Journal of Banking & Finance 40 (2014) 303–320

Contents lists available at ScienceDirect

Journal of Banking & Finance

journal homepage: www.elsevier.com/locate/jbf

The importance of the volatility risk premium for volatility forecasting ☆

Marcel Prokopczuk [a,b,*], Chardin Wese Simen [b]

[a] Zeppelin University, 88045 Friedrichshafen, Germany
[b] ICMA Centre, Henley Business School, University of Reading, Reading RG6 6BA, UK

ARTICLE INFO

Article history:
Received 23 May 2013
Accepted 1 December 2013
Available online 8 December 2013

JEL classification:
G13
G17

Keywords:
Volatility forecasting
Volatility risk premium
Implied volatility

ABSTRACT

In this paper, we study the role of the volatility risk premium for the forecasting performance of implied volatility. We introduce a non-parametric and parsimonious approach to adjust the model-free implied volatility for the volatility risk premium and implement this methodology using more than 20 years of options and futures data on three major energy markets. Using regression models and statistical loss functions, we find compelling evidence to suggest that the risk premium adjusted implied volatility significantly outperforms other models, including its unadjusted counterpart. Our main finding holds for different choices of volatility estimators and competing time-series models, underlying the robustness of our results.

1. Introduction

A plethora of academic publications compare option implied to time-series forecasts of realized volatility such as historical volatility and GARCH-type models.[1] Surprisingly, these studies pay little attention to the fact that implied volatility is obtained under the risk-neutral measure, Q, whereas the quantity to be forecasted, i.e. realized volatility, is observed under the physical measure, P. Thus, directly comparing option implied volatility to time-series models of volatility requires the assumption that the market price of volatility risk is zero. However, Carr and Wu (2009), Driessen et al. (2009), Trolle and Schwartz (2010), Mueller et al. (2011), and Prokopczuk and Wese Simen (2012) convincingly reject this assumption. They document a significant and time-varying volatility risk premium that effectively drives a wedge between the volatility forecasted under Q and subsequently realized under P. In light of this, it is natural to ask: can the forecasting performance of implied volatility be improved by adjusting for the volatility risk premium?

Our answer is "yes". In reaching this conclusion, we make two important contributions to the volatility forecasting literature. First, we build on the model-free implied volatility (MFIV) of Jiang and Tian (2005) to propose a simple and non-parametric adjustment to account for the market price of volatility risk. To the best of our knowledge, we are the first to study the role of the volatility risk premium for volatility forecasting in a model-free setting. This is in stark contrast with the approaches of Lamoureux and Lastrapes (1993) and Poteshman (2000), who rely on explicit option pricing models. Our approach also differs from that of Chernov (2007), in that we neither rely on an approximation nor on a small subset of option prices, specifically At-The-Money (ATM) options, as the author does.

Our second contribution consists of a thorough empirical assessment of the importance of our adjustment. To do this, we proceed in three stages. We begin by evaluating the relative information content of the volatility risk premium adjusted model-free implied volatility (RMFIV) vis-à-vis other models that include historical volatility (HIST) and GARCH-type models. We do this by estimating regressions of realized volatility on alternative forecasts of volatility. We then use four statistical loss functions, i.e. mean absolute errors (MAE), mean squared errors (MSE), mean absolute percentage errors (MAPE) and mean squared percentage errors (MSPE) to investigate the forecasting accuracy of each model. Lastly, we use the Diebold–Mariano and the non-parametric

☆ Marcel Prokopczuk gratefully acknowledges financial support from the British Academy. We thank two anonymous referees, Davide Avino, Christian Dorion, Olaf Korn, Ogonna Nneji and seminar participants at the FDIC Annual Derivatives Securities and Risk Management Conference (2013), the European Meeting of the Econometric Society (2013), the Meeting of the German Finance Association (2013), the Financial Management Association Meeting (2013), Zeppelin University and the University of Reading for valuable comments.

* Corresponding author at: Zeppelin University, 88045 Friedrichshafen, Germany. Tel.: +49 754160092231.

E-mail addresses: marcel.prokopczuk@zu.de (M. Prokopczuk), c.wesesimen@icmacentre.ac.uk (C. Wese Simen).

[1] See Poon and Granger (2003) for an excellent survey.

0378-4266/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jbankfin.2013.12.002

Table 1
Univariate and encompassing forecasts for crude oil's 30-day realized volatility.

	α	β_{HIST}	β_{GJR}	β_{ATM}	β_{MFIV}	β_{RMFIV}	Adj R^2	Wald	DW	Nobs
HIST	0.11	0.69					0.47	19.18	2.33	220
	(3.85)	**(7.52)**						**[0.00]**		
GJR	0.17		0.48				0.32	64.25	1.81	220
	(3.28)		**(3.19)**					**[0.00]**		
ATM IV	0.00			0.92			0.44	4.68	1.58	220
	(0.08)			**(7.84)**				**[0.01]**		
MFIV	−0.06				1.07		0.57	11.27	1.75	220
	(−1.41)				**(8.03)**			**[0.00]**		
RMFIV	−0.01					1.08	0.62	3.83	2.05	220
	(−0.28)					**(9.67)**		**[0.02]**		
HIST + ATM IV	0.01	0.48		0.45			0.53	44.78	2.30	220
	(0.46)	**(5.89)**		**(4.53)**				**[0.00]**		
HIST + MFIV	−0.04	0.20			0.83		0.58	6.69	2.04	220
	(−1.15)	**(2.12)**			**(6.37)**			**[0.01]**		
HIST + RMFIV	−0.01	−0.06				1.15	0.62	0.39	1.99	220
	(−0.35)	(−0.39)				**(4.87)**		[0.53]		
GJR + ATM IV	0.01		0.19	0.71			0.46	12.39	1.94	220
	(0.35)		(1.64)	**(5.19)**				**[0.00]**		
GJR + MFIV	−0.06		−0.01		1.07		0.57	0.01	1.74	220
	(−1.25)		(−0.07)		**(5.17)**			[0.93]		
GJR + RMFIV	−0.02		−0.19			1.34	0.64	10.71	1.80	220
	(−0.59)		(−1.85)			**(5.47)**		**[0.00]**		
ATM + MFIV	−0.06			−0.04	1.10		0.57	0.08	1.74	220
	(−1.47)			(−0.16)	**(3.77)**			[0.77]		
ATM + RMFIV	−0.01			0.02		1.06	0.62	0.05	2.06	220
	(−0.40)			(0.11)		**(4.50)**		[0.83]		

This table presents results from regressions of realized volatility on competing forecasts for the crude oil futures market. The dependent variable is realized volatility, estimated as follows:

$$RV_{t,T} = \sqrt{\frac{252}{T}\sum_{t=1}^{T}\left(\log\frac{F_{t,T}}{F_{t-1,T}}\right)^2}$$

where $RV_{t,T}$ refers to realized volatility between t and T. $F_{t,T}$ denotes the price at time t of the futures contract maturing at T. α and β_{HIST} denote the intercept and slope coefficients of historical volatility. Likewise, $\beta_{GJR}, \beta_{ATM}, \beta_{MFIV}$, and β_{RMFIV} refer to the slope coefficients of GJR, ATM IV, MFIV, and RMFIV, respectively. We report Newey–West t-statistics in brackets, computed with two lags. Adj R^2 reports the adjusted R^2 of the corresponding regression. Column "Wald" reports the Wald test statistic and associated p-value in square brackets. In univariate regressions, we restrict the intercept and slope estimates to be equal to zero and one, respectively. In multivariate regressions, we restrict the slope estimate of the model to the left to be equal to zero. DW and Nobs report the Durbin–Watson test statistic and the number of observations, respectively. Figures in bold indicate statistical significance at 5%.

signed rank tests to assess the statistical significance of differences between models.

In conducting our empirical analysis, we are careful to select three important markets, namely crude oil, heating oil and natural gas, that are purged of the host of data issues discussed in the volatility forecasting literature. These include asynchronous trading times, irregular expiration cycles, potentially imprecise dividend yield estimates, and limited range of strike prices to name but a few. We find compelling evidence to suggest that accounting for volatility risk premium significantly improves the volatility forecasting performance of MFIV. Typically, RMFIV yields the smallest average forecasting errors of all models. This is true for all loss functions and markets. More important, our formal statistical tests show that the difference between RMFIV and its competitors is not only economically large but also statistically significant. Our results are robust to alternative proxies for realized volatility and competing time-series models, further highlighting the importance of our findings.

The remainder of this paper proceeds as follows. Section 2 provides a brief overview of extant studies on volatility forecasting. Section 3 describes our dataset and empirical methodology. Section 4 discusses our main findings. Section 5 contains robustness checks. Finally, Section 6 concludes.

2. Literature

Arguably, one of the most controversial studies on the information content of option markets for volatility forecasting in equity markets is that by Canina and Figlewski (1993). The authors study the information content of option implied volatility and find that historical volatility forecasts are superior to option implied forecasts. Their findings cast doubt on the informational efficiency of options markets. In a subsequent study, Fleming (1998) reaches a different conclusion, reporting that forecasts based on option implied volatility outperform those based on historical volatility. Using a more refined econometric methodology based on non-overlapping data, Christensen and Prabhala (1998) corroborate this finding by showing that implied volatility outperforms historical forecasts.

Studying the US Dollar/Deutsche Mark and US Dollar/Yen markets, Guo (1996) documents the superior forecasting power of implied variance extracted from the Hull and White (1987) pricing formula. Similarly, Jorion (1995) examines the information content and predictive power of option implied volatility in the Deutsche Mark, Yen, and Swiss Franc markets. He reports that time-series models underperform option implied forecasts even when given the advantage of calibration over the whole sample. Relatedly, Martens and Zein (2004) compare the information content of implied volatility to time-series models that exploit high-frequency data in several markets, including the Yen/US Dollar market. They report that, forecasts based on intra day data sometimes outperform option implied forecasts. Finally, Charoenwong et al. (2009) assess the predictive power of implied volatility extracted from options traded on different venues namely the Philadelphia Stock Exchange, the Chicago Mercantile Exchange, and the over-the-counter (OTC) market. Their study concludes that, irrespective

Table 2
Univariate and encompassing forecasts for heating oil's 30-day realized volatility.

	α	β_{HIST}	β_{GJR}	β_{ATM}	β_{MFIV}	β_{RMFIV}	Adj R^2	Wald	DW	Nobs
HIST	0.20	0.40					0.15	44.19	1.93	216
	(7.11)	**(4.82)**						**[0.00]**		
GJR	0.19		0.42				0.14	35.42	1.86	216
	(5.95)		**(5.04)**					**[0.00]**		
ATM IV	0.15			0.54			0.18	17.87	1.62	216
	(5.64)			**(7.92)**				**[0.00]**		
MFIV	0.11				0.66		0.23	9.32	1.71	216
	(4.13)				**(9.23)**			**[0.00]**		
RMFIV	0.08					0.80	0.29	6.04	2.02	216
	(3.58)					**(9.90)**		**[0.00]**		
HIST + ATM IV	0.14	0.37		0.21			0.20	6.48	1.91	216
	(5.31)	**(3.29)**		**(2.23)**				**[0.01]**		
HIST + MFIV	0.11	0.11			0.55		0.23	1.72	1.86	216
	(4.15)	(1.37)			**(5.45)**			[0.19]		
HIST + RMFIV	0.08	−0.11				0.93	0.29	1.23	1.93	216
	(3.17)	(−1.83)				**(8.11)**		[0.27]		
GJR + ATM IV	0.13		0.17	0.41			0.19	3.51	1.82	216
	(4.97)		**(2.31)**	**(4.37)**				[0.06]		
GJR + MFIV	0.10		0.04		0.63		0.23	0.12	1.75	216
	(4.08)		(0.55)		**(7.42)**			[0.73]		
GJR + RMFIV	0.09		−0.27			1.10	0.30	5.65	1.84	216
	(3.22)		**(−2.45)**			**(5.94)**		**[0.02]**		
ATM + MFIV	0.10			−0.25	0.92		0.23	1.39	1.72	216
	(4.00)			(−0.74)	**(2.50)**			[0.24]		
ATM + RMFIV	0.09			−0.12		0.92	0.29	0.82	2.03	216
	(3.80)			(−0.48)		**(3.11)**		[0.37]		

This table presents results from regressions of realized volatility on competing forecasts for the heating oil futures market. The dependent variable is realized volatility, estimated as follows:

$$RV_{t,T} = \sqrt{\frac{252}{T}\sum_{t=1}^{T}\left(\log\frac{F_{t,T}}{F_{t-1,T}}\right)^2}$$

where $RV_{t,T}$ refers to realized volatility between t and T. $F_{t,T}$ denotes the price at time t of the futures contract maturing at T. α and β_{HIST} denote the intercept and slope coefficients of historical volatility. Likewise, β_{GJR}, β_{ATM}, β_{MFIV}, and β_{RMFIV} refer to the slope coefficients of GJR, ATM IV, MFIV, and RMFIV, respectively. We report Newey–West t-statistics in brackets, computed with two lags. Adj R^2 reports the adjusted R^2 of the corresponding regression. Column "Wald" reports the Wald test statistic and associated p-value in square brackets. In univariate regressions, we restrict the intercept and slope estimates to be equal to zero and one, respectively. In multivariate regressions, we restrict the slope estimate of the model to the left to be equal to zero. DW and Nobs report the Durbin–Watson test statistic and the number of observations, respectively. Figures in bold indicate statistical significance at 5%.

of the trading venue, implied volatility performs better than time-series forecasts. Szakmary et al. (2003) provide a comprehensive investigation of the forecasting ability of implied volatility for 20 commodity markets. They compare ATM implied volatility to GARCH and a simple moving average model. Their results point to the superiority of ATM implied forecasts compared to time-series models throughout the maturity of the contract. In a related study, Agnolucci (2009) considers a richer set of GARCH models that includes asymmetric specifications such as EGARCH and TGARCH and different distributions of the error term. Contrary to Szakmary et al. (2003), Agnolucci concludes that time-series models provide better forecasts than ATM IV in the crude oil market.

