Received: 13 December 2022 | Accepted: 5 March 2023

DOI: 10.1002/fut.22412

RESEARCH ARTICLE

THE JOURNAL OF FUTURES MARKETS WILEY

Less disagreement, better forecasts: Adjusted risk measures in the energy futures market

Ning Zhang[1] | **Yujing Gong**[2,3] | **Xiaohan Xue**[4]

[1]Department of Finance, International Business School Suzhou, Xi'an Jiaotong-Liverpool University, Suzhou, China

[2]Systemic Risk Centre, London School of Economics, London, UK

[3]Accounting and Finance Group, Management School, University of Liverpool, Liverpool, UK

[4]Finance Group, Norwich Business School, University of East Anglia, Norwich, UK

Correspondence
Yujing Gong, Systemic Risk Centre, London School of Economics, London, UK.
Email: Y.Gong19@lse.ac.uk

Abstract

This paper develops a generic adjustment framework to improve in the market risk forecasts of diverse risk forecasting models, which indicates the degree to which risk is under- and overestimated. In the context of the energy commodity market, a market in which tail risk management is of crucial importance, the empirical analysis shows that after this adjustment framework is applied, the forecasting performance of various risk models generally improves, as verified by a battery of backtesting methods. Additionally, our method also lessens the risk model disagreement among post-adjusted risk forecasts.

KEYWORDS

energy futures, expected shortfall, finance, model disagreement, value at risk

JEL CLASSIFICATION

C52, C53, G10

1 | INTRODUCTION

Interest in understanding the risks related to large price movements in the energy futures market has increased following the large influx of financial investors in this market after the financialization of the commodity market (Čech & Baruník, 2019; Qiao & Han, 2023; Xu & Lien, 2020). Recently, Ding et al. (2021) utilize artificial intelligence techniques to extend traditional financial econometric models for volatility risk forecasting allowing for liquidity effects. This highlights that the frequent sharp drops in prices and highly leveraged margin trading in the energy futures market, however, have made it challenging for energy producers, index investors, speculators, exchanges, and commodity market regulators to accurately evaluate the risks associated with potential extreme price movements. The accuracy of such risk forecasts constitutes a major concern in effective risk management practices. Risk measurement procedures often involve specification error in the underlying dynamic processes and estimation error in the model parameters—a concern invoked by numerous studies on risk estimation and forecast evaluation (for a comprehensive overview, see Christoffersen, 2012). In practice, then, when assessing market risk, relying on a single risk forecasting model across different market conditions and for various assets is implausible.[1] Lazar and Zhang (2019) show that commodity risk measures carry a higher model misspecification

[1]For a horse race of risk models, see Danielsson et al. (2016), Nolde and Ziegel (2017), Laporta et al. (2018), Taylor (2022), for example. In energy futures markets, several studies apply sophisticated models to capture the statistical properties of energy returns, see Miller and Liu (2006), Guo (2020) and Jang et al. (2020).

risk among other assets, because of the stylized facts of the commodity returns (heavy-tailedness and negative skewness). This difficulty has made risk managers and regulators more aware of the forecasting ability of particular risk models.

Another challenge in accurately evaluating risk is that financial institutions are required to allocate risk capital to shield themselves from unexpected negative shocks. Here, the amount of risk capital is determined by the risk model being used. However, the use of diverse risk models results in risk disagreement, leading to varying degrees of capital buffers. As the risk numbers given by diverse risk models are typically different, as shown in Danielsson et al. (2016), a regulatory arbitrage opportunity arises; that is, financial institutions can report less risk to regulators with respect to inadequate risk models to reduce their capital requirements (Liu & Stentoft, 2021). Hence, the model disagreement between numerous risk models makes risk management inefficient in allocating risk capital to cover potential losses quantified by risk measures. This inefficiency is even pronounced under Basel III (Basel Committee on Banking Supervision, 2019), as pointed out by Liu and Stentoft (2021). As such, regulators are incentivized to reduce the model disagreement. For example, the Basel III reforms (Basel Committee on Banking Supervision, 2021), also known as Basel IV, published revisions to the market risk capital requirements, showing more tendency towards the standardized approach, away from the internal model approach. Though the standardized approach is praised for its simplicity and consistency, it fails to allow for specific characteristics of individual institutions. In contrast, the internal model approach can overcome this issue. Motivated by these findings, our proposed adjustment methodology is aimed at not only providing adequate risk forecasts but also mitigating the market risk model disagreement while embracing the diversity in risk models.

To determine extreme risk exposures of an investment position in times of stress, standard risk measures are typically used. These measures are value-at-risk (VaR) and expected shortfall (ES) (Basel Committee on Banking Supervision, 2019): VaR measures the worst possible loss of an investment within a given probability level, and ES quantifies the average loss exceeding the VaR threshold. Our proposed methodology, an internal model approach, is built on the seminal work of Fissler and Ziegel (2016), which identifies the joint elicitability of VaR and ES measures. This desirable elicitability property enables the optimal risk estimates to be uniquely obtained by minimizing the average loss for a given loss function within the family of loss functions proposed by Fissler and Ziegel (2016) (hereafter, *FZ* loss function for simplicity). By exploiting this property, we are able to improve on the ex ante risk forecasts of risk models and reduce the model disagreement towards a recognized objective (i.e., the minimization of average loss over a time period). Specifically, given the time series of pairs of ex ante VaR and ES forecasts of a risk model at a predefined probability level and the associated return data, within a rolling-window scheme we obtain the time-varying adjustments for VaR and ES, respectively, via the minimization of average loss. If the value of the adjustment multiplier is smaller (resp. larger) than 1, this suggests that the risk forecast is overestimated (resp. underestimated). This adjustment methodology is applicable to different risk forecasting models, unlike the studies of Farkas et al. (2020) and Patra (2021), which modify risk estimates restricted to a certain model. In addition, it gives a clear indication of the inaccuracies of risk forecasts, rather than relying on the relative model performance within the model set (Barrieu & Scandolo, 2015; Danielsson et al., 2016).

Empirically, we test this proposed methodology by applying it to a battery of commonly used risk forecasting models for several futures in the energy futures market.[2] Our proposed adjustment methodology identifies the under- or over-estimated risk forecasts generated by these risk models. After making adjustments to the VaR and ES estimates, the forecasting ability of risk models generally improves, as verified via various VaR and ES backtesting methods. Additionally, we investigate eight energy crisis periods and the margin-level change date that signals a high-volatility market regime and find an abatement in risk disagreement among post-adjusted risk forecasts.

Our paper mainly relates to three strands of existing literature. First, this study focuses on the accuracy of risk measures. In general, the adequacy of VaR and ES forecasts is affected by different sources: (1) the model misspecification involving inappropriate assumptions about true data-generating processes that may lead to the choice of inadequate risk models (see Danielsson et al., 2016; Liu & Stentoft, 2021; Patra, 2021); and (2) the estimation errors in model parameters, as shown in Pitera and Schmidt (2018) and Farkas et al. (2020). This inevitably leads to varying

[2]Risk model specifications can be found in Section 3.

risk predictions depending on which risk estimation methodology is used. In practice, divergence in the risk forecasts of different risk models can induce speculative institutional investors to use models that produce inferior forecasts, with the aim of mitigating the risk capital required by regulators. Further, without knowing the statistical form of the underlying asset behavior, capturing model misspecification and parameter estimation errors is problematic. To address these issues, using the newly proposed methodology, we provide adjustments to quantify the extent to which one risk model underestimates or overestimates risk. Thus, we use these adjustments to build better risk estimates, which we subsequently evaluate with various backtesting methods. Our proposed methodology is a practical tool for making adjustments to the ex ante VaR and ES forecasts of a given model, and it facilitates the refinement of internal risk models for risk managers. It also helps to align the model performance across various models, thus disincentivizing financial institutions from reporting inadequate risk numbers to the regulatory authority.

Second, this paper relates to the abundant literature on market risk forecasting models. The VaR and ES forecasting methodologies can be classified into three main categories (Engle & Manganelli, 2004; Taylor, 2008): nonparametric, parametric, and semiparametric. Nonparametric approaches (see, e.g., Chen, 2007) rarely rely on assumptions about the conditional distribution of asset returns. Within this framework, VaR and ES are treated as quantiles of a selected sample of returns over a specific window at a given significance level. Nonparametric methods are model-free and easy to implement, but they are often criticized because of their sensitivity to window size selection. Conversely, before applying parametric models for VaR and ES forecasting (e.g., Escanciano & Olmo, 2010), we make a specific assumption on the asset returns distribution. The selection of the distribution function can cause differences in the forecasting performance of VaR and ES. Semiparametric models impose a parametric structure on the dynamics of VaR and ES but require no assumptions on the conditional distributions of financial returns (Engle & Manganelli, 2004; Patton et al., 2019). The parameters of semiparametric models are estimated by minimizing a specified loss (scoring) function instead of maximizing a distributional likelihood function described in the parametric techniques.[3] The forecasting accuracy of these risk models depends heavily on the data. Parsimonious models can easily outperform other benchmarks in a stable market, but they are often less efficient than highly sophisticated models during a turbulent period. While a myriad of studies propose VaR and ES models to fit the market data thoroughly, practitioners and academics prefer a limited number of them. Thus, in this study, we consider a group of commonly used VaR and ES models covering the three categories mentioned above.[4] Besides the risk models considered in this paper, our proposed methodology can also be applied to several sophisticated machine learning models incorporating liquidity and micro-structure information, for example, the LIQ-GARCH model in a genetic programming framework, proposed by Ding et al. (2019).

Last, our study also relates to the line of research on statistical backtesting methods, which are necessary to validate risk forecasting models. To thoroughly evaluate whether the forecasting accuracy of post-adjusted VaR and ES forecasts based on our methodology is improved, we consider traditional and comparative backtesting; the former is designed to directly test the forecasting ability of a given model, whereas the latter focuses on model performance comparisons among models. With respect to traditional backtesting, we adopt a series of commonly used and recently proposed traditional backtesting methods tailored to different desirable criteria. Specifically, the unconditional coverage (UC) test (Kupiec, 1995), the conditional coverage (CC) test (Christoffersen, 1998), and the dynamic quantile (DQ) test (Engle & Manganelli, 2004) are used to backtest VaR; the exceedance residual (ER) test (McNeil & Frey, 2000), the conditional calibration (CCA) test (Nolde & Ziegel, 2017), and the regression-based expected shortfall (ESR) test (Bayer & Dimitriadis, 2022) are used to backtest ES. With respect to comparative backtesting, we use the well-known Diebold-Mariano (Diebold & Mariano, 2002) and model confidence set tests (Hansen et al., 2011) to make model comparisons based on the joint *FZ* loss of VaR and ES.

The paper is structured as follows: Section 2 briefly introduces the proposed methodology to adjust risk forecasts based on the *FZ* loss function; Section 3 discusses the forecasting approaches for VaR and ES; the empirical data are presented in Section 4; Section 5 compares the risk model disagreement of pre- and

[3]In the extreme value theory framework, which is also classified as a semiparametric model, parameters are obtained by maximizing the generalized Pareto distribution (McNeil & Frey, 2000). However, they focus on the left tail of the conditional distribution instead of the whole conditional distribution.

