Improved Density Forecasts Using Mixed Frequency Data: A Focused Bayesian Approach

Rubén Loaiza-Maya*, Worapree Maneesoonthorn† and Andrew J. Patton‡

December 29, 2024

Abstract

We construct a modeling framework that allows for the flexible contribution of intraday information to predict the conditional distribution of asset returns. A flexible spline specification is proposed to capture the contribution of intraday returns inside a GARCH-type specification for the daily return, and the "focused Bayesian" prediction framework is adopted to allow the intraday contribution to be trained according to the region of the support that is of interest. An empirical analysis of three broad market indices and ten industrial indices demonstrates the gains from using (i) high-frequency data, (ii) incorporated flexibly into the low-frequency model, (iii) estimated using a focused Bayes approach.

Keywords: High-frequency data; MIDAS; Tail prediction
JEL Classification: C11, C22, C53, C58

*Department of Econometrics and Business Statistics, Monash University. Supported by the Australian Research Council Discovery Early Career Researcher Award DE230100029.

†Department of Econometric and Business Statistics, Monash University. Supported by the Australian Research Council Discovery Project DP200101414.

‡Corresponding author; University of New South Wales and Duke University.

1 Introduction

Building on the foundational work of Engle (1982) and Bollerslev (1987), the class of generalized autoregressive conditional heteroskedastic (GARCH) model has been extensively used in forecasting the distribution of financial returns. Since its introduction, this model has been extended in many directions including the addition of leverage effects (e.g., Glosten *et al.*, 1993, Zakoïan, 1994 and Chen and Ghysels, 2011), allowing for fat-tail innovations (e.g., Hall and Yao, 2003, Fan *et al.*, 2014) and accounting for long memory (e.g., Ding *et al.*, 1993; Baillie *et al.*, 1996). With the development of the realized volatility computed from intraday data, pioneered by Barndorff-Nielsen and Shephard (2002) and Andersen *et al.* (2003), this measure has been incorporated in the GARCH framework and shown to lead to improvements in predictive performance see Hansen *et al.* (2012), Engle *et al.* (2013) and more recently Liu *et al.* (2023).

The use of realized volatility as a predictor implicitly imposes that all intraday observations have equal predictive content. The class of "mixed data sampling" (MIDAS) models of Ghysels *et al.* (2006) provides a way to relax that constraint, by allowing intraday returns to have flexible, estimated, weights. We adopt this framework and propose a method to flexibly incorporate high-frequency data in a model for the conditional distribution of lower-frequency returns. We model the weights for the high-frequency observations using a flexible B-spline function, and we utilize regularized splines, so that the weights are (optimally) shrunk toward a flat weighting function. We name our model "High-Frequency GARCH" (HF-GARCH). We estimate the model either by maximizing the log-score (analogous to maximum likelihood), or maximizing a scoring rule that is "focused" on a particular region of the support of the predictive density, e.g. the left tail.

Our use of high-frequency returns in a model for lower-frequency returns builds on the work of Ghysels *et al.* (2006), Chen and Ghysels (2011) and Patton and Zhang (2022). Like those papers, we allow intradaily returns to flexibly enter the lower-frequency model, rather than imposing a fixed weight structure. However, distinct from these papers is our focus on modeling the entire predictive *distribution*, rather than focusing only on the conditional second moment. Also novel relative to these papers is our use of Bayesian estimation. Bayesian estimation for estimating predictive return densities goes back to Bawa *et al.* (1979), see also Jorion (1985) and Kandel and Stambaugh (1996) amongst others. The Bayesian approach provides a coherent framework for incorporating parameter estimation error, which can be sizable for flexible models such as ours, into the predictive distribution.

We adopt the "focused Bayesian" prediction framework introduced in Loaiza-Maya *et al.* (2021). This framework is particularly helpful when the model is misspecified, as it allows the researcher to optimize the fit of the model in the region(s) of greatest interest to the forecast user, by tailoring the scoring rule used in estimation. The concept of calibrating the model to a specific scoring rule has been investigated generically in the context of forecasting, see Gneiting *et al.* (2007), and more recently in Patton (2020), Hansen and Dumitrescu (2022) and Martin *et al.* (2022).

We apply our model to daily returns on three broad equity indices (the S&P 500, 400, and 600 indices, which track large, mid-, and small capitalization firms) and ten industrial indices. We find significant gains in predictive accuracy when using high-frequency data in the models. Interestingly, we find that calibrating the model based on the likelihood, thus prioritizing the overall fit of the predictive distribution, results in greater weight being placed on end-of-day observations, consistent with Patton and Zhang (2022) for volatility forecasting. However, when the focus is on the lower tail of the distribution of returns, the weights tend to be approximately flat throughout the trade day. The focused Bayesian estimation framework, coupled with the flexibility of the spline specified by the model, allows for the discovery of these features. Finally, consistent with the theory in Hansen and Dumitrescu (2022), we find that matching the scoring rule used for estimation to that used for forecast evaluation, i.e. adopting the focused Bayesian methodology for estimation almost always leads to forecast improvements.

The paper is organized as follows. In Section 2, we introduce our model for the predictive distribution of daily asset returns, the HF-GARCH model, and discuss its relationship with existing specifications. In Section 3 we outline the Bayesian prediction framework used in this paper. Section 4 presents our empirical analysis of 13 equity return indices, showing how the optimal weights for intradaily data vary, and comparing the competing models via out-of-sample forecast accuracy. Section 5 concludes, and a supplemental appendix contains additional details and analyses.

2 Density forecasting with mixed-frequency data

There are two main existing approaches for using high-frequency returns to construct a predictive distribution for daily returns. In the first approach, a volatility proxy for the second moment can be defined through some predetermined aggregation of the high-frequency returns, such as realized volatility or its variants. A model is then constructed for this proxy, and predictions of the proxy can be utilized as input for the conditional distribution of daily returns (e.g., see Corsi, 2009; Patton and Zhang, 2022). The second approach considers a model directly for the return distribution, where the dynamics of the second moment of the conditional distribution are based on past values of the daily returns (Engle, 1982; Bollerslev, 1987), exogeneous covariates (Glosten *et al.*, 1993; Brenner *et al.*, 1996), a volatility proxy constructed from high-frequency data (Hansen *et al.*, 2012), or high-frequency observations directly (Chen and Ghysels, 2011; Chen *et al.*, 2015; Dhaene and Wu, 2020). We build on the latter approach and incorporate high-frequency returns into the model for the conditional distribution of daily returns, allowing the data to dictate how past intraday returns are weighted.

2.1 A high-frequency GARCH model

We propose a variation on a Gaussian-GARCH(p, q) model for the conditional distribution of the daily asset return, y_t. We replace the p lagged squared returns by all $p \times M$ squared high-frequency returns over the last p lags:[1]

$$
\begin{aligned}
y_t | \mathcal{I}_{t-1} &\sim N(\mu, \sigma_t^2) \\
\sigma_t^2 &= \beta_0 + \sum_{i=1}^{q} \beta_i \sigma_{t-i}^2 + \mathbf{\Gamma}^\top \boldsymbol{\eta}_{t-p:t-1},
\end{aligned}
\tag{1}
$$

where $\boldsymbol{\eta}_{t-p:t-1} = \left(\boldsymbol{\eta}_{t-p}^\top, ..., \boldsymbol{\eta}_{t-1}^\top\right)^\top$ collects the squared high-frequency returns over the previous p trading days, with $\boldsymbol{\eta}_{t-i} = \left(y_{t-i,1}^2, \ldots, y_{t-i,M}^2\right)^\top$ being the M high-frequency returns on day $t - i$, and $\mathbf{\Gamma} = (\gamma_1, \ldots, \gamma_{Mp})^\top$. The quantity $y_{t,m}^2$ serves as a proxy for the volatility over the m^{th} trading period on day t, analogous to the squared daily return serving as a proxy for daily volatility. We refer to the model in equation (1) as the High Frequency-GARCH (HF-GARCH) model.

The main advantage of the model is its flexibility in aggregating high-frequency returns when constructing the predictive distribution of daily returns. However, a significant challenge is the high dimensionality of $\mathbf{\Gamma}$, which contains $p \times M$ free parameters. Even setting $p = q = 1$ and $M = 78$, the latter corresponding to a five-minute sampling frequency for stocks on the NYSE, the number of parameters involved in the model is 81 in total.

2.2 Flexible weights with B-splines

The most flexible form of the model in equation (1) is to allow all $M \times p$ elements of $\mathbf{\Gamma}$ to be freely estimated. However, the large number of parameters, which is unavoidable even for the smallest choice of p, remains an obstacle for inference. In addition, allowing the

[1] The model can be easily modified to accommodate an irregular number of high-frequency returns per day, as might arise when using event-time sampling. We focus on the simpler case for brevity.

elements of $\mathbf{\Gamma}$ to be freely estimated means that estimates of adjacent weights may differ more than is thought economically reasonable, due simply to estimation error.

To address this, we utilize the B-spline (see De Boor, 1978 for an early introduction and Eilers and Marx, 2010 for a review) to specify the elements of $\mathbf{\Gamma}$. The j^{th} element given by $\gamma_j = \exp(f(j))$, where $f(j)$ is a flexible function over the grid points $j = 1, \ldots, Mp$. We set $f(j) = \boldsymbol{w}_j^\top \boldsymbol{c}$, where $\boldsymbol{w}_j = (w_{1,j}, \ldots, w_{B,j})^\top$ is a vector of basis functions constructed from a uniform B-spline of order l, and whose knots are uniformly distributed across the interval $[1-(Mp-1)/2, Mp+(Mp-1)/2]$. The number of knots, k, along with the order of the spline, l, defines the number of basis functions, $B = k + l + 1$. Under this setting, $\boldsymbol{c}$ is a $B-$dimensional vector of B-spline coefficients, with $B << Mp$. The exponential transformation ensures that all elements of $\mathbf{\Gamma}$ are greater than zero, ensuring positivity of the conditional volatility.

The use of the B-spline offers several advantages over alternative specifications. The flexibility of the spline can be controlled by the number of basis functions B, and well-known regularization techniques in inference of the spline coefficients $\boldsymbol{c}$ assist with controlling the smoothness of the spline. In classical likelihood framework, a LASSO type penalty term can be applied, as in Eilers and Marx (1996). Under the Bayesian framework, a hierarchical prior that ensures smoothness can be applied; see Kneib *et al.* (2011) and Klein and Kneib (2016), for examples.