More recent work has improved the quality of option implied volatility forecasts by avoiding to rely on a specific option pricing model and a particular strike price. In a pioneering study, Jiang and Tian (2005) analyze the information content of MFIV in the S&P 500 index market and show that it subsumes ATM IV and historical volatility. Recent studies have investigated the robustness of this finding for international equity markets. Efforts in this direction include the work of Frijns et al. (2010), who investigate similar issues in the Australian stock market and report that MFIV outperforms GARCH and EWMA models.[2] In a similar vein, Cheng and Fung (2012) analyze the information content of MFIV in Hong Kong and report that ATM IV outperforms MFIV which in turn is superior to time-series models. Taylor et al. (2010) study the performances of ARCH, GJR GARCH, ATM IV and MFIV for individual equities between 1996 and 1999. They report that option implied forecasts outperform time-series models for the month ahead forecasting horizon. However, they find that MFIV is inferior to ATM IV. This result is most likely attributable to low liquidity and the small number of out-of-the-money (OTM) option prices available for individual equity options.

The aforementioned studies can be criticized on the grounds that they ignore the role of the volatility risk premium in their empirical investigations. This issue dates back to the work of Lamoureux and Lastrapes (1993), who examine the forecast quality of implied variance extracted from the option pricing model of Hull and White (1987) and reject the null hypothesis that *"available information cannot be used to improve the market's variance forecast embedded in observable prices."* This result leads them to conclude that *"one possible reason for the rejection of the null is that volatility risk is priced."* Alas, this conjecture remains largely unexplored.

In fact, since the seminal work of Lamoureux and Lastrapes (1993), only two studies have analyzed this hypothesis further. Favoring a fully parametric approach, Poteshman (2000) estimates implied volatility from the Heston (1993) model which, unlike the model of Hull and White (1987), allows for a volatility risk premium. Poteshman notes a reduction in the biasedness of implied volatility, suggesting that volatility risk premium might contribute to the reported bias. However, it is not entirely clear how robust this result is to model misspecification. This is particularly

[2] It is worth noticing, however, that the implied volatility index of Frijns et al. (2010) is based on the old rather than the new definition of VIX. Hence, their study should be viewed as a test of the information content of ATM options rather than the MFIV defined in Jiang and Tian (2005).

Table 3
Univariate and encompassing forecasts for natural gas's 30-day realized volatility.

	α	β_{HIST}	β_{GJR}	β_{ATM}	β_{MFIV}	β_{RMFIV}	Adj R^2	Wald	DW	Nobs
HIST	0.22	0.57					0.32	26.07	1.88	189
	(7.24)	**(8.56)**						**[0.00]**		
GJR	0.16		0.60				0.38	36.76	1.90	189
	(5.35)		**(10.02)**					**[0.00]**		
ATM IV	0.15			0.65			0.37	20.54	1.46	189
	(3.80)			**(8.46)**				**[0.00]**		
MFIV	0.08				0.78		0.48	15.02	1.58	189
	(2.38)				**(11.96)**			**[0.00]**		
RMFIV	0.10					0.85	0.44	7.69	1.58	189
	(3.20)					**(10.95)**		**[0.00]**		
HIST + ATM IV	0.12	0.45		0.29			0.42	14.93	1.78	189
	(3.28)	**(5.40)**		**(3.19)**				**[0.00]**		
HIST + MFIV	0.08	0.07			0.72		0.48	0.69	1.67	189
	(2.36)	(0.57)			**(6.56)**			[0.41]		
HIST + RMFIV	0.10	0.08				0.77	0.44	0.77	1.66	189
	(3.23)	(0.64)				**(5.52)**		[0.38]		
GJR + ATM IV	0.09		0.37	0.38			0.45	26.41	1.84	189
	(2.48)		**(3.43)**	**(3.53)**				**[0.00]**		
GJR + MFIV	0.06		0.20		0.61		0.50	6.32	1.81	189
	(1.87)		(1.16)		**(3.70)**			**[0.01]**		
GJR + RMFIV	0.09		0.22			0.62	0.46	6.65	1.79	189
	(2.74)		(1.23)			**(2.90)**		**[0.01]**		
ATM + MFIV	0.08			0.01	0.77		0.48	0.01	1.58	189
	(2.29)			(0.11)	**(9.08)**			[0.94]		
ATM + RMFIV	0.09			0.19		0.66	0.45	3.37	1.58	189
	(2.66)			**(2.14)**		**(5.99)**		[0.07]		

This table presents results from regressions of realized volatility on competing forecasts for the natural gas futures market. The dependent variable is realized volatility, estimated as follows:

$$RV_{t,T} = \sqrt{\frac{252}{T}\sum_{t=1}^{T}\left(\log\frac{F_{t,T}}{F_{t-1,T}}\right)^2}$$

where $RV_{t,T}$ refers to realized volatility between t and T. $F_{t,T}$ denotes the price at time t of the futures contract maturing at T. α and β_{HIST} denote the intercept and slope coefficients of historical volatility. Likewise, β_{GJR}, β_{ATM}, β_{MFIV}, and β_{RMFIV} refer to the slope coefficients of GJR, ATM IV, MFIV, and RMFIV, respectively. We report Newey–West t-statistics in brackets, computed with two lags. Adj R^2 reports the adjusted R^2 of the corresponding regression. Column "Wald" reports the Wald test statistic and associated p-value in square brackets. In univariate regressions, we restrict the intercept and slope estimates to be equal to zero and one, respectively. In multivariate regressions, we restrict the slope estimate of the model to the left to be equal to zero. DW and Nobs report the Durbin–Watson test statistic and the number of observations, respectively. Figures in bold indicate statistical significance at 5%.

important given the empirical evidence of Bakshi et al. (1997) and Eraker et al. (2003), among others, that the Heston model is misspecified.

Chernov (2007) devises a novel strategy which builds on an approximation linking the average implied volatilities of near-the-money (NTM) option contracts and the quadratic variation. He shows, both theoretically and empirically, that the existence of a volatility risk premium leads to biased volatility forecasts. However, by focusing on NTM options, he neglects potentially useful information embedded in other option prices.

3. Data and methodology

3.1. Data

Our survey of extant studies features mixed empirical evidence on the relative merits of option implied volatility forecasts. This lack of consensus could be due to potential measurement errors. First, in markets such as the S&P index option markets, the option and underlying markets close at different times. As a result, the non-synchronous closing times of stock and option markets introduce an error in the implementation of the option pricing model used to extract implied volatility. For example, back-of-the-envelope calculations in Jorion (1995) suggest that asynchronous trading times could bias implied volatility estimates by 1.2%. Second, extracting implied volatility from stock options requires dividend yield estimates. Unfortunately, these data are often difficult to obtain, leaving researchers with no choice but to make ad hoc assumptions about dividend yields. Third, the MFIV approach of Jiang and Tian (2005) requires integrating over an infinite range of strike prices. As these integrals need to be approximated over a finite number of discrete strike prices, they obviously depend on the existence of a sufficiently wide range of OTM option contracts. Finally, most existing studies rely on relatively short sample periods.[3] However, few options markets have monthly expiration cycles. As a result, using non-overlapping samples may significantly lower the power of statistical tests.

In light of these concerns, we view energy futures markets as the perfect testing ground for our study. They are well suited for our analysis, for several reasons. To begin with, Marshall et al. (2012) note that these markets are highly liquid. Moreover, more than 20 years of option data are available and wide ranges of strike prices are traded in these markets. Furthermore, the options are written on the corresponding futures contract, thus enabling us to avoid estimating storage costs and convenience yields (or, equivalently, dividends if studying equity markets). Finally, the futures and option contracts of these markets are traded on the same exchange and close at the same time, allaying the concerns related to non-synchronous closing times often encountered in equity indices.

Specifically, we consider futures and option settlement prices for crude oil, heating oil, and natural gas traded at NYMEX.[4] Our sample period extends from January 1989 to September 2011 and

[3] For example, Jiang and Tian (2005) consider data from 1988 to 1994. Likewise, Taylor et al. (2010) consider data from 1996 to 1999.

[4] In 2008, NYMEX was acquired by the CME Group; however, the name NYMEX still prevails.

Table 4
Forecasting errors: 30-day horizon.

	HIST (%)	GJR (%)	ATM IV (%)	MFIV (%)	RMFIV (%)
Panel A: Mean Absolute Errors (MAE)					
Crude oil	8.84	9.52	8.85	8.41	6.95
Heating oil	9.92	9.67	8.78	8.11	7.41
Natural gas	14.02	14.50	13.11	11.73	10.78
Panel B: Mean Squared Errors (MSE)					
Crude oil	1.90	3.32	1.81	1.47	1.20
Heating oil	2.98	2.86	2.38	2.08	1.87
Natural gas	3.44	3.41	3.02	2.36	2.37
Panel C: Mean Absolute Percentage Errors (MAPE)					
Crude oil	24.36	27.30	26.62	26.31	19.43
Heating oil	28.74	29.78	25.36	23.66	20.18
Natural gas	29.05	31.88	28.07	25.67	21.25
Panel D: Mean Squared Percentage Errors (MSPE)					
Crude oil	9.68	15.49	12.15	11.75	6.37
Heating oil	17.07	18.91	11.22	9.85	6.92
Natural gas	14.52	17.43	13.19	11.32	7.37

This table reports the out-of-sample forecast errors of each volatility model for realized volatility over a horizon of 30 days. Realized volatility is defined as:

$$RV_{t,T} = \sqrt{\frac{252}{T}\sum_{t=1}^{T}\left(\log\frac{F_{t,T}}{F_{t-1,T}}\right)^2}$$

where $RV_{t,T}$ refers to realized volatility between t and T. $F_{t,T}$ denotes the price at time t of the futures contract maturing at T. Panels A and B report the mean absolute errors (MAE) and mean squared errors (MSE) of individual models, respectively. Panels C and D report the mean absolute percentage errors (MAPE) and mean squared percentage errors (MSPE), respectively.

Table 5
Differences of forecasting errors: AE and SE.

	HIST (%)	GJR (%)	ATM IV (%)	MFIV (%)	RMFIV (%)
Panel A: Crude oil (AE)					
HIST		−0.69	−0.01	0.42	**1.89**
GJR	−0.02		0.68	1.11	**2.57**
ATM IV	0.21	−0.52		0.43	**1.90**
MFIV	0.53	−0.40	0.07		**1.46**
RMFIV	**−1.03**	**−1.28**	**−1.98**	**−2.23**	
Panel B: Crude oil (SE)					
HIST		−1.41	0.10	0.44	**0.70**
GJR	0.00		1.51	1.85	2.11
ATM IV	0.02	−0.02		0.34	**0.60**
MFIV	0.04	−0.03	0.00		**0.26**
RMFIV	**−0.07**	**−0.08**	**−0.15**	**−0.18**	
Panel C: Heating oil (AE)					
HIST		0.25	1.14	**1.81**	**2.51**
GJR	−0.37		0.89	**1.56**	**2.25**
ATM IV	−0.59	−0.53		**0.67**	**1.36**
MFIV	**−0.71**	−0.55	**−0.29**		**0.69**
RMFIV	**−1.71**	**−1.27**	**−1.02**	**−0.56**	
Panel D: Heating oil (SE)					
HIST		0.12	0.60	0.90	**1.10**
GJR	−0.03		0.48	**0.78**	**0.99**
ATM IV	−0.03	−0.03		**0.30**	**0.51**
MFIV	**−0.06**	−0.03	**−0.02**		0.21
RMFIV	**−0.13**	**−0.08**	**−0.07**	−0.04	
Panel E: Natural gas (AE)					
HIST		−0.49	0.91	**2.28**	**3.23**
GJR	0.31		1.39	**2.77**	**3.72**
ATM IV	−0.01	**−1.49**		**1.38**	**2.33**
MFIV	**−1.73**	**−2.55**	**−0.75**		0.95
RMFIV	**−2.74**	**−3.02**	**−2.91**	−1.99	
Panel F: Natural gas (SE)					
HIST		0.03	0.42	**1.08**	**1.07**
GJR	0.04		0.39	**1.04**	**1.04**
ATM IV	0.00	**−0.18**		**0.65**	**0.65**
MFIV	**−0.23**	**−0.35**	**−0.07**		0.00
RMFIV	**−0.39**	**−0.37**	**−0.40**	−0.26	

This table reports relative differences in the performance of competing models. The upper triangular matrices report the mean difference of absolute (AE) and squared (SE) forecasting errors, respectively. Similarly, the lower triangular matrices report the median difference of forecasting errors. We compute the difference between the errors of model [*name in row*] and those of model [*name in column*]. For example, the first row of Panel A presents the average difference in the AE of HIST vis-à-vis GJR, ATM IV, MFIV, and RMFIV, respectively. The numbers in bold indicate statistically significant differences at 5% as indicated by the Diebold–Mariano (computed with 2 lags) and the Wilcoxon signed rank tests for the upper and lower triangular matrices, respectively.

