



Financial density forecasts: A comprehensive comparison of risk-neutral and historical schemes

Ricardo Crisóstomo
Lorena Couso

Documentos de Trabajo
Nº 67



Financial density forecasts: A comprehensive comparison of risk-neutral and historical schemes

Ricardo Crisóstomo* and Lorena Couso**

Working paper

No. 67

November 2017

* Corresponding author. Comisión Nacional del Mercado de Valores (CNMV). Edison 4, 28006 Madrid. Email: rcayala@cnmv.es; and National Distance Education University (UNED).

** CaixaBank Asset Management.

We are grateful to MEFF for providing the market data. The opinions in this article are the sole responsibility of the authors and they do not necessarily coincide with those of the CNMV or CaixaBank.

The opinions in this Working Paper are the sole responsibility of the authors and they do not necessarily coincide with those of the CNMV.

The CNMV publishes this Working Paper Series to enhance research and contribute towards greater knowledge of the stock markets and their regulation.

The CNMV distributes its reports and publications via the internet at www.cnmv.es

© CNMV. The contents of this publication may be reproduced, subject to attribution.

Layout & printing: Composiciones Rali, S.A.

Abstract

We investigate the forecasting ability of the most commonly used benchmarks in financial economics. We approach the main methodological caveats of probabilistic forecasts studies –small samples, limited models and non-holistic validations– by performing a comprehensive comparison of 15 predictive schemes during a time period of over 21 years. All densities are evaluated in terms of their statistical consistency, local accuracy and forecasting errors. Through the development of a new indicator, the Integrated Forecast Score (IFS), we show that risk-neutral densities outperform historical-based predictions in terms of information content. We find that the Variance Gamma model generates the highest out-of-sample likelihood of observed prices and the lowest predictive errors, whereas the ARCH-based GJR-FHS delivers the most consistent forecasts across the entire density range. In contrast, lognormal densities, the Heston model or the non-parametric Breeden-Litzenberger formula yield biased predictions and are rejected in statistical tests.

Keywords: Probabilistic forecasts, risk-neutral densities, ARCH models, ensemble predictions, model validation.

JEL Classification: C14, G12, G13, C52, C53.

List of Contents

1	Introduction	11
2	Forecasting methods	13
2.1	Historical densities	13
2.2	Risk-neutral densities	14
3	Data	17
3.1	Option prices	17
3.2	Futures prices	18
3.3	Interest rates and dividends	18
4	Calibration	19
4.1	Historical densities	19
4.2	Risk-neutral densities	20
5	Density forecasts verification	21
5.1	Goodness-of-fit analyses	21
5.2	The Logarithmic Score	22
5.3	The Continuously Ranked Probability Score	22
6	Empirical results	25
6.1	PIT histograms visual inspection	25
6.2	Statistical consistency: Goodness-of fit tests	26
6.3	Local accuracy: Logarithmic Score	28
6.4	Forecasting errors: Continuous Ranked Probability Score	30
6.5	A holistic evaluation: The Integrated Forecast Score (IFS)	31
7	Conclusions	35
	References	37

List of Tables

TABLE 1	Summary statistics for the IBEX 35 option dataset	18
TABLE 2	Goodness-of-fit tests and statistical consistency	27
TABLE 3	Descriptive statistics of the transformed PIT sequences	28
TABLE 4	Out-of-sample log-likelihoods	29
TABLE 5	Out-of-sample Continuous Ranked Probability Score	31
TABLE 6	Comprehensive ranking of forecasting schemes	32

List of Figures

FIGURE 1	Histogram of PIT realizations	26
----------	-------------------------------	----

1 Introduction

Forecasting future asset prices is arguably one of the most relevant problems for risk managers, central bankers and investors. Historical and risk-neutral methods are the most widely used techniques in financial forecasting. Yet, when it comes to evaluate predictions across the entire density range, comprehensive comparisons are scarce and there is no consensus on which models perform better.

Historical methods generate future predictions based on past prices. These models are easy to implement and extensively used in risk management and stress testing. However, it is well-known that historical patterns do not repeat themselves, particularly in times of economic turmoil. Furthermore, historical models may yield different estimates depending on the length of the calibration window, introducing uncertainty and possible cherry-picking concerns.

Risk-neutral estimates, on the other hand, contain forward-looking expectations and react immediately to changing market conditions, thus being better suited conceptually for forecasting purposes. However, risk-neutral models do not explicitly consider the investors' risk preferences across different future states. Consequently, some agents rapidly dismiss risk-neutral models as the basis for financial predictions.

The previous literature on financial forecasts has been mainly devoted to volatility predictions. Poon and Granger (2003) compare the results from 18 academic papers showing that in 17 of them implied volatilities produce better forecasts than GARCH-based volatilities. Similarly, an extensive survey by Christoffersen, Jacobs, and Chang (2013) find that option-based volatilities beat historical forecasts in most empirical comparisons¹. Conversely, Canina and Figlewski (1993) find that implied volatilities do not accurately predict the future, providing an exception to the mainstream literature.

Much fewer studies consider entire density forecasts. While empirical analyses tend to find that risk-neutral densities (RNDs) outperform historical-based estimates², generalizations to other markets or time-periods are typically limited by three methodological reasons. First, data availability issues have led most researchers to work with relatively small option samples (e.g.: Anagnou et al., 2005 and Liu et al., 2007). Limited samples can significantly impact the evaluation of predictive densities, as the inability to reject a particular model can be due to the low statistical power of the testing procedures (Anagnou et al., 2003).

1 Some examples supporting the use of implied volatilities are: Busch, Christensen, and Nielsen (2011), Taylor, Yadav, and Zhang (2010); Giot and Laurent (2007), Jiang and Tian (2005) and Blair, Poon, and Taylor (2001), among others.

2 See Christoffersen, Jacobs, and Chang (2013).

Second, comparing density estimates from a wide range of schemes requires working with markedly different models and mathematical routines. As a result, most studies have contributed through vis-à-vis comparisons across particular model choices (e.g.: Silva and Kahl (1993), Melick and Thomas (1997) and Alonso, Blanco, and Rubio (2005)) or by surveying specific asset dynamics (Yun, 2014). However, empirical analyses covering a comprehensive range of risk-neutral and historical densities are scarce.

Third, the validation of financial density forecasts is typically performed through the so-called probability integral transforms (PIT), which assess the statistical consistency between the *ex-ante* densities and the *ex-post* realizations. However, several papers³ have shown that PIT-based analyses do not consider the forecasting accuracy of the competing methods or the magnitude of its errors, advocating for targeted scoring rules to supplement the PIT assessments.

We approach these methodological caveats –small samples, limited models and non-holistic validations– by performing a comprehensive analysis of 15 forecasting schemes during a period of over 21 years. Historical densities are generated using a wide variety of methods, spanning from returns bootstrapping or standard GARCH dynamics to asymmetric time-series models with filtered historical simulation. Similarly, we estimate RNDs using the most common benchmarks in financial economics, including lognormal densities, stochastic volatility, jump processes and non-parametric distributions.

All density forecasts are evaluated through a 3-tiered criteria. First, we consider a multi-factor goodness-of-fit analysis, assessing each PIT sequence by means of the Berkowitz, Kolmogorov-Smirnov and Jarque-Bera distributional tests. Second, we employ the logarithmic scoring rule, which evaluates the accuracy of each method in predicting the *ex-post* realizations. Third, we are the first to apply, to our knowledge, a return-based Continuous Ranked Probability Score (CRPS) to financial forecasts. The CRPS compares the realizations to the entire *ex-ante* densities, ranking all methods in terms of their prediction errors. Finally, we develop a new indicator, the Integrated Forecast Score (IFS), which aggregates all the statistical consistency, local accuracy and forecasting errors results into a single composite measure.

We calibrate our RNDs using market-derived option prices only. This approach contrasts with the use of exchange-reported settlement prices, which in many cases are theoretically estimated and already reflect specific modelling choices. Finally, we do not consider in this paper combinations of risk-neutral and historical methods; while this approach seems promising⁴, our aim is to shed light on the predictive ability of the most commonly used models in financial economics, thus leaving mixed densities for future research.