[4]More details are discussed in Section 3.

post-adjusted VaR and ES forecasts and validates the efficiency of this adjustment methodology via various backtesting methods; and Section 6 concludes.

2 | ADJUSTMENT METHODOLOGY FOR RISK FORECASTS

The theoretical property of the *FZ* loss function that optimal VaR and ES forecasts can result in the lowest average loss has motivated the applications of *FZ* loss functions in risk estimation and evaluation (Patton et al., 2019; Nolde & Ziegel, 2017). In this section, we propose a generic and empirical approach to jointly improve the accuracy of VaR and ES forecasts built by any given risk forecasting model via minimizing the *FZ* loss function.

Here, we would like to adjust risk forecasts generated by a risk model using past realized observations and 1-day ahead risk forecasts produced by the same model. A general setting is constructed to illustrate our methodology. Let r_t denote the asset return on day t, and correspondingly $\left(\hat{v}^{\alpha}_{t+1|t}, \hat{e}^{\alpha}_{t+1|t}\right)$, in which α will be suppressed for brevity, represent the pairs of 1-day-ahead VaR and ES forecasts for day $t+1$ based on the information available until t at α level.[5] Throughout this paper, the signs of VaR and ES are considered negative, which indicates the potential loss occurring in the left tail of the return distributions, following Nolde & Ziegel (2017). The pair of risk forecasts can be adjusted by

$$v_{t+1|t} = a_{1,t} \cdot \hat{v}_{t+1|t}, e_{t+1|t} = a_{2,t} \cdot \hat{e}_{t+1|t}, \tag{1}$$

where $(a_{1,t}, a_{2,t})$ are a set of arbitrary adjustment multipliers for VaR and ES forecasts formed at t, which satisfies that $a_{1,t}, a_{2,t} > 0$ to ensure the negativity of VaR and ES forecasts. The proposed form of adjustments, that is, the pair of multipliers, can provide direct evidence of under- and over-estimation risk and thus allow for model comparisons among diverse models.

The derived consistency of *FZ* loss functions implies that the true values of (VaR, ES) forecasts can minimize the expected loss (Fissler & Ziegel, 2016). When we use a consistent loss function, the optimal adjustment multipliers $\left(a^{*}_{1,t}, a^{*}_{2,t}\right)$ consequently minimize the expectation of the *FZ* loss function:

$$\left(a^{*}_{1,t}, a^{*}_{2,t}\right) = \underset{a_1,a_2}{\arg\min}\, \mathbb{E}\left[L_{FZ0}(r_t, a_1 \cdot \hat{v}_t, a_2 \cdot \hat{e}_t; \alpha)|\mathcal{F}_t\right], \tag{2}$$

where $\mathcal{F}_t$ denotes the information set up to time t, and $a_2 > a_1 \cdot \hat{v}_t/\hat{e}_t$ is set to ensure that for a pair of risk forecasts, ES values are always smaller than VaR values, showing a more extreme level of risk than VaR. L_{FZ0} is the *FZ*0 loss function considered in this paper, formulated as follows:

$$L_{FZ0}(r, v, e; \alpha) = -\frac{1}{\alpha e}\mathbb{1}\{r \le v\}(v - r) + \frac{v}{e} + \log(-e) - 1, \tag{3}$$

where $\mathbb{1}$ refers to the indicator function, which is 1 if $r \le v$, that is, a VaR exceedance occurs, and 0 otherwise. Of the broad *FZ* family, the aforementioned loss function *FZ*0 is the primary option in terms of risk estimation and backtesting; see related papers, Nolde and Ziegel (2017), Patton et al. (2019), and Merlo et al. (2021).

Then, we can obtain the estimators $(\hat{a}_{1,t}, \hat{a}_{2,t})$ of the optimal adjustment multipliers on day t via the minimization of average *FZ*0 loss over a training sample period with length M:[6]

$$(\hat{a}_{1,t}, \hat{a}_{2,t}) = \underset{a_1,a_2}{\arg\min}\, \frac{1}{M}\sum_{i=t-M}^{t-1} L_{FZ0}(r_{i+1}, a_1 \cdot \hat{v}_{i+1}, a_2 \cdot \hat{e}_{i+1}; \alpha), \text{ with } a_2 > a_1 \cdot \hat{v}_{i+1}/\hat{e}_{i+1}. \tag{4}$$

[5]Here, we follow Engle and Manganelli (2004) and use 1000 observations to estimate the parameter values in a risk model for (VaR, ES) forecasting.
[6]We use $M = 2000$ to avoid the outliers' effect, following the convention that the computation of average loss is based on 2000 data points in Nolde and Ziegel (2017).

Finally, we multiply optimized positive constants calibrated from the training sample to the next day VaR and ES forecasts $(\hat{v}_{t+1}, \hat{e}_{t+1})$ to adjust the forecasts given by the risk model:

$$\tilde{v}_{t+1|t} = \hat{a}_{1,t} \cdot \hat{v}_{t+1|t}, \tilde{e}_{t+1|t} = \hat{a}_{2,t} \cdot \hat{e}_{t+1|t}. \tag{5}$$

We use a rolling window scheme with the window length of M to obtain the time series of the pair of optimized multipliers for further analysis in this paper. The detailed procedure can be found in Algorithm 1.

Algorithm 1: Adjustments for Risk Forecasts

Input: $\{r_t\}_{t=1}^{T}$, $\{\hat{v}_t\}_{t=1}^{T}$, $\{\hat{e}_t\}_{t=1}^{T}$, M, α
Output: Adjusted risk measures $(\{\tilde{v}_t\}_{t=M+1}^{T}, \{\tilde{e}_t\}_{t=M+1}^{T})$
Initialization: $t = M$;
repeat
Select the time series of returns: $\{r_{t-M}, \ldots, r_{t-1}\}$;
Select the time series of pairs of risk forecasts: $\{\hat{v}_{t-M}, \ldots, \hat{v}_{t-1}\}$ and $\{\hat{e}_{t-M}, \ldots, \hat{e}_{t-1}\}$;
Estimate the adjustment parameters $(\hat{a}_{1,t}, \hat{a}_{2,t})$ by minimizing the average $FZ0$ loss function:

$$(\hat{a}_{1,t}, \hat{a}_{2,t}) = \arg\min_{a_1, a_2} \frac{1}{M} \sum_{i=t-M}^{t-1} L_{FZ0}\left(r_{i+1}, a_1 \cdot \hat{v}_{i+1}, a_2 \cdot \hat{e}_{i+1}; \alpha\right);$$

Compute the adjusted risk measures forecasts by multiplying the adjusted parameters:

$$\tilde{v}_{t+1} = \hat{a}_{1,t} \cdot \hat{v}_{t+1}, \qquad \tilde{e}_{t+1} = \hat{a}_{2,t} \cdot \hat{e}_{t+1};$$

$t = t + 1$;
until $t = T$;
return $\{\tilde{v}_t\}_{t=M+1}^{T}$, $\{\tilde{e}_t\}_{t=M+1}^{T}$.

3 | RISK FORECASTING MODELS

A large number of risk forecasting models are used in the academic, industry, and regulator system. The model choice highly depends on the preference and demand of end users. In this paper, we broadly select 10 commonly used risk forecasting models as our candidates, including nonparametric, parametric, and semiparametric models.

3.1 | Nonparametric methods

3.1.1 | Historical simulations (HS)

As a simple nonparametric approach, the standard HS is selected, which is implemented within an estimation window with the size $n = 1000$. The HS predicts the 1-day-ahead VaR by taking α quantile of past returns empirically. Additionally, we calculate the average left-tail returns beyond VaR to proxy ES.

3.1.2 | Weighted historical simulations (WHS)

We consider the WHS method, which is based on a geometrically declining scheme (see Boudoukh et al., 1998). Within this method, more recent observations are assigned with higher weights for forecasting. The weight of day $t^* = t - w + 1, ..., t$ is given as $\eta(t^*) = \eta^{t^*-1}(1-\eta)/(1-\eta^n)$, where $\eta = 0.99$.

3.1.3 | Cornish–Fisher (CF) approximation

Another nonparametric method for VaR and ES estimation is the CF approximation method, which is an extended normal quantile technique by considering the in-sample estimated skewness $\hat{\delta}_1$ and kurtosis $\hat{\delta}_2$:

$$\begin{aligned} v_t &= \hat{\mu}_{t-1} + \hat{\sigma}_{t-1} F_{CF}^{-1}(\alpha), \\ F_{CF}^{-1}(\alpha) &= \Phi_\alpha^{-1} + \left(\left(\Phi_\alpha^{-1}\right)^2 - 1\right)\frac{\hat{\delta}_1}{6} + \left(\left(\Phi_\alpha^{-1}\right)^3 - 3\Phi_\alpha^{-1}\right)\frac{\hat{\delta}_2 - 3}{24} - \left(2\left(\Phi_\alpha^{-1}\right)^3 - 5\Phi_\alpha^{-1}\right)\frac{\hat{\delta}_1^2}{36}, \end{aligned} \tag{6}$$

and Φ_α^{-1} denotes the inverse of the Gaussian cumulative distribution function, $\hat{\mu}_t$ and $\hat{\sigma}_t$ denote the in-sample mean and standard deviation, respectively.

Boudt et al. (2008) develop the CF approximation technique to estimate ES:

$$e_t = \hat{\mu}_{t-1} + \hat{\sigma}_{t-1}\mathbb{E}\left[z | z \le F_{CF}^{-1}(\alpha)\right], \tag{7}$$

where

$$\begin{aligned} \mathbb{E}\left[z | z \le F_{CF}^{-1}(\alpha)\right] = -\frac{1}{\alpha}\Bigg(&\phi\left(F_{CF}^{-1}(\alpha)\right) + \frac{\hat{\delta}_2}{24}\left(I^4 - 6I^2 + 3\phi\left(F_{CF}^{-1}(\alpha)\right)\right) + \frac{\hat{\delta}_1}{6}(I^3 - 3I^1) \\ &+ \frac{\hat{\delta}_1^2}{72}\left(I^6 - 15I^4 + 45I^2 - 15\phi\left(F_{CF}^{-1}(\alpha)\right)\right)\Bigg), \end{aligned}$$

$$I^p = \begin{cases} \sum_{i=1}^{p/2}\left(\frac{\prod_{j=1}^{p/2} 2j}{\prod_{j=1}^{i} 2j}\right) g_\alpha^{2i}\phi(g_\alpha) + \left(\prod_{j=1}^{p/2} 2j\right)\phi(g_\alpha), & \text{for } p \text{ is even}, \\ \sum_{i=0}^{p^*}\left(\frac{\prod_{j=0}^{p^*}(2j+1)}{\prod_{j=0}^{i}(2j+1)}\right) g_\alpha^{2i+1}\phi(g_\alpha) - \left(\prod_{j=0}^{p^*}(2j+1)\right)\phi(g_\alpha), & \text{for } p \text{ is odd}, \end{cases}$$

and $p^* = (p-1)/2$, $g_\alpha = F_{CF}^{-1}(\alpha)$. $\phi(\cdot)$ denotes the Gaussian probability density function.