November 1978 until September 2011 for the options and futures data, respectively.[5] All data have been obtained from the Commodity Research Bureau (CRB). As proxy for the risk-free rate, we employ three-month Treasury bill rates obtained from the Federal Reserve's website.

In order to mitigate problems due to stale option prices, we discard all observations with prices lower than five times the minimum tick size set by the exchange. These minimum values are \$0.01, \$0.0001, and \$0.001 for crude oil, heating oil and natural gas, respectively. The option dataset comprises American options. Therefore, we follow Trolle and Schwartz (2009) and convert them into European option prices by approximating the early exercise premium using the method developed by Barone-Adesi and Whaley (1987). Since futures contracts have specific maturity dates, particular caution must be exerted when computing returns around rollover dates. To mitigate problems due to spurious jumps around rollover dates, we discard all returns computed across different futures contracts.

Studies on forecasting realized volatility are invariably faced with the issue of overlapping observations bias induced by the rolling window used to estimate realized volatility. Canina and Figlewski (1993) and Jorion (1995) estimate realized volatility on a daily basis by employing rolling windows of returns, introducing overlapping observation biases in their analysis. Although these studies account for the overlapping periods by adjusting the standard errors, it is not entirely clear how effective this adjustment is. Therefore, Christensen and Prabhala (1998) recommend using non-overlapping data. We follow their advice. Whenever possible, we retain only options that mature in exactly 30 days. If this is not possible, we select the nearest trading day.[6] In these instances, a small adjustment to the implied volatility is required. For example, if the nearest time to maturity is 31 days, we adjust the corresponding implied volatility so as to reflect the volatility of 30 days and not 31 days. Since the option markets under consideration have monthly expiration cycles, our research design results in non-overlapping observations, making our analysis robust to the overlapping observations biases discussed by Christensen and Prabhala (1998).

3.2. Methodology

Realized volatility. Following Jorion (1995) among many others, realized volatility between t and T, $RV_{t,T}$, is computed in the usual way:

$$RV_{t,T} = \sqrt{\frac{252}{T}\sum_{t=1}^{T}\left(\log\frac{F_{t,T}}{F_{t-1,T}}\right)^2} \quad (1)$$

where $F_{t,T}$ denotes the price at time t of the futures contract maturing at T.[7]

[5] Specifically, the futures dataset begins from 03/30/1983, 11/14/1978, and 04/04/1990 in the crude oil, heating oil, and natural gas markets, respectively. Similarly, the options dataset is available from 01/11/1989, 01/11/1989, and 10/02/1992 for the crude oil, heating oil, and natural gas markets, respectively.

[6] In selecting the nearest trading day, we restrict ourselves to maturities between 28 and 32 days.

[7] In Section 5, we consider the alternative range estimator proposed by Garman and Klass (1980) and refined by Yang and Zhang (2000) as a robustness check.

Table 6
Differences of forecasting errors: APE and SPE.

	HIST (%)	GJR (%)	ATM IV (%)	MFIV (%)	RMFIV (%)
Panel A: Crude oil (APE)					
HIST		−2.94	−2.26	−1.95	**4.93**
GJR	−0.10		0.68	0.99	**7.87**
ATM IV	0.63	−2.14		0.31	**7.19**
MFIV	2.05	−1.27	0.17		**6.88**
RMFIV	**−3.14**	**−4.08**	**−6.62**	**−7.81**	
Panel B: Crude oil (SPE)					
HIST		−5.82	**−2.47**	−2.07	**3.31**
GJR	−0.02		3.34	3.75	**9.12**
ATM IV	0.20	−0.38		0.41	**5.78**
MFIV	0.52	−0.33	0.03		**5.38**
RMFIV	**−0.71**	**−0.99**	**−1.73**	**−2.21**	
Panel C: Heating oil (APE)					
HIST		−1.04	3.38	**5.07**	**8.55**
GJR	−1.26		**4.42**	**6.12**	**9.59**
ATM IV	−1.73	−1.59		**1.70**	**5.18**
MFIV	**−2.37**	−1.74	**−0.96**		**3.48**
RMFIV	**−5.14**	**−4.03**	**−3.83**	**−1.99**	
Panel D: Heating oil (SPE)					
HIST		−1.83	5.85	**7.22**	**10.16**
GJR	−0.20		**7.69**	**9.05**	**11.99**
ATM IV	−0.27	−0.27		**1.37**	**4.30**
MFIV	**−0.71**	−0.34	**−0.23**		**2.94**
RMFIV	**−1.38**	**−1.03**	**−0.97**	**−0.45**	
Panel E: Natural gas (APE)					
HIST		**−2.84**	0.98	**3.38**	**7.80**
GJR	0.69		**3.81**	**6.21**	**10.63**
ATM IV	−0.04	**−3.88**		**2.40**	**6.82**
MFIV	−3.74	**−5.96**	**−1.88**		**4.42**
RMFIV	**−6.05**	**−6.54**	**−6.31**	**−4.93**	
Panel F: Natural gas (SPE)					
HIST		−2.92	1.32	3.20	**7.15**
GJR	0.16		**4.24**	**6.12**	**10.06**
ATM IV	−0.01	**−1.33**		**1.88**	**5.82**
MFIV	−1.06	**−1.85**	**−0.43**		**3.95**
RMFIV	**−2.11**	**−1.48**	**−1.63**	**−1.25**	

This table reports relative differences in the performance of competing models. The upper triangular matrices report the mean difference of absolute percentage (APE) and squared percentage (SPE) forecasting errors, respectively. Similarly, the lower triangular matrices report the median difference of forecasting errors. We compute the difference between the errors of model [*name in row*] and those of model [*name in column*]. For example, the first row of Panel A presents the average difference in the APE of HIST vis-à-vis GJR, ATM IV, MFIV, and RMFIV, respectively. The numbers in bold indicate statistically significant differences at 5% as indicated by the Diebold–Mariano (computed with 2 lags) and the Wilcoxon signed rank tests for the upper and lower triangular matrices, respectively.

Time-series models. We consider two time-series models namely HIST and asymmetric GJR GARCH. HIST simply denotes the realized volatility defined in Eq. (1) over the preceding month. The GJR GARCH model is specified as follows:

$$\begin{aligned} y_t &= \mu + \epsilon_t; \quad \epsilon_t \sim N(0, \sigma_t^2); \\ \sigma_t^2 &= \omega + (\alpha + \gamma I_{t-1})\epsilon_{t-1}^2 + \beta h_{t-1}^2 \end{aligned} \tag{2}$$

where y_t is the daily log return of the underlying futures prices at date t. μ is the mean return, and ϵ_t represents the price innovation, which is normally distributed with mean zero and variance σ_t^2. I_{t-1} is an indicator function taking values one or zero if y_{t-1} is lower or greater than μ, respectively. Notice that the model in Eq. (2) nests the simple GARCH model, which can be obtained by constraining γ to be equal to zero.[8]

We recursively estimate the model parameters using futures returns and iteratively obtain 30 day ahead forecasts of realized variance.[9] Specifically, our initial parameter estimates are extracted from futures data until the first option observation.[10] With the parameter estimates at hand, we iteratively forecast variance for the next 30 days. Then, we expand our estimation window by 30 days and re-estimate the time-series model. Again, we make variance forecasts for the next 30 days and repeat our procedure until the end of the sample.

Option implied forecasts. We obtain ATM IV by averaging the Black (1976) implied volatilities of options with moneyness ranging from 0.97 to 1.03, with moneyness being defined as the ratio of the futures price over the strike price.[11] To obtain MFIV, we closely follow the steps outlined in Prokopczuk and Wese Simen (2012) and compute

$$MFIV_{t,T} = \sqrt{\frac{2e^{r_t(T-t)}}{T-t}\left[\int_0^{F_{t,T}} \frac{P(K)}{K^2}\,dK + \int_{F_{t,T}}^{+\infty} \frac{C(K)}{K^2}\,dK\right]} \tag{3}$$

where $MFIV_{t,T}$ refers to the MFIV between days t and T. r_t denotes the annualized risk-free rate. $P(K)$ and $C(K)$ denote European put and call options struck at K and expiring at T, respectively.

To obtain $MFIV_{t,T}$, we proceed as follows. We rank all out-of-the-money (OTM) options by time to maturity on a daily basis. Since energy options have a monthly expiration cycle, the first two maturities always span a period of 30 days. Hence, we retain options of the shortest (T_1) and second shortest maturities (T_2) only. Observations on trading days with less than two OTM put and two OTM call options per maturity are discarded. This step is important since the computation of MFIV requires several OTM options. We truncate the two integrals in Eq. (3) at the lower and upper bounds (strike prices) $K_l = F_{t,T}e^{-10\sigma T}$ and $K_u = F_{t,T}e^{10\sigma T}$, respectively. Here $F_{t,T}$ refers to the price at time t of the futures contract expiring at T, σ denotes the average implied volatility of all OTM options maturing at T (either T_1 or T_2).

We perform a linear interpolation of all implied volatilities across the market strike range for T_1 and T_2 (separately). For strikes outside of this range but between the truncation points, we assume constant implied volatility. Overall, this approach yields 1000 implied volatilities for each maturity, which we map into European option prices by applying the Black (1976) option pricing formula. Next, we implement the trapezoidal rule to numerically evaluate the integrands in Eq. (3) using the 1000 option prices. By doing so, we obtain the risk-neutral expectation of variance for each maturity. Lastly, we obtain the 30 day MFIV by performing a linear interpolation between the two risk-neutral expectations of variance.

Risk premium adjusted MFIV. The central theme of this study is the role of volatility risk premium for volatility forecasting. We introduce a non-parametric adjustment inspired by previous studies on variance risk premia. Specifically, Bollerslev et al. (2009) estimate the market price of variance risk as the difference between the risk-neutral and physical expectations of variance:

$$VRP_{t,T}^2 = \mathbb{E}_t^Q\left(V_{t,T}^2\right) - \mathbb{E}_t^P\left(V_{t,T}^2\right) \tag{4}$$

where $VRP_{t,T}^2$ refers to the variance risk premium between t and T. $\mathbb{E}_t^Q\left(V_{t,T}^2\right)$ is the ex ante forecast of variance under the risk-neutral measure. This is equivalent to $MFIV_{t,T}$, defined as above. $\mathbb{E}_t^P\left(V_{t,T}^2\right)$

8 As a robustness check, we also repeated the entire analysis using the EGARCH and GARCH (1,1) models. We also allow for an ARMA component in the return equation and obtain nearly identical results. See Section 5 for further details.

9 We adopt a recursive rather than a rolling window estimation because Lamoureux and Lastrapes (1993) compare the two methods and find that the recursively estimated GARCH model provides better performance.

10 This means that we obtain the first set of parameter estimates using 1546, 2501, and 637 returns in the crude oil, heating oil, and natural gas futures markets, respectively.

11 Our selection of the ATM range mirrors that of Bakshi et al. (1997).

Table 7
Univariate and encompassing forecasts for crude oil's 30-day realized volatility (range estimator).

	α	β_{HIST}	β_{GJR}	β_{ATM}	β_{MFIV}	β_{RMFIV}	Adj R^2	Wald	DW	Nobs
HIST	0.09	0.75					0.57	16.77	2.22	220
	(3.21)	**(8.53)**						**[0.00]**		
GJR	0.18		0.44				0.36	100.53	1.66	220
	(3.63)		**(3.04)**					**[0.00]**		
ATM IV	0.04			0.81			0.44	10.27	1.54	220
	(1.39)			**(8.23)**				**[0.00]**		
MFIV	−0.02				0.96		0.60	16.03	1.71	220
	(−0.63)				**(8.66)**			**[0.00]**		
RMFIV	0.02					0.97	0.64	1.04	1.93	220
	(0.70)					**(10.32)**		[0.35]		
HIST + ATM IV	0.03	0.31		0.57			0.60	89.77	2.29	220
	(1.34)	**(4.12)**		**(5.55)**				**[0.00]**		
HIST + MFIV	0.00	0.34			0.60		0.63	17.65	2.07	220
	(0.01)	**(2.48)**			**(3.80)**			**[0.00]**		
HIST + RMFIV	0.02	0.18				0.77	0.64	3.59	2.06	220
	(0.89)	(0.91)				**(3.22)**		[0.06]		
GJR + ATM IV	0.05		0.21	0.58			0.48	19.31	1.90	220
	(1.83)		(1.57)	**(4.25)**				**[0.00]**		
GJR + MFIV	−0.02		0.02		0.94		0.60	0.15	1.74	220
	(−0.50)		(0.20)		**(4.95)**			[0.70]		
GJR + RMFIV	0.01		−0.09			1.09	0.64	3.30	1.80	220
	(0.36)		(−1.08)			**(5.80)**		[0.07]		
ATM + MFIV	−0.02			−0.14	1.09		0.60	1.52	1.69	220
	(−0.57)			(−0.66)	**(4.19)**			[0.22]		
ATM + RMFIV	0.02			−0.06		1.01	0.64	0.38	1.91	220
	(1.00)			(−0.38)		**(5.36)**		[0.54]		

This table presents results from regressions of realized volatility on competing forecasts for the crude oil market. The dependent variable is realized volatility, estimated as follows:

$$RV^z_{t,T} = \sqrt{\frac{252}{T}\sum\nolimits_{t=1}^{T}(\log O_t - \log C_{t-1})^2 + \frac{1}{2}(\log H_t - \log L_t)^2 - (2\log 2 - 1)(\log C_t - \log O_t)^2}$$

where O_t, H_t, and L_t denote the opening, intra day high, and low prices of the underlying on trading day t, respectively. C_{t-1} and C_t refer to the previous and current closing prices, respectively. α and β_{HIST} denote the intercept and slope coefficients of historical volatility. Likewise, $\beta_{GJR}, \beta_{ATM}, \beta_{MFIV}$, and β_{RMFIV} refer to the slope coefficients of GJR, ATM IV, MFIV, and RMFIV, respectively. We report Newey–West t-statistics computed with 2 lags in brackets. Adj R^2 reports the adjusted R^2 of the corresponding regression. Column "Wald" reports the Wald test statistic and associated p-value in square brackets. In univariate regressions, we restrict the intercept and slope estimates to be equal to zero and one, respectively. In multivariate regressions, we restrict the slope estimate of the model to the left to be equal to zero. DW and Nobs report the Durbin–Watson test statistic and the number of observations, respectively. Figures in bold indicate statistical significance at 5%.

is the ex ante forecast of variance under the physical measure, proxied by the ex post realized variance.