The rest of the paper is organized as follows. Section 2 presents the competing models. Sections 3 and 4 contain the dataset and the calibration procedures. Section 5 explains the validation methods, followed by the empirical results in Section 6. Finally, Section 7 concludes.

3 See Bao, Lee, and Saltoğlu (2007), Amisano and Giacomini (2007) and Gneiting and Raftery (2007).

4 See Shackleton, Taylor, and Yu (2010), Høg and Tsiaras (2011), de Vincent-Humphreys and Noss (2012) and Ivanova and Puigvert Gutiérrez (2014).

2 Forecasting methods

2.1 Historical densities

We employ five specifications for the historical densities. The first assumes that future prices follow a geometric Brownian motion. The corresponding densities are then lognormal. Our second specification generates future price paths by a bootstrapping of past returns, thus randomly drawing returns from the empirical distribution function. For each observation date t , the one-day-ahead return is then given by:

$$r_{t+1} = \mu + z_{t+1}, \quad z_{t+1} \sim \{r^h\} \quad (1)$$

where $\{r^h\} = (r_1^h, \dots, r_t^h)$ denotes the set of historical returns and μ is the daily average return. Next, we consider two standard GARCH(1,1) models. This choice is supported by Hansen and Lunde (2005) which compare 330 ARCH models, finding no evidence that a simple GARCH(1,1) underperforms more sophisticated dynamics, with the exception of asymmetric models including a leverage effect. Under a GARCH(1, 1) future returns are given by

$$\begin{aligned} r_{t+1} &= \mu + e_{t+1} \\ e_{t+1} &= \sigma_{t+1} z_{t+1}, \quad z_{t+1} \sim f_p(0,1) \\ \sigma_{t+1}^2 &= \omega + \alpha e_t^2 + \beta \sigma_t^2 \end{aligned} \quad (2)$$

We consider two specifications for the standardized residuals z_t : (i) a Gaussian distribution (GARCH-N) and (ii) a leptokurtic Student's t (GARCH-t). Finally, we evaluate the Filtered Historical Simulation (FHS) approach introduced in Barone-Adesi, Engle, and Mancini (2008), which combines an asymmetric GJR-GARCH model⁵ with empirical innovations. Specifically, future returns in the GJR-FHS(1,1) model are computed as

$$\begin{aligned} r_{t+1} &= \mu + e_{t+1} \\ e_{t+1} &= \sigma_{t+1} z_{t+1}, \quad z_{t+1} \sim f_{np}(0,1) \\ \sigma_{t+1}^2 &= \omega + \alpha e_t^2 + \beta \sigma_t^2 + \gamma I_t e_t^2 \end{aligned} \quad (3)$$

where $I_t = 1$ when $e_t < 0$ and 0 otherwise, introducing a leverage effect. The innovations z_t are drawn from the scaled empirical distribution function, which is obtained by dividing each estimated return innovation \hat{e}_t by the estimated conditional volatility $\hat{\sigma}_t$ ⁶, thus reflecting the skewness and kurtosis observed in the historical calibration period.

⁵ Glosten, Jagannathan, and Runkle (1993).

⁶ Consequently, the set of scaled non-parametric innovations $f_{np}(0,1)$ is formed by a set of past returns $z_t^h = \hat{e}_t / \hat{\sigma}_t$.

Simulated price paths for the time-series models are generated by drawing from the assumed return distribution, calculating the conditional variance (where applicable), drawing another innovation, updating the conditional variance, and so on up to the forecasting horizon, denoted as t^* . Consequently, future prices at the forecasting horizon F_{t^*} are given by:

$$F_{t^*} = F_t \exp(\tau\mu + \sum_{i=1}^{\tau} \sigma_{t+i} z_{t+i}) \quad (4)$$

where F_t is the market-observed future price at date t , z_{t+i} denotes the daily innovations for each model, and τ is the number of business days between t and t^* .

2.2 Risk-neutral densities

Our simplest RND model is again a lognormal specification, but calibrated to current option prices. However, it is well-known that the volatility of financial assets is time-variant and stochastic. Therefore, we next consider the Heston (1993) model, which employs a mean-reverting stochastic volatility process that can generate skewed and leptokurtic distributions. Specifically, risk-neutral Heston dynamics are given by

$$dF_t = \sqrt{V_t} F_t dW_{t,1} \quad (5)$$

$$dV_t = a(\bar{V} - V_t)dt + \eta\sqrt{V_t} dW_{t,2} \quad (6)$$

where $dW_{t,1}$ and $dW_{t,2}$ are two correlated Wiener processes. The empirical literature suggests that a jump component can help explaining the observed equity returns, particularly in short term horizons⁷. Following Bates (1996), we complement the Heston volatility in (6) with a lognormal price jump, thus obtaining the dynamics:

$$dF_t = \sqrt{V_t} F_t dW_t^{(1)} + J_t F_t dN_t - \lambda\mu_j F_t dt \quad (7)$$

where N_t is a Poisson process with intensity λ and J_t are the jumps sizes, which are lognormally distributed with an average size μ_j and standard deviation ν_j . Leaving diffusion models, we also evaluate the purely discontinuous Variance Gamma (VG) model (Madan, Carr, and Chang, 1998), which combines frequent small moves with rare big jumps. The VG dynamics are:

$$\begin{aligned} F_{t^*} &= F_t e^{\lambda\tau + H(\tau; \sigma, \nu, \theta)} \\ \lambda &= \frac{1}{\nu} \ln\left(1 - \theta\nu - \frac{\sigma^2\nu}{2}\right) \\ H(\tau; \sigma, \nu, \theta) &= \theta G(\tau; \nu) + \sigma G(\tau; \nu) W_t \end{aligned} \quad (8)$$

where $G(t; \nu)$ a Gamma distribution and the parameters σ , ν and θ jointly control the volatility, asymmetry and kurtosis.

⁷ See Jones (2003), Bakshi, Cao, and Chen (1997), Cont and Tankov (2004), among others.

RNDs for all stochastic models are obtained through characteristic functions. Heston (1993) and Bakshi and Madan (2000) show that the cumulative density function of F_t can be directly obtained in terms of the characteristic function of $\ln(F_{t^*})$ as follows:

$$CDF_{F_{t^*}}(x) = \frac{1}{2} + \frac{1}{\pi} \int_0^{\infty} \operatorname{Re} \left[\frac{e^{-iw \ln(x)} \psi_{\ln F_{t^*}}(w)}{iw} \right] dw \quad (9)$$

where $\operatorname{Re}[\cdot]$ denotes the real operator⁸.

Finally, Breeden and Litzenberger (1978) show that given a continuous of non-arbitrable call prices, it is possible to obtain a unique risk-neutral distribution that replicates such option prices. Specifically, we employ the implementation in Malz (2014), which introduces a simple arbitrage-free approach based on cubic spline interpolations across the observed implied volatilities and a flat extrapolation at the endpoints. For each expiry t^* , the interpolated volatility function is used to compute the continuous call pricing function $C(x, t^*)$, and these prices are then numerically differentiated to obtain the CDF for all the strikes x as:

$$CDF_{F_{t^*}}(x) \approx 1 + e^{r\tau} \frac{1}{\Delta} \left[C\left(x - \frac{\Delta}{2}, t^*\right) - C\left(x + \frac{\Delta}{2}, t^*\right) \right] \quad (10)$$

where Δ denotes the step size used in the finite differentiation.

⁸ See Crisóstomo (2014, 2017) for further details.

3 Data

3.1 Option prices

Our option dataset is comprised of European-style contracts with underlying the IBEX 35 futures. Option prices are observed during a period of over 21 years, from November 1995 to December 2016. To construct our dataset, we follow the recommendations in Christoffersen, Jacobs, and Chang (2013), thus only employing liquid option contracts and minimizing the input modelling assumptions.

