Combining Density Forecast Accuracy Tests: An Application to Agricultural, Energy and Metal Commodities

Bernardina Algieri[a,b,*], Arturo Leccadito[a,c], Danilo Sicoli[d], Diana Tunaru[e]

[a] *Department of Economics, Statistics and Finance, University of Calabria, Ponte Bucci, 87030 Rende (CS), Italy*

[b] *Department of Economic and Technological Change, Zentrum für Entwicklungsforschung (ZEF), Universität Bonn, Genscherallee 3, 53113 Bonn, Germany*

[c] *LFIN/LIDAM, Université catholique de Louvain, Voie du Roman Pays 34, B-1348 Louvain-la-Neuve, Belgium*

[d] *Independent Researcher, Via Tommasone 54, 26025 Pandino, (CR), Italy*

[e] *University of Kent, Kent Business School, CT2 7FS Canterbury, UK*

Abstract

This study develops a new methodology for combining density forecast accuracy tests and assessing the relevance of psychological indicators in predicting commodity returns. Density forecasts provide a complete description of the uncertainty associated with a prediction and are highly requested by policy makers, central bankers and financial operators to define policy actions, manage financial risks and assess portfolio selection. The proposed methodology combines different tests and derives the p-value of the resulting test statistic by Monte Carlo simulations. To assess the power of the proposed methodology, we implement a set of experiments for several data-generating processes. Based on an empirical forecasting exercise applied to agricultural, energy and metal commodities, we find that sentiment variables and psychological factors improve the density forecasts of commodity futures returns, especially for agricultural commodities. Additionally, combinations of sentiment variables are more powerful in predicting returns than considering them separately.

Keywords: Density forecasts, commodity futures, behavioural finance, ARMAX -EGARCH-t model

JEL: C15, C53, G13, G17

*Corresponding author: b.algieri@unical.it. Co-author emails: arturo.leccadito@unical.it, danilosicoli@gmail.com, d.tunaru@kent.ac.uk.

Preprint submitted to Journal of the Royal Statistical Society. Series C: Applied Statistics

1. Introduction

How can an investor decide whether a set of explanatory variables or predictors do have explanatory power for forecasting asset returns? What influence do behavioural finance indicators have on the uncertainty of those forecasts? In this study, we aim to provide a possible answer to these important questions by focusing on the estimation and evaluation of density forecasts for commodity futures returns and by assessing the effect of sentiment factors on these predictions.

A density forecast of the realization of a random variable at a certain future time is an estimate of the probability distribution of the possible future values of that variable (Tay and Wallis, 2000). Compared to point and interval forecasts, density forecasts provide a complete picture of the uncertainty of a prediction and offer an evaluation of the quality of the forecast itself. Indeed, by enabling the estimation of the entire probability distribution of a target variable, density forecasts deliver a rich set of information regarding the possible future dynamics of the variable (Berkowitz, 2001; Niu and Harvey, 2022). Assessing and improving the out-of-sample performance of competing density forecasts are research topics with a continuous focus, given their importance in macroeconomics, finance and real-world applications. For instance, it is of great interest to policy makers and central banks to produce accurate density forecasts of GDP, exchange rates or inflation (see Clements, 2004; Mitchell and Hall, 2005; Gaglianone and Marins, 2017; Diebold et al., 2022). Investors and portfolio managers are keen to understand and predict the future behaviour of the distribution of asset prices/returns. Hua and Zhang (2008) suggested how to enhance density forecasts by capturing asymmetric features of the S&P500 returns. Also in other areas, Alexandridis et al. (2019) presented an application of density forecasts to the real estate market, while Alexandridis et al. (2021) showed their usefulness in pricing weather derivatives.

A field in which forecast techniques have attracted attention is commodity markets. Commodities play a crucial role both in the real economy, impacting macroeconomic variables such as inflation and unemployment (see Gokmenoglu and Fazlollahi, 2015), and in financial markets (see Adams and Glück, 2015; Ding et al., 2021). Therefore, accuracy in forecasting commodity price returns and volatility is extremely important to industrial producers and consumers on one side, and investors and traders on the other side. Possible sources for this predictive information are the spot (Plakandaras et al., 2019; Dichtl, 2020; Hollstein et al., 2021) and derivatives markets, such as commodity options (Hog and Tsiaras, 2011) and commodity futures markets (Luo and Chen, 2019; Algieri et al., 2021).

A relatively new academic debate has emerged around the impact of investors' sen-

timent on financial returns, with already substantial empirical evidence of such an effect on commodity markets (Wang, 2001; Wang et al., 2006; Bahloul and Bouri, 2016; Chen et al., 2021; Liu et al., 2023). The last two decades have also seen a growing strand of literature on combining forecasts across different models in order to obtain more accurate predictions. Many of the techniques and issues related to the theme of combining forecasts have been reviewed in Timmermann (2006), Wallis (2011) and Wang et al. (2022). Various methods of combining forecasts try to circumvent mis-specification biases and measurement errors in the data set, changes in parameters over time and the well-known fact that the data-generating process is unknown. Other substantial contributions to this topic include de Menezes et al. (2000), Amisano and Giacomini (2007), Hall and Mitchell (2007), Geweke and Amisano (2012), Billio et al. (2013), Hsiao and Wan (2014), Kapetanios et al. (2015), McAdam and Warne (2020), Koutsandreas et al. (2021) and Taylor and Taylor (2023).

Nevertheless, the vast majority of the previous studies combine different forecasts or models, followed by a validation procedure that would select the superior forecast combination. The decision is based on either a single measure of forecasting efficiency or several measures *in parallel*, and/or on using a single statistical test or several tests also in parallel.

In this study, we shift the focus from combining density forecasts to *combining tests* for evaluating the accuracy of density forecasts and assessing their statistical consistency. Combining tests allows researchers to overcome the drawback of conducting tests in parallel and reduces uncertainty when results across individual tests differ. We further investigate whether adding various behavioural factors helps to improve predictions of commodity futures returns. In doing so, we are able to contribute to the extant literature in three ways. First, we propose a new methodology to gauge the accuracy (goodness-of-fit) of density forecasts. Rather than using several tests in parallel, we pool a set of statistical tests in order to select the model with the best density predictions. To this end, we consider several previously known tests, including Berkowitz, Hong and Li, and Cramer-von Mises, and present a new combined test. Since a combination of tests does not always follow a standard distribution, it is difficult to calculate a p-value associated to that particular combination. We employ the Monte Carlo technique developed by Dufour (2006) to obtain the p-value of the combined test. Secondly, we found that our approach has a great ability to distinguish across very different data-generating processes. More specifically, to evaluate the performance of the proposed combination of tests, we conduct Monte Carlo simulations using different model specifications and sets of parameters for those models.

Finally, we provide important empirical evidence by applying our technique to futures contracts written on 15 commodities (7 agricultural, 4 energy and 4 metals), framed with the ARMAX-EGARCH-t family of models. A set of traditional predictors (Fed Fund interest rate, S&P500, the MSCI Emerging market index and exchange rates) and behavioural variables enter the mean and variance equations, respectively. The behavioural variables comprise the Thomson Reuters MarketPsych Index (TRMI), the commodity implied volatility (IV), the economic policy uncertainty (EPU) and the Google trends data (GGLE). Depending on how many behavioural variables are included, we estimate sixteen models for each commodity and compare them in terms of density forecasts.

This study is organised as follows. In Section 2, we present the methodology of combining tests for density forecast accuracy. Section 3 illustrates four Monte Carlo designs where we determine the rate of rejection of the null hypothesis for the standard and the proposed combined tests. The main empirical results are presented in Section 4. The last section summarizes our conclusions.

2. Combining Density Forecast Accuracy Tests

Let $F_t(\cdot)$ be the predictive cumulative distribution function (cdf) at time t from a given model using the information available up to time $t-1$. In line with Diebold et al. (1998), we consider the series of the probability integral transforms (PIT) $U_t = F_t(y_t)$, also called the generalised residuals, where y_t is the ex-post realized value of the target asset return to be predicted at time t. PIT is a tool to test whether the empirical predictive distribution of empirical models matches the true, unobserved distribution that generates the data. Under the null hypothesis that the model represented by F_t correctly forecasts the density, the PITs $\{U_t\}_{t\geq 0}$ are standard uniform and independent random variables (Rosenblatt, 1952). If at least one of these conditions is not satisfied, then we reject the density forecasting ability of the model. Testing the null of independence is important because if the PITs are predictable, then this predictability can be exploited to improve the model used to forecast the density and, hence, the model is not correctly specified (Christoffersen, 2012, chap. 13).

Tests of density forecast accuracy

This section presents several statistical tests previously proposed in the literature to assess the accuracy of density forecasts via dynamic PITs. These tests will be later considered to construct a new combined test in our study.