Carr and Wu (2009) and Prokopczuk and Wese Simen (2012) show that variance risk premia are significantly different from zero and, in absolute terms, increasing in variance. This level dependency has profound implications for our analysis. To see why, consider a two-period setting. Suppose periods one and two are characterized by high and moderate variance, respectively. Everything else equal, the level dependency implies a variance risk premium of larger magnitude in period one than in period two. Using this high market price of variance risk of period one to forecast realized variance in period two might lead to biased forecasts. Therefore, it is preferable to obtain *relative* estimates of variance risk premia.

We obtain these *relative* variance risk premia by computing the ratio (instead of the difference as in Eq. (4)) of the expected variance under the risk-neutral measure over the expectation of variance under the physical measure as follows:

$$RVRP^2_{t,T} = \frac{\mathbb{E}^Q_t\left(V^2_{t,T}\right)}{\mathbb{E}^P_t\left(V^2_{t,T}\right)} \tag{5}$$

where $RVRP^2_{t,T}$ refers to the relative variance risk premium between t and T. Carr and Wu (2009) and Trolle and Schwartz (2010) report that, contrary to the variance risk premia defined as in Eq. (4), relative variance risk premia are independent of the level of variance. Building on these insights, we estimate the average relative variance risk premium over a period of just under one year:

$$ARVRP^2_t = \frac{1}{252-\tau}\sum_{j=t-252}^{t-\tau} RVRP^2_{j,j+\tau} \tag{6}$$

$$= \frac{1}{252-\tau}\sum_{j=t-252}^{t-\tau} \frac{MFIV^2_{j,j+\tau}}{RV^2_{j,j+\tau}} \tag{7}$$

where $ARVRP^2_t$ is the average relative variance risk premium between $t-252$ and $t-\tau$. τ denotes the forecasting horizon. We refer to the square root of relative variance risk premium as the volatility risk premium.[12]

We obtain the risk-premium adjusted MFIV for the period t to T ($RMFIV_{t,T}$) by re-arranging Eq. (5)[13,14]:

$$RMFIV_{t,T} = \sqrt{\mathbb{E}^P_t\left(V^2_{t,T}\right)} = \frac{MFIV_{t,T}}{ARVRP_t}. \tag{8}$$

[12] Clearly, the length of the estimation window requires some trade-off. In particular, it must be large enough so that the volatility risk premium can be estimated with sufficient precision. Yet, it should not be too long in order to reflect recent market conditions. We view a period of just under one year (approx. 232 days) as a good trade-off. We experimented with an 18-month estimation window and obtained qualitatively similar results. These are available upon request.

[13] To obtain ex post estimates of RVRP, we assume that the ex ante forecast of realized volatility is unbiased. In other words, the ex post realized volatility equals the ex ante forecast of volatility made under the same probability measure. Similar assumptions are made in Carr and Wu (2009), for example.

[14] DeMiguel et al. (2012) employ a similar adjustment when studying the role of option implied moments in an asset allocation context.

Table 8
Univariate and encompassing forecasts for heating oil's 30-day realized volatility (range estimator).

	α	β_{HIST}	β_{GJR}	β_{ATM}	β_{MFIV}	β_{RMFIV}	Adj R^2	Wald	DW	Nobs
HIST	0.15	0.55					0.28	28.58	2.03	216
	(4.35)	**(5.09)**						**[0.00]**		
GJR	0.17		0.45				0.21	45.34	1.81	216
	(5.29)		**(5.02)**					**[0.00]**		
ATM IV	0.16			0.53			0.23	26.14	1.47	216
	(5.99)			**(7.81)**				**[0.00]**		
MFIV	0.10				0.67		0.32	12.67	1.59	216
	(4.62)				**(10.78)**			**[0.00]**		
RMFIV	0.09					0.76	0.39	8.60	1.96	216
	(4.86)					**(12.99)**		**[0.00]**		
HIST + ATM IV	0.12	0.25		0.38			0.30	23.12	1.99	216
	(4.64)	**(2.26)**		**(3.00)**				**[0.00]**		
HIST + MFIV	0.09	0.25			0.46		0.34	8.54	1.93	216
	(4.45)	**(2.34)**			**(4.26)**			**[0.00]**		
HIST + RMFIV	0.09	0.02				0.74	0.39	0.04	1.98	216
	(4.89)	(0.19)				**(6.08)**		[0.84]		
GJR + ATM IV	0.13		0.25	0.34			0.26	10.31	1.79	216
	(5.01)		**(2.85)**	**(3.40)**				**[0.00]**		
GJR + MFIV	0.10		0.09		0.59		0.32	1.33	1.71	216
	(4.47)		(1.94)		**(7.46)**			[0.25]		
GJR + RMFIV	0.09		−0.13			0.89	0.39	2.03	1.85	216
	(4.80)		(−1.41)			**(7.33)**		[0.16]		
ATM + MFIV	0.10			−0.42	1.10		0.33	5.85	1.63	216
	(4.40)			(−1.30)	**(3.20)**			**[0.02]**		
ATM + RMFIV	0.10			−0.17		0.92	0.39	2.51	1.98	216
	(4.49)			(−0.91)		**(4.75)**		[0.11]		

This table presents results from regressions of realized volatility on competing forecasts for the heating oil market. The dependent variable is realized volatility, estimated as follows:

$$RV^z_{t,T} = \sqrt{\frac{252}{T}\sum_{t=1}^{T}(\log O_t - \log C_{t-1})^2 + \frac{1}{2}(\log H_t - \log L_t)^2 - (2\log 2 - 1)(\log C_t - \log O_t)^2}$$

where O_t, H_t, and L_t denote the opening, intra day high, and low prices of the underlying on trading day t, respectively. C_{t-1} and C_t refer to the previous and current closing prices, respectively. α and β_{HIST} denote the intercept and slope coefficients of historical volatility. Likewise, $\beta_{GJR}, \beta_{ATM}, \beta_{MFIV}$, and β_{RMFIV} refer to the slope coefficients of GJR, ATM IV, MFIV, and RMFIV, respectively. We report Newey–West t-statistics computed with 2 lags in brackets. Adj R^2 reports the adjusted R^2 of the corresponding regression. Column "Wald" reports the Wald test statistic and associated p-value in square brackets. In univariate regressions, we restrict the intercept and slope estimates to be equal to zero and one, respectively. In multivariate regressions, we restrict the slope estimate of the model to the left to be equal to zero. DW and Nobs report the Durbin–Watson test statistic and the number of observations, respectively. Figures in bold indicate statistical significance at 5%.

4. Empirical evidence

In this section, we evaluate the information content and forecasting accuracy of competing forecasts. The former basically answers the question whether there is *some* useful information in the individual volatility forecasts. The latter addresses the question as to which of the forecasts is the most accurate and might be considered more relevant for practical application.

4.1. Information content

As is common in the forecasting literature, we estimate Mincer–Zarnowitz regressions. Specifically, we regress monthly realized volatility on our volatility forecasts as follows[15]:

$$RV_{t,T} = \alpha + \beta f_{t,T} + \epsilon_t \tag{9}$$

where $RV_{t,T}$ refers to the realized volatility from t to T and $f_{t,T}$ is either one volatility forecast or a vector containing several competing forecasts for volatility until T at time t. ϵ_t denotes the error term. These regressions are suitable for testing the unbiasedness and efficiency of individual forecasts. Briefly, we test for unbiasedness in univariate regressions (i.e. $f_{t,T}$ contains only one particular forecast) by imposing the restriction that α and β are jointly equal to zero and one, respectively. Testing for efficiency consists in ascertaining whether alternative models contain information beyond that of a baseline model. To do so, we constrain the slope of alternative forecasts to zero in encompassing regressions. Tables 1–3 present the results for the three markets considered.

Univariate regressions. If a volatility forecast contains some information for future volatility, then in univariate regressions, its slope coefficients must be statistically distinguishable from zero and, ideally close to one. The explanatory power should also be sizable.

The upper parts of Tables 1–3 present the results of univariate regressions for each market. We can observe statistically significant slope estimates in every instance, implying that each model's forecast is informative about next month's volatility. However, the coefficients of the time-series models are far away from one. Considering option implied forecasts, we can see that, as we move from ATM IV to RMFIV, the slope estimates rise steadily. For the heating oil and natural gas markets, this steady increase brings the coefficients much closer to one. Conversely, we notice a decrease in the magnitude of the intercepts in the crude oil and heating oil futures markets. These observations indicate that RMFIV is less biased than MFIV, which in turn is less biased than ATM IV.

These findings motivate us to investigate the unbiasedness of individual forecasts more formally. In doing so, we report in the column headed "Wald" the F-statistic testing the null hypothesis of α and β being jointly equal to zero and one, respectively. The

[15] Notice that some researchers prefer to estimate the model in logs. Our decision to estimate the model in level terms rather than in logs is motivated by Poteshman (2000), who notes that the biasedness of implied volatility might be due to the logarithmic transformation which introduces an upward bias. Repeating our analysis using log volatility does not affect our main findings.

Table 9
Univariate and encompassing forecasts for natural gas's 30-day realized volatility (range estimator).

	α	β_{HIST}	β_{GJR}	β_{ATM}	β_{MFIV}	β_{RMFIV}	Adj R^2	Wald	DW	Nobs
HIST	0.22	0.57					0.33	26.49	1.84	189
	(6.17)	**(7.46)**						**[0.00]**		
GJR	0.17		0.60				0.46	49.90	1.84	189
	(6.08)		**(11.30)**					**[0.00]**		
ATM IV	0.15			0.66			0.46	27.47	1.43	189
	(4.44)			**(9.73)**				**[0.00]**		
MFIV	0.10				0.75		0.55	21.95	1.62	189
	(3.61)				**(14.24)**			**[0.00]**		
RMFIV	0.10					0.83	0.52	6.32	1.67	189
	(3.94)					**(14.89)**		**[0.00]**		
HIST + ATM IV	0.11	0.50		0.24			0.49	13.13	1.70	189
	(3.64)	**(6.16)**		**(2.56)**				**[0.00]**		
HIST + MFIV	0.09	0.02			0.74		0.55	0.05	1.65	189
	(3.29)	(0.14)			**(7.60)**			[0.82]		
HIST + RMFIV	0.10	0.01				0.82	0.52	0.02	1.68	189
	(3.74)	(0.10)				**(6.86)**		[0.88]		
GJR + ATM IV	0.09		0.36	0.39			0.55	38.89	1.79	189
	(2.95)		**(3.66)**	**(4.08)**				**[0.00]**		
GJR + MFIV	0.08		0.24		0.54		0.58	13.79	1.87	189
	(2.83)		(1.57)		**(3.84)**			**[0.00]**		
GJR + RMFIV	0.08		0.26			0.57	0.55	14.22	1.88	189
	(3.35)		(1.57)			**(3.19)**		**[0.00]**		
ATM + MFIV	0.09			0.12	0.64		0.55	1.55	1.59	189
	(3.35)			(1.36)	**(7.19)**			[0.22]		
ATM + RMFIV	0.09			0.21		0.62	0.53	4.95	1.61	189
	(3.46)			**(2.21)**		**(6.20)**		**[0.03]**		

This table presents results from regressions of realized volatility on competing forecasts for the natural gas market. The dependent variable is realized volatility, estimated as follows:

$$RV^{z}_{t,T} = \sqrt{\frac{252}{T}\sum\nolimits_{t=1}^{T}(\log O_t - \log C_{t-1})^2 + \frac{1}{2}(\log H_t - \log L_t)^2 - (2\log 2 - 1)(\log C_t - \log O_t)^2}$$

where O_t, H_t, and L_t denote the opening, intra day high, and low prices of the underlying on trading day t, respectively. C_{t-1} and C_t refer to the previous and current closing prices, respectively. α and β_{HIST} denote the intercept and slope coefficients of historical volatility. Likewise, $\beta_{GJR}, \beta_{ATM}, \beta_{MFIV}$, and β_{RMFIV} refer to the slope coefficients of GJR, ATM IV, MFIV, and RMFIV, respectively. We report Newey–West t-statistics computed with 2 lags in brackets. Adj R^2 reports the adjusted R^2 of the corresponding regression. Column "Wald" reports the Wald test statistic and associated p-value in square brackets. In univariate regressions, we restrict the intercept and slope estimates to be equal to zero and one, respectively. In multivariate regressions, we restrict the slope estimate of the model to the left to be equal to zero. DW and Nobs report the Durbin–Watson test statistic and the number of observations, respectively. Figures in bold indicate statistical significance at 5%.

corresponding p-values are given in square brackets. Our results indicate a rejection of the null hypothesis at 5% in every instance, suggesting that all forecasts are biased predictors of future volatility. Nonetheless, it is instructive to examine the values of the test statistic more closely. The near monotonic decrease in the test statistic across all markets is noteworthy. Overall, RMFIV yields the smallest test statistic in every market. In the case of crude oil, the null hypothesis of unbiasedness cannot any longer be rejected at the 2% level. Taken together, these findings conform to the theoretical argument that the market price of volatility risk contributes to the biasedness of option implied forecasts.