2.3 Comparison with existing models

The model proposed in equation (1) nests existing models within the GARCH literature. When $M = 1$, the model reduces to the conventional GARCH(p,q) model of Bollerslev (1986), and no high-frequency information is used. When the parameter vector $\mathbf{\Gamma}$ exhibits a blocked structure, such that $\mathbf{\Gamma} = \left(\boldsymbol{\alpha}^\top \otimes \boldsymbol{\iota}_M^\top\right)^\top$, where $\boldsymbol{\alpha}$ is a $p \times 1$ vector of parameters and $\boldsymbol{\iota}_M$ is a $M \times 1$ vector of ones, the model corresponds to the GARCH-RV(p,q) model, which utilizes lagged realized volatility as past volatility proxy rather than lagged squared returns. The GARCH-RV model utilizes high-frequency information but imposes that all high-frequency returns on a given day have the same weight (α_i) in the model.

Our specification is related to the Hybrid GARCH proposed in Chen and Ghysels (2011), though we focus on flexible weights as a function of the time of day, while Chen and Ghysels (2011) allow the weights to vary as a function of the lagged return, thereby capturing a flexible form of leverage effect (Engle and Ng, 1993). Dhaene and Wu (2020) propose using intraday observations to model the conditional distributions of weekly returns, with weighting schemes taking various forms of exponentially decaying weights with an estimated decay parameter. With trading activities typically exhibiting diurnality patterns, as is captured in the context of the duration of trades in Engle and Russell (1998) and in the explicit modeling of intraday returns in Engle and Sokalska (2012), we allow for the information in the data to reveal how each intraday return contributes to future volatility. Finally, our model is in the spirit of Patton and Zhang (2022), who investigate the optimal intradaily return weights when forecasting realized volatility. In contrast, we focus on predicting the complete predictive distribution of daily returns, and we take into account the parameter estimation error in our model via our Bayesian estimation approach.

We highlight here that our aim is not to search for the best fitting *model* for conditional volatility with high-frequency data. Rather, we aim to utilize the mechanism of the model in equation (1), which allows for flexible weights on high-frequency data but still operates in a relatively simple GARCH structure, along with a focused Bayesian prediction framework to allow for potential differences in the optimal weights depending on the region(s) in the support of the density that are of interest. We discuss the focused Bayesian estimation framework in Section 3.2 below.

3 Model inference

We use Bayesian methods to estimate the model, which offer several advantages for our application. Most importantly, they allow us to incorporate the various sources and magnitudes of estimation error in our predictive density, see Bawa *et al.* (1979), Jorion (1985) and Kandel and Stambaugh (1996) for example. We also exploit the Bayesian structure and apply a hierarchical prior to the B-spline coefficients, which regularizes the estimated weights towards equal weights, and induces smoothness in the estimated weight function. Moreover, by employing the generalized Bayesian framework (see Bissiri *et al.*, 2016), we can generate predictions for the tails of return distributions that are robust to model misspecification.

In this section, we first outline the conventional likelihood-based Bayesian method, along with the prior specifications employed to assist in regularizing the B-splines in our model specification. We then describe the focused Bayesian prediction framework (see Loaiza-Maya *et al.*, 2021) that is used to produce tail-focused predictive distributions. The latter framework allows us to analyze the different optimal weight functions in the empirical analysis presented in Section 4.

3.1 Likelihood-based Bayesian prediction

Denote $\boldsymbol{y}_{1:T} = (y_1, \ldots, y_T)^\top$ to be the vector of T daily returns, and $\boldsymbol{\eta}_{1:T} = (\boldsymbol{\eta}_1^\top, \ldots, \boldsymbol{\eta}_T^\top)^\top$ to the corresponding vector of MT intraday high-frequency returns. The probabilistic model in equation (1) implies a density function of the form $p(y_t|\boldsymbol{\eta}_t, \boldsymbol{\varphi}) = \phi(y_t; \mu, \sigma_t^2)$ for y_t, where $\phi(x; \mu, \sigma^2)$ denotes the normal density function with mean μ and variance σ, evaluated at x. The model prescribes the likelihood

$$p(\boldsymbol{y}_{1:T}|\boldsymbol{\varphi}, \boldsymbol{\eta}_{1:T}) = \prod_{t=1}^{T} p(y_t|\boldsymbol{\eta}_{1:t-1}, \boldsymbol{\varphi}), \tag{2}$$

where $\boldsymbol{\varphi} = \left(\mu, \tilde{\beta}_0, \tilde{\beta}_1, \dots, \tilde{\beta}_p, \boldsymbol{c}^\top\right)^\top$ denotes the complete parameter vector with $\tilde{\beta}_0 = \log \beta_0$ and $\tilde{\beta}_i = \Phi^{-1}(\beta_i)$ for $i = 1, \dots, p$, where $\Phi(\cdot)$ is the univariate standard Gaussian distribution function. The transformations on the model parameters ensure that $\boldsymbol{\varphi} \in \mathbb{R}^d$ for ease of inference.

The key objective of Bayesian analysis is the computation of the posterior distribution, combining the likelihood with prior distributions of the model unknowns. To this end, we consider the uninformative priors $\mu \sim N(0, 10)$, $\beta_0 \sim \text{Lognormal}(0, 10)$ and $\beta_i \sim \text{Unif}(0, 1)$ for $i = 1, \dots, p$, inducing the implicit priors for the transformed parameter: $\tilde{\beta}_0 \sim N(0, 10)$ and $\tilde{\beta}_i \sim N(0, 1)$.

For the coefficient vector of the basis functions we employ the hierarchical prior structure $p(\boldsymbol{c}|\tau^2, \psi) = \phi_B(\boldsymbol{c}; \mathbf{0}_B, \frac{1}{\tau^2} P(\psi)^{-1})$, where $P^{-1}(\psi)$ denotes the banded precision matrix mimicking the precision matrix of an autoregressive process of order 1, with autoregressive coefficient ψ. This prior, proposed in Klein and Kneib (2016) and also considered in Ong *et al.* (2018), induces smoothness and regularizes the function $f(j)$ towards a flat line, with the hyperparameters τ and ψ, governing the deviation around the equal weights model. The hyper-parameters are assumed to have a prior density $p(\tau^2, \psi) = p(\tau^2) \frac{1}{0.99} I(0 < \psi < 0.99)$, where $(1/\tau^2) \sim \text{Weibull}(\lambda, k)$. We transform these two parameters to the real line as $\tilde{\tau} = \log \tau^2$ and $\tilde{\psi} = 0.99/(1 + \exp\{-\psi\})$, and their priors are also modified according to the Jacobian of the transformation. Under this hierarchical prior structure, the posterior distribution is given as

$$p(\boldsymbol{\varphi}, \boldsymbol{\xi}|\boldsymbol{\eta}_{1:T}, \boldsymbol{y}_{1:T}) \propto p(\boldsymbol{y}_{1:T}|\boldsymbol{\eta}_{1:T}, \boldsymbol{\varphi}) p(\boldsymbol{\varphi}|\boldsymbol{\xi}) p(\boldsymbol{\xi}), \tag{3}$$

where $\boldsymbol{\xi} = (\tilde{\psi}, \tilde{\tau})$, $p(\boldsymbol{\xi}) = p(\tilde{\psi}) p(\tilde{\tau})$ and $p(\boldsymbol{\varphi}|\boldsymbol{\xi}) = p(\boldsymbol{c}|\tau^2, \psi) p(\mu) \prod_{i=0}^{p} p(\tilde{\beta}_i)$. Such a hierarchical structure allows us to let the data inform the degree of regularization, with inference on the hyper-parameters being part of the inferential problem. As such, the resulting implied predictive distribution accounts for posterior uncertainties while at the same time regularizing towards the GARCH-RV model.[2]

The posterior in equation (3) is not of closed form, and thus its computation requires a simulation-based method. We employ a random walk Markov chain Monte Carlo (MCMC) with Metropolis-Hastings (MH) steps to generate samples for $\boldsymbol{\theta} = (\boldsymbol{\varphi}^\top, \boldsymbol{\xi}^\top)^\top$ from its posterior in equation (3). At the beginning of each MCMC iteration we randomly allocate the model parameters into K mutually exclusive and collectively exhaustive blocks. The k-th block is defined as $\boldsymbol{\theta}_k = \{\theta_s : s \in I_k\}$, where $I_k \subset \{1, \ldots, \dim(\boldsymbol{\theta})\}$ is a random subset of indexes such that $\bigcup_{k=1}^K I_k = \{1, \ldots, \dim(\boldsymbol{\theta})\}$ and $I_s \bigcap I_k = \emptyset$ for $s \neq k$. We sample one block of parameters at a time, conditional on its complement set, $\boldsymbol{\theta}_{\setminus k} = \{\theta_s : s \notin I_k\}$, using a Gaussian proposal of the form $q(\boldsymbol{\theta}_k) = \prod_{s \in I_k} \phi(\theta_s; \theta_s^*, \Sigma_s^*)$, where θ_s^* denotes the previous draw in the MCMC chain for the $s-$th element in $\boldsymbol{\theta}$, and Σ_s^* its corresponding step size, iteratively adapted to guarantee the MH acceptance rate between 30-70%. New proposal draws $\boldsymbol{\theta}_k^{'}$ are accepted with probability

$$\alpha = \min\left(1, \frac{p(\boldsymbol{y}_{1:T}|\boldsymbol{\eta}_{1:T}, \boldsymbol{\theta}_k^{'}, \boldsymbol{\theta}_{\setminus k}^*)p(\boldsymbol{\theta}_k^{'})}{p(\boldsymbol{y}_{1:T}|\boldsymbol{\eta}_{1:T}, \boldsymbol{\theta}^*)p(\boldsymbol{\theta}^*)}\right). \tag{4}$$

After adequate burnin and inference sampling periods, we collect J draws from the exact posterior denoted as $\{\boldsymbol{\theta}^{[j]}\}_{j=1}^J$.

Predictions from the model are computed from the posterior predictive density

$$p(y_{T+1}|\boldsymbol{y}_{1:T}, \boldsymbol{\eta}_{1:T}) = \int p(y_{T+1}|\boldsymbol{\varphi}, \boldsymbol{\eta}_{1:T})p(\boldsymbol{\varphi}|\boldsymbol{\eta}_{1:T}, \boldsymbol{y}_{1:T})d\boldsymbol{\varphi}. \tag{5}$$

[2]In the frequentist context, penalization terms can be introduced to the criterion function, akin to LASSO type penalty, see Eilers and Marx (1996). Such method relies on the selection of a penalty term to control the smoothness, typically done by optimizing some criteria function, such as the AIC, or via cross validation.