3.2 | Parametric methods

3.2.1 | GARCH(1,1)–Student's t model (G–t)

The generalized autoregressive conditional heteroskedasticity (GARCH) framework is proposed by Bollerslev (1986) to model conditional volatility with a specified distribution of innovations. In this paper, we consider the GARCH(1,1) model with the Student's t (GARCH–t) and skewed t (GARCH–*skt*) distributional innovations as our candidates. They are more accurate in terms of describing heavy tails and negative skewness (Patton et al., 2019). The (VaR, ES) forecasts can be obtained via the following specification:

$$\begin{aligned} v_t &= \hat{\mu}_t + \hat{\sigma}_t F_t^{-1}(\alpha, \nu), \\ e_t &= \hat{\mu}_t + \hat{\sigma}_t \frac{\nu + \left(F_t^{-1}(\alpha,\nu)\right)^2}{\nu - 1}\frac{t\left(F_t^{-1}(\alpha,\nu)\right)}{1-\alpha}, \end{aligned} \tag{8}$$

where ν denotes the estimated degree of freedom (DoF) parameter of the t distribution, and $t(\cdot)$ is the Student's t distribution probability density function.

3.2.2 | GARCH(1,1)–skewed t model (G–skt)

Next, we suppose that the innovations follow a skewed-t distribution. Thus, we forecast risk measures as in the following expression:

$$\begin{aligned} v_t &= \hat{\mu}_t + \hat{\sigma}_t F_{skt}^{-1}(\alpha, \nu, \lambda), \\ e_t &= \hat{\mu}_t + \hat{\sigma}_t \frac{1}{\alpha} \int_{-\infty}^{-v(\alpha)} x f_{skt}(x) dx, \end{aligned} \tag{9}$$

where $f_{skt}(\cdot)$ denotes the skewed-t distribution probability density function.

3.3 | Semiparametric methods

3.3.1 | Extreme value theory with peaks-over-threshold (EVT-POT)

The EVT provides a tool for estimating the tails based on available observations in the left tail of distribution. McNeil and Frey (2000) propose a semiparametric model applying EVT to describe the tail of the conditional distribution, which is developed by Samuel (2008). Following this, we use the POT method to consider the exceedances of past observations over a typically high threshold, where a generalized Pareto distribution (GPD) is employed to fit negative returns over this specified threshold:

$$v_t = \hat{\mu}_t + \hat{\sigma}_t F_{EVT}^{-1}(\alpha), \tag{10}$$

where the quantile $F_{EVT}^{-1}(\alpha)$ can be estimated as

$$F_{EVT}^{-1}(\alpha) = u + \frac{\hat{s}}{\hat{\xi}} \left(\left(\frac{1-\alpha}{n_u/n} \right)^{-\hat{\xi}} - 1 \right),$$

with both the scale parameter $\hat{s}$ and the shape parameter $\hat{\xi}$ estimated from the fitting of the GPD distribution. The term n_u denotes the number of observations exceeding the selected threshold u. Consequently, the predicted ES can be calculated as

$$e_t = \hat{\mu}_t + \hat{\sigma}_t F_{EVT}^{-1}(\alpha) \left(\frac{1}{1-\hat{\xi}} + \frac{\hat{s} - \hat{\xi} u}{(1-\hat{\xi}) \hat{z}_\alpha} \right).$$

3.3.2 | One-factor GAS model (GAS-1F)

Patton et al. (2019) propose a group of dynamic semiparametric models for VaR and ES forecasting. We select the one-factor generalized autoregressive score model (GAS-1F), where we use the scaled score of the $FZ0$ loss function to drive the time variation in the target parameter. The GAS-1F model is expressed as

$$\begin{aligned} v_t &= a \exp\{\kappa_t\}, \\ e_t &= b \exp\{\kappa_t\}, b < a < 0, \\ \kappa_t &= \omega + \beta \kappa_{t-1} + \gamma H_{t-1}^{-1} s_{t-1}, \end{aligned} \tag{11}$$

where the score variable s_t is defined as

$$s_t \equiv \frac{\partial L_{FZ0}(r_t, a\exp\{\kappa_t\}, b\exp\{\kappa_t\}; \alpha)}{\partial \kappa} = -\frac{1}{e_t}\left(\frac{1}{\alpha}\mathbb{1}\{r_t \le v_t\}r_t - e_t\right), \tag{12}$$

the Hessian factor H_t is set to one in our paper for simplicity. To estimate parameters in the framework, we minimize the proposed $FZ0$ loss function.

3.3.3 | Filtered historical simulations (FHS)

To estimate VaR and ES in a GARCH(1,1) framework without assuming the underlying conditional distribution, Barone-Adesi et al. (1999) propose a semiparametric model called FHS. In this framework, we randomly draw the historical standardized innovations (with replacement) B times to form a bootstrapped set: $\left\{z_i^*\right\}_{i=1}^{B} = z_1^*, ..., z_B^*$. The bootstrapped returns can be computed as follows:

$$r_t^{*(i)} = \hat{\mu}_t + \hat{\sigma}_t z_i^*,$$

where $\hat{\sigma}_t$ is estimated by GARCH(1,1) model. Next, we apply the HS method to the bootstrapped returns $\left\{r_t^{*(i)}\right\}_{i=1}^{B}$ to obtain the estimated VaR and ES at α significance level.

3.3.4 | CAViaR model with symmetric absolute value (CAViaR-SAV)

Taylor (2019) extends the ES estimation from the conditional autoregressive Value at Risk (CAViaR) models introduced by Engle and Manganelli (2004). In this study, we select CAViaR-SAV, which considers the symmetric absolute values of historical observations:

$$\begin{aligned} v_t &= \omega + \beta v_{t-1} + \gamma|r_{t-1}|, \\ e_t &= b \cdot v_t, b > 1. \end{aligned} \tag{13}$$

Here, the parameters of this model are estimated by maximizing the sum of logarithm of the asymmetric Laplace-likelihood function proposed by Koenker and Machado (1999):

$$f(r_t) = \frac{\alpha - 1}{e_t}\exp\left(\frac{(r_t - v_t)(\alpha - \mathbb{1}\{r_t \le v_t\})}{\alpha e_t}\right),$$

instead of minimizing the loss function.[7]

3.3.5 | CARE model with symmetric absolute value (CARE-SAV)

Newey and Powell (1987) define the expectile of a distribution as the tail expectation if values above it were more likely to occur than they actually are. Efron (1991) estimates VaR at α level by mapping it to the expectile at τ level, denoted by $q(\tau)$. Taylor (2008) proposes a set of conditional autoregressive expectile (CARE) models to forecast expectiles and ES. We consider the CARE-SAV, which is shown as:

[7]Taylor (2019) shows that maximizing the AL density function is the same as minimizing the FZ loss function under the assumption of zero-mean returns. Our estimation result is consistent when we apply the FZ loss function minimization.

$$q_t(\tau) = \omega + \beta q_{t-1} + \gamma |r_{t-1}|,$$
$$e_t(\alpha) = \left(1 + \frac{\tau}{(1-2\tau)\alpha}\right) q_t(\tau), \tag{14}$$

where the parameters are estimated by minimizing the loss function proposed by Newey and Powell (1987):

$$S(q_t, r_t; \tau) = |\tau - \mathbb{1}\{r_t \leq q_t\}|(r_t - q_t)^2.$$

4 | DATA

Our analysis focuses on four commodities in the energy sector, including WTI crude oil (CL), heating oil (HO), natural gas (NG), and RBOB/unleaded gasoline (XB), which are listed on the Chicago Mercantile Exchange (CME). We obtain daily prices for individual commodity's futures contracts and the corresponding open interests and volumes data from Bloomberg. The sample period used in this paper is between April 4, 1990, and December 31, 2021 (7971 days), since NG futures have a relatively short history, which is available starting from April 4, 1990. The XB futures contracts were replaced by the RBOB Gasoline futures contracts in 2006. Thus, we use XB before 2006 and RBOB Gasoline after 2006 to represent RBOB/XB.

We calculate daily excess returns of long positions based on the assumption of fully collateralized futures positions (e.g., Bakshi et al., 2019; Boons & Prado, 2019; Gorton et al., 2013; Koijen et al., 2018):

$$r^1_{t+1} = \frac{P^1_{t+1}}{P^1_t} - 1, \tag{15}$$

where P^1_t is the time-t closing price of the 1st front-month contract (the contract with the 1st shortest maturity at time t among all available contracts). Focusing on the 1st front-month contract could ensure sufficient liquidity. The front-month contract is rolled on the 7th day of the month or the next closest business day if 7th day of the month is not a business day (Kang et al., 2020).

In Table 1, we present the summary statistics of their daily excess returns, open interest, and trading volume. Over the time period studied, the means of all return series are close to zero, and the NG market is found to be the most volatile one. Non-zero skewness and high kurtosis reflect that all return series (especially of WTI CL and HO) are slightly skewed and display leptokurtic distributions. This indicates that the return distributions of energy commodities are not normally distributed and often have fat tails. It is also notable that WTI CL, as an important component of the downstream refined oil products and highly financialized commodity, has the highest open interest and daily trading volume in the energy sector.

TABLE 1 Summary statistics of energy commodity futures excess returns and their corresponding open interests and trading volumes

	CL	HO	NG	XB
Mean (%)	0.03	0.03	-0.05	0.06
SD	2.44	2.07	2.92	2.24
Min (%)	−43.37	−29.60	−17.46	−25.78
Max (%)	25.10	13.66	20.64	21.28
Skewness	−0.75	−0.34	0.21	−0.43
Kurtosis	31.50	11.61	5.91	14.15
AR(1)	−0.01	−0.02	−0.04	−0.01
OI	234,793	61,578	122,729	67,447
Volume	191,768	29,527	55,408	31,454

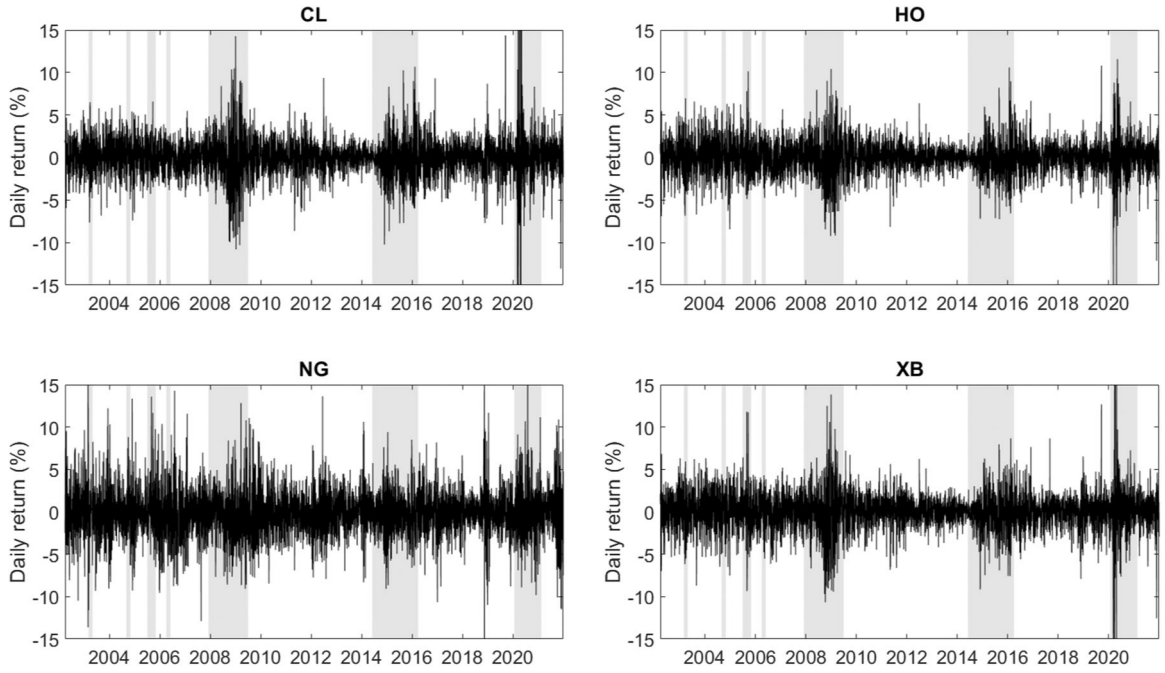

FIGURE 1 Daily energy commodity futures excess returns.