Berkowitz (2001) proposed using the transformed PIT $z_t = \Phi^{-1}(U_t)$ and testing whether the z_t is independent and standard normal, with Φ^{-1} being the standard normal quantile function. Under the alternative hypothesis, z_t follows an AR(1) model with mean μ:

$$z_t - \mu = \rho(z_{t-1} - \mu) + \epsilon_t, \qquad \epsilon_t \sim i.i.d.\ \ N(0, \sigma^2). \tag{1}$$

Therefore, under the null hypothesis of $z_t \sim i.i.d.$ and $N(0,1)$, the parameters of the above model are constrained to $\mu = 0, \rho = 0$ and $\sigma^2 = 1$. In other words, for a correctly specified density forecast model, the PIT follows an i.i.d. uniform distribution on the interval $(0,1)$, while its inverse normal transform follows an i.i.d. normal distribution. Denoting by $L(\mu, \sigma^2, \rho)$ the log-likelihood function of the model, the test statistic is given by the following Likelihood Ratio (LR):

$$LR_3 = -2[L(0,1,0) - L(\hat{\mu}, \hat{\sigma}^2, \hat{\rho})] \tag{2}$$

where $\hat{\mu}$, $\hat{\sigma}^2$, and $\hat{\rho}$ are the maximum-likelihood estimates from (1). Under the null, the test statistic has a chi-squared distribution with three degrees of freedom. The model given by (1) can be extended to include k lags:

$$z_t - \mu = \sum_{j=1}^{k} \rho_j (z_{t-j} - \mu) + \epsilon_t, \qquad \epsilon_t \sim i.i.d.\ \ N(0, \sigma^2). \tag{3}$$

The extended test statistic, denoted by LR_{2+k}, is computed similarly to (2) and its null distribution is chi-squared with $2 + k$ degrees of freedom.

As an alternative to the test based on (3), we consider the Ljung–Box (LB) test:

$$LB(k) = T\,(T+2) \sum_{j=1}^{k} \frac{\hat{\rho}_j^2}{T - j} \tag{4}$$

where T is the number of available PITs and $\hat{\rho}_j$ is the sample autocorrelation at lag j calculated on the time series of inverse transformed PITs $z_t = \Phi^{-1}(U_t)$.

Another useful test for density forecast accuracy has been developed by Hong and Li (2004) (HL hereafter), based on a kernel estimator of the joint density of U_t and U_{t-j}:

$$\hat{g}_j(a,b) = \sum_{t=j+1}^{T} K_h(a, U_t) K_h(b, U_{t-j}) \tag{5}$$

where $K_h(\cdot,\cdot)$ is a boundary-modified kernel and h is a bandwidth parameter. The

test statistic for the HL test, $Q(j)$, is the centered and scaled version of $M(j) = \int_0^1 \int_0^1 [\hat{g}_j(a,b) - 1]^2 \, \mathrm{d}a \, \mathrm{d}b$. Here, $M(j)$ measures the distance between the kernel estimator given by eq. (5) and the scalar 1, where 1 represents the resulting product of two standard uniform densities. Under the null, the test statistic $Q(j)$ follows a standard normal distribution for any fixed lag order $j > 0$. For later reference, we define $Q^*(j) = Q(j)^2$, which under the null has a chi-square distribution with one degree of freedom.

A simple and direct procedure that can be applied to test the null of i.i.d. $U(0,1)$ is the Cramer-von Mises test (CVM hereafter). The CVM test is based on the distance between the empirical cdf of (U_t, U_{t-j}) and the joint cdf of two independent $U(0,1)$ random variables. Computationally, this is given by:

$$\mathrm{CVM}(j) = \int_0^1 \int_0^1 (T-j)[\hat{F}_j(a,b) - a \times b]^2 \mathrm{d}\hat{F}_j(a,b) \tag{6}$$

where $\hat{F}_j(a,b) = \sum_{t=j+1}^{T} I(U_t \leq a, U_{t-j} \leq b)/(T-j)$ is the empirical cdf.

In a nutshell, we can assess a density forecast model by examining the departure of the PIT and transformed PIT from the fundamental property (independence and normality). The departure can be quantified by the CVM test, which measures the distance from a candidate model to the unknown true model.

Combining the Tests

Less attention has been devoted in the literature to combining the tests upon which one decides if a model produces accurate density predictions. There are different ways of pooling individual tests. A first technique consists in combining versions of the same test but across various lags (see Appendix A).

The second technique combines different tests.[1] In this study, we consider the combination of all the above tests and a Jarque-Bera test (JB) performed on z_t by adding them together:

$$C(k) = LR_{2+k} + JB + LB(k) + \widetilde{\mathrm{CVM}}(k) + \widetilde{Q^*}(k), \tag{7}$$

where the last two tests, $\widetilde{\mathrm{CVM}}(k)$ and $\widetilde{Q^*}(k)$, are presented in eqs.(A.1)-(A.2) (Appendix A).

[1] Dowd (2004) proposed pooling the Berkowitz (2001) test procedure with a test for normality such as the Jarque-Bera test. However, Dowd (2004) provides no explanation of how to establish the significance of the combined test.

The motivation for combining tests is that in the Monte Carlo study (Section 3) we observe that the power of the tests can change dramatically from one scenario to another, indicating a very high sensitivity of the individual tests to the data-generating process. Moreover, under the same design, for some particular parameter values, the power of a test can be very high, while for other values the rejection frequency can dramatically decrease. When all the individual tests seem to "agree", the decision on the accuracy of density forecasts provided by a given model is straightforward. However, there are numerous cases when the tests "disagree", and the researcher is faced with mixed results. Indeed, when applying the tests on real data, the researcher does not know ex-ante if a single test will prove powerful. By combining multiple individual tests, this problem of inconclusiveness could be circumvented as the decision on the accuracy of the density forecasts from a given model is based on a single test that takes into account all the individual tests. In addition, we show that this all-in-one test attains a satisfying level of power in all the considered designs.

The proposed combination test (C) has a non-standard probability distribution under the null and therefore it is not straightforward to calculate the p-values. To overcome this intractability issue, one can apply the Monte Carlo (MC) test technique of Dufour (2006) for computation of the p-values, since random draws can be easily generated by simulating from the null distribution. Since the joint distribution of the PITs is nuisance-parameter free under a correctly specified model and can be simulated with no further assumptions, the technique of Dufour (2006) can be applied for any statistic that depends on the data only through the PITs. Importantly, using the MC test technique produces exact tests, implying that size is controlled also in small samples.

To obtain a p-value for the above test statistic, we apply the following algorithm:

Step 1: Under the null, generate $M = 50,000$ time series of i.i.d. $U(0,1)$ random variables, each series of length T.

Step 2: For each $j = 1, \ldots, M$, calculate the test statistic and denote it by S_j.

Step 3: Compute the p-value as

$$\tilde{p}_M(S_0) = \frac{1}{M} \sum_{j=1}^{M} I(S_j \geq S_0) \tag{8}$$

where $I(\cdot)$ is the indicator function, and S_0 is the test statistic calculated from the original sample. For a given significance level α (for instance $\alpha = 5\%$), the null of accurate density forecasts is rejected if $\tilde{p}_M(S_0) < \alpha$. In other words, models with high p-values produce accurate density forecasts.

3. Monte Carlo Study

One way to check for the robustness of the newly proposed methodology of combining tests for density forecasts is to control the data-generating process. In this section we present the results of a range of Monte Carlo designs based on known distribution of the PITs. A total of 19 tests (or combined tests) are compared. We consider tests based on 1, 5 or 10 lags, which, in the case of daily financial data, correspond to one trading day, one trading week and two trading weeks, respectively. This choice is consistent with several studies (e.g. Hurlin and Tokpavi, 2006; Berkowitz et al., 2011; Du and Escanciano, 2017) in the related literature on backtesting the tails of the distribution. We check the size of the considered tests and report in Table I the rejection frequencies at the 5% level, based on 1,000 simulations of length T, with $T \in \{250, 500, 1000, 2500\}$. The results clearly show that size is controlled for the proposed combined test.

To assess the power of the proposed test, we select four scenarios based on the guidelines for combining forecasts proposed by de Menezes et al. (2000) to improve the quality of forecasts. More specifically, the author considers three criteria: the variance of the forecasts errors, the asymmetry of the forecast error distribution measured by skewness and the level of serial correlation in the forecast errors. The PITs are drawn in turn from a beta distribution, a correlated uniform distribution, a correlated beta distribution and a GARCH-t model. We consider these distributions since they represent a departure from uniformity and serial independence. It is worth noticing that we simulate under the alternative hypothesis; thus, tests with higher rejection frequencies are more powerful. In all the following designs we consider 1,000 simulations of length T, with $T \in \{250, 500, 1000, 2500\}$.