Encompassing regressions. We now turn to the issue of relative informational efficiency, which we address through encompassing regressions, i.e. we run regression (9) with $f_{t,T}$ containing more than one volatility forecast. If one forecast is more informative than another, then it will **(i)** exhibit a highly significant slope estimate in encompassing regressions and/or **(ii)** significantly improve the explanatory power of the restricted model.

The lower parts of Tables 1–3 present the results of these encompassing regressions. We begin by assessing the relative merits of option implied forecasts. One can observe that MFIV and RMFIV subsume ATM IV in all markets, i.e. the coefficient of MFIV or RMFIV is significant and that of ATM IV is not. This is evidenced by the highly robust test statistic of the slope estimate. The only exception to this pattern is the natural gas market. Furthermore, the adjusted R^2 of MFIV and RMFIV are unaffected by the addition of ATM IV to the baseline models. To further validate our findings, we check the efficiency of candidate models again through Wald tests. To this end, we restrict all slope estimates, excluding that of the model to the right, to be equal to zero. The results suggest that the null hypothesis cannot be rejected in all markets including natural gas, providing more evidence that MFIV and RMFIV are relatively efficient compared to ATM IV.

We now turn to the predictive power of the models. Our intuition is simple: if a model has high predictive power, then it should be able to explain variations in realized volatility. In this respect, the adjusted R^2 reported in Tables 1–3 are particularly enlightening. They show substantial variations in explanatory power across markets. For example, the adjusted R^2 fluctuates between 0.32 and 0.62 in the crude oil market. In contrast, we observe a smaller range, from 0.14 to 0.29, for the heating oil market.

Focusing on individual models, we see that RMFIV yields the highest explanatory power in the crude and heating oil markets. It is also worth highlighting that MFIV always exhibits higher explanatory power than ATM IV. We also find that time-series models do not explain as much variation in realized volatility as MFIV and RMFIV. Taking the crude oil market as an example, we notice that RMFIV achieves an adjusted R^2 of 62%. In contrast, GJR leads to a fit of 32%, almost 50% smaller than the corresponding figure for RMFIV. Augmenting RMFIV with GJR results in a negligible increase in goodness of fit from 62% to 64%.

In summary, all models are informative about next month's volatility. In line with previous studies, we find that all forecasts are biased, but to different degrees. Accounting for the volatility risk premium reduces the magnitude of the bias. We also show that RMFIV is highly efficient: it subsumes other forecasts. Finally, MFIV

Table 10
Forecast errors: 30-day horizon (range estimator).

	HIST (%)	GJR (%)	ATM IV (%)	MFIV (%)	RMFIV (%)
Panel A: Mean Absolute Errors (MAE)					
Crude oil	6.71	8.14	7.90	7.29	5.78
Heating oil	7.12	8.02	7.76	6.89	5.99
Natural gas	11.58	12.39	11.22	10.11	8.64
Panel B: Mean Squared Errors (MSE)					
Crude oil	1.17	2.92	1.45	1.08	0.86
Heating oil	1.72	2.11	1.80	1.43	1.24
Natural gas	2.76	2.66	2.26	1.80	1.65
Panel C: Mean Absolute Percentage Errors (MAPE)					
Crude oil	18.18	22.78	23.07	21.89	15.76
Heating oil	19.72	23.12	21.44	19.24	15.79
Natural gas	22.99	25.48	22.77	20.46	16.41
Panel D: Mean Squared Percentage Errors (MSPE)					
Crude oil	5.26	12.10	8.86	7.61	4.03
Heating oil	9.27	11.69	7.86	6.01	4.36
Natural gas	10.84	10.70	8.10	6.41	4.43

This table reports the out-of-sample forecast errors of each volatility model for realized volatility over a horizon of 30 days. We estimate realized volatility using the estimator proposed by Garman and Klass (1980) and refined by Yang and Zhang (2000):

$$RV^{z}_{t,T} = \sqrt{\frac{252}{T}\sum_{t=1}^{T}(\log O_t - \log C_{t-1})^2 + \frac{1}{2}(\log H_t - \log L_t)^2 - (2\log 2 - 1)(\log C_t - \log O_t)^2}$$

where O_t, H_t, and L_t denote the opening, intra day high, and low prices of the underlying on trading day t, respectively. C_{t-1} and C_t refer to the previous and current closing prices, respectively. Panels A and B report the mean absolute errors (MAE) and mean squared errors (MSE) of individual models, respectively. Panels C and D report the mean absolute percentage errors (MAPE) and mean squared percentage errors (MSPE), respectively.

Table 11
Forecasting errors: 30-day horizon (more GARCH models).

	EGARCH (%)	GJR (%)	GARCH (%)	ATM IV (%)	MFIV (%)	RMFIV (%)
Panel A: Mean Absolute Errors (MAE)						
Crude oil	10.37	9.52	9.44	8.85	8.41	6.95
Heating oil	8.76	9.67	9.72	8.78	8.11	7.41
Natural gas	12.42	14.50	14.39	13.11	11.73	10.78
Panel B: Mean Squared Errors (MSE)						
Crude oil	3.93	3.32	3.19	1.81	1.47	1.20
Heating oil	2.22	2.86	2.92	2.38	2.08	1.87
Natural gas	2.59	3.41	3.39	3.02	2.36	2.37
Panel C: Mean Absolute Percentage Errors (MAPE)						
Crude oil	31.49	27.30	27.13	26.62	26.31	19.43
Heating oil	27.42	29.78	29.92	25.36	23.66	20.18
Natural gas	27.43	31.88	31.92	28.07	25.67	21.25
Panel D: Mean Squared Percentage Errors (MSPE)						
Crude oil	20.36	15.49	15.29	12.15	11.75	6.37
Heating oil	14.16	18.91	19.28	11.22	9.85	6.92
Natural gas	12.51	17.43	17.62	13.19	11.32	7.37

This table reports the out-of-sample forecast errors of each volatility model for realized volatility over a horizon of 30 days. Realized volatility is defined as:

$$RV_{t,T} = \sqrt{\frac{252}{T}\sum_{t=1}^{T}\left(\log\frac{F_{t,T}}{F_{t-1,T}}\right)^2}$$

where $RV_{t,T}$ refers to realized volatility between t and T. $F_{t,T}$ denotes the price at time t of the futures contract maturing at T. Panels A and B report the mean absolute errors (MAE) and mean squared errors (MSE) of individual models, respectively. Panels C and D report the mean absolute percentage errors (MAPE) and mean squared percentage errors (MSPE), respectively.

dominates ATM IV and time-series models, confirming its theoretical potential.

4.2. Forecasting accuracy

We now turn to the question of out-of-sample forecasting accuracy, which might be considered more important in practice. To do this, we evaluate the forecasting errors by four commonly employed loss functions. Specifically, we use the MAE, MSE, MAPE and MSPE. These loss functions are defined as follows:

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|RV_{t,T} - f_{t,T}| \tag{10}$$

$$MSE = \frac{1}{n}\sum_{t=1}^{n}(RV_{t,T} - f_{t,T})^2 \tag{11}$$

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{RV_{t,T} - f_{t,T}}{RV_{t,T}}\right| \tag{12}$$

$$MSPE = \frac{1}{n}\sum_{t=1}^{n}\left(\frac{RV_{t,T} - f_{t,T}}{RV_{t,T}}\right)^2 \tag{13}$$

Table 12
Differences of forecasting errors: AE and SE (more GARCH models).

	EGARCH (%)	GJR (%)	GARCH (%)	ATM IV (%)	MFIV (%)	RMFIV (%)
Panel A: Crude oil (AE)						
EGARCH		**0.85**	**0.93**	1.52	1.96	**3.42**
GJR	−**1.28**		0.08	0.68	1.11	**2.57**
GARCH	−**1.03**	−0.02		0.59	1.03	**2.49**
ATM IV	−0.93	−0.52	−0.28		0.43	**1.90**
MFIV	−**1.11**	−0.40	−0.46	0.07		**1.46**
RMFIV	−**1.98**	−**1.28**	−**1.23**	−**1.98**	−**2.23**	
Panel B: Crude oil (SE)						
EGARCH		0.61	0.74	2.12	2.46	2.72
GJR	−**0.08**		0.13	1.51	1.85	2.11
GARCH	−**0.09**	0.00		1.38	1.72	1.98
ATM	−**0.06**	−0.02	−0.02		0.34	**0.60**
MFIV	−**0.11**	−0.03	−0.04	0.00		**0.26**
RMFIV	−**0.22**	−**0.08**	−**0.07**	−**0.15**	−**0.18**	
Panel C: Heating oil (AE)						
EGARCH		−**0.90**	−**0.95**	−0.02	0.66	**1.35**
GJR	0.18		−0.05	0.89	**1.56**	**2.25**
GARCH	0.15	0.01		0.94	**1.61**	**2.30**
ATM IV	−0.11	−0.53	−0.41		**0.67**	**1.36**
MFIV	−0.70	−0.55	−0.53	−**0.29**		**0.69**
RMFIV	−**1.09**	−**1.27**	−**1.35**	−**1.02**	−**0.56**	
Panel D: Heating oil (SE)						
EGARCH		−0.64	−0.70	−0.16	0.14	**0.34**
GJR	0.01		−0.06	0.48	**0.78**	**0.99**
GARCH	0.01	0.00		0.54	0.84	**1.04**
ATM IV	0.00	−0.03	−0.02		**0.30**	**0.51**
MFIV	−0.04	−0.03	−0.03	−**0.02**		0.21
RMFIV	−**0.08**	−**0.08**	−**0.07**	−**0.07**	−0.04	
Panel E: Natural gas (AE)						
EGARCH		−**2.08**	−**1.97**	−0.69	0.69	**1.64**
GJR	**1.07**		0.11	1.39	**2.77**	**3.72**
GARCH	**0.83**	−0.04		1.28	**2.66**	**3.61**
ATM IV	−0.29	−**1.49**	−1.45		**1.38**	**2.33**
MFIV	−1.16	−**2.55**	−**2.15**	−**0.75**		0.95
RMFIV	−**1.92**	−**3.02**	−**3.15**	−**2.91**	−1.99	
Panel F: Natural gas (SE)						
EGARCH		−**0.82**	−**0.80**	−0.43	0.23	0.22
GJR	**0.14**		0.02	0.39	**1.04**	**1.04**
GARCH	**0.08**	−0.01		0.38	**1.03**	**1.02**
ATM IV	−0.04	−**0.18**	−0.20		**0.65**	**0.65**
MFIV	−0.14	−**0.35**	−**0.28**	−**0.07**		0.00
RMFIV	−**0.23**	−**0.37**	−**0.34**	−**0.40**	−0.26	

This table reports relative differences in the performance of competing models. The upper triangular matrices report the mean difference of absolute (AE) and squared (SE) forecasting errors, respectively. Similarly, the lower triangular matrices report the median difference of forecasting errors. We compute the difference between the errors of model [*name in row*] and those of model [*name in column*]. For example, the first row of Panel A presents the average difference in the AE of HIST vis-à-vis GJR, ATM IV, MFIV, and RMFIV, respectively. The numbers in bold indicate statistically significant differences at 5% as indicated by the Diebold–Mariano (computed with 2 lags) and the Wilcoxon signed rank tests for the upper and lower triangular matrices, respectively.

where n and $RV_{t,T}$ denote the number of forecast windows and realized volatility, respectively. $f_{t,T}$ is a volatility forecast obtained from one of the following models: HIST, GJR, ATM IV, MFIV, and RMFIV. Obviously, a good model should minimize the forecasting error.

Table 4 summarizes the forecast errors. Starting with MAE in Panel A, we observe that option markets produce more accurate forecasts than time-series models. The only exception being that, in the crude oil futures market, ATM IV underperforms historical volatility by as little as 0.01%.[16] Interestingly, RMFIV yields the smallest forecast errors in each market. Similarly, MFIV provides the second-best forecast of volatility. Directly comparing MFIV to RMFIV sheds light on the importance of our volatility risk premium adjustment. The results indicate that accounting for the market price of volatility risk reduces the mean absolute forecast error of MFIV by a factor as high as 17.36% in the crude oil market.

We report our findings based on MSE in Panel B. The main findings are unchanged. Option markets provide more accurate volatility forecasts than time-series models. The overall ranking is broadly identical: RMFIV, MFIV, ATM IV, GJR, and HIST, in decreasing order of accuracy. There is, however, one exception in the natural gas futures market. RMFIV leads to pricing errors of 2.37%, which are slightly higher than the 2.36% of MFIV.[17] Again, the difference between ATM IV and MFIV is noticeable, especially in the more volatile natural gas market where, ATM IV and MFIV yield errors equal to 3.02% and 2.36%, respectively.