This predictive density is evaluated numerically as

$$\widehat{p}(y_{T+1}|\boldsymbol{y}_{1:T}, \boldsymbol{\eta}_{1:T}) = \frac{1}{J}\sum_{j=1}^{J} p(y_{T+1}|\boldsymbol{\varphi}^{[j]}, \boldsymbol{\eta}_{1:T}), \tag{6}$$

where $\boldsymbol{\varphi}^{[j]}$ are the relevant subset of elements of the draw $\boldsymbol{\theta}^{[j]}$ obtained via the MCMC scheme described above. We note that the Bayesian predictive is non-Gaussian even when the underlying model assumes conditionally Gaussian predictive, due to the numerical integration inducing a mixture of Gaussian as in equation (6).

3.2 Focused Bayesian prediction

If the model to be estimated is correctly specified, then maximizing the log-likelihood, as in equation (5), yields the most accurate estimator: the posterior distribution obtained in equation (3) will, via the required regularity conditions and the Bernstein-von Mises result, concentrate to the true parameter value $\boldsymbol{\varphi}_0$. Thus, heuristically, equation (5) will converge to the true predictive density $p(y_{T+1}|\boldsymbol{\varphi}_0, \boldsymbol{y}_{1:T}, \boldsymbol{\eta}_{1:T})$. However, if the model is misspecified, the predictive density in equation (5) is no longer optimal in this sense, and can only serve as an approximation to the true predictive distribution. Consequently, the Bayesian predictive distribution may fail to accurately capture important characteristics of the true predictive distribution of the returns, including its tails, which are often of primary interest in the context of risk management.

A misspecification-robust approach to Bayesian prediction is to construct a Bayesian posterior density tailored to the predictive task at hand, see Loaiza-Maya *et al.* (2021) and Frazier *et al.* (2024). This can be achieved by following the principles of generalized Bayesian inference (Bissiri *et al.*, 2016), in which a prior distribution can be updated to a posterior via a particular scoring function that reflects the relevant predictive task. Denote such a scoring function as $S(p(y_t|\boldsymbol{\varphi}, \boldsymbol{\eta}_{1:t-1}), y_t)$, which measures the quality of the quoted density $p(y_t|\boldsymbol{\varphi}, \boldsymbol{\eta}_{1:t-1})$ as a prediction of the realized value y_t. A focused posterior distribution of the hierarchical model parameters, including the uncertainty of the hierarchical priors described in the previous section, can be constructed as

$$p_w(\boldsymbol{\varphi}, \boldsymbol{\xi}|\boldsymbol{\eta}_{1:T}, \boldsymbol{y}_{1:T}) \propto \exp(wS_T(\boldsymbol{\varphi}))p(\boldsymbol{\varphi}|\boldsymbol{\xi})p(\boldsymbol{\xi}), \tag{7}$$

where

$$S_T(\boldsymbol{\varphi}) = \sum_{t=1}^{T} S(p(y_t|\boldsymbol{\varphi}, \boldsymbol{\eta}_{1:t-1}), y_t) \tag{8}$$

and w is scalar controlling the relative weight that the prior and the scoring function are given in the construction of the posterior. We can generate samples from the focused posterior by employing the same random walk Metropolis-Hastings algorithm described in Section 3.1.

The focused posterior distribution allocates higher posterior density to parameter values that lead to higher predictive accuracy according to the chosen scoring rule. Consequently, the focused Bayesian predictive density

$$p_w(y_{T+1}|\boldsymbol{y}_{1:T}, \boldsymbol{\eta}_{1:T}) = \int p(y_{T+1}|\boldsymbol{\varphi}, \boldsymbol{\eta}_{1:T})p_w(\boldsymbol{\varphi}|\boldsymbol{\eta}_{1:T}, \boldsymbol{y}_{1:T})d\boldsymbol{\varphi}, \tag{9}$$

is geared to the predictive task that is consistent with the prescribed scoring rule used in obtaining the focused posterior in equation (7).

We focus on strictly proper scoring rules (see Gneiting and Raftery, 2007 for a review of this topic) to ensure that if the predictive model happens to be correctly specified, then the focused Bayesian predictive density will converge to the true predictive density. Several proper scoring rules have been proposed to address various forecasting needs, such as evaluating prediction accuracy in the tails or other regions of return distributions, as well as assessing the calibration of predictive intervals. In the context of the HF-GARCH model, using focused inference and prediction allows us to analyze whether the weights associated with high-frequency returns vary over the trade day via the posterior inference of $\mathbf{\Gamma}$ in equation (1), even when the model is misspecified. Below, we discuss the scoring rules examined in this context. For all scores, the scalar w, which plays a key role in balancing prior and data information and thereby influencing the level of posterior uncertainty, is set as in Loaiza-Maya *et al.* (2021).

3.3 Scoring rules

The most common scoring rule for density estimation is the logarithmic score (LS), defined as $S_{LS}(p(y_t|\boldsymbol{\varphi}, \boldsymbol{\eta}_{1:t-1}), y_t) = \log p(y_t|\boldsymbol{\varphi}, \boldsymbol{\eta}_{1:t-1})$. This scoring rule evaluates the performance of the predictive model across the entire support of y_t. The inference in equation (7) coincides with the likelihood-based posterior in equation (3) when the logarithmic score is used as the updating rule and $w = 1$.

Another scoring rule that is widely used to gauge overall predictive performance is the continuously ranked probability score (CRPS). This scoring rule puts emphasis on the sharpness of the predictive distribution (Gneiting *et al.*, 2007), and is less sensitive to outliers than the logarithmic score. The CRPS is given as

$$S_{CRPS}(p(y_t|\boldsymbol{\varphi}, \boldsymbol{\eta}_{1:t-1}), y_t) = -\int_{-\infty}^{\infty} \big((P(y|\boldsymbol{\varphi}, \boldsymbol{\eta}_{1:t-1}) - I(y \geq y_t)\big)^2 \, dy, \tag{10}$$

where $P(y_t|\boldsymbol{\varphi}, \boldsymbol{\eta}_{1:t-1})$ is the cumulative distribution function associated with p.

Next we consider scoring rules that focus attention on specific region(s) of the support. Score-based optimization allows the model to adapt to the region of distribution that is of our key interest, as measured by the scoring rule, see Martin *et al.* (2022) in the frequentist framework and in Loaiza-Maya *et al.* (2021) in the Bayesian framework. If the model is correctly specified, the differences between the predictive distributions obtained using likelihood-based inference and those obtained using an alternative proper scoring rule vanish asymptotically. However, when the model is misspecified score-based estimation optimizes the fit according to the forecast user's preferences, as measured by the scoring rule.

In risk management applications, the region of interest is typically in the tails of the distribution, with the lower (upper) tail being relevant to risk of loss in long (short) positions in an asset. Diks *et al.* (2011) propose the censored log score (CS), which allows users to reward forecasts only on specific region(s) of interest. It is defined as

$$\begin{aligned} S_{CS}(p(y_t|\boldsymbol{\varphi}, \boldsymbol{\eta}_{1:t-1}), y_t) &= I(y_t \in \mathcal{A}) \log p(y_t|\boldsymbol{\varphi}, \boldsymbol{\eta}_{1:t-1}) \\ &\quad + I(y_t \in \mathcal{A}^c) \log \left(\int_{y \in A^c} p(y|\boldsymbol{\varphi}, \boldsymbol{\eta}_{1:t-1}) dy \right), \end{aligned} \tag{11}$$

where $\mathcal{A}$ is the region of interest and $\mathcal{A}^c$ is its complement. The CS has been applied in the evaluation of predictive distributions geared to risk management (Jensen and Maheu, 2013 and Chiang *et al.*, 2021), as well as in inferential procedures geared to certain data regions (Borowska *et al.*, 2020 and Loaiza-Maya *et al.*, 2021).

Finally, consider the case where one is interested in the accuracy of a $(1-\alpha)$ predictive interval with end points at the $\alpha/2$ and $(1-\alpha/2)$ quantiles. The interval score (IS) is an appropriate scoring function for this predictive task, and is defined as

$$S_{IS}(p(y_t|\boldsymbol{\varphi}, \boldsymbol{\eta}_{1:t-1}), y_t) = (u_t - l_t) + \frac{2}{\alpha}(l_t - y_t) I(y_t < l_t) + \frac{2}{\alpha}(y_t - u_t) I(u_t < y_t), \tag{12}$$

where $u_t \equiv P^{-1}(1-\alpha/2|\boldsymbol{\varphi}, \boldsymbol{\eta}_{1:t-1})$ and $l_t \equiv P^{-1}(\alpha/2|\boldsymbol{\varphi}, \boldsymbol{\eta}_{1:t-1})$.

Table 1: **Summary statistics.** This table presents the tickers of the indices studied in this paper, along with summary statistics for daily returns and daily realized variances for these indices. The first two indices are available from January 3, 2000, the third is available from October 30, 2002, and the remainder are available from November 3, 2006. The last day in our sample for all indices is December 2, 2022.

		Return statistics		RV statistics				
Symbol	Index Description	Skew	Kurt	Avg	5th	25th	75th	95th
SPY	Large cap	-0.220	10.186	0.988	0.080	0.211	0.977	3.123
MID	Mid cap	-0.426	10.532	0.956	0.113	0.244	0.847	3.046
SPCY	Small cap	-0.369	9.301	1.023	0.134	0.274	0.933	3.370
SPXEW10	Energy	-0.332	9.465	3.051	0.400	0.860	3.112	9.080
SPXEW15	Materials	-0.309	8.879	1.567	0.170	0.374	1.490	5.035
SPXEW20	Industrial	-0.231	8.696	1.183	0.122	0.261	1.090	3.848
SPXEW25	Cons Discretionary	-0.084	10.314	1.435	0.140	0.283	1.364	5.242
SPXEW30	Cons Staples	-0.428	10.879	0.636	0.092	0.176	0.566	1.817
SPXEW35	Health Care	-0.404	9.985	0.807	0.119	0.224	0.777	2.454
SPXEW40	Financial	-0.001	16.869	1.919	0.145	0.324	1.482	7.833
SPXEW45	I.T.	-0.177	7.695	1.283	0.148	0.316	1.287	4.147
SPXEW50	Telecom	-0.239	10.451	1.257	0.202	0.373	1.211	3.803
SPXEW55	Utilities	0.249	14.226	1.123	0.168	0.332	0.956	2.844

4 Empirical analysis

4.1 Data description

We study the predictive distribution of 13 equity indices. The first three indices are broad market indices: the SPY is an exchange-traded fund (ETF) that tracks the S&P 500 index, an index of large market capitalization stocks. MID and SPCY are ETFs that track the S&P 400 and S&P 600 indices, which are mid- and small-cap indices respectively. We additionally consider ten S&P industry indices. All the data comes from the New York Stock Exchange Trades and Quotes (TAQ) database. We sample high-frequency returns at the five-minute frequency, yielding $M = 78$ high-frequency returns per day, balancing the incremental information from finer sampling with the impact of market microstructure noise (Liu *et al.*, 2015). We omit the overnight return and any trading day with reduced trading hours. Summary statistics on our data are provided in Table 1.