Table II provides information on how many times a given test has the maximum rejection frequency across the different parameter scenarios that we consider under each of the four MC designs. For example, the number 7.317 in Panel B (Design II), corresponding to the LR_3 test when $T = 250$, means that the test has achieved a rejection frequency larger or equal than the remaining tests in 3 of the 41 cases under Design II (3/41=7.317%). In Appendix B, we report several descriptive statistics for the rejection frequency. From Tables B.1–B.4 (Supplementary Material), it is clear that, as expected, rejection frequencies increase with the length T of the simulated series of PITs.

Design I: Beta Distribution

We simulate the PITs from a beta distribution $U_t \sim i.i.d.\ \ beta(a, b)$, where the two parameters will take the values $a, b \in \{0.8, 0.85, 0.9, 0.95, 1, 1.05, 1.1, 1.15, 1.2, 1.5\}$.

Table I
Size

Test	$T = 250$	$T = 500$	$T = 1000$	$T = 2500$
LR_3	0.045	0.035	0.040	0.050
LR_7	0.050	0.040	0.060	0.040
LR_{12}	0.050	0.040	0.070	0.065
$LB(1)$	0.040	0.060	0.060	0.050
$LB(5)$	0.060	0.065	0.060	0.065
$LB(10)$	0.080	0.065	0.055	0.060
$Q^*(1)$	0.015	0.035	0.025	0.035
$Q^*(5)$	0.025	0.045	0.030	0.040
$Q^*(10)$	0.035	0.050	0.025	0.045
$\widetilde{Q^*}(5)$	0.020	0.040	0.035	0.040
$\widetilde{Q^*}(10)$	0.030	0.045	0.030	0.035
CVM(1)	0.055	0.040	0.030	0.060
CVM(5)	0.045	0.055	0.045	0.055
CVM(10)	0.050	0.050	0.055	0.050
$\widetilde{\text{CVM}}(5)$	0.050	0.045	0.040	0.075
$\widetilde{\text{CVM}}(10)$	0.045	0.045	0.050	0.075
$C(1)$	0.030	0.040	0.055	0.040
$C(5)$	0.035	0.045	0.035	0.050
$C(10)$	0.035	0.050	0.030	0.040

Notes: For each T we consider 1,000 independent simulations from $U(0,1)$. The table reports for each test the rejection frequencies at the 5% level.

For each of the resulting 99 pairs[2] (a, b), we report in Panel A of Table II how many times (out of 99) a given test achieves the maximum rejection frequency[3]. This means that the higher the percentage value, the better that test performs in comparison with the remaining tests.

Given the i.i.d assumption and hence the lack of any autocorrelation for this data-generating process, for this experiment, the LB tests is less powerful than the other tests and have little or no power in absolute terms (see Table B.1, Supplementary Material). The most powerful tests are LR_3 especially for large T ($T = 500$, $T = 1000$, and $T = 2500$) and the Cramer-von Mises test combined across lags for small T ($T = 250$).

Design II: 'Correlated Uniform' Distribution

We simulate the PITs using a normal AR(k) process:

[2]We discard the case $a = b = 1$ since $beta(1,1) \equiv U(0,1)$ and we are under the null.

[3]We do not report the results for the tests $\widetilde{Q^*}(1)$ and $\widetilde{\text{CVM}}(1)$ as they coincide with $Q^*(1)$ and CVM(1), respectively.

$$
\begin{aligned}
z_t &= \sum_{j=1}^{k} \beta_j z_{t-j} + \epsilon_t, \qquad \epsilon_t \sim i.i.d. N(0, \sigma^2) \\
U_t &= \Phi(z_t/\bar{\sigma}), \qquad \bar{\sigma} = \frac{\sigma}{\sqrt{1 - \sum_{j=1}^{k} \beta_j^2}}
\end{aligned}
\tag{9}
$$

Furthermore, we set $\sigma = 1$ and consider the following scenarios:

an AR(1) process with $\beta_1 \in \{0.05, 0.1, 0.15, 0.2, 0.5, 0.8, 0.9\}$

an AR(5) process with $\beta_j = 0$ for $j = 1, \ldots, 4$ and $\beta_5 \in \{0.05, 0.1, 0.15, 0.2, 0.5, 0.8, 0.9\}$

an AR(10) process with $\beta_j = 0$ for $j = 1, \ldots, 9$ and $\beta_{10} \in \{0.05, 0.1, 0.15, 0.2, 0.5, 0.8, 0.9\}$

an AR(5) process with $\beta_j = 0$ for $j = 2, \ldots, 4$, $\beta_1 \in \{0.05, 0.1\}$, and $\beta_5 \in \{0.05, 0.1\}$

an AR(10) process with $\beta_j = 0$ for $j = 2, \ldots, 9$, $\beta_1 \in \{0.05, 0.1\}$, and $\beta_{10} \in \{0.05, 0.1\}$

an AR(10) process with $\beta_j = 0$ for $j = 1, \ldots, 4$ and for $j = 6, \ldots, 9$, $\beta_5 \in \{0.05, 0.1\}$, and $\beta_{10} \in \{0.05, 0.1\}$

an AR(10) process with $\beta_j = 0$ for $j = 2, \ldots, 4$ and for $j = 6, \ldots, 9$, $\beta_1 \in \{0.05, 0.1\}$, $\beta_5 \in \{0.05, 0.1\}$, and $\beta_{10} \in \{0.05, 0.1\}$.

For each of the 41 different AR processes listed above, we give the results in Panel B of Table II. Due to the presence of serial correlation in the simulated PITs, the performance of LB at different lags improves substantially compared to Design I. In fact, LB(10) is the best performing test for high T ($T = 1000$ and $T = 2500$).

Design III: 'Correlated Beta' Distribution

Under this design, we simulate scenarios using correlated draws from the beta distribution. To this end, we simulate the PITs starting from a normal AR(k) process:

$$
\begin{aligned}
z_t &= \sum_{j=1}^{k} \rho_j z_{t-j} + \epsilon_t, \qquad \epsilon_t \sim i.i.d. \;\; N(0, \sigma^2) \\
U_t &= F^{-1}_{beta(a,b)} \left[\Phi(z_t) \right]
\end{aligned}
\tag{10}
$$

where $F^{-1}_{beta(a,b)}(\cdot)$ is the quantile function of a beta distribution.

We set $\sigma = 1$ and consider the following pairs for (a, b): $(0.95, 1)$, $(1, 0.95)$, $(1.05, 0.95)$, and $(1.05, 1.05)$. In conjunction with the above values for σ, a and b we take into account the following scenarios:

an AR(1) process with $\rho_1 = 0.1$

an AR(5) process with $\rho_j = 0$ for $j = 1, \ldots, 4$ and $\rho_5 = 0.1$

an AR(5) process with $\rho_1 = 0.1$, $\rho_j = 0$ for $j = 2, \ldots, 4$ and $\rho_5 = 0.1$

an AR(10) process with $\rho_j = 0$ for $j = 1, \ldots, 9$ and $\rho_{10} = 0.1$

an AR(10) process with $\rho_1 = 0.1$, $\rho_j = 0$ for $j = 2, \ldots, 9$ and $\rho_{10} = 0.1$

an AR(10) process with $\rho_j = 0$ for $j = 1, \ldots, 4$, $\rho_5 = 0.1$, $\rho_j = 0$ for $j = 6, \ldots, 9$, and $\rho_{10} = 0.1$

an AR(10) process with $\rho_1 = 0.1$, $\rho_j = 0$ for $j = 2, \ldots, 4$, $\rho_5 = 0.1$, $\rho_j = 0$ for $j = 6, \ldots, 9$, and $\rho_{10} = 0.1$.

For this set of simulations (28 different combinations of AR processes) the results are described in Panel C of Table II. In relative terms, the best performing tests in this design are the LB test for small T ($T < 1000$) and the Berkowitz test for large T.

Design IV: GARCH-t Model

As the fourth MC design, we consider the possibility that PITs are generated by a GARCH-t model as follows:

$$\begin{aligned} \varepsilon_t &= \sigma_t z_t, \quad z_t \sim stdt(\nu) \\ \sigma_t^2 &= \omega + \alpha \varepsilon_{t-1}^2 + \beta \sigma_{t-1}^2 \\ U_t &= \Phi(\varepsilon_t) \end{aligned} \tag{11}$$

where $stdt(\nu)$ denotes the standardized student-t distribution with ν degrees of freedom. For this experiment, we fix the α and ω parameters, and make β and ν vary. In particular, we set $\alpha = 0.05$, $\omega = 0.001$ and consider $\beta \in \{0.7, 0.75, 0.8, 0.85, 0.9\}$ and $\nu \in \{4, 5, 6, 7, 8, 9, 10, 20, 40\}$.

In Panel D of Table II, we report the results of this simulation experiment involving 45 different combinations of the β and ν parameters. For this Monte Carlo design, the LB and the combined test appear to be the most powerful testing procedure.

