Semicovariances. *Econometrica*, 88(4):1515–1551.

Bollerslev, T., Li, J., and Ren, Y. (2024b). Optimal inference for spot regressions. *American Economic Review*, 114(3):678–708.

Corsi, F. (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, 7(2):174–196.

Da, R. and Xiu, D. (2021). When Moving-Average Models Meet High-Frequency Data: Uniform Inference on Volatility. *Econometrica*, 89(6):2787–2825.

Diebold, F. X. and Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3):253–263.

Engle, R. F. and Sokalska, M. E. (2012). Forecasting intraday volatility in the us equity market. multiplicative component garch. *Journal of Financial Econometrics*, 10(1):54–83.

Foster, D. and Nelson, D. B. (1996). Continuous record asymptotics for rolling sample variance estimators. *Econometrica*, 64:139–174.

Granger, C. W. J. (1969). Prediction with a generalized cost of error function. *Journal of the Operational Research Society*, 20(2):199–207.

Hansen, P. R. and Lunde, A. (2006). Consistent ranking of volatility models. *Journal of Econometrics*, 131(1-2):97–121.

Jacod, J., Li, Y., Mykland, P. A., Podolskij, M., and Vetter, M. (2009). Microstructure noise in the continuous case: The pre-averaging approach. *Stochastic Processes and their Applications*, 119(7):2249–2276.

Jacod, J., Li, Y., and Zheng, X. (2019). Estimating the integrated volatility with tick observations. *Journal of Econometrics*, 208(1):80–100.

Jacod, J. and Protter, P. (2012). *Discretization of Processes*. Springer.

Jacod, J. and Rosenbaum, M. (2013). Quarticity and Other Functionals of Volatility: Efficient Estimation. *Annals of Statistics*, 118:1462–1484.

Kalnina, I. and Xiu, D. (2017). Nonparametric Estimation of the Leverage Effect: A Trade-Off Between Robustness and Efficiency. *Journal of the American Statistical Association*, 112(517):384–396.

Kohavi, R., Tang, D., and Xu, Y. (2020). *Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing*. Cambridge University Press.

Kristensen, D. (2010). Nonparametric filtering of the realized spot volatility: A kernel-based approach. *Econometric Theory*, 26(1):pp. 60–93.

Lewis, M. (2014). *Flash Boys: A Wall Street Revolt*. W. W. Norton & Company.

Li, J. (2013). Robust Estimation and Inference for Jumps in Noisy High Frequency Data: A Local-to-Continuity Theory for the Pre-Averaging Method. *Econometrica*, 81(4):1673–1693.

Li, J. and Patton, A. J. (2018). Asymptotic inference about predictive accuracy using high frequency data. *Journal of Econometrics*, 203(2):223–240.

Li, J., Wang, D., and Zhang, Q. (2024). Reading the Candlesticks: An OK Estimator for Volatility. *Review of Economics and Statistics*, 106(4):1114–1128.

Li, Y., Liu, G., and Zhang, Z. (2022). Volatility of volatility: Estimation and tests based on noisy high frequency data with jumps. *Journal of Econometrics*, 229(2):422–451.

Patton, A. J. (2011). Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics*, 160(1):246–256.

Saret, J. N. and Mitra, S. (2016). An AI approach to Fed watching. *Two Sigma Street View*.

Wang, C. D. and Mykland, P. A. (2014). The Estimation of Leverage Effect With High-Frequency Data. *Journal of the American Statistical Association*, 109(505):197–215.

West, K. D. (1996). Asymptotic inference about predictive ability. *Econometrica*, 64(5):1067–1084.

Zhang, L., Mykland, P. A., and Aït-Sahalia, Y. (2005). A tale of two time scales: Determining integrated volatility with noisy high-frequency data. *Journal of the American Statistical Association*, 100:1394–1411.