Our study focuses on strictly market-derived option prices. This choice deviates from the common practice of using exchange-reported settlement prices, which in some cases are theoretically estimated by the relevant exchange even if there is no real activity in the underlying contracts. For instance, prior to expiration, the settlement prices of IBEX 35 options are computed by MEFF assuming a linear relation in the implied volatility function for OTM and ITM options⁹. Therefore, these settlement prices reflect specific modelling choices and using them in the calibration would entail introducing an exogenously derived volatility shape into our models.

The option dataset is formed by front-month option contracts. This choice is justified by two reasons. First, near-to-expiry contracts exhibit the highest liquidity and thus availability of strictly market-derived prices. Second, the use of monthly cycles allows us to maximize the number of non-overlapping forecasts while minimizing the input modelling assumptions¹⁰. Observation dates are set 28 calendar days before each monthly expiry date. For such dates, we record the bid and ask prices of all the available call and put options. Since ITM options are less actively traded than OTM options, we build our dataset with OTM and ATM calls, whereas OTM put are converted into equivalent call prices using the Put-Call parity.

We only consider options exhibiting contemporaneous bid and ask quotes, while the consistency of each cross-section is ensured by removing contracts that do not respect non-arbitrage conditions¹¹. After filtering, we obtain a dataset of 6659 option prices distributed across 254 monthly cycles. The average number of strikes in

9 The linear slopes are different for ITM and OTM options and are calculated for each maturity. Prior to October 1996, settlement prices were computed using a constant implied volatility (Alonso, Blanco, and Rubio, 2005). MEFF is the official Spanish Futures and Options Exchange.

10 IBEX 35 options do not exhibit shorter than monthly expirations. Therefore, using shorter periods would lead to forecasting horizons that lack direct option quotes, requiring strong extrapolation assumptions. On the other hand, longer expiration cycles would entail both reducing the number of non-overlapping periods and relying on the less liquid back-month contracts.

11 Call and put-derived equivalent contracts whose market price is not a convex and decreasing function of the strike are removed from the dataset.

the cross-sections is 26, ranging from a minimum of 8 to a maximum of 72. Table 1 summarizes the main statistics of the option dataset.

Summary statistics for the IBEX 35 option dataset

TABLE 1

Option type	Total number	Average per day	Maximum per day	Minimum per day
Calls	3,151	12	38	1
Puts	3508	14	46	3
Overall	6659	26	72	8
Moneyness	F/K	No. of options	(%)	
Deep OTM put	>1.10	1755	26,36	
OTM put	1.03-1.10	1423	21,37	
Near the money	0.97-1-03	1496	22,47	
OTM call	0.90-0.97	1541	23,14	
Deep OTM call	<0.90	444	6,67	

A minimum of 8 options is required to calibrate the Bates model. Therefore, in seven observation dates we supplemented the cross-sections with at-the-money option contracts whose last traded price was consistent with the contemporaneous bid and ask prices. This resulted in an addition of 9 options (0.1% of the sample).

3.2 Futures prices

Daily front-month IBEX 35 futures prices are recorded from 19 November 1990 until 23 December 2016¹². Contrary to options, the daily settlement prices of front-month IBEX 35 futures are computed as the volume weighted average of market-derived transactions between 17:29 and 17:30. The minimum IBEX 35 future price is 1882, recorded on 6 October 1992, and the maximum is 15945.5, attained on 11 December 2007.

Separately, at each monthly expiration date, the final settlement price for each future is determined by MEFF by averaging the spot IBEX 35 index prices from 16:15 to 16:45, taking one observation per minute. These settlement prices constitute the underlying asset of the IBEX 35 options and futures used in our study. Consequently, these monthly prices are used to assess the predictive ability of all forecasting schemes for the 21-year sample.

3.3 Interest rates and dividends

For observation periods between January 1999 and December 2016, we employ the 1-month Euribor. In earlier dates, since the Euribor was not available, we employ the 1-month Mibor. All interest rates are consistently applied for each forecasting period using the corresponding act/360 day count convention. Using future contracts have the advantage of making dividend estimation irrelevant; therefore, dividend-related uncertainties do not affect our density forecasts.

¹² Prior to 21 July 1992 IBEX 35 futures were not yet traded. Therefore, its return is proxied by the spot IBEX 35, whose back-tested returns are available since 29 December 1989.

4 Calibration

Model's parameters are calibrated on a strictly *ex-ante* basis, considering only the current or historical information available at each observation date.

4.1 Historical densities

Historical models can be calibrated to any sufficiently long period of past prices. While this provides flexibility, different calibration windows generate different input values and thus different density predictions. To cope with this uncertainty, we consider two calibration periods: (i) a shorter 6-month window and (ii) a longer 5-year period.

The bootstrapping method does not require any statistical calibration, as it simply entails randomly selecting returns from the relevant historical periods. For the lognormal specification, the average return and standard deviation are obtained from the return set in each historical period $\{r^h\}$. Next, ARCH-based parameters are calibrated through maximum likelihood estimation (MLE). Even if the true innovations are not normal, Bollerslev and Wooldridge (1992) show that assuming a Gaussian density for the residuals provides consistent parameter values. Therefore, for the GARCH-N and GJR-FHS models, the parameters are estimated by maximizing the log-likelihood of observing each historical return sequence $\{r^h\}$ which is given by:

$$\ln L = \sum_{j=1}^t \left[-\ln(\sigma_j^2) - \left(\epsilon_j^2 / \sigma_j^2 \right) \right] \quad (11)$$

For the GARCH-t, we employ a two-step estimation process. First, the MLE values of all parameters except \hat{d} are obtained from (11). Then, following Christoffersen (2012), the MLE of \hat{d} is calculated to match the excess kurtosis κ of the GARCH residuals as $\hat{d} = 6 / \kappa + 4$.¹³

Finally, historical parameters are separately estimated for each observation date, performing 254 calibrations per model and observation window. For all time-series models, the calibrated dynamics are then used to generate 100,000 price paths, using equation (4).

13 We also tried to calibrate the GARCH parameters and \hat{d} simultaneously using the log-likelihood of the sample returns under a GARCH(1,1) with $t(d)$ residuals. However, this approach did not generate improved forecast results.

4.2 Risk-neutral densities

RNDs are obtained from the cross-section of option prices at each observation date. For the lognormal model, at-the-money volatilities are computed by linear interpolation of the two nearest-to-the-money implied volatilities in each cross-section. In the Heston, Bates and Variance Gamma models, parameters are calibrated by minimizing the sum of pricing errors. We choose to work with relative errors, which effectively assign a similar weight to all option contracts, thus generating more consistent results across different strike regions¹⁴. Denoting N_t the number of option prices are available at date t , we estimate the parameter set $\hat{\Theta}$ that minimizes the sum of relative errors for each stochastic model as

$$SRE_t = \sum_{i=1}^{N_t} \left| \frac{C_i - \hat{C}_i(\Theta)}{C_i} \right| \quad (12)$$

where C_i denote the mid-market price of each option in the cross-section and $\hat{C}_i(\Theta)$ is the model-dependent value obtained with the parameter set Θ . All natural constraints on the admissible values of the model's parameters are included in the calibration¹⁵. For the Variance Gamma process, we also consider the restriction $v^{-1} > \theta + \sigma^2 / 2$, which is required to avoid numerical blow-ups in the calibration process (see Itkin, 2010 and Crisóstomo, 2017). Following this procedure, risk-neutral parameters are estimated separately for each observation date, performing 254 calibrations per model.

Finally, for the Malz (2014) implementation of the Breeden-Litzenberger formula, we employ a step size $\Delta = 0.01F_t$, which avoids negative probabilities in our predictive densities.

14 Despite its popularity, the use of absolute errors entails overweighting the more expensive ITM options while underweighting the cheaper OTM contracts, thus leading to potential biases.

15 See Heston (1993), Bates (1996) and Madan, Carr, and Chang (1998) for a description of the admissible parameter values.