Figure 1 shows the dynamics of the daily returns of futures contracts of four energy commodities, with several energy crisis periods highlighted in the gray-shaded area. The energy crisis periods include (1) 2003 Iraq War (2003-03-01 to 2003-04-30), (2) 2004 Hurricane Ivan (2004-09-01 to 2004-10-31), (3) 2005 Hurricane Dennis (2005-07-01 to 2005-10-31), (4) 2006 Nigerian cuts (2006-04-01 to 2006-05-31), (5) 2008 Global Financial Crisis (2007-12-01 to 2009-06-30), (6) 2008 OPEC cuts in production (2008-12-01 to 2009-06-30), (7) 2014–2016 oil prices collapse (2014-06-01 to 2016-03-31) and (8) 2020 Covid pandemic (2020-01-23 to 2021-02-17).[8] As shown in Figure 1, high volatilities are noticeable around these crisis, indicating the high likelihood of incurring extreme losses. Among these crises, the 2003 Iraq War, the 2008 Global Financial Crisis, the 2008 OPEC cuts in production, the 2014–2016 oil prices collapse and 2020 Covid Crisis show a widespread impact on the four energy commodities.

5 | EMPIRICAL INVESTIGATION

Considering the widely used risk models in the current literature, we employ the proposed methodology to improve VaR and ES forecasts for the energy commodities. We evaluate 1-day-ahead VaR and ES forecasts for the four commodity futures returns and for the following three significance levels: 1%, 2.5%, and 5%, of most interest to financial regulators and institutions (Basel Committee on Banking Supervision, 2019). One-day-ahead VaR and ES forecasts are made with parameter values estimated within a fixed window of 1000 observations starting from April 4, 1990, for each model (except for nonparametric ones) and each probability level.[9] Then, the adjustment parameters are estimated using a rolling window of 2000 returns and risk forecasts, which moves forward by 1 day at a time. The out-of-sample period for each asset contains 4971 days, between March 27, 2002, and December 31, 2021.

[8]The sample period of the COVID-19 crisis starts on 2020-01-23, the day of the Wuhan lockdown in China, marking the beginning of the pandemic. We use 2021-02-17 as the end of Covid crisis, which is the first day when CL price was higher than its close price in 2019, following the impact of COVID-19.

[9]The fixed window scheme has been widely accepted for parameter estimation of parametric and semiparametric models among practitioners but has been criticized for its lack of model flexibility in adapting to market conditions. To illustrate the effectiveness of our methodology, risk forecasting models considered in this paper are mainly estimated using this fixed window strategy.

5.1 | Model disagreement of risk forecasts

Different characteristics of risk models lead to disagreement among their risk forecasts. One of the benefits of our risk adjustment method is to reduce these disagreements among diverse risk models; this method is not restricted to any model or model set used. To investigate the disagreement level of pre- and post-adjusted risk forecasts, we follow Danielsson et al. (2016) to calculate the ratio of the maximum to the minimum tail risk forecasts—referred to as the risk ratio—by utilizing all candidate models. At time $t + 1$, the risk ratio for one asset is defined as

$$RiskRatio_{t+1} = \frac{\max(TailRisk_{t+1})_{n=1}^{N}}{\min(TailRisk_{t+1})_{n=1}^{N}}, \tag{16}$$

where N is the number of candidate models, which is equal to 10 in this paper. According to the definition, the risk ratio must be greater than one. The risk ratio is approaching one, indicating less model disagreement.

Panel A in Table 2 displays average risk ratios from pre-adjusted 1% daily VaR forecasts by using 10 candidate models and by excluding one particular model over the full out-of-sample period and during seven energy crisis periods. Overall, risk model disagreement for oil-related products, including WTI CL, HO, and RBOB/Unleaded Gasoline, differs from NG. First, when all candidate models are considered, a larger disagreement between risk forecasts can be observed in oil-related products (risk ratio is more than 2.28) relative to NG products (risk ratio is 1.79). Second, for oil-related products, the risk ratio is relatively low when the CARE-SAV model is excluded, but the lowest risk ratio is observed after excluding the EVT-POT model in the NG market. It is interesting to note that both the CARE-SAV and EVT-POT models produce the highest tail risk in the (oil products) NG market, since a selection of threshold is required when implementing these procedures. In a crisis period, a biased selection of threshold may result in significant estimation errors. Last, during 2008 global financial crisis and 2020 global Covid crisis, risk disagreement for oil-related products is higher than NG, especially for WTI CL. This is because CL is a highly financialized energy commodity and can be significantly influenced by shocks.

Market conditions potentially have impacts on risk ratios—for example, a change in political environment (i.e., wars), supply/demand disruptions (i.e., hurricanes), and so on. Inspired by the key geopolitical and economic events that affect oil prices, we consider eight events as energy crises, covering wars, natural disasters, a financial crisis, OPEC supply disruptions, and Covid global pandemic.[10] The mean of risk ratios during the crisis period in Panel A in Table 2 shows that risk disagreement is relatively high during the crisis period. For oil-related products, risk disagreement is extremely high during the 2003 Iraq War, 2008 Global Financial Crisis, 2008 OPEC cuts in production, 2014–2016 oil prices collapse, and 2020 Covid crisis. However, a high risk ratio for NG is observed during Hurricane Ivan and Hurricane Dennis (in 2004 and 2005, respectively), which resulted in a reduction in the production of NG.

Panel B in Table 2 shows risk ratios obtained from post-adjusted VaR forecasts. First, compared to Panel A in Table 2, the risk ratios of VaR forecasts calculated from all candidate models witness a drop after utilizing an adjustment method to 1% daily VaR over the out-of-sample period.[11] For example, the risk ratio decreases from 3.06 and 1.79 to 2.54 and 1.56 in the WTI CL market and NG market, respectively. Second, the decrease in risk disagreement after adjusting VaR forecasts can also be observed in crisis periods. In the 2008 financial crisis period, VaR forecasts from different risk models are highly divergent in the WTI CL market (average risk ratio in this period is 3.08). This disagreement reduces to 2.24 after applying the risk adjustment method. Last, after the VaR adjustment, the risk ratios are more stable even if one risk model is excluded from the candidate models. This means that the risk adjustment method leads to the VaR level becoming less sensitive to one particular model.

Panels A and B in Table 3 show the risk ratios calculated from pre- and post-adjusted ES, respectively. Consistent with the VaR disagreement, the ES disagreement also lessens after applying the adjustment method in the full sample and most of the crisis periods. Thus, the risk adjustment method not only helps reduce the VaR disagreement but also improves the ES disagreement.

In addition, the change in margin requirements often happens in the energy commodity futures market, as regulators intend to adjust the margin level to stabilize the market in the face of high fluctuating commodity futures prices. For example, CME Clearing uses a variety of VaR-based models to determine their benchmark margin levels.[12]

[10]See https://www.eia.gov/finance/markets/crudeoil/spot_prices.php for more details.
[11]For more details on the 2.5% and 5% VaR and ES, see Tables C1 and C2, respectively.
[12]See https://www.cmegroup.com/clearing/risk-management/futures-and-options-margin-model.html for more details.

TABLE 2 Risk ratio sensitivity: 1% VaR.