Discussion

To assess the overall reliability of the individual tests across each design, we refer to the "average power" of a test (the number in the middle of each cell in Tables B.1– B.4, Supplementary Material). For Design I across all the 99 scenarios considered, we find that the test that fails to reject a false null most of the time is the LB test, while the other individual tests do rather well overall; for Design II across 41 scenarios, the LB test seems to be the best test with the highest average frequency of correct rejections of the null; for Design III, the results across the 28 scenarios are very mixed, while for Design IV across 45 scenarios, the Q^* and the CVM tests are the most powerful.

Table II
Results of Monte Carlo Designs I-IV

	Panel A: Design I				Panel B: Design II			
Test	$T = 250$	$T = 500$	$T = 1000$	$T = 2500$	$T = 250$	$T = 500$	$T = 1000$	$T = 2500$
LR_3	34.343	62.626	74.747	91.919	7.317	9.756	9.756	12.195
LR_7	10.101	25.253	42.424	75.758	14.634	14.634	17.073	26.829
LR_{12}	4.040	22.222	39.394	68.687	21.951	21.951	21.951	48.780
$LB(1)$	0.000	0.000	0.000	0.000	17.073	17.073	19.512	26.829
$LB(5)$	0.000	0.000	0.000	0.000	17.073	34.146	36.585	46.341
$LB(10)$	0.000	0.000	0.000	0.000	29.268	39.024	56.098	80.488
$Q^*(1)$	6.061	20.202	35.354	58.586	7.317	7.317	7.317	7.317
$Q^*(5)$	8.081	25.253	43.434	62.626	9.756	9.756	14.634	14.634
$Q^*(10)$	8.081	25.253	45.455	67.677	12.195	14.634	14.634	21.951
$\widetilde{Q^*}(5)$	8.081	27.273	43.434	62.626	9.756	12.195	14.634	14.634
$\widetilde{Q^*}(10)$	8.081	27.273	45.455	64.646	12.195	17.073	17.073	21.951
CVM(1)	19.192	36.364	51.515	66.667	0.000	0.000	4.878	7.317
CVM(5)	23.232	34.343	52.525	73.737	0.000	0.000	7.317	9.756
CVM(10)	21.212	33.333	52.525	68.687	0.000	0.000	4.878	12.195
$\widetilde{\text{CVM}}(5)$	31.313	49.495	52.525	73.737	2.439	4.878	9.756	14.634
$\widetilde{\text{CVM}}(10)$	49.495	51.515	58.586	74.747	0.000	7.317	14.634	21.951
$C(1)$	10.101	26.263	44.444	77.778	7.317	7.317	12.195	12.195
$C(5)$	10.101	27.273	43.434	66.667	17.073	14.634	17.073	21.951
$C(10)$	10.101	27.273	45.455	68.687	14.634	21.951	21.951	21.951

	Panel C: Design III				Panel D: Design IV			
Test	$T = 250$	$T = 500$	$T = 1000$	$T = 2500$	$T = 250$	$T = 500$	$T = 1000$	$T = 2500$
LR_3	0.000	3.571	0.000	39.286	0.000	0.000	0.000	0.000
LR_7	0.000	0.000	0.000	50.000	0.000	0.000	0.000	0.000
LR_{12}	0.000	0.000	0.000	64.286	0.000	0.000	0.000	0.000
$LB(1)$	7.143	7.143	7.143	35.714	0.000	0.000	0.000	0.000
$LB(5)$	7.143	7.143	3.571	28.571	0.000	0.000	0.000	0.000
$LB(10)$	7.143	10.714	21.429	60.714	0.000	0.000	0.000	0.000
$Q^*(1)$	0.000	0.000	0.000	0.000	0.000	0.000	11.111	22.222
$Q^*(5)$	0.000	0.000	0.000	0.000	0.000	0.000	11.111	33.333
$Q^*(10)$	0.000	0.000	0.000	0.000	0.000	0.000	11.111	44.444
$\widetilde{Q^*}(5)$	0.000	0.000	0.000	0.000	0.000	0.000	11.111	33.333
$\widetilde{Q^*}(10)$	0.000	0.000	0.000	0.000	0.000	0.000	11.111	33.333
CVM(1)	0.000	0.000	0.000	0.000	0.000	0.000	0.000	22.222
CVM(5)	0.000	0.000	0.000	0.000	0.000	0.000	0.000	22.222
CVM(10)	3.571	0.000	0.000	0.000	0.000	0.000	0.000	22.222
$\widetilde{\text{CVM}}(5)$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	22.222
$\widetilde{\text{CVM}}(10)$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	22.222
$C(1)$	0.000	0.000	3.571	28.571	0.000	11.111	44.444	77.778
$C(5)$	0.000	0.000	0.000	17.857	0.000	11.111	22.222	44.444
$C(10)$	0.000	0.000	0.000	0.000	0.000	0.000	22.222	44.444

Notes: For each simulation of length T, the table reports the percentage of cases (out of 99 for Design I, out of 41 for Design II, out of 28 for Design III, and out of 45 for Design IV) for which the test has the maximum rejection frequency.

These overall findings highlight the difficulty of deciding on the general reliability of an individual test.

On a more granular level, we observe that the power of the tests can vary notably from one scenario to another, indicating a very high sensitivity of the individual tests to the data-generating process, even under the same design. There are numerous cases when the tests "disagree", and the researcher is faced with mixed results. For example, under Design I, for the scenario corresponding to the parameters values of $a = 1.15$ and $b = 0.85$, $T = 250$, the power of LB test (for all lag versions) is low at 0.035 (35/1000 correct rejections of the null), while the CVM(2) test has a high power of 1 (1000/1000 correct rejections of the null). Similarly, under Design IV, for the scenario corresponding to the parameters values of $\beta = 0.9$, $\nu = 0.4$, $T = 1000$, the individual tests of Berkowitz with 1 lag and LB(1) have low power of 0.035 and 0.255, respectively; while the $Q^*(1)$ and CVM(1) individual tests have high power of 1 and 0.995, respectively. These cases remain inconclusive, as there is no unanimous decision regarding the accuracy of a set of density forecasts.

An important remark regarding the power of the combined test stems from the results of Table II. The combined test may inherit the same power as the most powerful test among the ones we combine. Nevertheless, there are cases, particularly in Designs III and IV, in which the newly proposed test is even more powerful than any individual test that enters the combination. Regarding the choice of the number of lags for our combined tests, some indications can be drawn from the Monte Carlo experiments. For example, in Design I (Table B.1), capturing only departures from the uniform distribution and not from independence, the combined test performs in the same way across the three lag-lengths (1, 5 and 10). In Designs II and III, where several forms of departures from the independence assumptions are considered, on average the $C(1)$ and $C(5)$ tests dominate in terms of power the $C(10)$ test (Tables B.2 and B.3). In Design IV, based on simulating directly daily returns rather than PITs, the $C(1)$ test dominates the remaining two tests. Based on this observation, a possible recommendation is to use, with daily data, 1 lag for the combined test. However, in case of conflicting results across various lags, the researcher could consider a second layer of combinations, i.e., similarly to the tests in Appendix A, the sum of the proposed test at different lags.

4. Empirical Application

A major aim of this study is to apply the combined testing procedure, detailed in Section 2, to assess if behavioural factors help to improve the predictions of commodity futures returns. To be more precise, the objective is to use the proposed combination test

to establish which model specifications, different in terms of the explanatory variables in the variance equation, produce accurate density forecasts. We analyse 15 commodities, including major metals, energy and agricultural products, as listed in Table III.

Table III
Commodity futures (Bloomberg Tickers)

Agricultural	
Ticker	Description
C1 Comdty	Generic 1st Corn No. 2 Yellow futures, US$
S1 Comdty	Generic 1st Soybean No. 2 Yellow futures, US$ (CBOT)
W1 Comdty	Generic 1st Wheat futures, US$ (CBOT)
KC1 Comdty	Generic 1st Coffee futures contract
SB1 Comdty	Generic 1st Sugar No. 11 (raw) futures
RR1 Comdty	Generic 1st Rice futures
CC1 Comdty	Generic 1st Cocoa
Energy	
Ticker	Description
CO1 Comdty	Brent oil
CL1 Comdty	WTI crude oil
HO1 Comdty	Heating oil
NG1 Comdty	Natural Gas
Metals	
Ticker	Description
GC1 Comdty	Gold
SI1 Comdty	Silver
PL1 Comdty	Platinum
HG1 Comdty	Copper

For each commodity i, we first calculate the daily log-returns, $r_{i,t}$, as the difference in log futures prices between time t and $t-1$, and then we fit an ARMAX-EGARCH-t model specified as:

$$\begin{aligned} r_{i,t} &= \mu_i + \phi_i r_{i,t-1} + \theta_i \varepsilon_{i,t-1} + \varphi_i' x_{t-1} + \varepsilon_{i,t} \\ \varepsilon_{i,t} &= \sigma_{i,t}\eta_{i,t}, \qquad \eta_{i,t} \sim i.i.d \quad stdt(\nu) \\ \log\left(\sigma_{i,t}^2\right) &= \omega_i + \beta_i \log\left(\sigma_{i,t-1}^2\right) + \gamma_i \frac{\varepsilon_{i,t-1}}{\sqrt{\sigma_{i,t-1}^2}} + \alpha_i \left[\frac{|\varepsilon_{i,t-1}|}{\sqrt{\sigma_{i,t-1}^2}} - \sqrt{\frac{2}{\pi}}\right] + \delta_i' s_{t-1}, \end{aligned} \tag{12}$$

where $r_{i,t}$ is the time t log-return calculated on the futures price of commodity i. Commodity log-returns are based on the first generic futures contracts series extracted from Bloomberg. Specifically, we consider, at each date, the price of the contract closest to maturity. When a given contract approaches the expiration date, Bloomberg provides a

smooth transition between two consecutive contracts by computing a weighted average of the prices of the futures contracts. The data sample covers the period from 2 January 2004 to 31 December 2019 for a total of 5,843 observations for each commodity.