4.2 Choosing the lag orders for the HF-GARCH(p,q) model

The HF-GARCH(p, q) model in equation (1) can be expressed as a form of ARCH(∞) model:

$$
\begin{aligned}
\sigma_t^2 &= \beta_0 \sum_{i=1}^{q} \beta_i \sigma_{t-i}^2 + \boldsymbol{\Gamma}^\top \boldsymbol{\eta}_{t-p:t-1} \\
&= \frac{\beta_0}{1 - B(L)} + \frac{1}{1 - B(L)} \boldsymbol{\Gamma}^\top \boldsymbol{\eta}_{t-p:t-1}, \\
&= \frac{\beta_0}{1 - B(L)} + \sum_{j=1}^{\infty} (B(L))^{j-1} \boldsymbol{\Gamma}^\top \boldsymbol{\eta}_{t-p:t-1},
\end{aligned} \tag{13}
$$

where $B(L) = \sum_{i=1}^{q} \beta_i L^i$ and L is the lag operator such that $L^d x_t = x_{t-d}$ for a generic time series x_t, and the third line follows from the expansion of the term $1/(1 - B(L))$ as the sum of an infinite geometric series. From the above, it is clear that even with $p = q = 1$ the HF-GARCH allows us to incorporate intraday information from beyond just a single lag.

Using the structure in equation (13), we extract the model-implied weights from the HF-GARCH(1,1) specification for SPY, fitted with $B = 5$ basis functions, to study the impact of the intraday information over the 21 lags.[3] We compare these with the model-implied weights from the HF-GARCH(21,1), constructed with $B = 15$ basis functions. The HF-GARCH(21,1), when constructed using the five-minute sampling interval, incorporates 1638 observations of squared intraday returns for each evaluation of σ_t^2, with the spline to fit $\boldsymbol{\Gamma}$ requiring a greater number of basis functions, resulting in a greater computational cost.

Panel A of Figure 1 depicts the fitted conditional volatility and reveals the fitted volatilities are very similar. Panel B depicts the model-implied weights of intraday information, with those produced by the HF-GARCH(1,1) plotted in blue and the HF-GARCH(21,1)

[3] The choice of 21 lags here is motivated by the lag choice in the construction of the HAR model of Corsi (2009).

Panel A: Fitted σ_t^2 Panel B: Model implied intraday weights

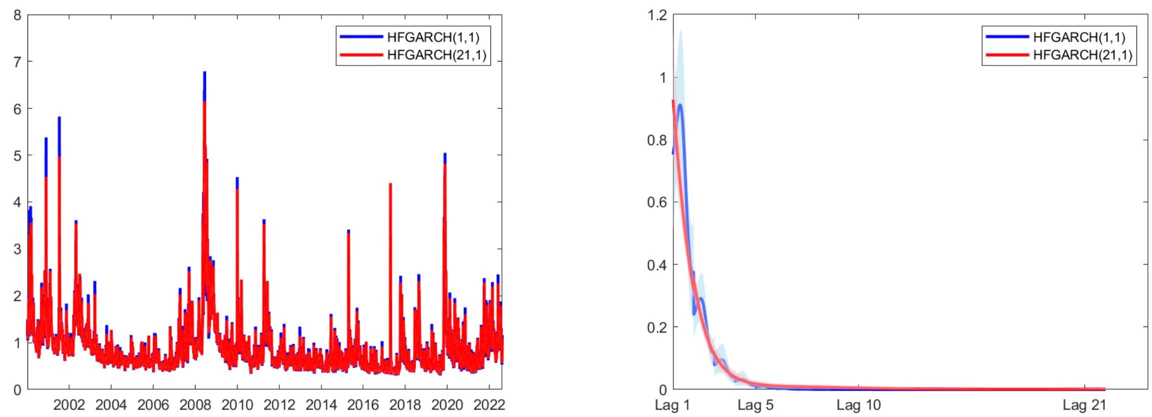

Figure 1: **Comparison of HF-GARCH(1,1) and HF-GARCH(21,1) for SPY.** Fitted σ_t^2 are depicted in Panel A, and the model-implied intraday weights of squared returns on conditional volatility in Panel B.

in red. The implied intraday weights, in general, decay at an almost identical rate over the trading days, but the HF-GARCH(1,1) captures more nuances in terms of the patterns across a trading day. We observe such similarities across all series considered, and further results, including an assessment of the comparative predictive performance of the models calibrated using maximum likelihood estimation, are provided in Supplemental Appendix C. Given the similarity in the information structure and fitted volatility, we adopt the HF-GARCH(1,1) for computational speed.

4.3 Out-of-sample predictive performance

We now consider the out-of-sample performance of the competing models and estimation methods. The model is estimated on the initial period of 1500 trading days via the MCMC sampling scheme described in Section 3.1. We consider a burnin sample of 5000 and collect an inference sample of 20000 draws, with every 10^{th} draw used to construct the predictive distribution. One-step-ahead predictive distributions are then obtained and evaluated for the trading day following the estimation period. We then expand the estimation window by one trading day, and repeat this procedure until predictions for the last observation in the sample are produced. To model the weights in the HF-GARCH model, we consider a B-spline of order three and a total number of $B = 5$ basis functions.[4] In addition to the HF-GARCH model, we also consider the GARCH-RV model, which incorporates high frequency data but imposes that all intradaily returns are equally weighted, and the GARCH model, which does not use intradaily data.

4.3.1 The gains from Focused Bayesian estimation

Before comparing all models, we first analyze the gains from matching the scoring rule used in estimation to that used for forecast evaluation. Work by Gneiting *et al.* (2007), Patton (2020), Hansen and Dumitrescu (2022) and Martin *et al.* (2022) suggests that this improves forecasts when the model is misspecified. In Table 2, we consider the HF-GARCH model (equation 1) estimated using one of seven scoring rules. We first consider the log-score, analogous to maximum likelihood.[5] We then consider the continuously ranked probability score (CRPS), given in equation (10); the interval score (IS), given in equation (12); and four censored log-scores, given in equation (11). Two censored scores are focused on the lower 10% and 20% tails of the returns, and are denoted as CS10 and CS20, and two focused on the upper 10% and 20% tails, denoted as CS90 and CS80. In each estimation window, the region $\mathcal{A}$ required in the evaluation of the censored score is defined via the marginal empirical quantiles.

Table 2 reports the average out-of-sample scores of the HF-GARCH model when estimated using each of the seven scoring rules given in the row labels and evaluated using each of the rules in the column labels. All scores are reported in positive orientation, so

[4] We also considered $B \in (3, 5, 10)$ and the results are robust to this choice.

[5] We also provide a comparison of the Bayesian likelihood-based prediction and the predictions constructed using estimates from conventional maximum likelihood estimation (MLE) with penalized spline in Supplemental Appendix B. While the two methods produce predictive distributions that are very similar, we adopt the Bayesian predictive here due to its ability to incorporate parameter estimation error, including any error related to spline regularizations.

Table 2: **Out-of-sample scores for SPY density forecasts**. This table shows the average out-of-sample scores for density forecasts of daily SPY returns using the HF-GARCH model in equation (1), estimated using the scoring rule given in the row label and evaluated using the scoring rule given in the column label. Within each column, the estimation method with the highest average score is given in bold and the models included in the 95% model confidence set of Hansen *et al.* (2011) are shaded green.

	Evaluation scoring rule						
	LS	CS10	CS20	CS80	CS90	CRPS	IS
Estimation rule							
LS	-1.069	-0.312	-0.513	-0.478	-0.265	-0.448	-4.129
CS10	-1.210	**-0.303**	-0.502	-0.612	-0.362	-0.485	-4.676
CS20	-1.171	**-0.303**	**-0.501**	-0.577	-0.334	-0.473	-4.514
CS80	-1.071	-0.311	-0.514	-0.476	-0.265	-0.450	**-4.090**
CS90	-1.081	-0.311	-0.518	-0.478	-0.265	-0.452	-4.125
CRPS	**-1.064**	-0.320	-0.518	**-0.475**	**-0.264**	**-0.446**	-4.130
IS	-1.100	-0.311	-0.528	-0.478	-0.265	-0.459	-4.108

that larger values are preferred. Within each column, the model with the highest score is presented in bold, and models contained in the 95% model confidence set (MCS) of Hansen *et al.* (2011) are shaded green. The diagonal of this matrix represents the cases where the estimation and evaluation scoring rules match, and we observe that in three of the seven cases the diagonal achieves the highest score, and in all seven cases the diagonal is included in the MCS.[6] Conversely, models not calibrated to the evaluation scoring rule are less frequently included in the MCS, as reflected by the off-diagonal values. This is strong support for matching the scoring rule used in estimation to that used for evaluation.

4.3.2 Comparison with alternative models

We next compare the HF-GARCH model with the GARCH and GARCH-RV models. Given the results supporting focused estimation presented in Table 2, we consider two estimation methods for each of the models: one using the log-score, and the other matching the scoring rule used in evaluation, i.e., focused Bayesian estimation. Models estimated using

[6]Estimation by CRPS achieves the highest out-of-sample scores in four out of seven cases, including the case where it is also used for evaluation, but it is included in the MCS in only five out of seven cases.

the latter method are appended with "-F" below. Further details on the implementation of these approaches can be found in Supplemental Appendix A.

Tables 3, 4a and 4b report the predictive scores for the out-of-sample predictive assessment for the three broad indices and the ten industrial indices, respectively. In each panel, the seven column labels indicate the scoring rule used to measure predictive performance and the six row labels indicate the model and estimation method used for prediction. As in Table 2, within each column the model with the highest score is presented in bold, and models contained in the 95% model confidence set of Hansen *et al.* (2011) are shaded green. Note that for the LS column, the likelihood-based method and the focused method coincide.

A striking finding from Tables 3, 4a and 4b is that our proposed HF-GARCH model, when estimated using focused Bayesian methods (i.e., the method labeled "HF-GARCH-F"), is in the 95% model confidence set (MCS) for all $13 \times 7 = 91$ combinations of assets and scoring rules, and has the absolute highest out-of-sample average score in 78 of these cases. This is strong evidence in support of our flexible approach for incorporating high frequency data into a density forecast for lower-frequency returns, when combined with an estimation method that acknowledges the underlying model may be misspecified.