5 Density forecasts verification

Evaluating financial densities is not particularly straightforward. While in classical inference experimental outcomes can be repeated under similar conditions to assess whether they come from a particular distribution, in financial forecasts only one realization is available to evaluate each entire density prediction. One way to tackle this issue is working with ensemble predictions, thus jointly assessing a sequence of predictive densities and their corresponding sequence of realizations.

However, even with ensembles, the limited availability of liquid options prices makes it difficult to generate a high number of non-overlapping density forecasts. To cope with potential sample issues, we evaluate the forecasting ability of all density schemes during a period of over 21 years, hence analyzing 254 non-overlapping monthly cycles. To our knowledge, this is the highest available in comparable research.

Prior to 2007, most predictive densities were evaluated through PIT-based goodness-of-fit analyses. However, PIT sequences are not informative about the accuracy of the competing methods or the magnitude of their errors. Therefore, to supplement PIT analyses, we compute two additional scoring rules. First, the logarithmic score, which evaluates each model accuracy in predicting the final realizations. Second, a return-based CRPS, which ranks all density schemes in terms of their prediction errors.

5.1 Goodness-of-fit analyses

Diebold, Gunther, and Tay (1998) show that the statistical consistency between a sequence of probabilistic forecasts and the corresponding sequence of realizations can be assessed through PIT-based analyses. For a given date t , the PIT represents the quantile of the *ex-ante* distribution at which the *ex-post* realization is observed. Thus,

$$PIT_t = \int_{-\infty}^{x_t} f_t(x) dx \quad (13)$$

Intuitively, in a well-specified model the realizations should be indistinguishable from random draws from the predictive distributions, and the sequence of PIT values should be therefore uniformly distributed in the $(0, 1)$ range. However, statistical tests based on uniform distributions are typically not powerful enough for small samples (Mitchell and Hall, 2005). Consequently, Berkowitz (2001) proposes a reformulation that transforms the PIT values into a new sequence (T-PIT) that should be formed by i.i.d. $N(0,1)$ variables if the predictive densities are correctly specified¹⁶. The Berkowitz

16 This reformulation brings about the more powerful tests associated with Gaussian variables.

test is carried out by first computing the T-PIT values as $T-PIT_t = \Phi^{-1}(PIT_t)$ and next the AR(1) model

$$T-PIT_t - \mu = \rho(T-PIT_{t-1} - \mu) + \varepsilon_t \quad (14)$$

is estimated to jointly assess their mean, variance and serial correlation using the likelihood ratio test $LR_3 = -2(L(0,1,0) - L(\hat{\mu}, \hat{\sigma}^2, \hat{\rho}))$, which compares the likelihood of the restricted model, where $\mu = 0$, $\text{var}(\varepsilon_t) = 1$ and $\rho = 0$, with that of an unrestricted one.

However, the LR_3 test does not directly address the normality of the T-PIT values. Consequently, if a T-PIT sequence exhibits $\mu = 0$ and $\text{var}(\varepsilon_t) = 1$, but it is non-normal in its higher moments, Berkowitz's test will fail to detect such failures (Dowd, 2004). We thus complement the LR_3 with the Kolmogorov-Smirnov (KS) and Jarque-Bera (JB) tests. The KS test examines whether the maximum distance between the T-PIT distribution and a $N(0,1)$ variable is statistically significant, whereas the JB test considers the skewness and kurtosis of the T-PIT values, hence assessing the higher moments not covered in the LR_3 test.

5.2 The Logarithmic score

The accuracy of a forecasting scheme can be assessed by comparing the likelihood of the *ex-post* realizations evaluated with the *ex-ante* predictive densities. Following, Liu et al. (2007), Shackleton, Taylor, and Yu (2010) and Høg and Tsiaras (2011) we compute the log-likelihood of the realizations for each predictive scheme as

$$L = \sum_{t=1}^N \log(f_t(x_{t^*})) \quad (15)$$

Where f_t denotes the *ex-ante* density computed at observation date t and x_{t^*} denotes the *ex-post* realization at time t^* . Therefore, for each method, the logarithmic rule assigns a loss score to each actual realization depending on its *ex-ante* probability of occurrence, and aggregating these scores over the entire sample, we rank density predictions in terms of their out-of-sample accuracy. Furthermore, when all models are possibly misspecified (as it is the case in financial forecasts), the model with maximum L generates the predictive densities which are nearest to the true generating densities, according to the Kullback-Leibler Information Criterion (KLIC)¹⁷ (Bao, Lee, and Saltoğlu, 2007).

5.3 The Continuous Ranked Probability Score

The logarithmic score considers the likelihood of the *ex-post* realizations, but ignores any other probability masses. In contrast, the CRPS considers the entire predictive distribution, measuring the statistical distance between the actual realization and all other probabilistic outcomes (Matheson and Winkler (1976)). As a result, the CRPS gives good scores to densities that assign high probabilities to future val-

¹⁷ Kullback and Leibler (1951).

ues that are close, but are not identical, to the one materializing (Gneiting and Raftery (2007)). Denoting by CDF^m and CDF^r the cumulative distributions of the forecasting model and the realization, the CRPS is given by:

$$CRPS_t = \int_{-\infty}^{\infty} (CDF^m(x) - CDF^r(x))^2 dx \quad (16)$$

where:

$$CDF^r(x) = \begin{cases} 0 & \text{for } x_{t^*} < x \\ 1 & \text{for } x_{t^*} \geq x \end{cases} \quad (17)$$

Hersbach (2000) shows that the CRPS has the dimension of the parameter x (i.e.: IBEX 35 future points), which enters in the calculus through dx , thus facilitating the CRPS interpretation as a generalization of the mean absolute error for the entire density forecast.

However, in our empirical sample, the settlement prices of IBEX 35 futures range from a minimum of 1882 to a maximum of 15945.5. Therefore, a CRPS of e.g. 200 index points may have markedly different interpretations depending on the observation date, hindering the comparability of the individual CRPS values across different time periods. Consequently, we slightly modify the CRPS to consider return deviations instead of index points. Finally, to aggregate the CRPS values over the entire sample, we compute the average return-based CRPS as

$$CRPS^{rb} = \frac{1}{N} \sum_{i=1}^N \sqrt{\int_{-\infty}^{\infty} (CDF^m(x) - CDF^r(x))^2 dx} \quad (18)$$

For a given forecasting scheme, a $CRPS^{rb}$ of 0.05 can be interpreted as an average return deviation of 5% between the *ex-post* realizations and all the *ex-ante* probability masses, thus providing a direct way to rank competing forecasts.

6 Empirical results

Due to the high number of variants tested, we first summarize the different forecasting schemes and their naming conventions. We first consider two simple historical methods: a lognormal density (LN-HIS) and a bootstrapping of historical return (BTS). We also evaluate two standard GARCH models, either with normal or Student's t innovations (GARCH-N / GARCH-t). Finally, we employ an asymmetric GJR-GARCH model with filtered historical simulation (GJR-FHS). The required parameters for all historical models are calibrated using either a 6-month period (6m) or a 5-year history (5y).

For the RNDs, we consider again a simple lognormal model (LN-ATM). We then evaluate the density obtained with a stochastic volatility model (HESTON) and a Jump-diffusion process (BATES). Finally, we consider the purely discontinuous Variance Gamma process (VG) and the Malz (2014) implementation of the Breeden-Litzenberger formula (BL-MALZ).

6.1 PIT histograms visual inspection

The consistency between an *ex-ante* density scheme and the observed realizations can be intuitively assessed by a simple inspection of the PIT histograms. In reliable forecasts, the histogram of PIT values should resemble a uniform distribution, with departures from the flat line indicating regions where the frequency of realizations is higher or lower than in the *ex-ante* predictions. Figure 1 presents the PIT histograms for all forecasting schemes.

Three main aspects can be highlighted. First, all historical methods calibrated to 6-month significantly understate the probability of large losses. This can be seen in the higher *ex-post* occurrence of severe downside movements (i.e.: left-most histogram bar) compared to the *ex-ante* probabilities (i.e.: horizontal dashed line).