So far, our analysis has been concerned with the level of forecast errors. Interestingly, the MAE reported for the natural gas futures market are an order of magnitude higher than those of the crude oil and heating oil futures markets. This is not too surprising given that the natural gas market counts among the most volatile commodity markets. Against this backdrop, it is prudent to assess competing forecasts based on relative forecast errors. Panel C of Table 4 reports forecast errors based on MAPE. Notice that the errors reported for different markets are of the same order of magnitude.

[16] As one might suspect, further analysis shows that this difference is not statistically significant.

[17] As we shall see in Table 5, this difference is not statistically significant.

Table 13
Differences of forecasting errors: APE and SPE (more GARCH models).

	EGARCH (%)	GJR (%)	GARCH (%)	ATM IV (%)	MFIV (%)	RMFIV (%)
Panel A: Crude oil (APE)						
EGARCH		**4.19**	**4.36**	**4.87**	**5.18**	**12.06**
GJR	−**3.59**		0.17	0.68	0.99	**7.87**
GARCH	−**3.59**	−0.07		0.51	0.82	**7.70**
ATM IV	−**2.69**	−2.14	−0.99		0.31	**7.19**
MFIV	−**3.24**	−1.27	−1.30	0.17		**6.88**
RMFIV	−**6.40**	−**4.08**	−**4.33**	−**6.62**	−**7.81**	
Panel B: Crude oil (SPE)						
EGARCH		**4.86**	**5.06**	**8.21**	**8.61**	**13.99**
GJR	−**0.83**		0.20	3.34	3.75	**9.12**
GARCH	−**0.95**	−0.02		3.14	3.55	**8.92**
ATM IV	−**0.68**	−0.38	−0.36		0.41	**5.78**
MFIV	−**0.94**	−0.33	−0.35	0.03		**5.38**
RMFIV	−**2.09**	−**0.99**	−**0.69**	−**1.73**	−**2.21**	
Panel C: Heating oil (APE)						
EGARCH		−**2.36**	−**2.50**	2.06	**3.76**	**7.24**
GJR	0.50		−0.14	**4.42**	**6.12**	**9.59**
GARCH	0.48	0.03		**4.56**	**6.26**	**9.74**
ATM IV	−0.36	−1.59	−0.94		**1.70**	**5.18**
MFIV	−1.93	−1.74	−1.56	−**0.96**		**3.48**
RMFIV	−**3.35**	−**4.03**	−**4.20**	−**3.83**	−**1.99**	
Panel D: Heating oil (SPE)						
EGARCH		−**4.74**	−**5.12**	2.94	**4.31**	**7.25**
GJR	0.07		−0.37	**7.69**	**9.05**	**11.99**
GARCH	0.06	0.01		**8.06**	**9.43**	**12.36**
ATM IV	−0.03	−0.27	−0.12		**1.37**	**4.30**
MFIV	−0.42	−0.34	−0.32	−**0.23**		**2.94**
RMFIV	−**0.77**	−**1.03**	−**0.92**	−**0.97**	−**0.45**	
Panel E: Natural gas (APE)						
EGARCH		−**4.45**	−**4.48**	−0.64	1.76	**6.18**
GJR	**1.78**		−0.03	**3.81**	**6.21**	**10.63**
GARCH	**1.89**	−0.12		**3.85**	**6.24**	**10.67**
ATM IV	−0.85	−**3.88**	−**3.68**		**2.40**	**6.82**
MFIV	−1.95	−**5.96**	−**4.58**	−**1.88**		**4.42**
RMFIV	−**3.34**	−**6.54**	−**6.90**	−**6.31**	−**4.93**	
Panel F: Natural gas (SPE)						
EGARCH		−**4.92**	−**5.10**	−0.68	1.20	**5.14**
GJR	**0.57**		−0.18	**4.24**	**6.12**	**10.06**
GARCH	**0.46**	−0.05		**4.42**	**6.30**	**10.25**
ATM IV	−0.14	−**1.33**	−1.26		**1.88**	**5.82**
MFIV	−0.60	−**1.85**	−**1.26**	−**0.43**		**3.95**
RMFIV	−**1.16**	−**1.48**	−**2.09**	−**1.63**	−**1.25**	

This table reports relative differences in the performance of competing models. The upper triangular matrices report the mean difference of absolute percentage (APE) and squared percentage (SPE) forecasting errors, respectively. Similarly, the lower triangular matrices report the median difference of forecasting errors. We compute the difference between the errors of model [*name in row*] and those of model [*name in column*]. For example, the first row of Panel A presents the average difference in the APE of HIST vis-à-vis GJR, ATM IV, MFIV, and RMFIV, respectively. The numbers in bold indicate statistically significant differences at 5% as indicated by the Diebold–Mariano (computed with 2 lags) and the Wilcoxon signed rank tests for the upper and lower triangular matrices, respectively.

Again, this loss function leads to a ranking similar to the one emanating from Panel A. Specifically, MFIV and RMFIV outperform ATM IV and all the time-series models. The crude oil market represents however a notable exception: historical volatility outperforms ATM IV and MFIV. Notwithstanding this exception, RMFIV remains the best forecast. Adjusting for the volatility risk premium substantially reduces the errors of MFIV from 26.31%, 23.66% and 25.67% to 19.43%, 20.18%, and 21.25% in the crude oil, heating oil, and natural gas futures markets, respectively. Finally, Panel D reports results based on MSPE. These results are broadly similar to those obtained from MAPE. Briefly, RMFIV dominates its competitors: its forecast errors are an order of magnitude smaller than those of all other forecasts, including MFIV.

In order to assess these important results in greater detail, we investigate whether the observed differences in performance are statistically significant. Tables 5 and 6 present the mean differences in the absolute errors (AE), the squared errors (SE), the absolute percentage errors (APE), and the squared percentage errors (SPE) in the upper triangular matrices. As a robustness check, we present the median differences in forecast errors in the lower triangular matrices. We compute the differences between the forecast errors of model [*name in row*] and those of model [*name in column*]. For example, looking at Panel A of Table 5, we can see that on average absolute forecast errors of RMFIV are 1.46% smaller than those of MFIV. The median figures show that the absolute forecast errors of RMFIV are 2.23% smaller than those of MFIV. The figures in bold indicate statistical significance at the 5% level. The mean differences are tested with the Diebold–Mariano statistic calculated with 2 lags. The median differences are assessed through the non-parametric Wilcoxon signed rank test.

Several observations are in order. First, the differences in performance between time-series models are not statistically significant. Second, with the exception of the crude oil market, there is a statistically significant difference between MFIV and ATM IV. Third, and most importantly, RMFIV is statistically superior to all other forecasts. This is true for all markets and loss functions. The AE and SE of the natural gas futures market, where no significant differences between MFIV and RMFIV are observed, are the only exceptions to this pattern. In most cases, the improvement in forecast accuracy achieved after adjusting for the volatility risk

Table 14
Forecasting errors: 30-day horizon (ARMA-GARCH models).

	ARMA–EGARCH (%)	ARMA–GJR (%)	ARMA–GARCH (%)	ATM IV (%)	MFIV (%)	RMFIV (%)
Panel A: Mean Absolute Errors (MAE)						
Crude oil	10.36	9.63	9.46	8.85	8.41	6.95
Heating oil	8.80	9.71	9.76	8.78	8.11	7.41
Natural gas	12.56	14.55	14.55	13.11	11.73	10.78
Panel B: Mean Squared Errors (MSE)						
Crude oil	3.97	3.48	3.30	1.81	1.47	1.20
Heating oil	2.23	2.88	2.94	2.38	2.08	1.87
Natural gas	2.68	3.46	3.47	3.02	2.36	2.37
Panel C: Mean Absolute Percentage Errors (MAPE)						
Crude oil	31.46	27.46	27.18	26.62	26.31	19.43
Heating oil	27.57	29.91	30.08	25.36	23.66	20.18
Natural gas	27.58	32.00	32.24	28.07	25.67	21.25
Panel D: Mean Squared Percentage Errors (MSPE)						
Crude oil	20.45	15.85	15.58	12.15	11.75	6.37
Heating oil	14.36	19.01	19.40	11.22	9.85	6.92
Natural gas	12.61	17.44	17.76	13.19	11.32	7.37

This table reports the out-of-sample forecast errors of each volatility model for realized volatility over a horizon of 30 days. Realized volatility is defined as:

$$RV_{t,T} = \sqrt{\frac{252}{T}\sum_{t=1}^{T}\left(\log\frac{F_{t,T}}{F_{t-1,T}}\right)^2}$$

where $RV_{t,T}$ refers to realized volatility between t and T. $F_{t,T}$ denotes the price at time t of the futures contract maturing at T. Panels A and B report the mean absolute errors (MAE) and mean squared errors (MSE) of individual models, respectively. Panels C and D report the mean absolute percentage errors (MAPE) and mean squared percentage errors (MSPE), respectively.

premium is of substantial magnitude, showing that the results are also economically significant.

5. Robustness checks

In this section, we study the robustness of our results. We follow a three-pronged approach. First, we investigate the robustness of our findings using the range estimator of realized volatility proposed by Garman and Klass (1980) and refined by Yang and Zhang (2000). Second, we expand the pool of time-series models to include EGARCH and the simple GARCH (1,1). We also investigate the effect of alternative specifications of the return equation by adding an ARMA component. Third, we analyze the robustness of our main results to an alternative approach to estimate the volatility risk premium.

5.1. Alternative estimator of realized volatility

Accurately measuring realized volatility has been the focus of several studies.[18] In a pioneering study, Andersen and Bollerslev (1998) discuss the effect of noisy volatility proxies on the forecasting performance of GARCH models. The authors demonstrate, both theoretically and empirically, the "*importance of proper ex post evaluation criteria when assessing volatility forecasts*".

In light of these considerations, we repeat our analysis after replacing the classical volatility estimator described in Eq. (1) with the more efficient estimator of volatility developed by Garman and Klass (1980) and refined by Yang and Zhang (2000). This estimator is given as[19]

$$RV^z_{t,T} = \sqrt{\frac{252}{T}\sum_{t=1}^{T}(\log O_t - \log C_{t-1})^2 + \frac{1}{2}(\log H_t - \log L_t)^2 - (2\log 2 - 1)(\log C_t - \log O_t)^2} \quad (14)$$

where O_t, H_t, and L_t denote the opening, the daily high, and the daily low prices of the underlying on trading day t, respectively; C_{t-1} and C_t refer to the previous and current closing prices, respectively. This estimator can be loosely described as a high-frequency estimator. We specifically select this estimator because, in addition to capturing the highest and lowest intra day prices, it also contains information from overnight returns.

Tables 7–9 repeat the regression analyses. It can be readily seen that the results do not change significantly. If anything, they are more supportive of our main finding that RMFIV provides the best volatility forecasts. Briefly, the intercept and slope estimates converge toward zero and one as we progress from ATM IV to RMFIV, confirming that RMFIV is less biased than MFIV and ATM IV. In fact, the Wald test p-value of 0.35 for the crude oil futures market indicates that the RMFIV forecasts are unbiased. Although the hypothesis is rejected in the other two markets, we still observe a sharp decline of the F-statistic. Taking the natural gas futures market as an example, the F-statistic falls from 21.95 to 6.32. Turning to the multivariate regressions, we observe that RMFIV subsumes all models. This result is corroborated by the Wald statistics

[18] See Andersen et al. (2010) for an excellent survey.

[19] Alternatively, one could employ the realized volatility estimator described in Andersen and Bollerslev (1998) and account for overnight returns as in Jiang and Tian (2005). Unfortunately, it requires intra day data for a 20-year period which are not available. We view the range estimator presented in Eq. (14) as a viable alternative. Its good performance has been demonstrated in multiple studies including Garman and Klass (1980), Yang and Zhang (2000) and Shu and Zhang (2006).

Table 15
Differences of forecasting errors: AE and SE (ARMA-GARCH models).