A second key finding from Tables 3, 4a and 4b is the value of flexibly incorporating high frequency data: we find that the HF-GARCH model has higher average out-of-sample scores than the GARCH-RV and GARCH models (estimated either using the log-score or a focused loss function) in 100% of combinations of assets, scoring rules, and estimation methods. That is, optimally weighting the high-frequency returns significantly improves out-of-sample forecast performance relative to imposing equal weights (as in traditional realized variance) or ignoring high frequency information altogether (as in the GARCH model).

It is noteworthy that in some cases the HF-GARCH-F method is the *only* model in the

Table 3: **Out-of-sample scores for various density forecasts: Large-, mid-, and small-cap indices**. This table shows the average out-of-sample scores for density forecasts of daily returns on the SPY, MID and SPCY indices. Three models are considered (GARCH, GARCH-RV and HF-GARCH) and each model is estimated either to maximize the log-score, or to maximize the scoring rule used in forecast evaluation (given in the column labels), the latter are marked "-F" in the row labels. When forecasts are evaluated using the log-score (first column) the two estimation methods coincide. Within each column, the estimation method with the highest average score is given in bold and the models included in the 95% model confidence set of Hansen *et al.* (2011) are shaded green.

	Out-of-Sample predictive score						
	LS	CS10	CS20	CS80	CS90	CRPS	IS
				Panel A: Large caps (SPY)			
GARCH	-	-0.339	-0.544	-0.492	-0.274	-0.454	-4.414
GARCH-F	-1.125	-0.314	-0.515	-0.490	-0.273	-0.452	-4.366
GARCH-RV	-	-0.326	-0.529	-0.485	-0.271	-0.451	-4.276
GARCH-RV-F	-1.096	-0.312	-0.512	-0.484	-0.271	-0.450	-4.247
HF-GARCH	-	-0.312	-0.513	-0.478	**-0.265**	-0.448	-4.129
HF-GARCH-F	**-1.069**	**-0.303**	**-0.501**	**-0.476**	**-0.265**	**-0.446**	**-4.108**
				Panel B: Mid caps (MID)			
GARCH	-	-0.353	-0.586	-0.511	-0.291	-0.552	-5.083
GARCH-F	-1.312	-0.342	-0.574	-0.505	-0.287	-0.552	-5.065
GARCH-RV	-	-0.344	-0.577	-0.512	-0.291	-0.552	-5.015
GARCH-RV-F	-1.303	-0.340	-0.571	-0.506	-0.289	-0.551	-4.937
HF-GARCH	-	-0.340	-0.572	-0.507	-0.288	-0.550	-4.941
HF-GARCH-F	**-1.292**	**-0.334**	**-0.565**	**-0.501**	**-0.285**	**-0.549**	**-4.831**
				Panel C: Small caps (SPCY)			
GARCH	-	-0.338	-0.602	-0.549	-0.318	-0.582	-5.332
GARCH-F	-1.412	-0.330	-0.593	-0.545	-0.315	-0.582	-5.307
GARCH-RV	-	-0.330	-0.593	-0.547	-0.315	-0.581	-5.226
GARCH-RV-F	-1.400	-0.327	-0.589	-0.544	-0.313	-0.581	-5.180
HF-GARCH	-	-0.325	-0.587	-0.541	-0.310	-0.578	-5.083
HF-GARCH-F	**-1.386**	**-0.321**	**-0.583**	**-0.538**	**-0.308**	**-0.577**	**-5.012**

MCS, while in other cases the MCS contains all of the models, indicating a lack of power to discriminate between the methods. Tables 4a and 4b show that this is more common when using the censored log-score (CS) or interval score (IS) as the scoring rule. When using the log-score, in contrast, the MCS contains a single model in all but two cases, consistent

Table 4a: **Out-of-sample scores for various density forecasts: Industry indices (I)**. This table shows the average out-of-sample scores for density forecasts of daily returns on the SPXEW10, SPXEW15, SPXEW20, SPXEW25 and SPXEW30 indices. See the notes to Table 3 for a detailed description of this table.

	Out-of-Sample predictive score						
	LS	CS10	CS20	CS80	CS90	CRPS	IS
Panel A: Energy (SPXEW10)							
GARCH	-	-0.382	-0.667	-0.637	-0.385	-0.800	-7.007
GARCH-F	-1.715	-0.375	-0.662	-0.636	-0.385	-0.800	-7.016
GARCH-RV	-	-0.374	-0.659	-0.636	-0.386	-0.799	-6.993
GARCH-RV-F	-1.706	-0.371	-0.656	-0.636	-0.387	-0.799	-6.985
HF-GARCH	-	-0.370	-0.655	**-0.631**	**-0.382**	**-0.796**	**-6.877**
HF-GARCH-F	**-1.696**	**-0.369**	**-0.653**	-0.631	-0.382	-0.796	-6.885
Panel B: Materials (SPXEW15)							
GARCH	-	-0.277	-0.553	-0.472	-0.263	-0.507	-4.894
GARCH-F	-1.294	-0.264	-0.539	-0.470	-0.264	-0.506	-4.885
GARCH-RV	-	-0.278	-0.554	-0.465	-0.259	-0.506	-4.914
GARCH-RV-F	-1.285	-0.264	-0.538	-0.465	-0.259	-0.505	-4.840
HF-GARCH	-	-0.263	-0.537	-0.461	**-0.255**	-0.501	-4.697
HF-GARCH-F	**-1.259**	**-0.254**	**-0.527**	**-0.460**	-0.256	**-0.501**	**-4.649**
Panel C: Industrials (SPXEW20)							
GARCH	-	-0.314	-0.577	-0.474	-0.245	-0.458	-4.533
GARCH-F	-1.197	-0.292	-0.555	-0.468	-0.245	-0.457	-4.515
GARCH-RV	-	-0.309	-0.571	-0.468	-0.243	-0.457	-4.487
GARCH-RV-F	-1.183	-0.291	-0.553	-0.465	-0.243	-0.456	-4.417
HF-GARCH	-	-0.294	-0.554	-0.461	-0.236	-0.453	-4.289
HF-GARCH-F	**-1.155**	**-0.283**	**-0.543**	**-0.456**	**-0.236**	**-0.453**	**-4.244**
Panel D: Consumer discretionary (SPXEW25)							
GARCH	-	-0.313	-0.574	-0.467	-0.256	-0.500	-4.872
GARCH-F	-1.247	-0.300	-0.559	-0.459	-0.256	-0.499	-4.857
GARCH-RV	-	-0.307	-0.568	-0.461	-0.253	-0.499	-4.762
GARCH-RV-F	-1.233	-0.297	-0.556	-0.457	-0.251	-0.499	-4.670
HF-GARCH	-	-0.295	-0.553	-0.455	-0.247	**-0.496**	-4.572
HF-GARCH-F	**-1.210**	**-0.289**	**-0.546**	**-0.452**	**-0.245**	-0.496	**-4.502**
Panel E: Consumer staples (SPXEW30)							
GARCH	-	-0.360	-0.568	-0.497	-0.314	-0.370	-3.639
GARCH-F	-0.986	-0.344	-0.551	-0.498	-0.322	-0.369	-3.632
GARCH-RV	-	-0.359	-0.567	-0.486	-0.303	-0.369	-3.590
GARCH-RV-F	-0.973	-0.342	-0.549	-0.484	-0.302	-0.369	-3.548
HF-GARCH	-	-0.350	-0.558	-0.478	-0.296	-0.368	-3.451
HF-GARCH-F	**-0.954**	**-0.337**	**-0.543**	**-0.476**	**-0.295**	**-0.367**	**-3.402**

Table 4b: **Out-of-sample scores for various density forecasts: Industry indices (II)**. This table shows the average out-of-sample scores for density forecasts of daily returns on the SPXEW35, SPXEW40, SPXEW45, SPXEW50 and SPXEW55 indices. See the notes to Table 3 for a detailed description of this table.

	Out-of-Sample predictive score						
	LS	CS10	CS20	CS80	CS90	CRPS	IS
	Panel F: Health care (SPXEW35)						
GARCH	-	-0.395	-0.632	-0.521	-0.323	-0.449	-4.376
GARCH-F	-1.172	-0.371	-0.604	-0.521	-0.323	-0.448	-4.338
GARCH-RV	-	-0.398	-0.634	-0.519	-0.322	-0.448	-4.343
GARCH-RV-F	-1.170	-0.374	-0.608	-0.519	-0.322	-0.448	-4.273
HF-GARCH	-	-0.388	-0.624	**-0.515**	**-0.318**	**-0.447**	-4.246
HF-GARCH-F	**-1.154**	**-0.369**	**-0.602**	-0.515	-0.318	-0.447	**-4.191**
	Panel G: Financials (SPXEW40)						
GARCH	-	-0.252	-0.529	-0.450	-0.207	-0.497	-4.796
GARCH-F	-1.258	-0.237	-0.511	-0.447	-0.207	-0.496	-4.787
GARCH-RV	-	-0.250	-0.526	-0.451	-0.209	-0.498	-4.886
GARCH-RV-F	-1.254	-0.240	-0.513	-0.447	-0.205	-0.498	-4.818
HF-GARCH	-	-0.236	-0.511	-0.443	-0.204	-0.494	-4.669
HF-GARCH-F	**-1.227**	**-0.232**	**-0.504**	**-0.441**	**-0.201**	**-0.493**	**-4.650**
	Panel H: Information technology (SPXEW45)						
GARCH	-	-0.349	-0.605	-0.504	-0.272	-0.520	-4.992
GARCH-F	-1.300	-0.328	-0.637	-0.674	-0.271	-0.519	-4.945
GARCH-RV	-	-0.347	-0.602	-0.498	-0.267	-0.518	-4.942
GARCH-RV-F	-1.288	-0.327	-0.581	-0.498	-0.267	-0.518	-4.832
HF-GARCH	-	-0.332	-0.584	-0.490	**-0.262**	-0.514	-4.766
HF-GARCH-F	**-1.257**	**-0.319**	**-0.570**	**-0.488**	-0.262	**-0.514**	**-4.659**
	Panel I: Telecommunications (SPXEW50)						
GARCH	-	-0.364	-0.614	-0.564	-0.337	-0.528	-4.985
GARCH-F	-1.349	-0.353	-0.604	-0.563	-0.336	-0.527	-4.980
GARCH-RV	-	-0.356	-0.605	-0.558	-0.332	-0.526	-4.894
GARCH-RV-F	-1.331	-0.350	-0.601	-0.559	-0.332	-0.526	-4.901
HF-GARCH	-	-0.348	-0.596	-0.551	-0.326	**-0.523**	-4.783
HF-GARCH-F	**-1.313**	**-0.346**	**-0.594**	**-0.550**	**-0.325**	-0.523	**-4.767**
	Panel J: Utilities (SPXEW55)						
GARCH	-	-0.356	-0.608	-0.555	-0.353	-0.480	-4.550
GARCH-F	-1.239	-0.348	-0.599	-0.558	-0.355	-0.480	-4.551
GARCH-RV	-	-0.357	-0.610	-0.550	-0.347	-0.479	-4.403
GARCH-RV-F	-1.233	-0.346	-0.597	-0.552	-0.348	-0.478	-4.407
HF-GARCH	-	-0.351	-0.604	**-0.544**	**-0.342**	-0.477	**-4.322**
HF-GARCH-F	**-1.220**	**-0.343**	**-0.595**	-0.547	-0.343	**-0.476**	-4.330

with this being the most informationally efficient scoring rule when the underlying model is correctly specified.