Second, although all models calibrated to 5-year periods improve the left-tail fit, they conversely tend to allocate too much probability to significant upside movements. The most prominent bias is observed in the GARCH-t(5y), where almost no realizations are observed in the right tail compared to the expected probabilities. The GJR-FHS(5y), in contrast, generates a reasonably flat PIT histogram, suggesting that this model produces consistent results across the entire density range.

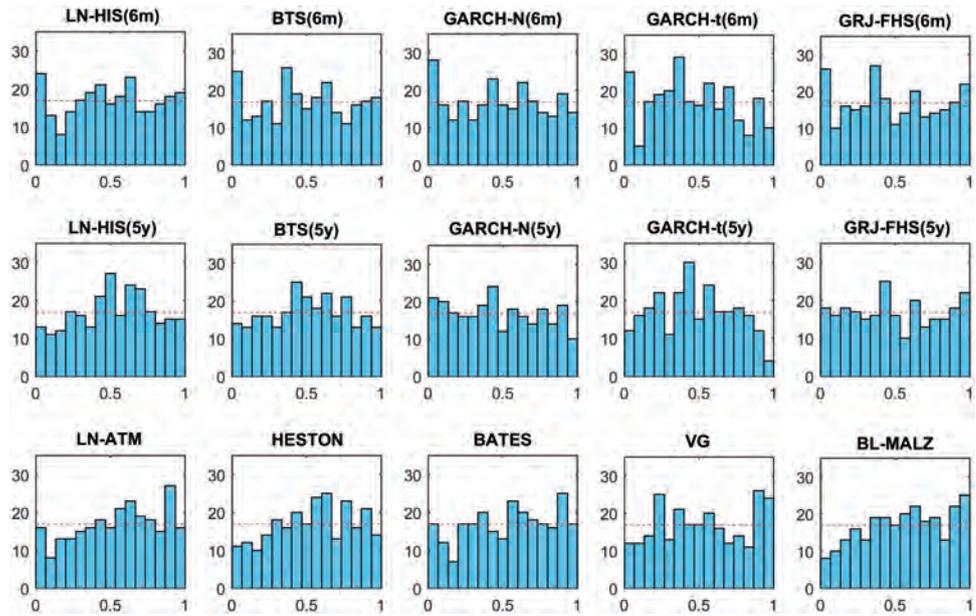
Third, despite being calibrated to the same inputs, RNDs produce notably different outcomes depending on the model's dynamics. Where the Bates model generates a fairly consistent histogram, the HESTON and BL-MALZ models show substantial biases, both allocating too much weight to the left tail while understating either the center or the right tail of the distribution.

6.2 Statistical consistency: Goodness-of fit tests

A well-specified model should simultaneously pass the Berkowitz, JB and KS statistical tests, as rejection in any of these indicates a departure from the T-PIT normality assumption in either: (i) the mean, sigma or autocorrelation, (ii) the asymmetry and kurtosis or (iii) the distance between the theoretical and the observed CDF. Tables 2 and 3 summarize the results from the goodness-of-fit tests and the T-PIT descriptive statistics.

Histogram of PIT realizations

FIGURE 1



The horizontal line represents the expected number of realizations for each density region, while vertical bars show the actual number of observations.

6.2.1 Historical methods

Of all 10 historical schemes, only the GJR-FHS(5y) and the GARCH-N(5y) simultaneously pass the Berkowitz, JB and KS tests at a 5% significance level. Furthermore, the GJR-FHS(5y) exhibits more satisfactory p -values than the GARCH-N(5y) in all goodness-of-fit tests, thus becoming the most statistically reliable historical model.

The rejection in most historical methods (6 out of 10) comes from failures in the JB test, highlighting the importance of assessing the asymmetry and kurtosis of the T-PIT distribution. For the rejected 6-month variants, Table 3 shows that the failure stems from both negative skewness and a lack of kurtosis. Moreover, where a stand-

ard Normal variable should exhibit ± 1.64 critical values for the 5th and 95th quantiles, the T-PIT distribution for the rejected variants show values of (-1.71, -2.22) and (1.38, 1.78) respectively, exposing (i) a clear understatement of extreme losses in the *ex-ante* predictions and (ii) a significant asymmetry of results.

The LN-HIS(5y) and BTS(5y) failures also stem from a left-skewed and leptokurtic T-PIT distribution. Furthermore, the GARCH-t(5y) performs particularly bad, failing

the Berkowitz test. Although this rejection can be attributed to the low variance of the GARCH-t(5y) T-PIT distribution (0.66), the most striking bias occurs in the right tail, where the fat-tailed Student's t innovations generate an expectation of significant upside movements that is not validated by the *ex-post* realizations. In contrast, all historical methods pass the KS test at a 5% level.

Goodness-of-fit tests and statistical consistency

TABLE 2

Model	Berkowitz				Jarque Bera		Kolmogorov-Smirnov		
	Mu	Variance	Rho	LR3	Statistic	p-value (%)	Statistic (%)	p-value (%)	
Historical 6-months									
LN-HIS(6m)	0,002	1,205	-0,011	4,733	19,25	7,24	3,02	5,23	47,51
BTS(6m)	-0,045	1,217	-0,012	5,770	12,33	12,54	1,79	4,10	75,07
GARCH-N(6m)	-0,120	1,200	0,024	8,309	4,00	8,48	2,32	5,94	32,81
GARCH-t (6m)	-0,118	0,995	0,019	3,646	30,24	6,21	4,22	6,71	19,36
GJR-FHS(6m)	-0,010	1,284	0,036	9,091	2,81	0,85	50,00	4,22	74,09
Historical 5-years									
LN-HIS(5y)	0,041	0,978	0,060	1,388	70,82	57,11	0,10	8,27	5,84
BTS(5y)	0,001	0,948	0,065	1,375	71,14	35,95	0,10	6,50	22,41
GARCH-N(5y)	-0,129	1,001	0,061	5,200	15,77	3,20	16,23	6,12	28,51
GARCH-t (5y)	-0,101	0,662	0,052	21,999	0,01	4,18	9,77	8,41	5,19
GJR-FHS(5y)	0,027	1,094	0,094	3,726	29,26	1,76	36,88	3,83	83,68
Risk-neutral									
LN-ATM	0,134	0,933	0,064	6,106	10,66	3,65	12,74	9,67	1,61
HESTON	0,114	0,865	0,056	6,501	8,96	43,19	0,10	10,35	0,80
BATES	0,108	1,020	0,079	4,589	20,45	1,24	50,00	7,10	14,69
VG	0,117	1,032	0,064	4,637	20,04	0,88	50,00	7,45	11,33
BL-MALZ	0,221	0,889	0,075	15,278	0,16	0,99	50,00	10,31	0,84

6.2.2 Risk-neutral methods

Two risk-neutral schemes, BATES and VG, simultaneously pass all three goodness-of-fit tests, exposing the key role of discontinuous jumps in producing consistent density estimates. Specifically, the Bates model exhibits the highest *p*-values in all goodness-of-fit tests, becoming the most reliable RND scheme in terms of statistical consistency.

Densities obtained from risk-neutral methods perform better than historical models in the JB test. As shown in Tables 2 and 3, except for the Heston model, the T-PIT distributions generated from all RNDs exhibit asymmetry and kurtosis levels that are not statistically different from that of a standard Normal variable at a 5% level.

In contrast, the LN-ATM, HESTON and BL-MALZ are rejected in the KS test. For these schemes, the maximum deviation between the empirical and theoretical dis-

tributions occurs at CDF points between 0.34 and 0.39, and for distances ranging from 9.37% to 10.35%. Therefore, contrary to most historical methods, these RNDs allocate too much probability to significant losses compared to the actual realizations. Moreover, the BL-MALZ fails the Berkowitz test (due to its biased 0.22 T-PIT distribution mean), whereas the Heston model is rejected in the JB test (due to negative skewness and a lack of kurtosis in the observed T-PIT values).