		Panel A: Pre-adjusted										
Energy	Sample/Events	None	HS1000	WHS	CF	G-s	G-skt	EVT-POT	GAS-1F	FHS	CAViaR-SAV	CARE-SAV
CL	Full Sample	3.06	2.96	3.04	2.52	3.05	3.06	3.06	2.91	3.06	3.00	2.53
	2003 Iraq War	3.17	2.88	3.10	3.17	3.16	3.17	3.17	3.10	3.17	3.17	2.01
	2004 Hurricane Ivan	2.37	2.37	2.34	2.31	2.37	2.37	2.37	2.27	2.37	2.37	1.87
	2005 Hurricane Dennis	2.61	2.61	2.61	2.55	2.61	2.61	2.61	2.19	2.61	2.61	1.95
	2006 Nigerian Cuts	2.32	2.32	2.27	2.32	2.31	2.32	2.32	2.26	2.32	2.32	1.63
	2008 Global Financial Crisis	3.08	2.98	3.06	3.08	3.07	3.08	3.08	3.02	3.07	3.07	2.07
	2008 OPEC Cuts Production	3.08	2.95	3.06	3.08	3.08	3.08	3.08	3.08	3.08	3.06	2.11
	2014–2016 Oil Prices Collapse	2.86	2.70	2.82	2.81	2.86	2.86	2.86	2.85	2.86	2.85	1.98
	2020 Covid Crisis	7.81	6.63	7.81	4.56	7.81	7.81	7.80	7.69	7.81	7.55	7.12
HO	Full Sample	2.28	2.22	2.25	2.18	2.27	2.28	2.26	2.18	2.27	2.26	1.95
	2003 Iraq War	2.65	2.50	2.65	2.65	2.64	2.65	2.64	2.47	2.65	2.65	1.90
	2004 Hurricane Ivan	1.95	1.95	1.87	1.90	1.95	1.95	1.92	1.95	1.95	1.95	1.65
	2005 Hurricane Dennis	2.52	2.52	2.51	2.50	2.51	2.52	2.52	2.20	2.52	2.52	2.03
	2006 Nigerian Cuts	2.00	1.98	1.91	2.00	2.00	2.00	2.00	1.91	2.00	2.00	1.63
	2008 Global Financial Crisis	2.63	2.56	2.54	2.63	2.62	2.63	2.63	2.56	2.63	2.63	2.04
	2008 OPEC Cuts Production	2.73	2.60	2.66	2.73	2.73	2.73	2.72	2.65	2.73	2.73	2.12
	2014–2016 Oil Prices Collapse	2.50	2.30	2.47	2.45	2.49	2.50	2.50	2.48	2.46	2.49	2.05
	2020 Covid Crisis	3.23	2.81	3.15	2.88	3.22	3.23	3.20	3.04	3.23	3.22	2.82
NG	Full Sample	1.79	1.78	1.77	1.76	1.79	1.78	1.70	1.72	1.77	1.78	1.79
	2003 Iraq War	2.14	2.14	2.04	2.14	2.14	2.14	2.02	2.01	2.14	2.11	2.14
	2004 Hurricane Ivan	1.94	1.94	1.92	1.93	1.94	1.94	1.85	1.94	1.94	1.94	1.94
	2005 Hurricane Dennis	1.93	1.93	1.91	1.90	1.93	1.93	1.80	1.91	1.93	1.93	1.93
	2006 Nigerian Cuts	1.70	1.70	1.67	1.70	1.70	1.69	1.56	1.57	1.70	1.69	1.69
	2008 Global Financial Crisis	1.69	1.69	1.68	1.68	1.69	1.67	1.57	1.59	1.68	1.68	1.69
	2008 OPEC Cuts Production	1.87	1.86	1.87	1.87	1.87	1.87	1.69	1.65	1.87	1.87	1.87
	2014–2016 Oil Prices Collapse	1.68	1.68	1.66	1.68	1.68	1.68	1.58	1.68	1.62	1.68	1.66
	2020 Covid Crisis	1.88	1.87	1.87	1.84	1.88	1.88	1.74	1.76	1.88	1.87	1.88
XB	Full Sample	2.36	2.31	2.34	2.12	2.36	2.36	2.36	2.14	2.36	2.35	2.18
	2003 Iraq War	2.02	1.93	2.01	2.02	1.99	2.02	2.02	1.87	2.02	2.02	1.77
	2004 Hurricane Ivan	1.71	1.71	1.69	1.67	1.70	1.71	1.71	1.62	1.71	1.71	1.55
	2005 Hurricane Dennis	2.37	2.30	2.33	2.34	2.33	2.37	2.37	1.98	2.37	2.37	2.11
	2006 Nigerian Cuts	1.88	1.88	1.85	1.83	1.87	1.88	1.88	1.72	1.88	1.88	1.65
	2008 Global Financial Crisis	2.11	2.07	2.07	2.09	2.10	2.11	2.11	1.90	2.11	2.10	1.89
	2008 OPEC Cuts Production	2.19	2.16	2.11	2.19	2.18	2.19	2.19	1.96	2.19	2.18	2.00
	2014–2016 Oil Prices Collapse	2.21	2.12	2.18	2.21	2.20	2.21	2.21	1.92	2.18	2.20	2.01
	2020 Covid Crisis	4.63	3.99	4.62	3.55	4.62	4.63	4.62	4.24	4.63	4.61	4.35

Note: The mean of the risk ratio of the highest to the lowest daily 1% pre-adjusted (post-adjusted) VaR forecasts in Panel A (in Panel B) for four energy commodities futures indicated in the first column. The second column indicates the sample/event period used for risk ratio calculation, including both full out-of-sample period and eight energy crisis periods. Risk ratios presented in the third column ("None") are calculated from daily VaR forecasts by using all 10 candidate models. The rest of columns display risk ratios when one specific model is excluded to avoid the effect of outlying forecasts. The excluded model is indicated by the column name.

Panel B: Post-adjusted

None	HS1000	WHS	CF	G-s	G-skt	EVT-POT	GAS-1F	FHS	CAViaR-SAV	CARE-SAV
2.54	2.44	2.50	2.14	2.54	2.54	2.54	2.48	2.53	2.53	2.15
2.30	2.15	2.21	2.28	2.30	2.30	2.30	2.30	2.30	2.30	1.63
1.86	1.86	1.69	1.75	1.86	1.86	1.86	1.86	1.86	1.86	1.57
1.78	1.78	1.76	1.74	1.78	1.78	1.78	1.70	1.78	1.78	1.50
1.74	1.70	1.58	1.74	1.74	1.74	1.74	1.74	1.74	1.74	1.46
2.24	2.24	2.21	2.11	2.24	2.24	2.24	2.22	2.23	2.23	1.73
2.18	2.18	2.11	2.04	2.18	2.18	2.18	2.17	2.16	2.17	1.74
2.38	2.32	2.29	2.31	2.38	2.38	2.38	2.37	2.37	2.37	1.74
6.26	5.14	6.26	4.16	6.26	6.26	6.26	6.11	6.26	6.23	5.45
2.12	2.06	2.08	2.07	2.12	2.12	2.12	2.06	2.09	2.11	1.71
2.62	2.57	2.60	2.62	2.62	2.62	2.62	2.54	2.60	2.62	1.70
1.89	1.89	1.70	1.88	1.89	1.89	1.89	1.88	1.89	1.89	1.46
2.27	2.26	2.16	2.27	2.27	2.27	2.27	2.19	2.27	2.27	1.64
1.84	1.79	1.74	1.84	1.84	1.84	1.84	1.79	1.84	1.84	1.55
2.18	2.16	2.10	2.16	2.18	2.18	2.18	2.16	2.18	2.18	1.73
2.26	2.23	2.18	2.21	2.26	2.26	2.26	2.22	2.26	2.25	1.87
2.44	2.30	2.42	2.39	2.44	2.44	2.44	2.42	2.39	2.44	1.78
2.99	2.68	2.92	2.81	2.99	2.99	2.99	2.82	2.99	2.99	2.37
1.56	1.54	1.51	1.54	1.56	1.56	1.56	1.53	1.55	1.54	1.55
1.78	1.78	1.69	1.78	1.78	1.78	1.78	1.77	1.77	1.72	1.77
1.70	1.69	1.48	1.70	1.70	1.70	1.70	1.70	1.70	1.70	1.70
1.51	1.51	1.33	1.49	1.51	1.50	1.51	1.51	1.50	1.50	1.51
1.34	1.34	1.30	1.34	1.34	1.34	1.34	1.34	1.31	1.33	1.32
1.41	1.40	1.40	1.39	1.41	1.41	1.41	1.38	1.41	1.39	1.40
1.34	1.33	1.33	1.34	1.34	1.34	1.34	1.34	1.34	1.32	1.33
1.50	1.49	1.45	1.50	1.50	1.50	1.50	1.47	1.46	1.49	1.49
1.52	1.51	1.50	1.46	1.52	1.52	1.52	1.47	1.52	1.50	1.51
2.29	2.23	2.24	2.12	2.29	2.29	2.29	2.05	2.28	2.28	2.07
1.93	1.90	1.90	1.93	1.93	1.93	1.93	1.83	1.93	1.93	1.58
1.75	1.75	1.67	1.75	1.75	1.75	1.75	1.63	1.75	1.75	1.52
2.26	2.25	2.21	2.24	2.26	2.26	2.26	1.95	2.26	2.25	1.90
1.78	1.78	1.69	1.78	1.78	1.78	1.78	1.72	1.78	1.78	1.49
2.16	2.16	2.11	2.12	2.16	2.16	2.16	1.92	2.15	2.16	1.87
2.22	2.22	2.19	2.15	2.22	2.22	2.22	1.95	2.21	2.22	1.99
2.32	2.24	2.30	2.32	2.32	2.32	2.32	1.94	2.28	2.32	2.08
4.33	3.67	4.18	3.61	4.33	4.33	4.33	3.91	4.33	4.33	4.04

TABLE 3 Risk ratio sensitivity: 1% ES.

		Panel A: Pre-adjusted										
Energy	Sample/Events	None	H-S1000	WHS	CF	G-s	G-skt	EVT-POT	GAS-1F	FHS	CAViaR-SAV	CARE-SAV
CL	Full Sample	3.17	3.10	3.14	2.32	3.16	3.17	2.94	3.14	3.16	3.17	2.86
	2003 Iraq War	2.34	2.20	2.25	2.32	2.33	2.34	2.33	2.34	2.33	2.34	1.78
	2004 Hurricane Ivan	2.08	2.08	1.94	1.94	2.08	2.08	2.00	2.08	2.08	2.08	1.84
	2005 Hurricane Dennis	1.96	1.96	1.92	1.87	1.95	1.96	1.78	1.96	1.96	1.96	1.72
	2006 Nigerian Cuts	2.00	2.00	1.82	2.00	1.99	2.00	1.98	1.99	2.00	2.00	1.66
	2008 Global Financial Crisis	2.72	2.60	2.69	2.72	2.70	2.72	2.72	2.69	2.71	2.72	2.18
	2008 OPEC Cuts Production	2.79	2.53	2.79	2.79	2.79	2.79	2.78	2.72	2.78	2.79	2.26
	2014–2016 Oil Prices Collapse	2.61	2.43	2.61	2.53	2.61	2.61	2.58	2.57	2.61	2.61	2.06
	2020 Covid Crisis	9.79	9.27	9.79	4.24	9.79	9.79	8.74	9.78	9.79	9.79	9.40
HO	Full Sample	2.30	2.26	2.28	2.16	2.29	2.30	2.14	2.24	2.29	2.30	2.09
	2003 Iraq War	2.22	2.15	2.22	2.22	2.21	2.22	2.16	2.20	2.22	2.22	1.79
	2004 Hurricane Ivan	1.99	1.99	1.96	1.99	1.99	1.99	1.79	1.91	1.99	1.99	1.85
	2005 Hurricane Dennis	2.11	2.11	2.01	2.11	2.09	2.11	1.96	2.09	2.11	2.11	1.88
	2006 Nigerian Cuts	1.92	1.92	1.87	1.92	1.91	1.92	1.76	1.88	1.92	1.92	1.76
	2008 Global Financial Crisis	2.38	2.35	2.29	2.38	2.35	2.38	2.34	2.35	2.38	2.38	2.07
	2008 OPEC Cuts Production	2.44	2.39	2.43	2.44	2.43	2.44	2.35	2.39	2.44	2.44	2.13
	2014–2016 Oil Prices Collapse	2.44	2.26	2.44	2.40	2.43	2.44	2.41	2.32	2.39	2.44	2.12
	2020 Covid Crisis	3.26	2.96	3.26	2.70	3.26	3.26	2.93	3.24	3.26	3.26	3.07
NG	Full Sample	1.80	1.80	1.80	1.72	1.80	1.79	1.79	1.69	1.77	1.80	1.80
	2003 Iraq War	2.12	2.12	2.12	2.09	2.12	2.11	2.10	1.87	2.12	2.11	2.10
	2004 Hurricane Ivan	2.06	2.06	2.06	1.95	2.06	2.06	2.05	1.76	2.06	2.06	2.06
	2005 Hurricane Dennis	1.96	1.96	1.96	1.88	1.96	1.96	1.95	1.77	1.96	1.96	1.96
	2006 Nigerian Cuts	1.70	1.70	1.70	1.66	1.70	1.70	1.68	1.54	1.70	1.70	1.70
	2008 Global Financial Crisis	1.68	1.68	1.68	1.64	1.68	1.67	1.67	1.56	1.67	1.68	1.68
	2008 OPEC Cuts Production	1.86	1.85	1.86	1.86	1.86	1.86	1.83	1.60	1.86	1.86	1.86
	2014–2016 Oil Prices Collapse	1.65	1.64	1.65	1.64	1.65	1.65	1.64	1.65	1.55	1.65	1.64
	2020 Covid Crisis	2.02	2.01	2.02	1.80	2.02	2.02	2.00	1.72	2.01	2.02	2.01
XB	Full Sample	2.59	2.56	2.59	2.14	2.59	2.59	2.55	2.39	2.59	2.58	2.40
	2003 Iraq War	1.87	1.83	1.85	1.86	1.83	1.87	1.87	1.75	1.87	1.87	1.69
	2004 Hurricane Ivan	1.73	1.73	1.71	1.70	1.71	1.73	1.68	1.62	1.73	1.73	1.58
	2005 Hurricane Dennis	2.21	2.17	2.19	2.21	2.17	2.21	2.21	1.89	2.21	2.21	2.04
	2006 Nigerian Cuts	1.82	1.82	1.76	1.77	1.80	1.82	1.81	1.75	1.82	1.82	1.64
	2008 Global Financial Crisis	2.10	2.02	2.09	2.06	2.09	2.10	2.08	1.89	2.10	2.10	1.92
	2008 OPEC Cuts Production	2.17	2.07	2.17	2.14	2.15	2.17	2.15	1.94	2.17	2.17	2.02
	2014–2016 Oil Prices Collapse	2.21	2.11	2.20	2.20	2.20	2.21	2.20	1.87	2.18	2.21	2.03
	2020 Covid Crisis	5.75	5.54	5.75	3.24	5.75	5.75	5.58	5.52	5.75	5.75	5.40