In eq. (12), x denotes the vector of explanatory financial variables entering the mean equation. The components of the x vector include the Fed Fund interest rate, the log-returns of the S&P500 index, the MSCI Emerging Market index and the dollar effective exchange rate collected from Bloomberg[4]. These variables have been widely used in the empirical literature as predictors of commodity price returns and volatilities (e.g. Koch, 2014; Benkraiem et al., 2018; Algieri and Leccadito, 2020). In the variance equation, the vector s includes all the possible combinations of the following behavioural variables related to sentiment and investors' psychology: the Thomson Reuters MarketPsych (TRMI) index, commodity-specific implied volatility (IV), economic policy uncertainty (EPU) and Google trends data.

The TRMI index measures a number of emotional and topical items across global news and social media sites that could drive the operator's behaviour in commodity markets. Data have been directly provided by Refinitiv. The commodity implied volatility is a measure of fear and uncertainty in the commodity market. Typically, its value increases when the specific commodity market is falling, while it lessens when the market surges. The IV data have been collected from Bloomberg.[5] EPU is a sentiment index that reflects the degree of uncertainty and insecurity in the economy. This index, proposed by Baker et al. (2016), has been used in several empirical analyses to capture the mood and psychology of the market's operators (e.g. Shahzad et al., 2017; Algieri, 2021; Dew-Becker et al., 2021; Dou et al., 2022). It is a composed index constructed using three components: 1) the main newspaper coverage frequency regarding economic policy insecurity, 2) the total federal tax code provisions planned for the future and 3) the economic forecasts discrepancies across professional forecasters concerning public expenditures and consumer price indices. EPU data have been extracted from Bloomberg.[6] Google Trends data on commodities provide information about market-level sentiment with respect to commodities' prices by constructing a Search Volume Index. An increase in the Search Volume Index is associated with a temporary positive pressure on commodity prices (Da et al., 2015). We construct the Google trend commodity index by extracting the search activity data over time for the following words: "Gold, Silver,

[4]The tickers are FDFD Index, SPX, MXEF Index and DXY, respectively.
[5]The ticker for each commodity is HIST_CALL_IMP_VOL.
[6]The ticker is EPUCNUSD Index.

Oil, Corn, Coffee, Platinum, Copper, Palladium, Rough Rice, Sugar, Cocoa, Soybeans and Wheat". Since not everyone searching for words, such as "Coffee" or "Corn", intends to trade these commodities, we tried to reduce the word trap bias by restricting the Google search to categories linked with financial markets[7]. For example, the search for the word "Gold" was filtered through ten categories. Then, since only for some of the sub-samples, daily data was available, we scraped monthly and weekly data and rescaled them via min-max normalization. Finally, daily data were obtained by linear interpolation. We formed the sentiment measure in the last step by adding the top ten positively correlated sub-categories with the corresponding commodities market price log returns[8].

To assess the impact of the four behavioural factors mentioned above, we consider all possible variance specifications based on their combinations, including the empty set. This gives 16 different ARMAX-EGARCH-t models for each commodity and the corresponding forecasts of the daily log-returns $r_{i,t}$. For each estimation, we use a rolling window such that 2,500 density forecasts (and PITs) are available for out-of-sample comparisons.

For each commodity, we compare the performance of the 16 different model specifications using our proposed combination test, eq. (7), with $k \in \{1, 5, 10\}$. Table IV reports for each commodity and for each of the 16 models (eq. 12) the p-values of the $C(1)$, $C(5)$, and $C(10)$ tests. The results show that sentiment variables significantly help ameliorate predictions for several commodities, compared to models without their inclusion (NoExternal). For example, for corn, coffee, soybean (only in the case of the $C(10)$ test), sugar, and natural gas, models including combinations such as IV+GGLE, IV+GGLE+EPU, or IV+GGLE+TRMI or all sentiment factors together, display the best predictive performance. The importance of implied volatility as a trustworthy predictor of the short-term volatility of the underlying asset and, therefore, of its price, is in line with the pioneering study by Taylor and Xu (1997) and Blair et al. (2001). We further highlight that adding other factors to implied volatilities improves the pre-

[7]Categories include Business Finance, Business News, Commodities and Futures Trading, Company News, Business and Industrial, Advertising and Marketing, Investment Banking, Risk Management, Company Earnings, Economy News, Financial Markets, Fiscal Policy News, Oil and Gas, Renewable and Alternative Energy, Fuel Economy and Gas Prices, Finance, Banking, Currencies Foreign Exchange, Financial and Planning, Insurance, Investing, Retirement and Pension, Academic Conferences and Publications, Resumes and Portfolios, Newspapers, Economics.

[8]For some of the commodities such as Gold, Rough Rice and Soybeans, our methodology returned fewer than 10 categories. In this case, we just summed all the categories that returns a number Google trend.

dictive performance. For the gold, platinum and Brent oil commodities, all 16 model specifications are instead rejected by the tests. Within the energy commodities group, the improvement is less pronounced for crude and heating oil. Interestingly, sentiment variables tend to enhance predictions when combined rather than used in isolation. The importance of sentiment factors was first signalled by John Maynard Keynes in 1936. The economist argued how sentiments and impulses, the well-renowned "animal spirits", drive human behaviours. Afterwards, other authors (e.g. Shiller, 2015; Bekiros et al., 2018) pinpointed how unfounded market optimism could generate irrational enthusiasm, resulting in a market bubble and likely market panic. On the other hand, high uncertainty levels can harm economic activity: investors may postpone their investment decisions until the uncertainties dissipate. Our analysis corroborates the behavioural strand of the literature and confirms that including sentiment factors in the investigation could ameliorate predictions.

As an additional check, in Table V, we report for each model and each commodity, the log predictive scores (LPS) defined as $\frac{1}{S}\sum_{t=1}^{S}\log f_t(r_t)$, where S is the number of predictions, $f_t(\cdot)$ the predicted density for time t, and r_t the realised commodity return. Using this indicator to compare several forecasts, models with larger LPS should be preferred over models with smaller LPS. For each of the 15 commodities we report the largest LPS in bold (Table V). The findings point to some similarities with Table IV. For instance, all sentiment variables together ameliorate the predictive performance for corn, coffee, rice, natural gas and copper. Differently from Table IV, the combination IV+GGLE is less performative. However, one must underline that while our proposed tests account for both the hypotheses of uniformity and independence of the PITs, the log predictive scores overlook the hypothesis of independence.

For three commodities, one from each category (namely soybean, silver and natural gas), we identify one of the best-performing models based on Table IV and depict their corresponding predicted densities at various times (Figure C.1, Supplementary Material). The chosen model considers the sentiment factor combination TRMI+IV+GGLE for soybeans (S1), IV+EPU for silver (SI1), and TRMI+EPU+GGLE for natural gas (NG1). We plot the predicted densities starting with the 22nd of June 2014, when, due to adverse meteorological conditions, soybeans experienced a large increase in price. The effect on the density has been an increase in the mean of distribution (compared to two of the remaining densities) and thickness of its tails, since the kurtosis is higher compared to the other three densities. The next date we choose is the 14th of December 2015, when silver reached the lowest value of about \$13 after the peak of about \$50 per ounce recorded on the 29th of April 2011 when economic uncertainty and inflation con-

Table IV

P-values of the combined tests $C(1)$, $C(5)$, and $C(10)$.