Descriptive statistics of the transformed PIT sequences

TABLE 3

Model	Mean	5th percentil	Median	95th percentil	Std	Skewness	Kurtosis	AR(1)
Historical 6-months								
LN-HIS(6m)	0,002	-1,972	0,011	1,779	1,100	-0,302	2,396	-0,011
BTS(6m)	-0,044	-1,975	-0,034	1,695	1,106	-0,288	2,231	-0,012
GARCH-N(6m)	-0,119	-2,225	-0,059	1,519	1,099	-0,416	2,700	0,025
GARCH-t (6m)	-0,133	-1,708	-0,154	1,328	0,965	-0,265	2,251	0,038
GJR-FHS(6m)	-0,009	-1,957	-0,096	1,858	1,139	0,113	2,749	0,036
Historical 5-years								
LN-HIS(5y)	0,040	-1,537	0,066	1,584	0,993	-0,621	0,969	0,060
BTS(5y)	-0,001	-1,686	0,033	1,492	0,980	-0,554	1,365	0,065
GARCH-N(5y)	-0,129	-1,892	-0,131	1,455	1,004	-0,284	2,902	0,063
GARCH-t (5y)	-0,102	-1,399	-0,108	1,226	0,818	-0,290	2,883	0,055
GJR-FHS(5y)	0,026	-1,621	-0,060	1,790	1,052	0,182	2,900	0,094
Risk-neutral								
LN-ATM	0,134	-1,649	0,202	1,613	0,970	-0,294	2,913	0,063
HESTON	0,113	-1,346	0,148	1,582	0,933	-0,492	1,173	0,056
BATES	0,107	-1,732	0,148	1,691	1,015	-0,153	2,815	0,078
VG	0,116	-1,455	0,065	1,806	1,020	0,139	3,059	0,064
BL-MALZ	0,220	-1,303	0,203	1,815	0,948	0,115	2,769	0,074

6.3 Local accuracy: Log-likelihood score

Table 4 presents the log-likelihood comparisons for all density schemes. Following Shackleton, Taylor, and Yu (2010) we define the benchmark log-likelihood as the value for the simplest historical method, namely the LN-HIS(6m). For all other models, Table 4 shows the log-likelihood values in excess of the benchmark level. The sample is divided in two subperiods. The first comprises observations dates from November 1995 to December 2006, whereas the second covers the most recent 10 years (January 2007 to December 2016).

6.3.1 Historical methods

The GJR-FHS(5y) and GARCH-N(5y) deliver the highest log-likelihoods among the historical methods (+28.74 and +26.29 over the benchmark level). Hence, these methods exhibit the greatest accuracy in predicting the *ex-post* realizations, also generating the densities which are closer to the true densities according to the KLIC.

Table 4 also reveals two notable findings. First, in all historical schemes, the use of 5-year calibration periods results in higher log-likelihoods than employing 6-month windows. Second, for any given calibration length, the log-likelihoods of conditional volatility models are higher than those of lognormal densities or bootstrapped returns, showing that the use of ARCH models pays off in terms of accuracy.

Most of these results are also robust across the subperiods; in all subsamples: (i) the GJR-FHS(5y) achieve the maximum log-likelihood among historical methods and (ii) all models calibrated to a 5-year history improve the log-likelihoods of 6-month-based counterparts. In contrast, the period from 2007 to 2016 is challenging for the two GARCH-t variants. While both perform well when significant upside or downside movements are observed, the GARCH-t is a poor predictor of the bulk of *ex-post* realizations where more modest movements occur. Out of the 59 monthly dates in 2007-2016 where the IBEX 35 returns remain within a $\pm 4\%$ range, in 31 of them either the GARCH-t(5y) or the GARCH-t(6m) assign the smallest probability to the actual realizations. Conversely, in none of such dates these variants are the best performers.

Out-of-sample log-likelihoods

TABLE 4

Model	Observation period		Entire Sample
	Nov 1995 - Dec 2006	Jan 2007 - Dec 2016	
Historical 6-months			
LN-HIS(6m)	-971,46	-872,70	-1844,16
BTS(6m)	4,54	-2,20	2,34
GARCH-N(6m)	12,41	0,28	12,68
GARCH-t (6m)	10,23	-2,59	7,63
GJR-FHS(6m)	10,10	-0,37	9,73
Historical 5-years			
LN-HIS(5y)	4,21	3,08	7,29
BTS(5y)	6,30	5,82	12,12
GARCH-N(5y)	17,82	8,47	26,29
GARCH-t (5y)	18,39	0,99	19,37
GJR-FHS(5y)	19,45	9,30	28,74
Risk-neutral			
LN-ATM	15,86	5,62	21,48
HESTON	19,87	6,88	26,75
BATES	14,62	9,41	24,03
VG	19,65	11,63	31,28
BL-MALZ	16,20	11,49	27,69

Log-likelihoods are computed as the value in excess of the LN-HIS(6m) benchmark.

6.3.2 Risk-neutral methods

Table 4 shows that RND schemes typically exhibit higher log-likelihoods than historical methods. The simplest risk-neutral model, the LN-ATM, achieves a log-

likelihood score that is higher than in 8 out of 10 historical methods. Furthermore, the VG model delivers the highest log-likelihood of all schemes (+31.28 points over the LN-HIS(6m)), whereas the Bates model also achieves a good likelihood score (+24.03).

Despite failing in the goodness-of-fit tests, HESTON and ML-MALZ also deliver notably good log-likelihoods. These contradictory results highlight the importance of multi-factor verifications in probabilistic forecasts, exposing how partial evaluations may fail to detect seriously misspecified models. For instance, although the BL-MALZ ranks particularly high in the log-likelihood score, we previously concluded that this method was inconsistent in statistical terms, thus being unable to produce reliable forecasts across the entire density space.

The analysis by subperiods confirms these results and reveals an additional finding. While statistically consistent methods tend to show a similar log-likelihood performance across different subperiods, non-consistent methods (i.e.: GARCH-t(5Y) or HESTON) exhibit a higher variability. This finding suggests that the statistical consistency could be associated with a higher stability of the log-likelihoods in different subperiods.

6.4 Forecasting errors: Continuous Ranked Probability Score

Table 5 summarizes the prediction errors for all forecasting methods. CRPS values are calculated as the average return-based CRPS for the 21-year sample and the two subperiods, and expressed as the value in excess over the benchmark LN-HIS(6m) level.

RNDs generally exhibit lower forecasting errors than historical-based counterparts. Of all 15 density schemes, the VG model achieves the best CRPS value (-0.286% compared to the benchmark level) followed by the Bates model (-0.274%). Furthermore, the RND scheme exhibiting the worst CRPS value, the BL-MALZ, outperforms 8 out of 10 historical methods.

The GARCH-N(5y) and GJR-FHS(5y) produce the lowest CRPS among the historical methods (-0.260% and -0.257%, respectively). Furthermore, Table 5 also confirms the patterns observed in the log-likelihood analyses. First, in all historical schemes, the use of 5-year calibration periods results in significantly lower prediction errors. Second, density forecasts generated through ARCH-based models typically exhibit lower CRPS values than either lognormal distributions or a bootstrapping of historical returns.

In contrast, the GARCH-t(5y) again provides an exception. In all 2007-16 monthly dates where the IBEX 35 futures move up by more than 10%, the GARCH-t(5y) outperforms other historical methods. However, out of the 42 dates where the IBEX 35 exhibit modest losses or small gains (i.e.: -5% to +1%), in 81% of them the GARCH-t(5y) produces the highest forecasting errors, making this method the worst-performing among the 5-year historical schemes.

Model	Observation period		
	Nov 1995 - Dec 2006	Jan 2007 - Dec 2016	Entire Sample
Historical 6-months			
LN-HIS(6m)	3,672	3,801	3,737
BTS(6m)	0,029	0,004	0,012
GARCH-N(6m)	-0,051	-0,005	-0,034
GARCH-t (6m)	-0,090	0,065	-0,021
GJR-FHS(6m)	-0,057	-0,002	-0,036
Historical 5-years			
LN-HIS(5y)	-0,210	-0,209	-0,214
BTS(5y)	-0,187	-0,200	-0,198
GARCH-N(5y)	-0,313	-0,192	-0,260
GARCH-t (5y)	-0,268	-0,079	-0,184
GJR-FHS(5y)	-0,281	-0,220	-0,257
Risk-neutral			
LN-ATM	-0,272	-0,235	-0,259
HESTON	-0,282	-0,184	-0,241
BATES	-0,293	-0,242	-0,274
VG	-0,309	-0,250	-0,286
BL-MALZ	-0,210	-0,215	-0,217

CRPS figures are computed as the value in excess of the LN-HIS(6m) benchmark and expressed in percentage.