4.4 Understanding the predictive performance improvements

As discussed above, optimally weighting high-frequency returns in the density forecasting model improves out-of-sample performance. We now turn our attention to the features of the optimal weights and how they lead to better predictions.

4.4.1 Heterogeneity in intraday weights across series and time

Firstly, we analyze the structure of the estimated weights, $\mathbf{\Gamma}$, imposing the standardization that the mean weight equals one, so we study $\tilde{\mathbf{\Gamma}}$ defined as:

$$\tilde{\mathbf{\Gamma}} = Mp \left(\boldsymbol{\iota}^{\top} \mathbf{\Gamma} \right)^{-1} \mathbf{\Gamma} \tag{14}$$

With the scale impact removed, we gain insights into differences in the structure of the intraday weights among the thirteen series over time. Figure 2 depicts the standardized weights for the six index series where the HF-GARCH model significantly improves predictive accuracy, namely the three broad market indices (SPY, MID, and SPCY) and the Materials, Industrials, and Consumer Staples indices (SPXEW15, SPXEW20, and SPXEW30). Each panel depicts the mean of the likelihood-based posterior estimate obtained using the data from the beginning of the sample up to the end of the year specified by the axis. Note that the corresponding intraday weight implied by the GARCH-RV specification is a constant weight equal to one.

The six panels presented in Figure 2 reveal heterogeneity in the weights across indices as well as time. For the three broad stock indices depicted in panels A to C, we observe that the information from the end-of-day trading period tends to have a larger impact on

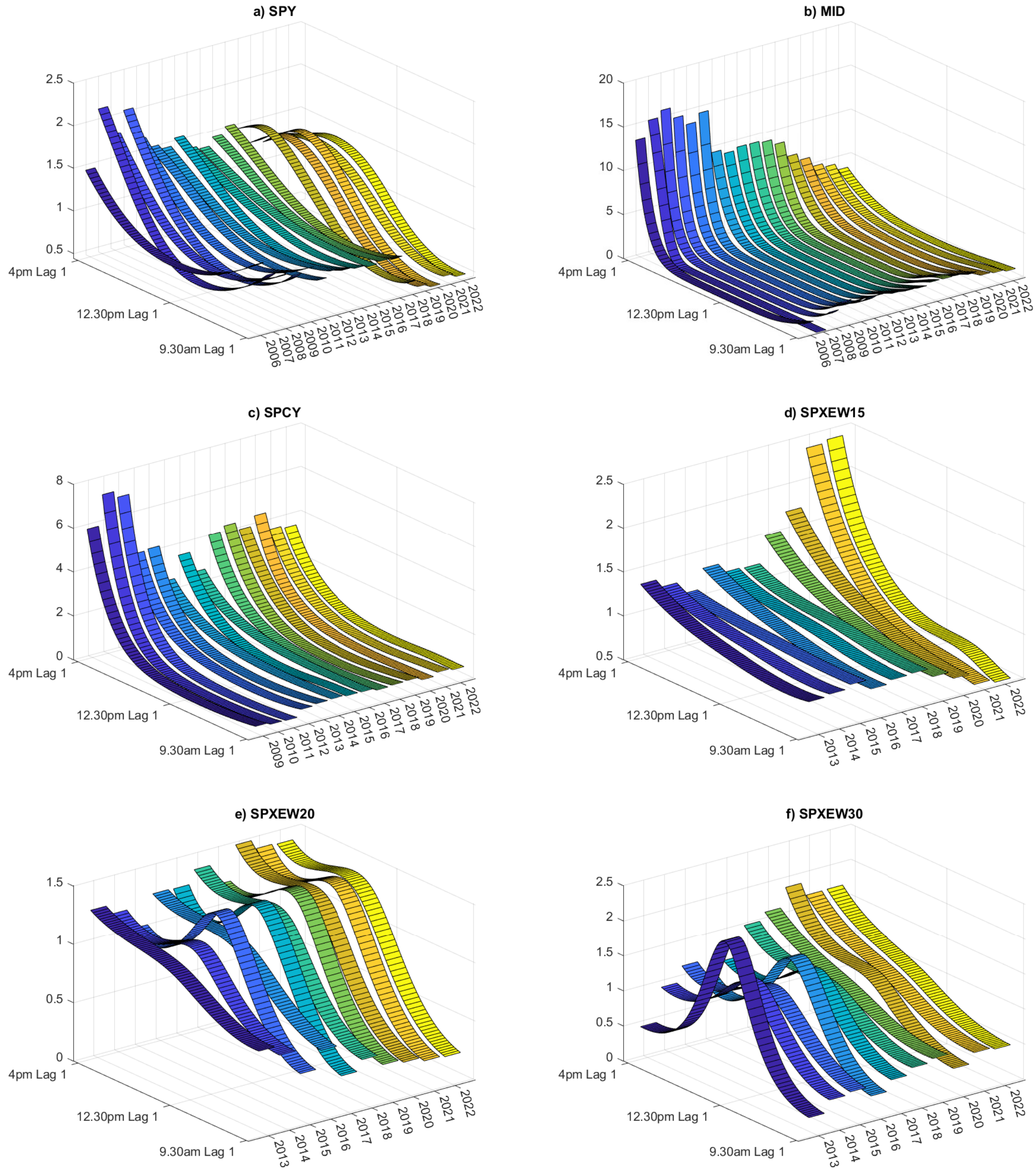

Figure 2: **Weights on intraday returns as a function of time of day, by year.** The weights are from the the HF-GARCH(1,1) estimated using the log-score. Dark blue ribbons are represent weights that are older in time, and yellow ribbons represent the most recent weights.

the return conditional distribution of the following day, broadly in line with the pattern discovered in Patton and Zhang (2022). In the early part of the sample we also observe an increase at the beginning of the day, which matches with the U-shaped diurnality pattern typically associated with the volume of trading activities, however this is only present for the SPY index.

There is a greater degree of heterogeneity observed in the industrial sub-indices, presented in panels D to F. In the case of the Materials index (SPXEW15), the model implied weights are relatively flat at the start of the sample, while towards the end of the sample the intraday weights gradually shift to patterns that align with the broad market indices. The weights for both the Industrials (SPXEW20) and Consumer Staples (SPXEW30) indices in panels E and F exhibit a pattern showing a larger contribution towards the middle of the day, at least for the early part of the sample period, with the blue ribbons peaking in the middle of the time-of-day dimension. However, towards the end of the sample period there is a shift to larger weights on the end-of-day observations. For the Industrials index (SPXEW20), the returns after noon exhibit relatively equal weights near 2022. In contrast, for Consumer Staples (SPXEW30), the weights become monotonically decreasing. These results highlight the flexibility of the spline structure used to capture the intraday information in our proposed HF-GARCH model.

4.4.2 Differences in weights from focused prediction

We now turn our attention to the differences in the estimated intraday weights when we focus the predictive distribution to a specific scoring rule. Figure 3 summarizes the standardized weights, $\tilde{\mathbf{\Gamma}}$ (defined in equation 14) for the three major indices. We depict the log-score based posterior median weight along with its 90% posterior interval, which is constructed using posterior quantiles and is shaded in blue, along with the tail-focused counterparts in red. A comparison of the log-score based intervals and the CS20 focused intervals are depicted in the left column in panels A, C, and E, while comparisons between the log-score based intervals and the CS80 focused intervals are depicted in the right column in panels B, D and F. In all plots, the posterior intervals are obtained from the model