Model / Commodity	NoExternal	TRMI	IV	GGLE	EPU	TRMI + IV	TRMI+ GGLE	TRMI+ EPU	IV+ GGLE	IV+ EPU	GGLE+ EPU	EPU+ IV+ GGLE	TRMI+ EPU+ GGLE	TRMI+ IV+ EPU	TRMI+ IV+ GGLE	All
Panel A: $C(1)$ Test																
C1-Corn	0.000	0.000	0.011	0.000	0.000	0.011	0.000	0.000	**0.981**	0.013	0.000	**0.981**	0.000	0.000	**0.981**	**0.980**
S1-Soybean	0.000	0.0000	0.000	0.000	0.000	0.000	0.0000	0.000	0.003	0.000	0.000	0.003	0.000	0.000	0.002	0.002
W1-Wheat	0.005	0.011	**0.107**	0.011	0.010	**0.084**	0.014	0.012	0.031	**0.114**	0.015	0.038	0.021	0.011	0.022	0.028
KC1-Coffee	0.033	**0.635**	**0.324**	**0.608**	**0.636**	**0.353**	**0.648**	**0.671**	**0.219**	**0.461**	**0.644**	**0.445**	**0.545**	**0.636**	**0.457**	**0.633**
SB1-Sugar	0.011	0.006	**0.121**	0.001	0.017	**0.132**	0.001	0.030	**0.147**	**0.146**	0.011	**0.152**	0.006	0.006	**0.167**	**0.165**
RR1-Rice	0.003	0.001	0.012	0.001	0.001	0.014	0.001	0.001	0.003	0.024	0.001	0.017	0.004	0.001	0.004	0.020
CC1-Cocoa	**0.283**	**0.361**	**0.151**	**0.662**	**0.460**	**0.154**	0.001	**0.451**	**0.666**	**0.133**	**0.665**	**0.573**	**0.638**	**0.361**	**0.510**	**0.380**
CO1-Brent Oil	0.0000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
CL1-WTI Oil	**0.243**	**0.185**	0.000	**0.416**	**0.246**	0.000	**0.251**	**0.344**	0.000	0.000	**0.405**	0.000	**0.296**	**0.185**	0.000	0.000
H01-Heating Oil	0.042	**0.053**	0.001	**0.062**	0.040	0.001	0.041	**0.067**	0.001	0.001	**0.076**	0.002	**0.053**	**0.053**	0.001	0.003
NG1-Natural Gas	**0.054**	**0.214**	**0.398**	**0.259**	**0.080**	**0.334**	**0.344**	0.049	**0.378**	**0.360**	**0.262**	**0.372**	**0.368**	**0.214**	**0.373**	**0.366**
GC1-Gold	0.0000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
SI1-Silver	0.000	0.000	**0.238**	0.000	0.000	**0.243**	0.000	0.000	**0.088**	**0.310**	0.000	**0.100**	0.000	0.000	**0.110**	**0.127**
PL1-Platinum	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
HG1-Copper	0.024	0.009	0.032	0.017	**0.054**	0.040	0.004	0.038	**0.146**	0.032	0.031	**0.149**	0.009	0.009	**0.163**	**0.165**
Panel B: $C(5)$ Test																
C1-Corn	0.000	0.000	0.002	0.000	0.000	0.002	0.000	0.000	**0.239**	0.002	0.000	**0.231**	0.000	0.000	**0.251**	**0.241**
S1-Soybean	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.037	0.000	0.000	0.032	0.000	0.000	0.034	0.036
W1-Wheat	0.008	0.014	0.034	**0.072**	0.011	0.026	**0.087**	0.013	**0.167**	0.035	**0.094**	**0.193**	**0.128**	0.014	**0.130**	**0.155**
KC1-Coffee	0.014	**0.103**	**0.108**	**0.852**	**0.112**	**0.107**	**0.854**	**0.129**	**0.505**	**0.139**	**0.870**	**0.694**	**0.818**	**0.103**	**0.686**	**0.779**
SB1-Sugar	0.050	0.036	**0.255**	0.001	**0.088**	**0.264**	0.001	**0.141**	**0.067**	**0.284**	0.008	**0.070**	0.008	0.036	**0.074**	**0.075**
RR1-Rice	0.032	0.024	**0.055**	0.003	0.016	**0.065**	0.006	0.008	0.008	**0.105**	0.016	0.042	0.034	0.024	0.009	0.045
CC1-Cocoa	**0.306**	**0.375**	**0.164**	**0.984**	**0.431**	**0.166**	0.034	**0.430**	**0.968**	**0.158**	**0.979**	**0.941**	**0.970**	**0.375**	**0.915**	**0.850**
CO1-Brent Oil	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
CL1-WTI Oil	**0.386**	**0.263**	0.000	**0.570**	**0.384**	0.000	**0.407**	**0.460**	0.000	0.000	**0.576**	0.001	**0.449**	**0.263**	0.000	0.002
H01-Heating Oil	**0.080**	**0.074**	0.002	0.048	**0.079**	0.002	0.040	**0.099**	0.002	0.005	**0.057**	0.007	0.045	**0.074**	0.003	0.008
NG1-Natural Gas	0.044	**0.114**	0.043	**0.688**	**0.058**	0.032	**0.740**	0.045	**0.543**	0.039	**0.686**	**0.547**	**0.750**	**0.114**	**0.534**	**0.540**
GC1-Gold	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.002	0.000	0.000	0.002	0.000	0.000	0.003	0.003
SI1-Silver	0.000	0.000	0.002	0.000	0.000	0.003	0.000	0.000	0.027	0.003	0.000	0.042	0.000	0.000	0.027	0.041
PL1-Platinum	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.000
HG1-Copper	0.012	0.005	0.003	0.002	0.025	0.004	0.001	0.018	0.006	0.003	0.003	0.006	0.001	0.005	0.007	0.007
Panel C: $C(10)$ Test																
C1-Corn	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	**0.476**	0.000	0.000	**0.464**	0.000	0.000	**0.505**	**0.492**
S1-Soybean	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	**0.134**	0.000	0.001	**0.131**	0.001	0.000	**0.148**	**0.153**
KC1-Coffee	0.012	**0.232**	**0.163**	**0.811**	**0.233**	**0.166**	**0.780**	**0.256**	**0.445**	**0.189**	**0.804**	**0.547**	**0.746**	**0.232**	**0.543**	**0.591**
SB1-Sugar	0.041	0.028	**0.118**	0.002	**0.071**	**0.122**	0.002	**0.115**	**0.051**	**0.128**	0.012	**0.054**	0.013	0.028	**0.060**	**0.060**
RR1-Rice	**0.054**	0.048	**0.056**	0.037	0.038	**0.064**	**0.060**	0.025	0.040	**0.091**	**0.112**	**0.147**	**0.189**	0.048	0.043	**0.156**
CC1-Cocoa	**0.425**	**0.445**	**0.168**	**0.993**	**0.550**	**0.173**	**0.135**	**0.532**	**0.989**	**0.163**	**0.992**	**0.982**	**0.988**	**0.445**	**0.970**	**0.946**
CO1-Brent Oil	0.000	0.000	0.000	0.001	0.000	0.000	0.002	0.000	0.000	0.000	0.001	0.000	0.003	0.000	0.000	0.000
CL1-WTI Oil	**0.720**	**0.605**	0.001	**0.538**	**0.718**	0.001	**0.450**	**0.789**	0.000	0.004	**0.540**	0.004	**0.477**	**0.605**	0.001	0.010
H01-Heating Oil	**0.154**	**0.149**	0.009	0.021	**0.151**	0.010	0.014	**0.191**	0.002	0.022	0.019	0.006	0.014	**0.149**	0.002	0.006
NG1-Natural Gas	0.041	**0.086**	0.039	**0.835**	0.049	0.029	**0.876**	0.035	**0.560**	0.033	**0.819**	**0.563**	**0.882**	**0.086**	**0.554**	**0.558**
GC1-Gold	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.017	0.000	0.000	0.019	0.000	0.000	0.021	0.023
SI1-Silver	0.000	0.000	0.026	0.000	0.000	0.030	0.000	0.000	**0.053**	0.039	0.000	**0.076**	0.000	0.000	0.050	**0.073**
PL1-Platinum	0.000	0.000	0.000	0.000	0.000	0.000	0.002	0.000	0.001	0.000	0.001	0.001	0.002	0.000	0.006	0.003
HG1-Copper	0.005	0.002	0.001	0.002	0.009	0.001	0.000	0.006	0.005	0.001	0.003	0.005	0.001	0.002	0.005	0.005

Notes: This table reports, for all the commodities, the p-values for the combined tests of density forecast accuracy across all different specifications in terms of the s_t vector of behavioural variables in the variance equation of the model (12). P-values larger than 5% are indicated in bold.

Table V
Log predictive scores.