Note: The mean of the risk ratio of the highest to the lowest daily 1% pre-adjusted (post-adjusted) ES forecasts in Panel A (in Panel B) for four energy commodities futures indicated in the first column. The second column indicates the sample/event period used for risk ratio calculation, including both full out-of-sample period and eight energy crisis periods. Risk ratios presented in the third column ("None") are calculated from daily ES forecasts by using all 10 candidate models. The rest of columns display risk ratios when one specific model is excluded to avoid the effect of outlying forecasts. The excluded model is indicated by the column name.

Panel B: Post-adjusted

None	HS1000	WHS	CF	G-s	G-skt	EVT-POT	GAS-1F	FHS	CAViaR-SAV	CARE-SAV
2.86	2.82	2.84	2.32	2.86	2.86	2.86	2.80	2.84	2.84	2.39
2.33	2.33	2.30	2.32	2.33	2.33	2.33	2.28	2.30	2.33	1.39
1.81	1.81	1.77	1.80	1.81	1.81	1.81	1.80	1.81	1.80	1.53
1.83	1.83	1.83	1.83	1.83	1.83	1.83	1.68	1.83	1.82	1.49
1.71	1.67	1.66	1.71	1.71	1.71	1.71	1.71	1.71	1.71	1.35
2.41	2.38	2.41	2.30	2.41	2.41	2.41	2.39	2.40	2.41	1.72
2.22	2.12	2.22	2.21	2.22	2.22	2.22	2.20	2.17	2.22	1.66
2.56	2.50	2.51	2.52	2.56	2.56	2.56	2.55	2.55	2.55	1.69
7.65	7.28	7.65	4.01	7.65	7.65	7.65	7.51	7.65	7.55	7.08
2.21	2.18	2.19	2.15	2.21	2.21	2.21	2.13	2.19	2.20	1.76
2.37	2.37	2.36	2.33	2.37	2.37	2.37	2.23	2.35	2.37	1.53
1.79	1.77	1.70	1.79	1.79	1.79	1.79	1.78	1.79	1.79	1.44
2.15	2.13	2.13	2.15	2.15	2.15	2.15	1.98	2.15	2.15	1.58
1.83	1.78	1.77	1.83	1.83	1.83	1.83	1.78	1.83	1.83	1.50
2.26	2.26	2.21	2.24	2.26	2.26	2.26	2.24	2.26	2.26	1.65
2.20	2.20	2.18	2.16	2.20	2.20	2.20	2.15	2.20	2.19	1.66
2.57	2.47	2.57	2.53	2.57	2.57	2.57	2.54	2.52	2.56	1.71
3.21	3.04	3.18	2.95	3.21	3.21	3.21	2.94	3.21	3.21	2.60
1.60	1.58	1.58	1.56	1.60	1.60	1.60	1.56	1.58	1.58	1.59
1.83	1.83	1.80	1.83	1.83	1.83	1.83	1.83	1.83	1.73	1.79
1.65	1.60	1.60	1.65	1.65	1.65	1.65	1.65	1.65	1.65	1.65
1.50	1.50	1.41	1.49	1.50	1.50	1.50	1.50	1.50	1.49	1.50
1.33	1.33	1.31	1.33	1.33	1.33	1.33	1.32	1.26	1.32	1.33
1.44	1.43	1.44	1.44	1.44	1.44	1.44	1.39	1.44	1.43	1.43
1.37	1.37	1.37	1.36	1.37	1.37	1.37	1.36	1.37	1.36	1.36
1.45	1.44	1.43	1.44	1.45	1.45	1.45	1.43	1.39	1.43	1.44
1.68	1.68	1.68	1.53	1.68	1.68	1.68	1.59	1.68	1.65	1.67
2.44	2.42	2.41	2.18	2.44	2.44	2.44	2.23	2.42	2.43	2.18
2.00	2.00	1.95	2.00	2.00	2.00	2.00	1.87	1.99	2.00	1.56
1.74	1.72	1.70	1.74	1.74	1.74	1.74	1.66	1.74	1.74	1.48
2.34	2.33	2.28	2.33	2.34	2.34	2.34	2.03	2.34	2.34	1.90
1.88	1.88	1.77	1.88	1.88	1.88	1.88	1.82	1.88	1.88	1.52
2.25	2.22	2.21	2.24	2.25	2.25	2.25	2.04	2.24	2.25	1.86
2.26	2.22	2.26	2.24	2.26	2.26	2.26	2.03	2.24	2.25	1.92
2.39	2.37	2.37	2.38	2.39	2.39	2.39	2.15	2.34	2.39	1.96
5.01	4.90	5.01	3.19	5.01	5.01	5.01	4.68	5.01	5.01	4.74

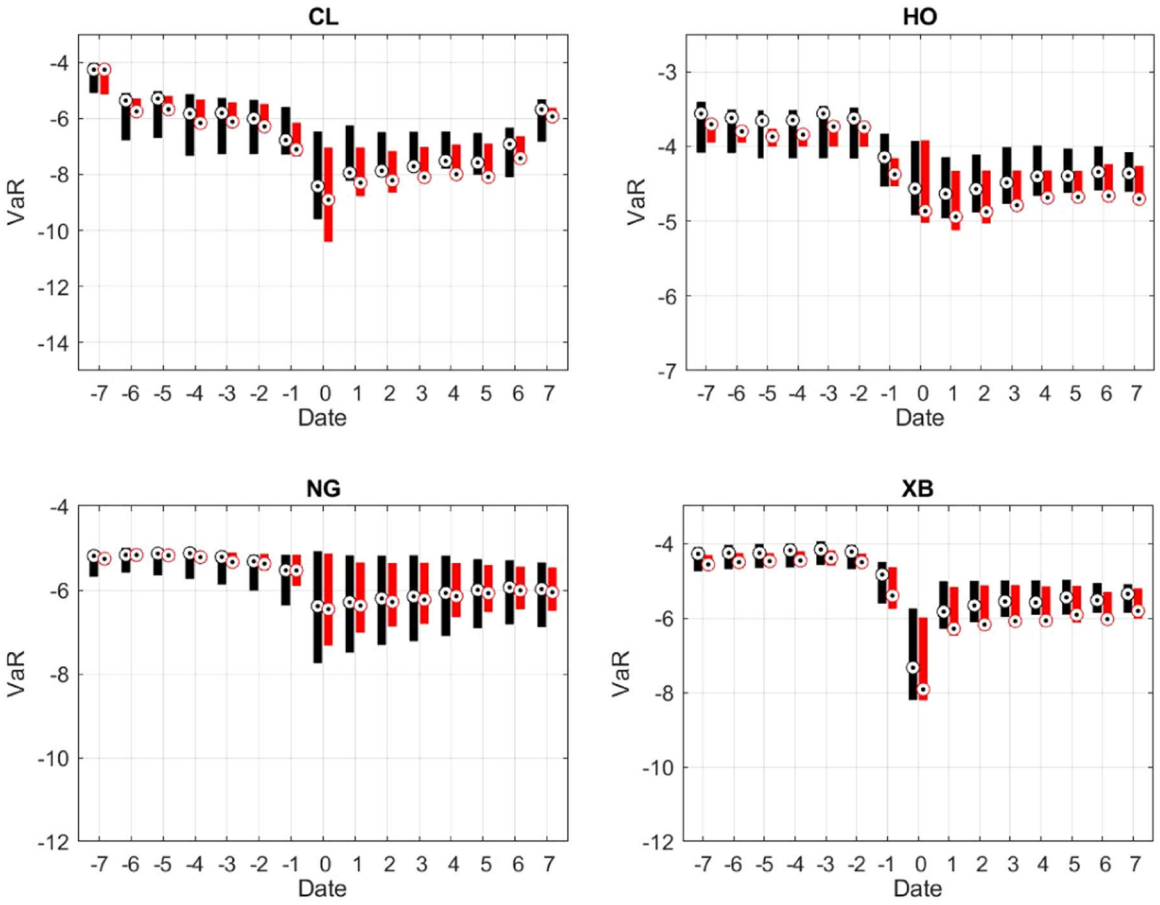

FIGURE 2 The range of average pre- and post-adjusted 2.5% VaR 7 days before and after the increase in margin requirements for CL, HO, NG and XB. This boxplot represents the range of average VaR forecasts from 10 risk models around the days of margin requirements increase. We first take the average of VaR forecasts from each risk model on (and specific days before and after) the event day. Then plot the range of these average VaR forecasts across 10 different models. The black box is for pre-adjusted 2.5% VaR, and the red box is for post-adjusted 2.5% VaR. The dot in the middle of each box is the median of the sample. The bottom and top of each box are the 25th and 75th percentiles of the sample. On the *x*-axis, zero indicates the day that margin requirements increase, and a negative (positive) number means the number of days before (after) the day of that margin increase. [Color figure can be viewed at wileyonlinelibrary.com]

As an additional exercise, we would like to explore the efficiency of this adjustment methodology around the margin level change date that characterizes the market state as highly volatile (Hedegaard, 2014; Park & Abruzzo, 2016).

Figure 4 shows the maintenance margin of the front-month contract for WTI CL, NG, HO, and RBOB/Unleaded Gasoline between 2009 and 2021.[13,14] Large margin variation could be observed over time for all four energy commodities. Oil products share common factors that have impacts on their supply/demand side, which results in high risk in the market. Regulators could change the margin level across oil products to stabilize the market. Thus, the changes in the margin level for WTI CL, HO, and RBOB/Unleaded Gasoline have a similar trend. As the most traded products, the margin level of WTI CL changes more frequently than the other oil products.