6.5 A holistic evaluation: The Integrated Forecast Score (IFS)

To provide a comprehensive ranking of the competing density schemes, we develop a novel scoring system that jointly considers the results from the statistical consistency, local accuracy and forecasting error analyses. Following Gneiting and Katzfuss (2014), the IFS assign better scores to predictive methods that deliver a high forecasting accuracy subject to statistical consistency. In order to consistently aggregate all the statistical scores, we first normalize the results obtained in the three previous criteria into standardized $[0, 1]$ scales.

The normalized score for statistical consistency simultaneously considers both the number of rejections and the specific p -values achieved in the three goodness-of-fit tests. First, a 0.25 score is allocated to each variant for each non-rejected test. Next, the remainder 0.25 is assigned by averaging the position of each model p -values measured in an empirical $[0, 1]$ scale. Specifically, the best and worst p -values in each test are assigned 1 and 0 values, whereas other variants are allocated in the $[0, 1]$ scale through linear interpolation. This scoring rule effectively ranks all forecasting methods in terms of the number of tests passed while further discriminating among competing densities by their specific scores in the Berkowitz, JB and KS tests.

To calculate the normalized local accuracy and forecasting errors scores, we consider that the overall log-likelihood and CRPS values for the 21-year sample are normally

distributed. This can be justified by the central limit theorem: since the overall log-likelihood and CRPS values are computed by a sum of 254 independent and similarly distributed observations, they should converge approximately to a normal distribution. Specifically, we first obtain the mean and variance of the observed overall values and then each method is ranked in the $[0, 1]$ scale according to their quantile position in the assumed distribution. The IFS is finally obtained by averaging, for each method, the three standardized scores.

Table 6 shows the normalized scores for all density schemes. The first aspect to highlight is that RNDs deliver better IFS than historical models. Although we could argue that the performance of the risk-neutral schemes stands out due to the low IFS values of the 6-month historical variants, a comparison of RNDs and densities calibrated to 5-year periods confirms the outperformance of RNDs, with average IFS of 0.72 versus 0.64 in the 5-year variants. Of all forecasting methods, the VG model achieves the highest IFS (0.880), being the top-rated in both local accuracy and forecasting errors, whereas the GJR-FHS(5y) ranks second in the IFS (0.864), and it is the best performer in statistical consistency.

Comprehensive ranking of forecasting schemes

TABLE 6

Model	IFS	Normalized Scores			Calibration
		Statistical Consistency	Local Accuracy	Forecasting Errors	
Consistent schemes					
VG	0,880	0.867 [3]	0.914 [1]	0.859 [1]	Risk-neutral
GJR-FHS(5y)	0,864	0.929 [1]	0.868 [2]	0.793 [5]	5-year history
BATES	0,817	0.871 [2]	0.747 [6]	0.833 [2]	Risk-neutral
GARCH-N(5y)	0,812	0.823 [4]	0.811 [5]	0.802 [3]	5-year history
Non-consistent schemes					
LN-ATM	0,665	0.534 [11]	0.662 [7]	0.800 [4]	Risk-neutral
BL-MALZ	0,619	0.334 [13]	0.846 [3]	0.679 [7]	Risk-neutral
HESTON	0,612	0.260 [15]	0.823 [4]	0.751 [6]	Risk-neutral
GARCH-t(5y)	0,558	0.521 [12]	0.585 [8]	0.567 [10]	5-year history
BTS(5y)	0,511	0.605 [6]	0.312 [10]	0.615 [9]	5-year history
LN-HIS(5y)	0,476	0.588 [8]	0.170 [13]	0.670 [8]	5-year history
GJR-FHS(6m)	0,341	0.660 [5]	0.234 [11]	0.127 [11]	6-month history
GARCH-t(6m)	0,280	0.561 [10]	0.178 [12]	0.103 [13]	6-month history
GARCH-N(6m)	0,249	0.291 [14]	0.332 [9]	0.123 [12]	6-month history
BTS(6m)	0,242	0.592 [7]	0.075 [14]	0.059 [15]	6-month history
LN-HIS(6m)	0,232	0.574 [9]	0.048 [15]	0.072 [14]	6-month history

The IFS ranking also validate our previous findings regarding historical methods. First, in all predictive models the use of 5-year calibration periods results in higher IFS than employing 6-month windows. Second, for any given calibration period, all ARCH-based models achieve better IFS values than either lognormal densities or a bootstrapping of historical returns.

Given the distinct focus of the log-likelihood score and the CRPS (i.e.: local accuracy vs. global errors), it is remarkable that most density schemes simultaneously achieve either a good or a bad ranking in both. The rationale stems from the relation between the probability assigned to a given observation and its distance to other probability masses in mound-shaped distributions¹⁸.

However, this is not always the case. For instance, while the addition of jumps significantly improves the IFS value of the Bates model compared to Heston dynamics¹⁹, its impact in the CRPS and log-likelihood scores goes in opposite directions. The difference can be attributed to the excessively fat-tailed *ex-ante* distributions generated by the Heston model, which assign too much weight to extreme events, thus performing well in the most likelihood-sensitive observation dates, but it conversely results in higher forecasting errors when the entire *ex-ante* distribution is considered in the computations.

18 In such distributions, realizations falling near the tails (center) of the distributions tend to exhibit (i) a low (high) log-likelihood score and (ii) a high (low) distance to the other probability masses, thus leading to either good or bad ranking in both CRPS and log-likelihood scores.

19 The IFS gains are driven by the improvement in statistical consistency: through the added jumps, the Bates model simultaneously passes all goodness-of-fit tests, whereas the Heston model fails the JB and KS tests.

7 Conclusions

This paper presents a comprehensive analysis of the most commonly used density schemes in financial economics. Through the development of a novel Integrated Forecasting Score (IFS), we show that risk-neutral densities outperform historical-based predictions in terms of information content. The IFS is constructed by aggregating the statistical consistency, local accuracy and forecasting errors results into a single normalized measure.

Using an option dataset covering from 1995 to 2016, we find that the Variance Gamma model simultaneously delivers the largest out-of-sample log-likelihood and the lowest forecasting errors, thus ranking first in the IFS. In contrast, the ARCH-based GJR-FHS achieves the best score in statistical consistency, generating the most reliable forecasts across the entire density range.

We also find two strong patterns regarding historical models. First, in all density schemes the use of 5-year calibration periods outperforms the forecasting ability of 6-month calibration windows. Second, densities obtained from ARCH-type models are more informative than those generated with lognormal methods or a bootstrapping of historical returns. Conversely, frequently used benchmarks like the Heston model or the non-parametric Breeden-Litzenberger formula yield biased predictions and are rejected in statistical tests.

Looking forward, optimally mixing the information content of risk-neutral and historical schemes, and exploring the use of machine learning algorithms to calibrate such models is worthy of research. Moreover, while the IFS provides a simple solution to a complex verification problem, applying the IFS in other datasets or testing its performance in real trading strategies could help to validate the usefulness of this measure as a new tool in financial forecasting. These items remain in our agenda for future research.