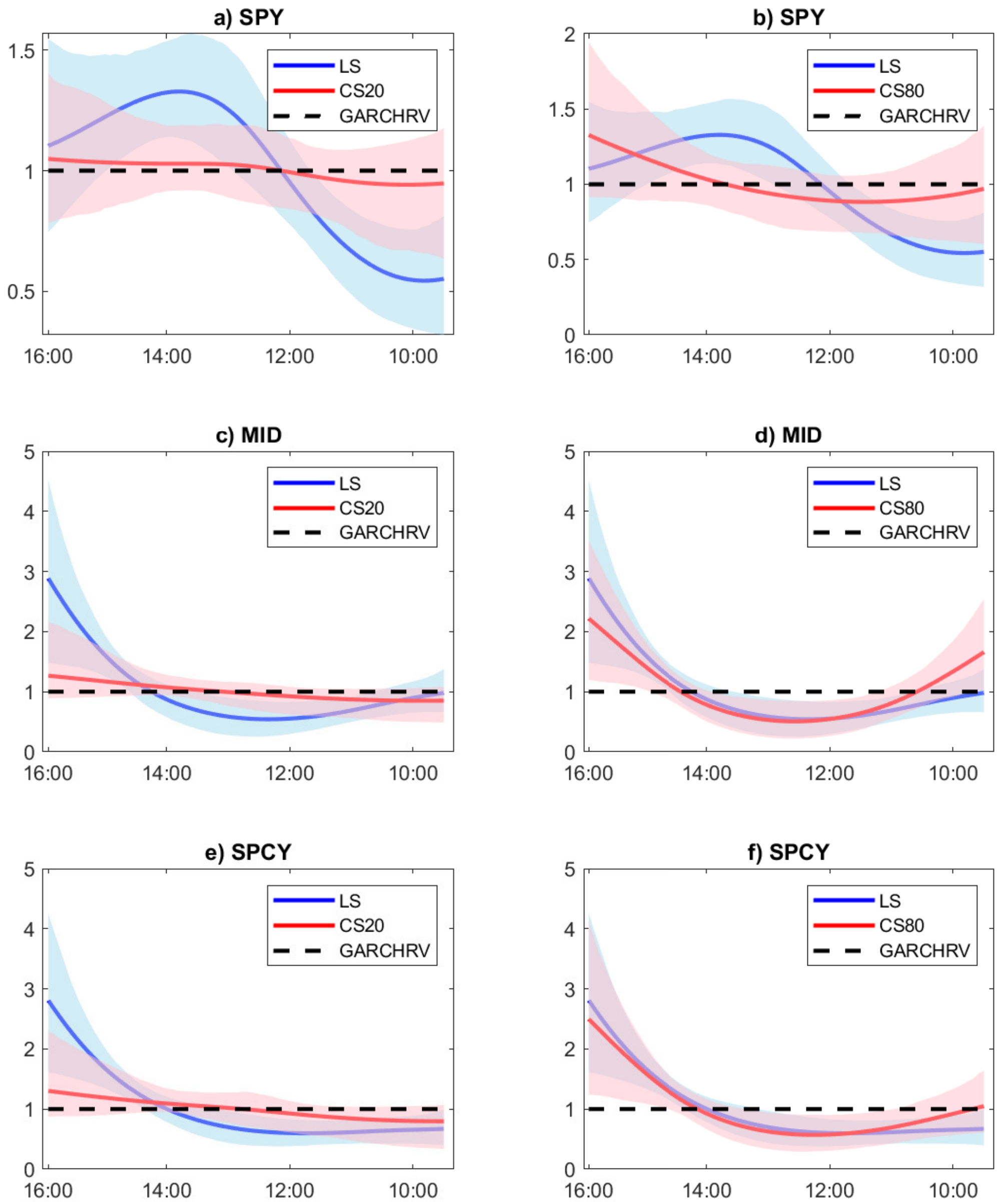

Figure 3: **Optimal weights on intraday returns by estimation method.** This plot shows the optimal weights on intradaily returns as a function of time of day, when estimated using different scoring rules, for the three broad equity indices.

fitted using the entire dataset, and the constant GARCH-RV weights are plotted as the horizontal black dashed line for comparison. The results of the remaining industrial indices are presented in Supplemental Appendix C.

The results from Figure 3 reveal that the CS20-focused intraday weights in the right column tend to be flatter in structure than the log-score based intraday weights for the three broad market indices. This implies that in predicting the tail of the return distribution, earlier information in the trading day also plays a role in improving the predictive accuracy. It is important to note that while a flatter information structure is observed, there is still a tendency for a higher weight given to observations later in the day, with the weight declining as we move from 4pm back to 9.30am. On the other hand, the CS80 focused information flow in the right column exhibits a pattern similar to that of the LS based information flow for the three indices inspected here. This general pattern, for both CS20 and CS80 focused predictions, is also observed for the Industrial, Consumer Staples and Information Technology indices (SPXEW20, SPXEW30 and SPXEW45), presented in Supplemental Appendix D, where the HF-GARCH-F model yields a significant predictive gain compared to the log-score prediction.

4.4.3 Focused prediction and the tails of density forecasts

Finally, we present a comparison of the predictive distributions produced from log-score based HF-GARCH model with the tail-focused counterparts on a representative day when the realized volatility of each respective index is at its median value. Figures 4 and 5 present the predictive densities for the three broad market indices and the three industrial indices where the HF-GARCH-F model produced significant predictive gain in the left tail, namely the SPXEW20, SPXEW30 and SPXEW45 indices.[7] When the upper tail is the focus, we observe little difference between the log-score predictive density and the prediction obtained using the upper-tail focused scoring rule, CS80, see the right columns of Figures 4 and 5. In contrast, when the lower tail is the focus, we see that the tail-focused predictive density

[7]The predictive density plots of the remaining indices are provided in Supplemental Appendix D.

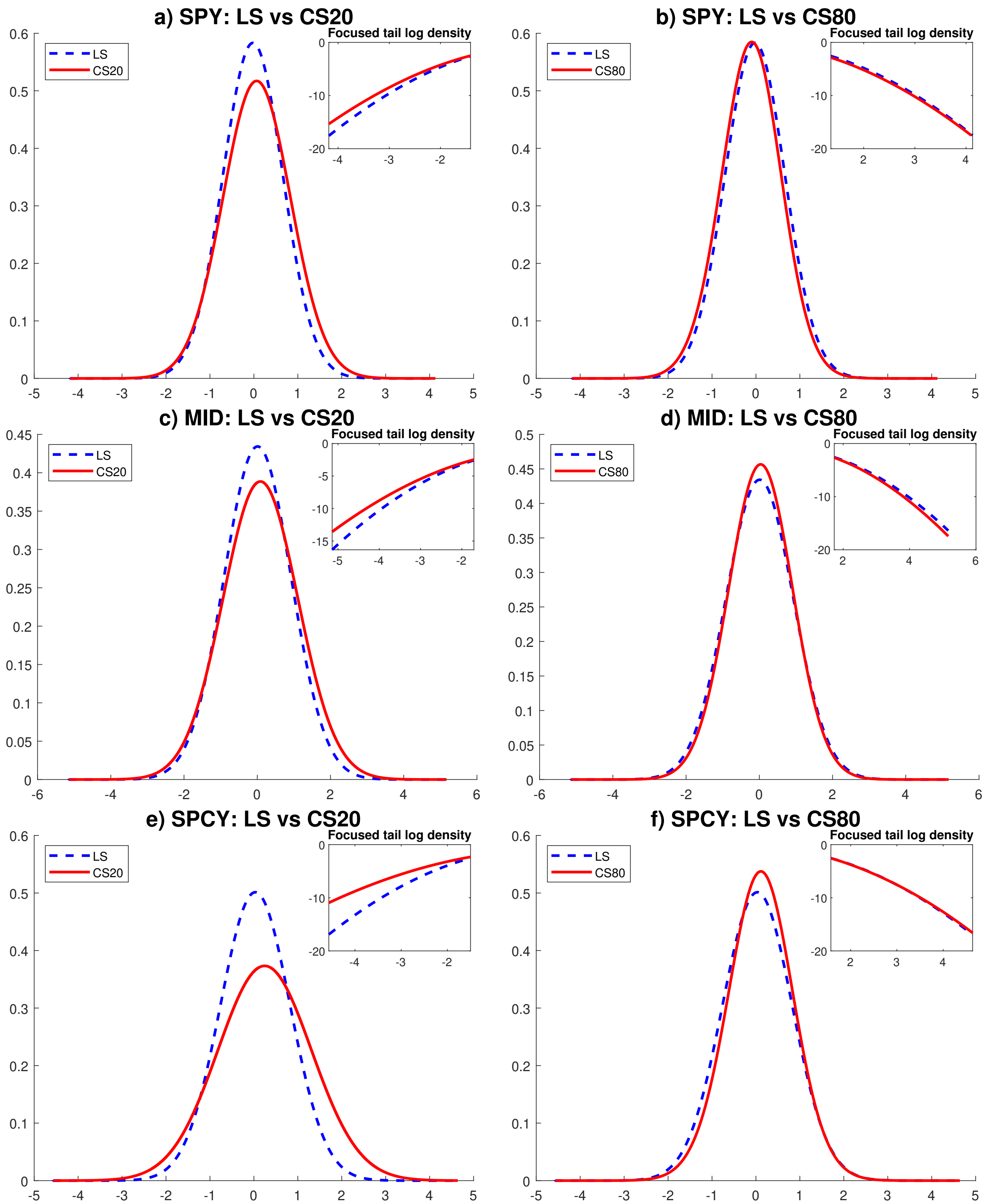

Figure 4: **Representative predictive densities for three broad equity indices.** This plot shows predictive densities for the three broad equity indices on a day where the realized volatility is at its median value, when estimated by different scoring rules. The left column compares the log-score and censored log-score at the 20% threshold, focusing on the left tail, and the right column shows analogous results for the right tail.

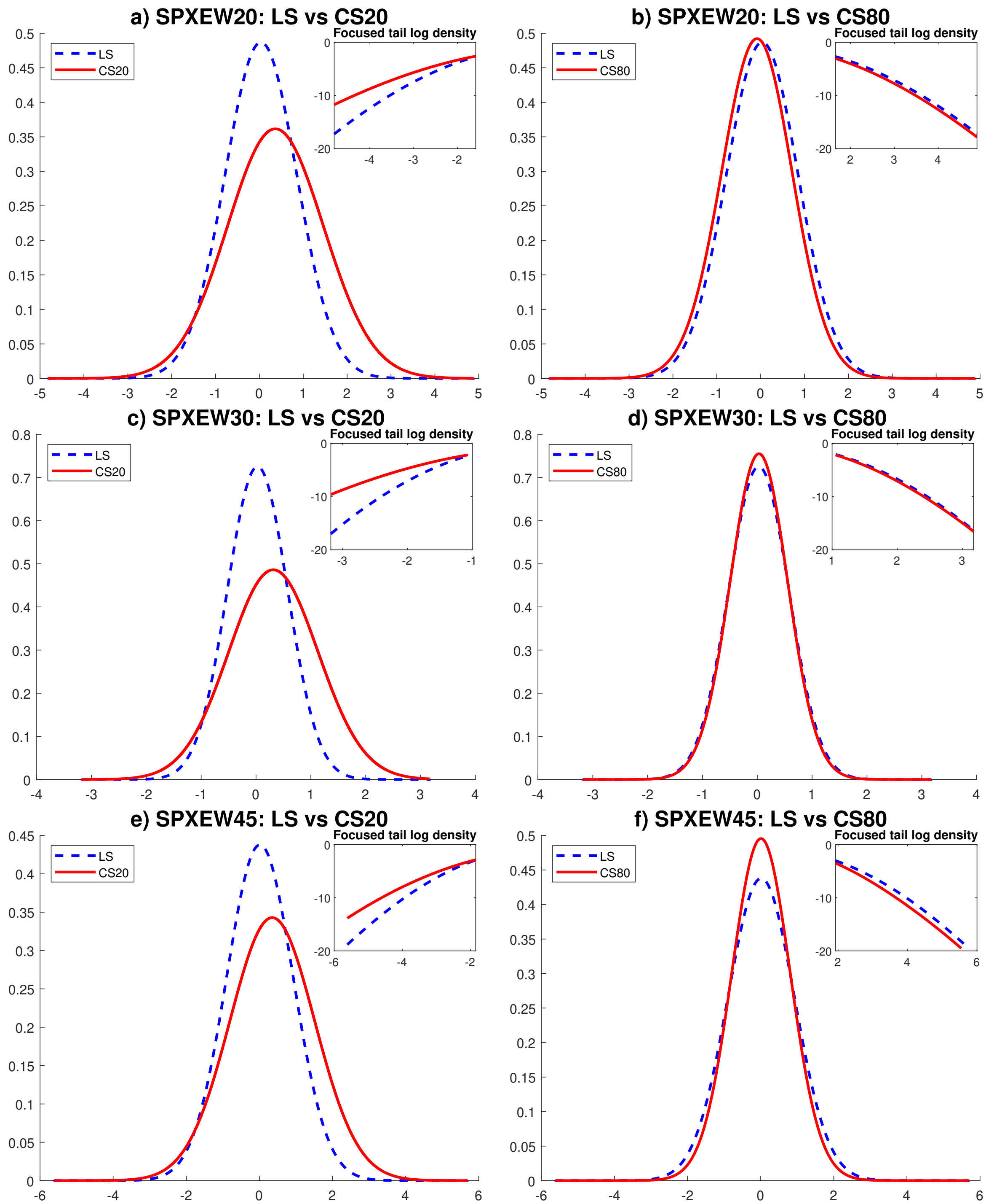

Figure 5: **Representative predictive densities for the SPXEW20, SPXEW30 and SPXEW45 indices.** This plot shows predictive densities for three industry indices on a day where the realized volatilityis at its median value, when estimated by different scoring rules. The left column compares the log-score and censored log-score at the 20% threshold, focusing on the left tail, and the right column shows analogous results for the right tail.