Model / Commodity	NoExternal	TRMI	IV	GGLE	EPU	TRMI + IV	TRMI+ GGLE	TRMI+ EPU	IV+ GGLE	IV+ EPU	GGLE+ EPU	EPU+ IV+ GGLE	TRMI+ EPU+ GGLE	TRMI+ IV+ EPU	TRMI+ IV+ GGLE	All
C1-Corn	3.0992	3.0956	3.1390	3.0429	3.0992	3.1382	3.0417	3.0989	3.1202	3.0391	3.0445	3.1370	3.0432	3.0956	3.1400	**3.1697**
S1-Soybean	3.2661	3.2655	3.2877	3.2100	3.2683	3.2857	3.2096	3.2656	3.2625	3.2877	3.2113	3.2226	3.2107	3.2655	**3.2915**	3.2714
W1-Wheat	2.9218	2.9235	2.9497	2.8318	2.9227	2.9499	2.8324	2.9204	2.8472	2.9496	2.8303	2.8472	2.8317	2.9235	**2.9966**	2.8467
KC1-Coffee	2.7648	2.7640	2.7754	2.6957	2.7649	2.7744	2.6975	2.7641	2.7047	2.7756	2.6957	2.7053	2.6973	2.7641	2.7063	**2.8066**
SB1-Sugar	2.9190	2.9183	2.9569	2.8647	2.9211	2.9569	2.8666	2.9211	2.8910	2.9570	2.8671	2.8908	2.8668	2.9183	2.8910	**2.9808**
RR1-Rice	3.2396	3.2400	3.2629	3.1754	3.2468	**3.2631**	3.1738	3.2486	3.1836	3.2628	3.1734	3.1851	3.1753	3.2400	3.1842	3.1856
CC1-Cocoa	2.9812	2.9815	**2.9990**	2.9212	2.9837	2.9984	2.8840	2.9832	2.9240	2.9988	2.9210	2.9219	2.9243	2.9815	2.9227	2.9234
CO1-Brent Oil	2.8453	2.8492	2.8939	2.8064	2.8451	2.8931	2.8092	**2.9165**	2.8247	2.8953	2.8060	2.8272	2.8093	2.8492	2.8291	2.8302
CL1-WTI Oil	2.7919	2.8012	2.8323	2.7492	2.7920	2.8330	2.7543	**2.8938**	2.7691	2.8344	2.7488	2.7703	2.7550	2.8012	2.7732	2.7745
H01-Heating Oil	2.9343	2.9353	2.9765	2.8853	2.9325	2.9775	2.8907	**2.9820**	2.9152	2.9776	2.8924	2.9144	2.8913	2.9353	2.9152	2.9146
NG1-Natural Gas	2.4917	2.4907	2.5205	2.4198	2.4912	2.5209	2.4200	2.4928	2.4499	2.5201	2.4209	2.4499	2.4201	2.4907	2.4998	**2.5497**
GC1-Gold	3.6217	3.6231	3.6692	3.5785	3.6241	3.6690	3.5787	3.6231	3.6111	**3.6698**	3.5818	3.6130	3.5834	3.6231	3.6111	3.6130
SI1-Silver	3.1144	3.1116	3.1427	3.0415	3.1217	3.1388	3.0415	3.1211	3.0736	3.1437	3.0506	**3.1741**	3.0526	3.1117	3.0715	3.0717
PL1-Platinum	3.0648	3.1592	3.1617	3.0919	3.1633	3.1641	3.0926	3.1610	3.0963	3.1581	3.0930	3.0948	3.1920	3.1587	**3.1945**	3.1924
HG1-Copper	3.2475	3.2484	3.2775	3.2169	3.2480	3.2775	3.2148	3.2483	3.2342	3.2777	3.2160	3.2343	3.2135	3.2484	3.2793	**3.2834**

Notes: This table reports the Log predictive scores, defined as $\frac{1}{S}\sum_{t=1}^{S}\log f_t(r_t)$, where r_t is the realized commodity return at time t.

cerns pushed investors towards the white metal as a hedge against rising prices. Indeed, in panel (b) of Figure C.1, we observe that the predicted density is shifted to the left (compared to the other densities we report) exhibiting large leptokurtosis, which implies a greater likelihood of extreme events. Another date of interest, at least for natural gas, is the 7th of March 2016, when the price fell due to an unusually warm winter, which led to lower heating demand, resulting in a surplus of natural gas supply. From panel (c), we can see that the density has associated a similar mean and variance, compared with the one for the 31st December 2019, the last day in the sample.

Based on the same three identified models, we plot, for each of the three representative commodities and for each day in the out-of-sample period, the 99%-confidence bands along with the realized returns (Figure C.2, Supplementary Material). Although there are a few realized returns that are outside these bands, the vast majority belong to the predicted confidence interval. Hence, the method of combining various tests of density forecasts accuracy can be used as a reliable tool in selecting "accurate" models in terms of forecasting performance.

5. Conclusions

Deciding which model to use for forecasting asset prices is highly relevant for financial and economic operators, institutional investors and policy makers. In this study, we develop a methodology for combining density forecast accuracy tests. This technique is applied to a range of well-known tests including Berkowitz test, Hong and Li test and Cramer-von Mises test, in order to enhance decisions. Considering multiple individual tests, in fact, can often lead to a inconclusive set of results, and hence the difficult task of assessing whether a set of density forecasts from a given model is accurate. The ability of creating an all-in-one test has the potential to circumvent this problem.

We further illustrate how to apply our methodology to a set of 15 commodity futures contracts and how to improve forecasts of futures returns by adding sentiment variables into the models. The empirical results show that including different behavioural variables in the volatility equation improves the density forecasts of different commodity futures returns. This holds true, especially for agricultural commodities. For energy commodities and metals, the futures return predictions also improve, but to a lesser extent than the other commodity class. This result could be explained by the fact that the impact of sentiment variables might be more pronounced in less efficient markets. Therefore, it is likely that sentiment variables have a more significant role in agricultural markets compared to energy and metal markets, given also the specific characteristics

of agricultural commodities (weather dependence, growth and harvest cycles, trade policy) that make them more susceptible to sentiment-driven fluctuations. Further, the combinations of sentiment variables are more powerful in predicting returns than considering sentiment variables separately. Our results are robust across both tests and model specifications.

Overall, the methodology presented in this study could be applied to other financial and real markets (including the assessment of forecasting models of inflation, output, unemployment, and exchange rates as a measure of international competitiveness) and could be valuable to support validation teams and risk management departments which need to frequently screen through a wide range of models that traders may use.

References

Adams, Z., Glück, T., 2015. Financialization in commodity markets: A passing trend or the new normal? Journal of Banking and Finance 60, 93–111. doi:https://doi.org/10.1016/j.jbankfin.2015.07.008.

Alexandridis, A.K., Gzyl, H., ter Horst, E., Molina, G., 2021. Extracting pricing densities for weather derivatives using the maximum entropy method. Journal of the Operational Research Society 72, 2412–2428.

Alexandridis, A.K., Karlis, D., Papastamos, D., Andritsos, D., 2019. Real estate valuation and forecasting in non-homogeneous markets: A case study in Greece during the financial crisis. Journal of the Operational Research Society 70, 1769–1783.

Algieri, B., 2021. Fast & furious: Do psychological and legal factors affect commodity price volatility? The World Economy 44, 980–1017.

Algieri, B., Leccadito, A., 2020. Extreme price moves: an INGARCH approach to model coexceedances in commodity markets. European Review of Agricultural Economics 48, 878–914.

Algieri, B., Leccadito, A., Tunaru, D., 2021. Risk premia in electricity derivatives markets. Energy Economics 100. doi:10.1016/j.eneco.2021.1053.

Amisano, G., Giacomini, R., 2007. Comparing density forecasts via weighted likelihood ratio tests. Journal of Business Economics and Statistics 25, 177–190.

Bahloul, W., Bouri, A., 2016. The impact of investor sentiment on returns and conditional volatility in u.s. futures markets. Journal of Multinational Financial Management 36, 89 – 102.

Baker, S.R., Bloom, N., Davis, S.J., 2016. Measuring Economic Policy Uncertainty. The Quarterly Journal of Economics 131, 1593–1636.

Bekiros, S., Jlassi, M., Naoui, K., Uddin, G.S., 2018. Risk perception in financial markets: On the flip side. International Review of Financial Analysis 57, 184–206. URL: https://ideas.repec.org/a/eee/finana/v57y2018icp184-206.html.

Benkraiem, R., Lahiani, A., Miloudi, A., Shahbaz, M., 2018. New insights into the us stock market reactions to energy price shocks. Journal of International Financial Markets, Institutions and Money 56, 169–187.

Berkowitz, J., 2001. Testing density forecasts, with applications to risk management. Journal of Business and Economic Statistics 19, 465–74.

Berkowitz, J., Christoffersen, P., Pelletier, D., 2011. Evaluating Value-at-Risk models with desk-level data. Management Science 57, 2213–2227.

Billio, M., Casarin, R., Ravazzolo, F., van Dijk, H., 2013. Time-varying combinations of predictive densities using nonlinear filtering. Journal of Econometrics 177, 213–232.

Blair, B.J., Poon, S.H., Taylor, S.J., 2001. Forecasting s&p 100 volatility: the incremental information content of implied volatilities and high-frequency index returns. Journal of Econometrics 105, 5–26. URL: https://www.sciencedirect.com/science/article/pii/S0304407601000689, doi:https://doi.org/10.1016/S0304-4076(01)00068-9. forecasting and empirical methods in finance and macroeconomics.