References

- Alonso, Francisco, Roberto Blanco, and Gonzalo Rubio, 2005, Testing the Forecasting Performance of Ibox 35 Option-Implied Risk-Neutral Densities, *Bank of Spain Working Papers* 0504.
- Amisano, G, and R Giacomini, 2007, Comparing density forecasts via weighted likelihood ratio tests, *Journal of Business & Economic Statistics* 25, 177–190.
- Anagnou, I, M. Bedendo, Stewart Hodges, and Robert G. Tompkins, 2005, Forecasting accuracy of implied and GARCH-based probability density functions, *Review of Futures Markets* 11, 41–66.
- Anagnou, Iliana, Mascia Bedendo, Stewart Hodges, and Robert Tompkins, 2003, The Relation Between Implied and Realised Probability Density Functions, *Financial Options Research Centre, University of Warwick*.
- Bakshi, Gurdip, Charles Cao, and Zhiwu Chen, 1997, Empirical Performance of Alternative Option Pricing Models, *The Journal of Finance* 52, 2003–2049.
- Bakshi, Gurdip, and Dilip B. Madan, 2000, Spanning and derivative-security valuation, *Journal of Financial Economics* 55, 205–238.
- Bao, Yong, Tae Hwy Lee, and Burak Saltoğlu, 2007, Comparing density forecast models, *Journal of Forecasting* 26, 203–225.
- Barone-Adesi, Giovanni, Robert F. Engle, and Lorian Mancini, 2008, A GARCH option pricing model with filtered historical simulation, *Review of Financial Studies* 21, 1223–1258.
- Bates, David S., 1996, Jumps and Stochastic Volatility: Exchange Rate Processes Implicit in Deutsche Mark Options, *Review of Financial Studies* 9, 69–107.
- Berkowitz, Jeremy, 2001, Testing Density Forecasts, With Applications to Risk Management, *Journal of Business & Economic Statistics* 19, 465–474.
- Blair, Bevan J., Ser Huang Poon, and Stephen J. Taylor, 2001, Forecasting S&P 100 volatility: The incremental information content of implied volatilities and high-frequency index returns, *Journal of Econometrics* 105, 5–26.
- Bollerslev, Tim, and Jeffrey M. Wooldridge, 1992, Quasi-maximum likelihood estimation and inference in dynamic models with time-varying covariances, *Econometric Reviews* 11, 143–172.

- Breedon, Douglas T., and Robert H. Litzenberger, 1978, Prices of State-Contingent Claims Implicit in Option Prices, *The Journal of Business* 51, 621–651.
- Busch, Thomas, Bent Jesper Christensen, and M. Ø Nielsen, 2011, The role of implied volatility in forecasting future realized volatility and jumps in foreign exchange, stock, and bond markets, *Journal of Econometrics* 160, 48–57.
- Canina, Linda, and Stephen Figlewski, 1993, The informational content of implied volatility, *The Review of Financial Studies* 6, 659–681.
- Christoffersen, Peter, Kris Jacobs, and Bo-Young Chang, 2013, Forecasting with option-implied information, *Handbook of Economic Forecasting* 2, 581–656.
- Christoffersen, Peter., 2012, *Elements of Financial Risk Management, Second Edition (Elsevier)*.
- Cont, Rama, and Peter Tankov, 2004, *Financial Modelling with Jump Processes (Chapman & Hall/CRC)*.
- Crisóstomo, Ricardo, 2014, An Analysis of the Heston Stochastic Volatility Model: Implementation and Calibration using Matlab, *CNMV Working papers No. 58*.
- Crisóstomo, Ricardo, 2017, Speed and biases of Fourier-based pricing choices: a numerical analysis, *International Journal of Computer Mathematics*, 1–18.
- de Vincent-Humphreys, Rupert, and Joseph Noss, 2012, Estimating probability distributions of future asset prices: empirical transformations from option-implied risk-neutral to real-world density functions, *Bank of England Working Papers* 455.
- Diebold, Francis X., Todd Gunther, and Anthony Tay, 1998, Evaluating density forecasts with applications to financial risk management, *International Economic Review* 39, 863–883.
- Dowd, Kevin, 2004, A modified Berkowitz back-test, *Risk magazine* 17, 86–87.
- Giot, Pierre, and Sébastien Laurent, 2007, Information content of implied volatility in light of the jump / continuous decomposition of realized volatility, *The Journal of Futures Markets* 27, 337–359.
- Glosten, Lawrence R., Ravi Jagannathan, and David E. Runkle, 1993, On the Relation between the Expected Value and the Volatility of the Nominal Excess Return on Stocks, *Journal of Finance* 48, 1779–1801.
- Gneiting, Tilmann, and Matthias Katzfuss, 2014, Probabilistic Forecasting, *Annual Review of Statistics and Its Application* 1, 125–151.
- Gneiting, Tilmann, and Adrian E Raftery, 2007, Strictly Proper Scoring Rules, Prediction, and Estimation, *Journal of the American Statistical Association* 102, 359–378.
- Hansen, Peter R., and Asger Lunde, 2005, A forecast comparison of volatility models: Does anything beat a GARCH(1,1)?, *Journal of Applied Econometrics* 20, 873–889.

- Hersbach, Hans, 2000, Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems, *Weather and Forecasting* 15, 559–570.
- Heston, Steven, 1993, A Closed-Form Solution for Options with Stochastic Volatility with Applications to Bond and Currency Options, *The Review of Financial Studies* 6, 327–343.
- Høg, Esben, and Leonidas Tsiaras, 2011, Density forecasts of crude-oil prices using option-implied and ARCH-type models, *Journal of Futures Markets* 31, 727–754.
- Itkin, Andrey, 2010, Pricing options with VG model using FFT, *arXiv working paper*.
- Ivanova, Vesela, and Josep Maria Puigvert Gutiérrez, 2014, Interest rate forecasts, state price densities and risk premium from Euribor options, *Journal of Banking and Finance* 48, 210–223.
- Jiang, George J., and Yisong S. Tian, 2005, The model-free implied volatility and its information content, *Review of Financial Studies* 18, 1305–1342.
- Jones, Christopher S., 2003, The dynamics of stochastic volatility: Evidence from underlying and options markets, *Journal of Econometrics* 116, 181–224.
- Kullback, S., and R. A. Leibler, 1951, On Information and Sufficiency, *The Annals of Mathematical Statistics* 22, 79–86.
- Liu, Xiaoquan, Mark B. Shackleton, Stephen J. Taylor, and Xinzhong Xu, 2007, Closed-form transformations from risk-neutral to real-world distributions, *Journal of Banking and Finance* 31, 1501–1520.
- Madan, Dilip B., Peter P. Carr, and Eric C. Chang, 1998, The Variance Gamma Process and Option Pricing, *Review of Finance* 2, 79–105.
- Malz, Allan M, 2014, A Simple and Reliable Way to Compute Option-Based Risk-Neutral Distributions, *Federal Reserve Bank of New York Staff Reports* 677.
- Matheson, J. E., and R. L. Winkler, 1976, Scoring Rules for Continuous Probability Distributions, *Management Science* 22, 1087–1096.
- Melick, William R, and Charles P Thomas, 1997, Recovering an Asset's Implied PDF from Option Prices : An Application to Crude Oil during the Gulf Crisis, *The Journal of Financial and Quantitative Analysis* 32, 91–115.
- Mitchell, James, and Stephen G. Hall, 2005, Evaluating, comparing and combining density forecasts using the KLIC with an application to the bank of england and NIESR fan charts of inflation, *Oxford Bulletin of Economics and Statistics* 67, 995–1033.
- Poon, S H, and C Granger, 2003, Forecasting financial market volatility: A review, *Journal of Economic Literature* 41, 478–539.

Shackleton, Mark B., Stephen J. Taylor, and Peng Yu, 2010, A multi-horizon comparison of density forecasts for the S&P 500 using index returns and option prices, *Journal of Banking and Finance* 34, 2678–2693.

Silva, E. M. d. S., and K. H. Kahl, 1993, Reliability of soybean and corn option-based probability assessments, *Journal of Futures Markets* 13, 765–779.

Taylor, Stephen J., Pradeep K. Yadav, and Yuanyuan Zhang, 2010, The information content of implied volatilities and model-free volatility expectations: Evidence from options written on individual stocks, *Journal of Banking and Finance* 34, 871–881.

Yun, Jaeho, 2014, Out-of-sample density forecasts with affine jump diffusion models, *Journal of Banking and Finance* 47, 74–87.