(produced using the CS20 scoring rule) has fatter tails than the forecast from the log-score based model. This is consistent with the negative skew in the returns on these indices (see Table 1) being incompatible with the Gaussian density used in the model (see equation 1). The focused scoring rule enables the misspecified model to better fit the region of interest, in this case the left tail, at the expense of fitting other parts of the distribution.

5 Concluding remarks

Forecasting models that incorporate information from realized volatility measures have found widespread success, see Hansen *et al.* (2012) and Engle *et al.* (2013) for examples, however such models implicitly impose that all high-frequency returns are of equal predictive value. This paper introduces a novel density forecasting model that relaxes this assumption and allows for a flexible and automated aggregation of high-frequency information, in the spirit of the class of "mixed data sampling" (MIDAS) models of Ghysels *et al.* (2006) and the "bespoke realized variance" model of Patton and Zhang (2022). Instead of employing a pre-constructed measure of realized volatility, we model the weights on the high-frequency observations using a flexible B-spline function, and we adopt a regularizing hierarchical prior on the B-spline basis coefficients so that the weights deviate from equal weights only when supported by the data. We estimate the model either by maximizing the log-score (analogous to maximum likelihood), or via "focused Bayesian prediction" (Loaiza-Maya *et al.*, 2021) where we maximize a scoring rule that is "focused" on a particular region of the support of the predictive density, e.g. the left tail.

We apply the model to daily returns on 13 equity indices over the period 2000 to 2022. Consistent with Andersen *et al.* (2006) and many others, we find significant gains from incorporating high frequency data into the predictive model. We also find that allowing for heterogeneous weights over the trade day leads to additional significant gains, across all assets and estimation methods, and these gains are particularly strong when the model is estimated using the same scoring rule as used in forecast evaluation (i.e., using focused Bayesian methods). The optimal intradaily weights tend to be higher towards market closing time if overall fit is of interest, while they tend to be more even across the time of day if the prediction of downside risk is of interest. We also find evidence of variation in the optimal weights across time. These findings highlight the flexibility of our proposed model, with the heteroeneity and time variation in the weights pointing to potential research direction that incorporate these features in a model with richer time series dynamics and/or a model of multivariate dimension.

References

Andersen, T. G., Bollerslev, T., Christoffersen, P. F., and Diebold, F. X. (2006). Volatility and correlation forecasting. In Elliott, G., Granger, C. W. J., and Timmermann, A., editors, *Handbook of Economic Forecasting*, volume 1, pages 777–878. Elsevier.

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

Baillie, R. T., Bollerslev, T., and Mikkelsen, H. O. (1996). Fractionally integrated generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 74(1):3–30.

Barndorff-Nielsen, O. E. and Shephard, N. (2002). Econometric analysis of realized volatility and its use in estimating stochastic volatility models. *Journal of the Royal Statistical Society Series B*, 64(2):253–280.

Bawa, V. S., Brown, S. J., and Klein, R. W. (1979). *Estimation Risk and Optimal Portfolio Choice*. North-Holland.

Bissiri, P. G., Holmes, C. C., and Walker, S. G. (2016). A general framework for updating belief distributions. *Journal of the Royal Statistical Society Series B*, 78(5):1103–1130.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3):307–327.

Bollerslev, T. (1987). A conditionally heteroskedastic time series model for speculative prices and rates of return. *Review of Economics & Statistics*, 69(3):542–547.

Borowska, A., Hoogerheide, L., Koopman, S. J., and van Dijk, H. K. (2020). Partially censored posterior for robust and efficient risk evaluation. *Journal of Econometrics*, 217(2):335–355.

Brenner, R. J., Harjes, R. H., and Kroner, K. F. (1996). Another look at models of the short-term interest rate. *Journal of Financial and Quantitative Analysis*, 31(1):85–107.

Chen, X. and Ghysels, E. (2011). News—good or bad—and its impact on volatility predictions over multiple horizons. *Review of Financial Studies*, 24(1):46–81.

Chen, X., Ghysels, E., and Wang, F. (2015). Hybrid-GARCH: A generic class of models for volatility predictions using high frequency data. *Statistica Sinica*, 25(2):759–786.

Chiang, I.-H. E., Liao, Y., and Zhou, Q. (2021). Modeling the cross-section of stock returns using sensible models in a model pool. *Journal of Empirical Finance*, 60:56–73.

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

De Boor, C. (1978). *A Practical Guide to Splines*, volume 27. Springer.

Dhaene, G. and Wu, J. (2020). Incorporating overnight and intraday returns into multivariate GARCH volatility models. *Journal of Econometrics*, 217(2):471–495.

Diks, C., Panchenko, V., and Van Dijk, D. (2011). Likelihood-based scoring rules for comparing density forecasts in tails. *Journal of Econometrics*, 163(2):215–230.

Ding, Z., Granger, C. W., and Engle, R. F. (1993). A long memory property of stock market returns and a new model. *Journal of Empirical Finance*, 1(1):83–106.

Eilers, P. H. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, 11(2):89–121.

Eilers, P. H. and Marx, B. D. (2010). Splines, knots, and penalties. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(6):637–653.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, 50(4):987–1007.

Engle, R. F., Ghysels, E., and Sohn, B. (2013). Stock market volatility and macroeconomic fundamentals. *Review of Economics & Statistics*, 95(3):776–797.

Engle, R. F. and Ng, V. K. (1993). Measuring and testing the impact of news on volatility. *Journal of Finance*, 48(5):1749–1778.

Engle, R. F. and Russell, J. R. (1998). Autoregressive conditional duration: A new model for irregularly spaced transaction data. *Econometrica*, 66(5):1127–1162.

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.

Fan, J., Qi, L., and Xiu, D. (2014). Quasi-maximum likelihood estimation of GARCH models with heavy-tailed likelihoods. *Journal of Business & Economic Statistics*, 32(2):178–191.

Frazier, D. T., Loaiza-Maya, R., Martin, G. M., and Koo, B. (2024). Loss-based variational Bayes prediction. *Journal of Computational and Graphical Statistics*, forthcoming.

Ghysels, E., Santa-Clara, P., and Valkanov, R. (2006). Predicting volatility: getting the most out of return data sampled at different frequencies. *Journal of Econometrics*, 131(1-2):59–95.

Glosten, L. R., Jagannathan, R., and Runkle, D. E. (1993). On the relation between the expected value and the volatility of the nominal excess return on stocks. *Journal of Finance*, 48(5):1779–1801.

Gneiting, T., Balabdaoui, F., and Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society Series B*, 69(2):243–268.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Hall, P. and Yao, Q. (2003). Inference in ARCH and GARCH models with heavy–tailed errors. *Econometrica*, 71(1):285–317.

Hansen, P. R. (2005). A test for superior predictive ability. *Journal of Business & Economic Statistics*, 23(4):365–380.

Hansen, P. R. and Dumitrescu, E.-I. (2022). How should parameter estimation be tailored to the objective? *Journal of Econometrics*, 230(2):535–558.

Hansen, P. R., Huang, Z., and Shek, H. H. (2012). Realized GARCH: A joint model for returns and realized measures of volatility. *Journal of Applied Econometrics*, 27(6):877–906.

Hansen, P. R., Lunde, A., and Nason, J. M. (2011). The model confidence set. *Econometrica*, 79(2):453–497.

Jensen, M. J. and Maheu, J. M. (2013). Bayesian semiparametric multivariate GARCH modeling. *Journal of Econometrics*, 176(1):3–17.

Jorion, P. (1985). International portfolio diversification with estimation risk. *Journal of Business*, 58(3):259–278.

Kandel, S. and Stambaugh, R. F. (1996). On the predictability of stock returns: An asset-allocation perspective. *Journal of Finance*, 51(2):385–424.

Klein, N. and Kneib, T. (2016). Scale-dependent priors for variance parameters in structured additive distributional regression. *Bayesian Analysis*, 11(4):1071–1106.

Kneib, T., Konrath, S., and Fahrmeir, L. (2011). High dimensional structured additive regression models: Bayesian regularization, smoothing and predictive performance. *Journal of the Royal Statistical Society Series C*, 60(1):51–70.

Liu, C., Wang, C., Tran, M.-N., and Kohn, R. (2023). Deep learning enhanced realized GARCH. *arXiv preprint arXiv:2302.08002*.

Liu, L. Y., Patton, A. J., and Sheppard, K. (2015). Does anything beat 5-minute RV? A comparison of realized measures across multiple asset classes. *Journal of Econometrics*, 187(1):293–311.

Loaiza-Maya, R., Martin, G. M., and Frazier, D. T. (2021). Focused Bayesian prediction. *Journal of Applied Econometrics*, 36(5):517–543.

Martin, G. M., Loaiza-Maya, R., Maneesoonthorn, W., Frazier, D. T., and Ramírez-Hassan, A. (2022). Optimal probabilistic forecasts: When do they work? *International Journal of Forecasting*, 38(1):384–406.

Ong, V. M.-H., Nott, D. J., and Smith, M. S. (2018). Gaussian variational approximation with a factor covariance structure. *Journal of Computational and Graphical Statistics*, 27(3):465–478.

Patton, A. J. (2020). Comparing possibly misspecified forecasts. *Journal of Business & Economic Statistics*, 38(4):796–809.

Patton, A. J. and Zhang, H. (2022). Bespoke realized volatility: Tailored measures of risk for volatility prediction. *Available at SSRN 4315106*.

Zakoïan, J.-M. (1994). Threshold heteroskedastic models. *Journal of Economic Dynamics & Control*, 18(5):931–955.