When market uncertainty, particularly the level of extreme losses, is high, regulators can raise the margin levels to calm down the market. In Figures 2 and 3, we plot the range of pre- and post-adjusted 2.5% daily VaR and ES forecasts calculated from 10 candidate models around the days when maintenance margins are increased for WTI CL, NG, HO, and RBOB/Unleaded Gasoline.[15] First, before CME Group raises the maintenance margins, the median of both pre- and post-adjusted 2.5% daily VaR forecasts falls significantly. This implies that CME Group steps into the market

[13]CME Group usually sets the maintenance margin at first and then sets initial margins by adjusting the maintenance margin with a specific ratio, which almost always remains the same for a contract.
[14]See https://www.cmegroup.com/clearing/risk-management/historical-margins.html for more details.
[15]We use 2.5% significant level since the Basel III regulations propose a shift from a 1% VaR framework to a 2.5% ES.

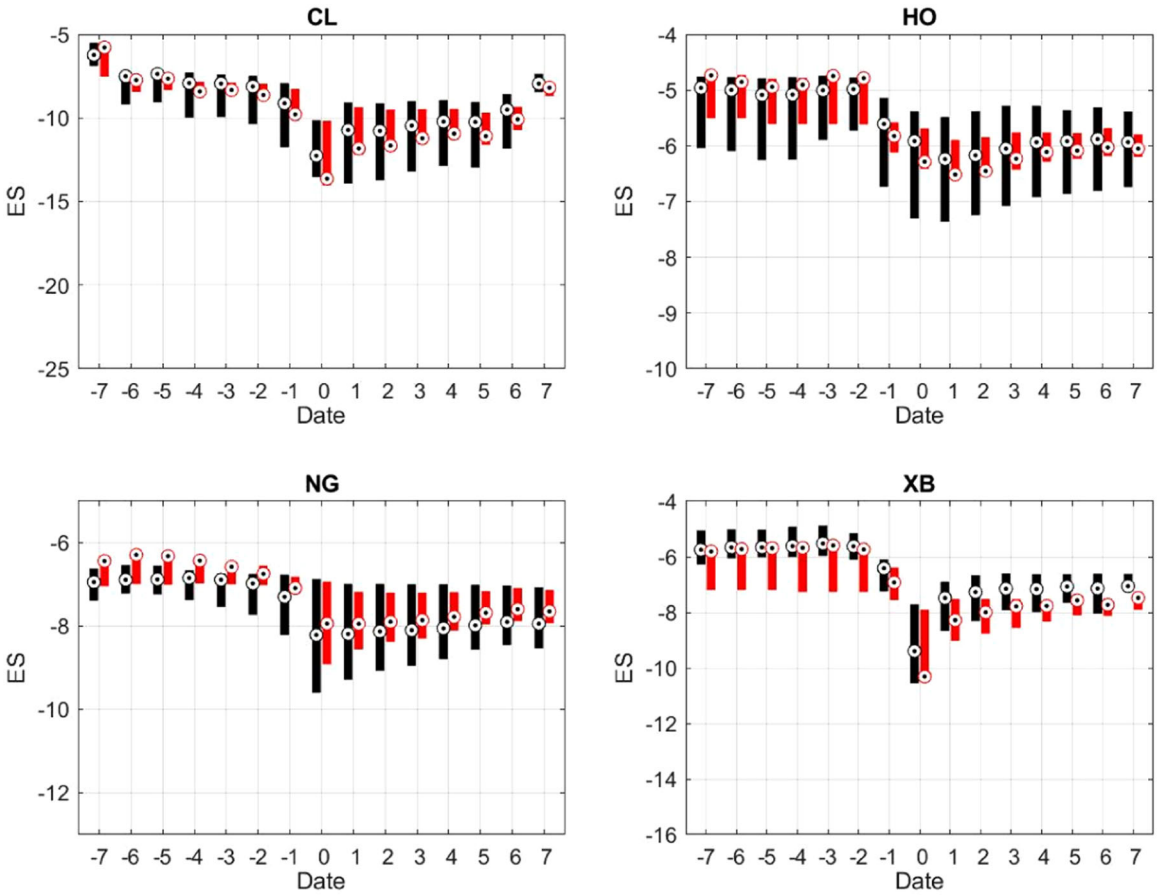

FIGURE 3 The range of pre- and post-adjusted 2.5% ES 7 days before and after the increase in margin requirements for CL, HO, NG, and XB. This boxplot represents the range of average ES forecasts from 10 risk models. We first take the average of ES forecasts from each risk model on (and specific days before and after) the event day. Then plot the range of these average ES forecasts across 10 different models. The black box is for pre-adjusted 2.5% ES, and the red box is for post-adjusted 2.5% ES. The dot in the middle of each box is the median of the sample. The bottom and top of each box are the 25th and 75th percentiles of the sample. On the *x*-axis, zero indicates the day that margin requirements increase, and a negative (positive) number means the number of days before (after) the day of that margin increase. [Color figure can be viewed at wileyonlinelibrary.com]

with a margin increase policy only when it observes the tail risk measurements reaching an extremely high level. After the increase in margin levels, the downward trend in both pre- and post-adjusted VaR and ES has stopped; thus, a margin increase policy is an effective way to reduce tail risk and calm down the market. Hence, the risk model adjustment methodology could successfully capture the trend of tail risk movements. Second, the range of post-adjusted VaR (ES) forecasts around the time with a margin level increase is significantly narrower than that of the pre-adjusted VaR (ES) forecasts in general. This suggests that our adjustment methodology to the individual risk model is effective at reducing tail risk disagreements between different models, even in super-chaotic periods.

5.2 | VaR and ES backtesting

To justify the efficiency of adjustments made to risk estimates, it is necessary to investigate the forecasting ability of adjusted VaR and ES forecasts. At first glance, Table 4 compares the actual VaR exceedance rates before (resp. after) applying optimized adjustment multipliers to improve on the out-of-sample raw risk estimates at $\alpha = 1\%, 2.5\%$, and 5% with the expected VaR exceedance rates (i.e., three α levels), for various risk models and several energy futures. This highlights that the improved VaR forecasts can achieve a desirable level of actual VaR exceedances, which generally matches the expected level. Tables 5 and 6 compare the average $FZ0$ loss of pre- and post-adjusted risk forecasts estimated at 1%, 2.5%, and 5% for 10 candidate models and four energy futures over the full out-of-sample period and during energy crisis periods, respectively. The lower the average loss, the better forecasting performance the pair of risk

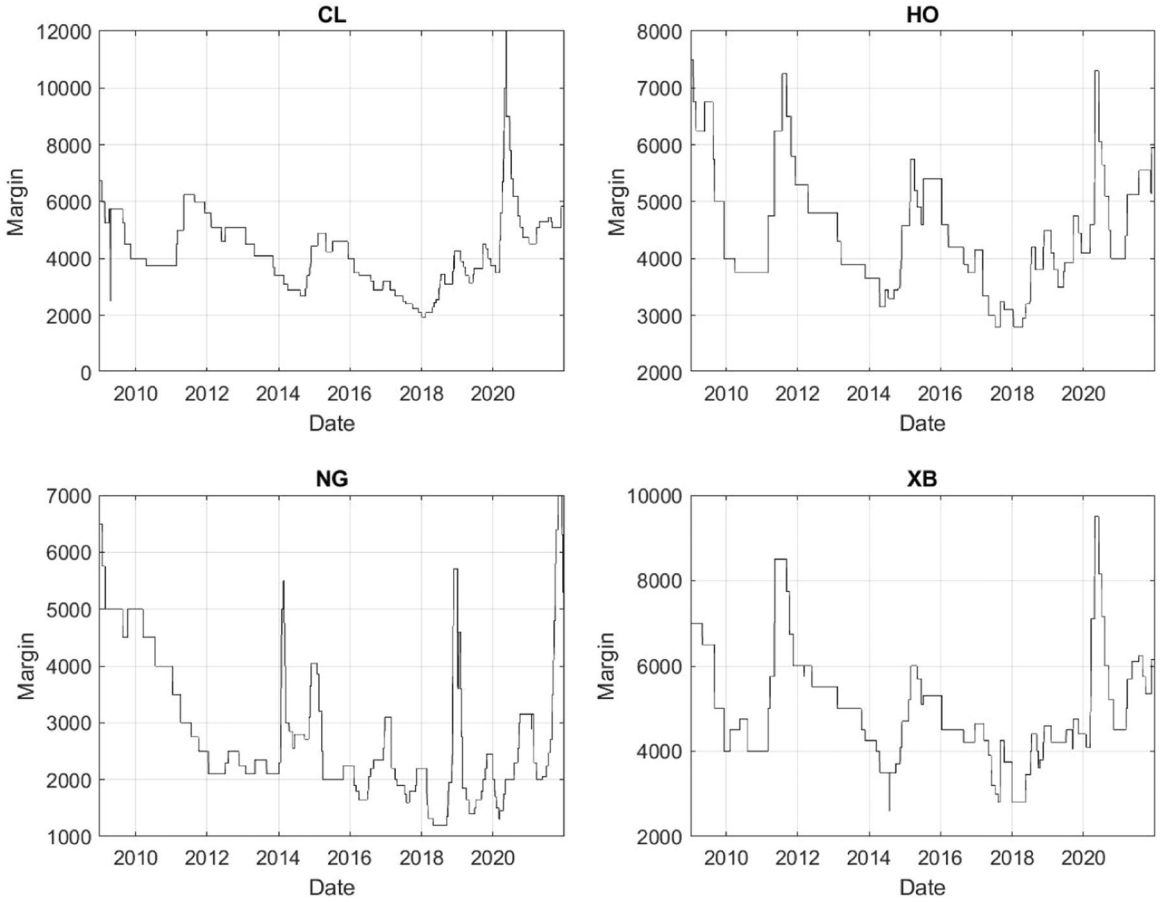

FIGURE 4 Historical margin requirements from Chicago Mercantile Exchange (CME) Group for first front contracts of CL, HO, NG, and XB between January 2, 2009, and December 31, 2021.

forecasts have. The results show that our adjustment method indeed improves the forecasting ability of raw risk estimates.

In the following, we adopt the formal statistical backtesting procedures including traditional tests and comparative tests. The traditional backtesting methods are designed to test whether the forecasting model is correctly specified and provides appropriate forecasts by comparing the realized returns, VaR and ES data, whereas the comparative backtesting methods focus on making model comparisons. By examining the backtesting performance of pre- and post-adjusted risk forecasts via these two types of backtesting methodologies in an exhaustive manner, we check whether the risk model performance is improved after applying the proposed adjustment method.

Regarding the traditional backtesting, we organize well-known and recently developed tests into two broad categories: (1) VaR backtests and (2) ES backtests.[16] We backtest VaR individually via the unconditional coverage (UC) test of Kupiec (1995), the conditional coverage (CC) test of Christoffersen (1998), and the dynamic quantile (DQ) test of Engle and Manganelli (2004). For the evaluation of the VaR forecasting accuracy, the UC test simply considers the exceedance frequency and is deficient in detecting clustered exceedances, which is made up by CC and DQ tests. Both CC and DQ tests are designed to jointly test for the frequency and independence of exceedances.