	ARMA–EGARCH (%)	ARMA–GJR (%)	ARMA–GARCH (%)	ATM IV (%)	MFIV (%)	RMFIV (%)
Panel A: Crude Oil (AE)						
ARMA–EGARCH		**0.73**	**0.90**	1.52	1.95	**3.41**
ARMA–GJR	**−1.20**		0.17	0.79	1.22	**2.68**
ARMA–GARCH	**−1.13**	−0.01		0.61	1.05	**2.51**
ATM IV	−0.95	−0.49	−0.19		0.43	**1.90**
MFIV	**−1.20**	−0.56	−0.41	0.07		**1.46**
RMFIV	**−1.87**	**−1.33**	**−1.22**	**−1.98**	**−2.23**	
Panel B: Crude Oil (SE)						
ARMA–EGARCH		0.49	0.67	2.16	2.50	2.77
ARMA–GJR	**−0.08**		0.18	1.67	2.01	2.27
ARMA–GARCH	**−0.10**	0.00		1.49	1.83	2.09
ATM IV	−0.05	−0.03	−0.01		0.34	**0.60**
MFIV	**−0.11**	−0.03	−0.04	0.00		**0.26**
RMFIV	**−0.20**	**−0.09**	**−0.07**	**−0.15**	**−0.18**	
Panel C: Heating Oil (AE)						
ARMA–EGARCH		**−0.91**	**−0.96**	0.02	**0.69**	**1.39**
ARMA–GJR	0.18		−0.05	0.93	**1.60**	**2.30**
ARMA–GARCH	0.17	0.03		0.98	**1.65**	**2.35**
ATM IV	−0.10	−0.50	−0.32		**0.67**	**1.36**
MFIV	−0.74	−0.43	−0.46	**−0.29**		**0.69**
RMFIV	**−1.17**	**−1.23**	**−1.40**	**−1.02**	**−0.56**	
Panel D: Heating Oil (SE)						
ARMA–EGARCH		−0.65	−0.71	−0.15	0.15	**0.36**
ARMA–GJR	**0.01**		−0.06	0.50	0.80	**1.01**
ARMA–GARCH	0.01	0.00		0.56	0.86	**1.07**
ATM IV	0.00	−0.03	−0.02		**0.30**	**0.51**
MFIV	−0.03	−0.04	−0.03	**−0.02**		0.21
RMFIV	**−0.09**	**−0.09**	**−0.07**	**−0.07**	−0.04	
Panel E: Natural Gas (AE)						
ARMA–EGARCH		**−1.99**	**−1.99**	−0.55	0.83	**1.78**
ARMA–GJR	**0.86**		0.00	1.44	**2.82**	**3.76**
ARMA–GARCH	**0.99**	0.03		1.44	**2.82**	**3.77**
ATM IV	−0.28	**−1.67**	**−1.77**		**1.38**	**2.33**
MFIV	−0.95	**−2.38**	**−2.52**	**−0.75**		0.95
RMFIV	**−2.08**	**−3.25**	**−3.42**	**−2.91**	−1.99	
Panel F: Natural Gas (SE)						
ARMA–EGARCH		**−0.78**	**−0.79**	−0.34	0.32	0.31
ARMA–GJR	**0.15**		−0.02	0.44	**1.09**	**1.09**
ARMA–GARCH	**0.15**	0.00		0.46	**1.11**	**1.10**
ATM IV	−0.05	**−0.22**	−0.26		**0.65**	**0.65**
MFIV	−0.17	**−0.33**	**−0.39**	**−0.07**		0.00
RMFIV	**−0.28**	**−0.43**	**−0.38**	**−0.40**	−0.26	

This table reports relative differences in the performance of competing models. The upper triangular matrices report the mean difference of absolute (AE) and squared (SE) forecasting errors, respectively. Similarly, the lower triangular matrices report the median difference of forecasting errors. We compute the difference between the errors of model [*name in row*] and those of model [*name in column*]. For example, the first row of Panel A presents the average difference in the AE of HIST vis-à-vis GJR, ATM IV, MFIV, and RMFIV, respectively. The numbers in bold indicate statistically significant differences at 5% as indicated by the Diebold–Mariano (computed with 2 lags) and the Wilcoxon signed rank tests for the upper and lower triangular matrices, respectively.

which cannot reject the null of efficiency. These findings are true for all markets except natural gas, where GJR remains statistically significant.

Table 10 repeats the tests of forecasting accuracy, i.e. it reports forecast errors employing the range-based estimator of volatility. Comparing these figures to those reported in Table 4 sheds light on the importance of efficient volatility estimates. Overall, the errors based on the range estimator are smaller than those obtained using the classical estimator. This striking difference in magnitude highlights the importance of efficient volatility estimates. Most importantly, our major findings are broadly unchanged. In particular, RMFIV still dominates all rival forecasts with improvements of substantial magnitude that should be considered economically significant. We have, of course, also evaluated the statistical significance and the results are largely unchanged, i.e. the forecast errors of RMFIV are significantly smaller than those of MFIV. We do not report these results to save space but they are available upon request.

5.2. Different GARCH models

Our study may be criticized on the grounds that the GJR–GARCH is only one representative from a large family of potential models. Hence, it is important to analyze other GARCH-type models to ensure our results are robust. Guided by this idea, we consider the simple GARCH (1,1) and the EGARCH model in addition to the GJR–GARCH model. Table 11 reports the results of this analysis. Panels A, B, C and D report the MAE, MSE, MAPE and MSPE, respectively. Columns 2 through 7 relate to the EGARCH, GJR, GARCH, ATM IV, MFIV and RMFIV models, respectively.

We observe some differences across GARCH-type models. For example, EGARCH, GJR and the simple GARCH yield MSPE equal to 20.36%, 15.49% and 15.29% in the crude oil market, respectively. More importantly, the RMFIV yields the smallest MSPE (6.37%). Overall, we find that RMFIV is superior to all other models, including EGARCH, GJR and GARCH. This is true for the MAE, MSE, MAPE and MSPE, lending more credence to our main finding. Tables 12

Table 16
Differences of forecasting errors: APE and SPE (ARMA-GARCH models).

	ARMA–EGARCH (%)	ARMA–GJR (%)	ARMA–GARCH (%)	ATM IV (%)	MFIV (%)	RMFIV (%)
Panel A: Crude oil (APE)						
ARMA–EGARCH		**4.00**	**4.28**	**4.83**	**5.15**	**12.03**
ARMA–GJR	**−3.30**		0.28	0.84	1.15	**8.03**
ARMA–GARCH	**−3.54**	−0.06		0.56	0.87	**7.75**
ATM IV	**−2.87**	−1.99	−0.85		0.31	**7.19**
MFIV	**−3.44**	−1.63	−1.27	0.17		**6.88**
RMFIV	**−6.42**	**−3.93**	**−4.69**	**−6.62**	**−7.81**	
Panel B: Crude oil (SPE)						
ARMA–EGARCH		**4.60**	**4.87**	**8.30**	**8.71**	**14.09**
ARMA–GJR	**−0.76**		0.27	3.70	4.11	**9.49**
ARMA–GARCH	**−0.84**	−0.01		3.43	3.83	**9.21**
ATM IV	**−0.61**	−0.46	−0.35		0.41	**5.78**
MFIV	**−1.13**	−0.41	−0.38	0.03		**5.38**
RMFIV	**−2.00**	**−1.14**	**−0.74**	**−1.73**	**−2.21**	
Panel C: Heating oil (APE)						
ARMA–EGARCH		**−2.34**	**−2.50**	2.21	**3.91**	**7.39**
ARMA–GJR	0.48		−0.17	**4.55**	**6.25**	**9.72**
ARMA–GARCH	0.49	0.10		**4.72**	**6.42**	**9.89**
ATM IV	−0.40	−1.54	−0.91		**1.70**	**5.18**
MFIV	−1.86	**−1.60**	−1.45	**−0.96**		**3.48**
RMFIV	**−3.64**	**−4.06**	**−4.01**	**−3.83**	**−1.99**	
Panel D: Heating oil (SPE)						
ARMA–EGARCH		**−4.65**	**−5.04**	3.14	**4.51**	**7.45**
ARMA–GJR	0.07		−0.39	**7.79**	**9.15**	**12.09**
ARMA–GARCH	0.12	0.01		**8.18**	**9.55**	**12.49**
ATM IV	−0.04	−0.26	−0.16		**1.37**	**4.30**
MFIV	−0.36	**−0.29**	−0.33	**−0.23**		**2.94**
RMFIV	**−0.95**	**−1.11**	**−0.94**	**−0.97**	**−0.45**	
Panel E: Natural gas (APE)						
ARMA–EGARCH		**−4.42**	**−4.66**	−0.49	1.91	**6.33**
ARMA–GJR	**1.83**		−0.24	**3.93**	**6.33**	**10.76**
ARMA–GARCH	**2.07**	0.06		**4.17**	**6.57**	**10.99**
ATM IV	−0.67	**−3.92**	**−3.63**		**2.40**	**6.82**
MFIV	−2.57	**−5.82**	**−5.27**	**−1.88**		**4.42**
RMFIV	**−3.49**	**−6.75**	**−7.08**	**−6.31**	**−4.93**	
Panel F: Natural gas (SPE)						
ARMA–EGARCH		**−4.83**	**−5.15**	−0.59	1.29	**5.24**
ARMA–GJR	**0.62**		−0.32	**4.24**	**6.12**	**10.07**
ARMA–GARCH	**0.85**	0.01		**4.57**	**6.44**	**10.39**
ATM IV	−0.29	**−1.32**	**−1.14**		**1.88**	**5.82**
MFIV	−0.57	**−1.92**	**−1.44**	**−0.43**		**3.95**
RMFIV	**−1.31**	**−1.85**	**−1.91**	**−1.63**	**−1.25**	

This table reports relative differences in the performance of competing models. The upper triangular matrices report the mean difference of absolute percentage (APE) and squared percentage (SPE) forecasting errors, respectively. Similarly, the lower triangular matrices report the median difference of forecasting errors. We compute the difference between the errors of model [*name in row*] and those of model [*name in column*]. For example, the first row of Panel A presents the average difference in the APE of HIST vis-à-vis GJR, ATM IV, MFIV, and RMFIV, respectively. The numbers in bold indicate statistically significant differences at 5% as indicated by the Diebold–Mariano (computed with 2 lags) and the Wilcoxon signed rank tests for the upper and lower triangular matrices, respectively.

and 13 also formally show that the RMFIV is not only economically but also statistically superior to all three GARCH specifications. This result is supported by both the Diebold–Mariano (computed with 2 lags) and the Wilcoxon signed rank tests.

One may also question the specification of the return equation in the GARCH-type models. Since commodity prices exhibit price trends and mean reverting patterns, it may be interesting to include an autoregressive moving average (ARMA) component in the return equation to capture these features. We modify the return equation of all three time-series models, i.e. EGARCH, GJR and GARCH, to include an ARMA (1,1) component. Specifically, we analyze the ARMA–EGARCH, the ARMA–GJR and the ARMA–GARCH model.

Table 14 summarizes our findings. We can directly compare these results to those of Table 11, which deals with time-series models that feature a constant return. We do not discern big differences between the baseline models with a constant return and their extensions with an ARMA component in the return equation. Taking the MAE of the crude oil market for example, the baseline (with a constant return) EGARCH, GJR and GARCH models yield 10.37%, 9.52% and 9.44%, respectively. Augmenting the return equation with an ARMA process results in roughly similar MAE 10.36%, 9.63% and 9.46% for the ARMA–EGARCH, ARMA–GJR and ARMA–GARCH models, respectively. More important for our analysis, RMFIV outperforms all time-series models, indicating that our results are robust to alternative specifications of the return process of GARCH-type models. We also report the results of the Diebold–Mariano (computed with 2 lags) and the Wilcoxon signed rank tests in Tables 15 and 16. These tests unequivocally show that RMFIV provides significantly better forecasts than its competitors.

5.3. Modeling the volatility risk premium

Finally, we also analyze the robustness of our key findings to the approach used to estimate the volatility risk premium.[20] Each

[20] We thank a referee for this suggestion.

Table 17
Forecasting errors: 30-day horizon (ARMA model for RVRP).

	HIST (%)	GJR (%)	ATM IV (%)	MFIV (%)	RMFIV (%)
Panel A: Mean Absolute Errors (MAE)					
Crude oil	8.36	9.00	8.01	7.66	6.18
Heating oil	8.57	8.39	7.39	6.88	6.07
Natural gas	14.22	14.66	13.31	11.84	10.43
Panel B: Mean Squared Errors (MSE)					
Crude oil	1.54	2.87	1.09	0.93	0.75
Heating oil	1.79	1.61	0.99	0.82	0.71
Natural gas	3.53	3.50	3.11	2.41	2.27
Panel C: Mean Absolute Percentage Errors (MAPE)					
Crude oil	24.46	27.45	26.23	26.06	18.84
Heating oil	27.15	28.82	24.15	22.92	18.86
Natural gas	29.07	31.71	28.45	25.87	20.21
Panel D: Mean Squared Percentage Errors (MSPE)					
Crude oil	9.75	15.72	11.65	11.73	6.16
Heating oil	14.22	17.68	10.19	9.29	6.01
Natural gas	14.46	17.29	13.62	11.61	6.63

This table reports the out-of-sample forecast errors of each volatility model for realized volatility over a horizon of 30 days. Realized volatility is defined as:

$$RV_{t,T} = \sqrt{\frac{252}{T}\sum_{t=1}^{T}\left(\log\frac{F_{t,T}}{F_{t-1,T}}\right)^2}$$

where $RV_{t,T}$ refers to realized volatility between t and T. $F_{t,T}$ denotes the price at time t of the futures contract maturing at T. Panels A and B report the mean absolute errors (MAE) and mean squared errors (MSE) of individual models, respectively. Panels C and D report the mean absolute percentage errors (MAPE) and mean squared percentage errors (MSPE), respectively.

Table 18
Differences of forecasting errors: AE and SE (ARMA model for RVRP).