Chen, Z., Liang, C., Umar, M., 2021. Is investor sentiment stronger than vix and uncertainty indices in predicting energy volatility? Resources Policy 74, 102391. doi:https://doi.org/10.1016/j.resourpol.2021.102391.

Christoffersen, P., 2012. Elements of Financial Risk Management. second ed., Elsevier Science.

Clements, M., 2004. Evaluating the Bank of England density forecasts for inflation. The Economic Journal 114, 844–866.

Da, Z., Engelberg, J., Gao, P., 2015. The sum of all fears investor sentiment and asset prices. The Review of Financial Studies 28, 1–32.

Dew-Becker, I., Giglio, S., Kelly, B., 2021. Hedging macroeconomic and financial uncertainty and volatility. Journal of Financial Economics 142, 23–45.

Dichtl, H., 2020. Forecasting excess returns of the gold market: Can we learn from stock market predictions? Journal of Commodity Markets 19, 100106. doi:https://doi.org/10.1016/j.jcomm.2019.100106.

Diebold, F.X., Gunther, T.A., Tay, A.S., 1998. Evaluating density forecasts with applications to finacial risk management. International Economic Review 39, 863–883.

Diebold, F.X., Shin, M., Zhang, B., 2022. On the aggregation of probability assessments: Regularized mixtures of predictive densities for eurozone inflation and real interest rates. Journal of Econometrics doi:https://doi.org/10.1016/j.jeconom.2022.06.008.

Ding, S., Cui, T., Zheng, D., Du, M., 2021. The effects of commodity financialization on commodity market volatility. Resources Policy 73, 102220. doi:https://doi.org/10.1016/j.resourpol.2021.102220.

Dou, Y., Li, Y., Dong, K., Ren, X., 2022. Dynamic linkages between economic policy uncertainty and the carbon futures market: Does covid-19 pandemic matter? Resources Policy 75, 102455. doi:https://doi.org/10.1016/j.resourpol.2021.102455.

Dowd, K., 2004. A modified Berkowitz back-test. Risk Magazine 4, 36–36.

Du, Z., Escanciano, J.C., 2017. Backtesting expected shortfall: Accounting for tail risk. Management Science 63, 940–958.

Dufour, J.M., 2006. Monte Carlo tests with nuisance parameters: A general approach to finite-sample inference and nonstandard asymptotics. Journal of Econometrics 133, 443–477.

Gaglianone, W.P., Marins, J.T.M., 2017. Evaluation of exchange rate point and density forecasts: An application to brazil. International Journal of Forecasting 33, 707–728. doi:https://doi.org/10.1016/j.ijforecast.2016.12.002.

Geweke, J., Amisano, G., 2012. Prediction with misspecified models. AEA Papers and Proceedings 102, 482–486.

Gokmenoglu, K.K., Fazlollahi, N., 2015. The interactions among gold, oil, and stock market: Evidence from S&P500. Procedia Economics and Finance 25, 478 – 488. 16th Annual Conference on Finance and Accounting, ACFA Prague 2015, 29th May 2015.

Hall, S., Mitchell, J., 2007. Combining density forecasts. International Journal of Forecasting 23, 1–13.

Hog, E., Tsiaras, L., 2011. Density forecasts of crude-oil prices using option-implied and arch-type models. The of Futures Markets 31, 727 –754. doi:10.1002/fut.20487.

Hollstein, F., Prokopczuk, M., Tharann, B., Wese Simen, C., 2021. Predictability in commodity markets: Evidence from more than a century. Journal of Commodity Markets 24, 100171. doi:https://doi.org/10.1016/j.jcomm.2021.100171.

Hong, Y., Li, H., 2004. Nonparametric Specification Testing for Continuous-Time Models with Applications to Term Structure of Interest Rates. The Review of Financial Studies 18, 37–84.

Hsiao, C., Wan, S.K., 2014. Is there an optimal forecast combination? Journal of Econometrics 178, 249–309.

Hua, Z., Zhang, B., 2008. Improving density forecast by modeling asymmetric features: An application to S&P500 returns. European Journal of Operational Research 185, 716–725.

Hurlin, C., Tokpavi, S., 2006. Backtesting value-at-risk accuracy: a simple new test. Journal of Risk 9, 19–37.

Kapetanios, G., Mitchell, J., Price, S., Fawcett, N., 2015. Generalised density forecast combinations. Journal of Econometrics 188, 150–165.

Koch, N., 2014. Tail events: A new approach to understanding extreme energy commodity prices. Energy Economics 43, 195–205.

Koutsandreas, D., Spiliotis, E., Petropoulos, F., Assimakopoulos, V., 2021. On the selection of forecasting accuracy measures. Journal of the Operational Research Society 0, 1–18. doi:10.1080/01605682.2021.1892464.

Liu, W., Yang, J., Chen, J., Xu, L., 2023. How social-network attention and sentiment of investors affect commodity futures market returns: New evidence from china. SAGE Open 13, 21582440231152131. doi:10.1177/21582440231152131.

Luo, J., Chen, L., 2019. Multivariate realised volatility forecasts of agricultural commodity futures. Journal of Futures Markets 39, 1565–1586. doi:10.1002/fut.22052.

McAdam, P., Warne, A., 2020. Density Forecast Combinations: The Real-Time Dimension. Technical Report 2378. European Central Bank.

de Menezes, L., Bunn, D., Taylor, J., 2000. Review of guidelines for the use of combined forecasts. European Journal of Operational Research 120, 190–204.

Mitchell, J., Hall, S., 2005. Evaluating, comparing and combining density forecasts using KLIC with an application to the Bank of England and NIESR 'fan' charts of inflation. Oxford Bulletin of Economics and Statistics 67, 995–1033.

Niu, X., Harvey, N., 2022. Point, interval, and density forecasts: Differences in bias, judgment noise, and overall accuracy. Futures & Foresight Science 4. doi:https://doi.org/10.1002/ffo2.124.

Plakandaras, V., Gupta, R., Wong, W.K., 2019. Point and density forecasts of oil returns: The role of geopolitical risks. Resources Policy 62, 580–587.

Rosenblatt, M., 1952. Remarks on a multivariate transformation. The Annals of Mathematical Statistics 23, 470–472.

Shahzad, S.J.H., Raza, N., Balcilar, M., Ali, S., Shahbaz, M., 2017. Can economic policy uncertainty and investors sentiment predict commodities returns and volatility? Resources Policy 53, 208–218.

Shiller, R.J., 2015. Irrational Exuberance. Number 10421 in Economics Books, Princeton University Press. URL: https://ideas.repec.org/b/pup/pbooks/10421.html.

Tay, A.S., Wallis, K.F., 2000. Density forecasting: a survey. Journal of Forecasting 19, 235–254.

Taylor, J.W., Taylor, K.S., 2023. Combining probabilistic forecasts of covid-19 mortality in the united states. European Journal of Operational Research 304, 25–41. doi:https://doi.org/10.1016/j.ejor.2021.06.044.

Taylor, S.J., Xu, X., 1997. The incremental volatility information in one million foreign exchange quotations. Journal of Empirical Finance 4, 317–340. URL: https://ideas.repec.org/a/eee/empfin/v4y1997i4p317-340.html.

Timmermann, A., 2006. Forecast combinations, in: Elliott, G., Granger, C., Timmermann, A. (Eds.), Handbook of Economic Forecasting. Elsevier, Amsterdam. volume 1, pp. 135–196.

Wallis, K.F., 2011. Combining forecasts – forty years later. Applied Financial Economics 21, 33–41.

Wang, C., 2001. Investor sentiment and return predictability in agricultural futures markets. Journal of Futures Markets 21, 929–952. doi:10.1002/fut.2003.

Wang, X., Hyndman, R.J., Li, F., Kang, Y., 2022. Forecast combinations: An over 50-year review. International Journal of Forecasting doi:https://doi.org/10.1016/j.ijforecast.2022.11.005.

Wang, Y.H., Keswani, A., Taylor, S.J., 2006. The relationships between sentiment, returns and volatility. International Journal of Forecasting 22, 109–123.

List of Figures

Data Availability

The dataset analysed during the current study is available from Bloomberg and Thomson Reuters but restrictions apply to the availability of these data, which were used under license for the current study, and hence are not publicly available. Data are, however, available from the authors upon reasonable request and with permission of Bloomberg and Thomson Reuters.

Conflict of Interest

The authors declare that they have no conflict of interest.

Funding Statement

Bernardina Algieri gratefully acknowledges financial support from the Italian Ministry of University via research grant [DD 967 30.06.2023], PRIN 2022 - Project number 20229NB2MT: "Green Competitiveness For A Stronger And More Sustainable Italian Economy - GREENGO".

Acknowledgments

We wish to thank the Editor, an Associate Editor and two anonymous referees for very useful comments which have helped to improve the article. A. Leccadito would like to thank Professor Lynda Khalaf for her valuable suggestions.