	HIST (%)	GJR (%)	ATM IV (%)	MFIV (%)	RMFIV (%)
Panel A: Crude oil (AE)					
HIST		−0.63	0.35	0.71	**2.18**
GJR	−0.04		0.98	1.34	**2.81**
ATM IV	0.09	−0.61		0.36	**1.83**
MFIV	0.40	−0.49	0.08		**1.47**
RMFIV	**−1.25**	**−1.40**	**−2.06**	**−2.57**	
Panel B: Crude oil (SE)					
HIST		−1.33	0.45	0.61	**0.79**
GJR	0.00		1.77	1.93	2.12
ATM IV	0.01	−0.04		0.16	**0.34**
MFIV	0.02	−0.04	0.00		0.18
RMFIV	**−0.10**	**−0.08**	**−0.14**	**−0.18**	
Panel C: Heating oil (AE)					
HIST		0.18	1.18	**1.69**	**2.50**
GJR	−0.32		1.00	**1.51**	**2.32**
ATM IV	−0.72	−0.74		**0.51**	**1.32**
MFIV	**−0.76**	**−0.75**	**−0.28**		**0.81**
RMFIV	**−1.55**	**−1.40**	**−0.95**	**−0.89**	
Panel D: Heating oil (SE)					
HIST		0.18	0.80	0.97	1.08
GJR	−0.02		0.62	0.79	**0.90**
ATM IV	−0.06	−0.05		0.17	**0.28**
MFIV	**−0.07**	**−0.05**	**−0.01**		0.11
RMFIV	**−0.13**	**−0.09**	**−0.08**	**−0.06**	
Panel E: Natural gas (AE)					
HIST		−0.44	0.91	**2.38**	**3.80**
GJR	0.29		1.35	**2.82**	**4.24**
ATM IV	−0.07	**−1.49**		**1.48**	**2.89**
MFIV	**−1.86**	**−2.55**	**−0.83**		**1.41**
RMFIV	**−2.82**	**−2.71**	**−4.12**	**−3.77**	
Panel F: Natural gas (SE)					
HIST		0.03	0.42	**1.12**	**1.26**
GJR	0.03		0.39	**1.09**	**1.23**
ATM IV	−0.01	−0.18		**0.70**	**0.84**
MFIV	**−0.24**	**−0.35**	**−0.09**		0.14
RMFIV	**−0.43**	**−0.41**	**−0.46**	**−0.32**	

This table reports relative differences in the performance of competing models. The upper triangular matrices report the mean difference of absolute (AE) and squared (SE) forecasting errors, respectively. Similarly, the lower triangular matrices report the median difference of forecasting errors. We compute the difference between the errors of model [*name in row*] and those of model [*name in column*]. For example, the first row of Panel A presents the average difference in the AE of HIST vis-à-vis GJR, ATM IV, MFIV, and RMFIV, respectively. The numbers in bold indicate statistically significant differences at 5% as indicated by the Diebold–Mariano (computed with 2 lags) and the Wilcoxon signed rank tests for the upper and lower triangular matrices, respectively.

Table 19
Differences of forecasting errors: APE and SPE (ARMA model for RVRP).

	HIST (%)	GJR (%)	ATM IV (%)	MFIV (%)	RMFIV (%)
Panel A: Crude oil (APE)					
HIST		−2.99	−1.78	−1.60	**5.62**
GJR	−0.15		1.21	1.39	**8.61**
ATM IV	0.32	−2.28		0.17	**7.39**
MFIV	1.52	−1.40	0.25		**7.22**
RMFIV	**−4.31**	**−4.32**	**−7.41**	**−9.41**	
Panel B: Crude oil (SPE)					
HIST		−5.97	−1.90	−1.97	**3.59**
GJR	−0.04		4.07	4.00	**9.56**
ATM IV	0.17	−0.45		−0.08	**5.49**
MFIV	0.36	−0.48	0.04		**5.57**
RMFIV	**−1.14**	**−0.96**	**−1.58**	**−2.48**	
Panel C: Heating oil (APE)					
HIST		−1.67	3.00	**4.23**	**8.29**
GJR	−1.16		**4.67**	**5.90**	**9.96**
ATM IV	−2.21	−2.24		**1.23**	**5.29**
MFIV	**−2.78**	**−3.19**	**−0.88**		**4.06**
RMFIV	**−5.37**	**−4.48**	**−3.13**	**−2.97**	
Panel D: Heating oil (SPE)					
HIST		**−3.46**	4.03	**4.93**	**8.21**
GJR	−0.19		**7.49**	**8.39**	**11.67**
ATM IV	−0.62	−0.40		**0.90**	**4.18**
MFIV	**−0.76**	**−0.54**	**−0.10**		**3.28**
RMFIV	**−1.24**	**−0.98**	**−1.10**	**−0.79**	
Panel E: Natural gas (APE)					
HIST		**−2.64**	0.62	3.20	**8.86**
GJR	0.27		3.26	**5.84**	**11.50**
ATM IV	−0.15	**−3.88**		**2.58**	**8.24**
MFIV	−4.02	**−5.96**	**−1.95**		**5.66**
RMFIV	**−5.86**	**−5.60**	**−7.72**	**−7.61**	
Panel F: Natural gas (SPE)					
HIST		−2.82	0.85	2.85	**7.83**
GJR	0.04		3.67	**5.68**	**10.65**
ATM IV	−0.03	**−1.33**		**2.01**	**6.98**
MFIV	−1.10	**−1.85**	**−0.47**		**4.98**
RMFIV	**−2.09**	−1.84	−2.29	−1.47	

This table reports relative differences in the performance of competing models. The upper triangular matrices report the mean difference of absolute percentage (APE) and squared percentage (SPE) forecasting errors, respectively. Similarly, the lower triangular matrices report the median difference of forecasting errors. We compute the difference between the errors of model [*name in row*] and those of model [*name in column*]. For example, the first row of Panel A presents the average difference in the APE of HIST vis-à-vis GJR, ATM IV, MFIV, and RMFIV, respectively. The numbers in bold indicate statistically significant differences at 5% as indicated by the Diebold–Mariano (computed with 2 lags) and the Wilcoxon signed rank tests for the upper and lower triangular matrices, respectively.

day, we estimate the average relative variance risk premium over a period of just under one year (see Eq. (7)). We then fit an ARMA (1,1) model to the time-series of average relative variance risk premia observed from the beginning of our sample right until time $t-\tau$, where τ denotes the 30-day forecasting horizon. We then use these parameter estimates to forecast the 30-day ahead risk premium, which we use to adjust the MFIV.

Table 17 reports the forecasting errors of individual models. Panels A through D present the MAE, MSE, MAPE and MSPE, respectively. Columns 2–6 report the results of HIST, GJR, ATM IV, MFIV and RMFIV. Looking at Panels A through D, we can observe that RMFIV always yields the smallest forecasting errors, confirming the importance of the volatility risk premium for volatility forecasting.

It is worth noticing that the magnitude of the forecasting errors reported in Tables 17–19 are comparable but slightly different from those shown in Tables 4–6. The main difference stems from the fact that we use several observations (at the beginning of the sample) to estimate the parameters of the ARMA process. As a result, our sample changes slightly. In our baseline analysis, we have 220, 216 and 189 data points for the crude oil, heating oil and natural gas markets, respectively. When we explicitly model the dynamics of the risk premium, we obtain a sample of 212, 203 and 177 observations for the crude oil, heating oil and natural gas markets, respectively.

To further appreciate the importance of explicitly modeling the dynamics of the volatility risk premium, we look at the differences between pairs of models. Tables 18 and 19 presents these results. We can see that the difference between RMFIV and MFIV is statistically significant, indicating that explicitly modeling the volatility risk premium improves the predictive power of implied volatility. The bottom right entry of each panel, which reports the median of the difference between RMFIV and MFIV is particularly revealing. We can see that the entries are always more negative in Tables 18 and 19 than in Tables 5 and 6. This suggests that explicitly modeling the dynamics of the volatility risk premium leads to further improvements in the forecasting power of the baseline RMFIV.

6. Conclusion

This paper analyzes the role of the volatility risk premium for volatility forecasting. Specifically, we investigate the extent to which the biasedness of option implied volatility forecasts might be attributable to the wedge between the risk-neutral and the physical measures. We propose a simple model-free adjustment to account for the market price of volatility risk. Empirically examining the effect of our adjustment, the evidence convincingly shows that accounting for the volatility risk premium results in superior volatility forecasting performance. We also analyze the extent to which MFIV is superior to ATM IV. Generally, MFIV provides the second-best volatility forecast. In particular, it subsumes ATM IV and exhibits higher predictive power than ATM IV.

Our study can be extended in several directions. First, a natural extension would consist in applying our adjustment to other markets. Second, our arguments could be extended to forecasting covariance matrices. In particular, adjusting both implied correlation and volatility by their respective risk premia might improve covariance forecasts. Asset allocation would be a potential application. Finally, although our study shows that the volatility risk premium attenuates the biasedness of option implied volatility, it does not entirely eliminate such bias. Thus, further research is needed. Exploring the role of trading frictions, as suggested by Figlewski (1997), could prove a fruitful avenue.

References

Agnolucci, P., 2009. Volatility in crude oil futures: a comparison of the predictive ability of GARCH and implied volatility models. Energy Economics 31 (2), 316–321.

Andersen, T., Bollerslev, T., 1998. Answering the skeptics: yes, standard volatility models do provide accurate forecasts. International Economic Review 39 (4), 885–905.

Andersen, T., Bollerslev, T., Diebold, F., 2010. Parametric and nonparametric volatility measurement. In: Aït-Sahalia, Y., Hansen, L. (Eds.), Handbook of Financial Econometrics: Tools and Techniques. Elsevier Science.

Bakshi, G., Cao, C., Chen, Z., 1997. Empirical performance of alternative option pricing models. Journal of Finance 52 (5), 2003–2049.

Barone-Adesi, G., Whaley, R., 1987. Efficient analytic approximation of American option values. Journal of Finance 42 (2), 301–320.

Black, F., 1976. The pricing of commodity contracts. Journal of Financial Economics 3 (1), 167–179.

Bollerslev, T., Tauchen, G., Zhou, H., 2009. Expected stock returns and variance risk premia. Review of Financial Studies 22 (11), 4463–4492.

Canina, L., Figlewski, S., 1993. The informational content of implied volatility. Review of Financial Studies 6 (3), 659–681.

Carr, P., Wu, L., 2009. Variance risk premiums. Review of Financial Studies 22 (3), 1311–1341.

Charoenwong, C., Jenwittayaroje, N., Low, B., 2009. Who knows more about future currency volatility? Journal of Futures Markets 29 (3), 270–295.

Cheng, X., Fung, J., 2012. The information content of model-free implied volatility. Journal of Futures Markets 32 (8), 792–806.
Chernov, M., 2007. On the role of risk premia in volatility forecasting. Journal of Business & Economic Statistics 25 (4), 411–426.
Christensen, B., Prabhala, N., 1998. The relation between implied and realized volatility. Journal of Financial Economics 50 (2), 125–150.
DeMiguel, V., Plyakha, Y., Uppal, R., Vilkov, G., 2012. Improving portfolio selection using option-implied volatility and skewness. Journal of Financial and Quantitative Analysis (forthcoming).
Driessen, J., Maenhout, P., Vilkov, G., 2009. The price of correlation risk: evidence from equity options. Journal of Finance 64 (3), 1377–1406.
Eraker, B., Johannes, M., Polson, N., 2003. The impact of jumps in volatility and returns. Journal of Finance 58 (3), 1269–1300.
Figlewski, S., 1997. Forecasting volatility. Financial Markets, Institutions & Instruments 6 (1), 1–88.
Fleming, J., 1998. The quality of market volatility forecasts implied by S&P 100 index option prices. Journal of Empirical Finance 5 (4), 317–345.
Frijns, B., Tallau, C., Tourani-Rad, A., 2010. The information content of implied volatility: evidence from Australia. Journal of Futures Markets 30 (2), 134–155.
Garman, M.B., Klass, M.J., 1980. On the estimation of security price volatilities from historical data. Journal of Business 53 (1), 67–78.
Guo, D., 1996. The predictive power of implied stochastic variance from currency options. Journal of Futures Markets 16 (8), 915–942.
Heston, S., 1993. A closed-form solution for options with stochastic volatility with applications to bond and currency options. Review of Financial Studies 6 (2), 327–343.
Hull, J., White, A., 1987. The pricing of options on assets with stochastic volatilities. Journal of Finance 42 (2), 281–300.
Jiang, G., Tian, Y., 2005. The model-free implied volatility and its information content. Review of Financial Studies 18 (4), 1305–1342.
Jorion, P., 1995. Predicting volatility in the foreign exchange market. Journal of Finance 50 (2), 507–528.
Lamoureux, C., Lastrapes, W., 1993. Forecasting stock-return variance: toward an understanding of stochastic implied volatilities. Review of Financial Studies 6 (2), 293–326.
Marshall, B., Nguyen, N., Visaltanachoti, N., 2012. Commodity liquidity measurement and transaction costs. Review of Financial Studies 25 (2), 599–638.
Martens, M., Zein, J., 2004. Predicting financial volatility: high-frequency time-series forecasts vis-à-vis implied volatility. Journal of Futures Markets 24 (11), 1005–1028.
Mueller, P., Vedolin, A., Yen, Y., 2011. Bond Variance Risk Premia. LSE Working Paper.
Poon, S., Granger, C., 2003. Forecasting volatility in financial markets: a review. Journal of Economic Literature 41 (2), 478–539.
Poteshman, A., 2000. Forecasting Future Volatility From Option Prices. Working Paper.
Prokopczuk, M., Wese Simen, C., 2012. Variance Risk Premia in Commodity Markets. ICMA Centre Discussion Paper.
Shu, J., Zhang, J., 2006. Testing range estimators of historical volatility. Journal of Futures Markets 26 (3), 297–313.
Szakmary, A., Ors, E., Kyoung Kim, J., Davidson, W., 2003. The predictive power of implied volatility: evidence from 35 futures markets. Journal of Banking & Finance 27 (11), 2151–2175.
Taylor, S., Yadav, P., Zhang, Y., 2010. The information content of implied volatilities and model-free volatility expectations: evidence from options written on individual stocks. Journal of Banking & Finance 34 (4), 871–881.
Trolle, A., Schwartz, E., 2009. Unspanned stochastic volatility and the pricing of commodity derivatives. Review of Financial Studies 22 (11), 4423–4461.
Trolle, A., Schwartz, E., 2010. Variance risk premia in energy commodities. Journal of Derivatives 17 (3), 15–32.
Yang, D., Zhang, Q., 2000. Drift-independent volatility estimation based on high, low, open, and close prices. Journal of Business 73 (3), 477–492.