Compared with VaR backtesting, ES backtesting is not straightforward as it often needs more information (such as VaR estimates at multiple levels required in Emmer et al., 2015) relative to the ES per se. To backtest ES, we employ the exceedance residual (ER) test of McNeil and Frey (2000), the conditional calibration (CCA) test based on moment conditions (Nolde & Ziegel, 2017), and several specifications of regression based expected shortfall (ESR) tests

[16]For more details of VaR and ES backtesting methods, see Appendices A and B, respectively.

TABLE 4 Actual VaR exceedance rates of pre- and post-adjusted risk forecasts over full sample.

	Panel A: CL						Panel B: HO					
	1%		2.5%		5%		1%		2.5%		5%	
	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After
HS	1.53%	1.01%	2.90%	2.57%	5.45%	4.73%	1.41%	0.99%	2.90%	2.45%	5.39%	4.71%
WHS	1.35%	1.03%	2.72%	2.51%	5.11%	4.87%	1.25%	1.05%	2.57%	2.43%	5.13%	4.79%
CF	0.76%	1.01%	2.45%	2.49%	5.37%	4.63%	0.80%	0.87%	2.49%	2.41%	5.41%	4.87%
G-t	0.42%	1.11%	1.33%	2.45%	2.88%	5.11%	0.34%	1.07%	1.23%	2.55%	3.00%	5.03%
G-skt	1.17%	1.11%	2.82%	2.45%	6.18%	5.11%	1.17%	1.05%	3.00%	2.53%	6.04%	5.07%
EVT-POT	1.41%	1.07%	2.92%	2.43%	6.22%	5.09%	1.81%	1.03%	3.70%	2.51%	6.38%	5.07%
GAS-1F	2.13%	0.99%	4.77%	2.53%	6.08%	4.93%	2.03%	1.03%	3.42%	2.41%	6.64%	4.91%
FHS	1.13%	1.11%	2.39%	2.60%	4.89%	5.03%	1.05%	1.07%	2.55%	2.49%	5.19%	4.95%
CAViaR-SAV	1.77%	1.17%	2.72%	2.55%	4.49%	5.31%	1.87%	0.99%	3.38%	2.45%	5.47%	5.21%
CARE-SAV	0.64%	1.31%	1.97%	2.70%	4.33%	5.21%	1.23%	1.11%	2.35%	2.76%	4.63%	5.47%
	Panel C: NG						Panel D: XB					
	1%		2.5%		5%		1%		2.5%		5%	
	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After
HS	0.99%	1.11%	2.51%	2.55%	4.93%	4.95%	1.41%	0.91%	2.92%	2.37%	5.29%	4.95%
WHS	1.19%	1.07%	2.74%	2.66%	5.27%	5.11%	1.19%	1.03%	2.53%	2.49%	4.97%	4.93%
CF	0.70%	0.99%	2.45%	2.53%	5.19%	5.07%	0.80%	0.76%	2.72%	2.41%	5.15%	4.85%
G-t	0.32%	0.99%	1.13%	2.37%	3.14%	4.97%	0.95%	1.15%	2.29%	2.68%	4.20%	5.19%
G-skt	0.78%	0.97%	2.66%	2.39%	5.69%	4.97%	1.49%	1.15%	3.52%	2.68%	6.36%	5.19%
EVT-POT	0.22%	0.97%	0.99%	2.39%	4.18%	4.97%	1.65%	1.15%	3.50%	2.68%	6.04%	5.13%
GAS-1F	1.73%	0.99%	5.53%	2.29%	6.04%	5.15%	1.93%	1.09%	3.26%	2.70%	5.41%	5.09%
FHS	1.05%	0.99%	2.49%	2.43%	4.85%	4.87%	1.09%	1.03%	2.62%	2.43%	5.11%	5.13%
CAViaR-SAV	0.78%	1.01%	1.69%	2.43%	4.73%	4.99%	1.75%	1.17%	3.36%	2.64%	5.59%	5.13%
CARE-SAV	0.74%	0.93%	2.74%	2.21%	6.80%	4.77%	1.59%	1.11%	2.74%	2.68%	4.97%	5.19%

Note: This table presents the actual VaR exceedance rates of pre- and post-adjusted risk forecasts based on the *FZ*0 loss function across four energy futures and three significance levels. Results based on pre- and post-adjusted VaR forecasts are labeled as column "before" and "after", respectively.

including the strict ESR (ESR Strict) test, the auxiliary ESR (ESR AUX) test, and the intercept ESR (ESR INT) test, introduced by Bayer and Dimitriadis (2022).

In the format of colormaps, Figures 5, C1, and C2 visually display the *p*-values of backtests for VaR and ES forecasts at the 1%, 2.5%, and 5% level, respectively, over the out-of-sample time period (27-Mar-2002 to 31-Dec-2021) and for various models and four futures, before (shown in the left column) and after (shown in the right column) adjustments are made based on our proposed methodology. For the sake of brevity and clarity, the backtests are labeled 1–11, representing the following tests: UC, CC, DQ, two-sided ER, one-sided ER, two-sided CCA, one-sided CCA, ESR Strict, ESR AUX, two-sided ESR INT, and one-sided ESR INT, accordingly; the risk models are numbered 1–10, denoting HS, WHS, CF, G-*t*, G-*skt*, EVT-POT, GAS-1F, FHS, CAViaR-SAV, and CARE-SAV, respectively. If the *p*-value is smaller than 0.05, the cell is colored red; if the value falls between 0.05 and 0.1, the cell is in yellow; otherwise, it is in green. The red and yellow colors suggest that the model fails the test at the 5% and 10% significance levels, respectively. Regarding all the aforementioned backtests, the forecasting performance of risk models considered in this paper is indeed improved after our adjustment methodology is applied to improve on raw risk forecasts. Notably, the UC test (backtest 1) with respect to the frequency of VaR exceedances and the ER test focused on the magnitude of ES forecasts (backtests 4–5) benefit most from adjustments made to VaR and ES forecasts. The UC backtesting results indicate that the actual VaR exceedance rate becomes closer to the nominal α level after making the optimized adjustments. And the

TABLE 5 Average loss values of pre- and post-adjusted risk forecasts over the full sample.

	Panel A: CL						Panel B: HO					
	1%		2.5%		5%		1%		2.5%		5%	
	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After
HS	2.29	**2.28**	1.99	**1.99**	1.75	**1.75**	2.05	**2.03**	1.78	**1.77**	1.58	**1.57**
WHS	2.03	2.03	1.79	1.79	1.61	1.61	1.85	**1.85**	1.63	1.63	1.45	1.45
CF	2.27	2.31	2.00	2.03	1.77	1.78	1.98	1.98	1.75	**1.75**	1.57	**1.57**
G-*s*	2.02	**1.97**	1.79	**1.75**	1.59	**1.55**	1.88	**1.80**	1.65	**1.60**	1.47	**1.43**
G-*skt*	1.97	**1.97**	1.75	1.75	1.56	**1.55**	1.80	1.81	1.60	1.60	1.44	**1.43**
EVT-POT	2.01	**1.97**	1.77	**1.75**	1.57	**1.55**	1.88	**1.80**	1.64	**1.60**	1.45	**1.43**
GAS-1F	2.03	**1.97**	1.78	**1.73**	1.55	**1.55**	1.88	**1.82**	1.62	**1.62**	1.44	**1.43**
FHS	1.96	1.97	1.74	1.74	1.55	1.55	1.81	1.81	1.58	1.59	1.42	1.42
CAViaR-SAV	2.03	**1.99**	1.76	**1.76**	1.57	**1.56**	1.85	**1.82**	1.61	**1.60**	1.43	**1.43**
CREA-SAV	2.33	**2.30**	1.95	**1.94**	1.61	**1.61**	2.15	2.16	1.99	**1.96**	1.61	**1.61**
# imp.	7		6		7		6		7		8	
	Panel C: NG						**Panel D: XB**					
	1%		2.5%		5%		1%		2.5%		5%	
	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After
HS	2.22	2.23	2.03	2.03	1.85	1.86	2.23	**2.20**	1.96	**1.94**	1.73	**1.72**
WHS	2.24	**2.24**	2.00	2.00	1.82	1.82	2.01	2.01	1.78	1.79	1.61	1.61
CF	2.21	**2.21**	2.02	2.02	1.85	1.86	2.17	2.17	1.93	**1.93**	1.72	1.72
G-*s*	2.14	**2.06**	1.94	**1.90**	1.78	**1.75**	1.93	1.94	1.73	1.73	1.56	**1.56**
G-*skt*	2.06	2.07	1.89	1.90	1.75	1.75	1.96	**1.94**	1.76	**1.73**	1.57	**1.56**
EVT-POT	2.21	**2.07**	1.94	**1.90**	1.75	**1.75**	1.99	**1.94**	1.77	**1.73**	1.58	**1.56**
GAS-1F	2.25	**2.19**	2.09	**1.97**	1.82	**1.82**	2.08	**2.02**	1.79	**1.77**	1.61	**1.60**
FHS	2.05	2.06	1.88	1.89	1.74	1.74	1.93	1.93	1.73	1.74	1.55	1.56
CAViaR-SAV	2.06	**2.06**	1.91	**1.90**	1.75	**1.74**	2.01	**1.96**	1.77	**1.74**	1.57	**1.56**
CREA-SAV	2.16	2.16	1.88	1.88	1.75	**1.74**	2.30	**2.21**	1.99	**1.96**	1.73	**1.71**
# imp.	6		4		5		6		7		7	

Note: Each panel presents average *FZ*0 loss values of 10 candidate models before and after optimized adjustments are made to VaR and ES estimated at 1%, 2.5%, and 5%, for a given asset. Results based on pre- and post-adjusted VaR and ES forecasts are labeled as columns "before" and "after," respectively. The average loss values indicating the outperformance of post-adjusted forecasts than pre-adjusted forecasts are highlighted in bold. # imp. denotes the number of models that experience improvements after adjustment.

ER backtesting results show that the magnitude of ES forecasts becomes reasonable (i.e., neither overestimating nor underestimating risks) and highlights that our proposed adjustment methodology can facilitate market risk measures in capturing the appropriate size of tail losses. It is expected that, in terms of the CC and DQ tests (backtests 2–3) the backtesting performance of risk models is slightly affected by our adjustment methodology, which is limited in removing the endogenous issue of volatility clustering. Additionally, we find that G-*t*, G-*skt*, EVT-POT, GAS-1F, CAViaR-SAV, and CARE-SAV are unable to produce adequate risk forecasts at different levels for various energy futures considered, as shown by the high rejection rates of backtests. After making the refinements to these models based on our methodology, the rejection rates sharply decline.

As seen from Figure 1, which shows the dynamics of daily returns of futures contracts, the market regimes are clearly changed during crisis periods, except for NG futures in a consistently high-volatility regime. To further test the efficiency of our adjustment methodology, in similar fashion we present the backtesting results of risk models over the global financial crisis period ranging from 01-Dec-2007 to 30-Jun-2009 in Figures 6, C3 and C4. As expected, the

TABLE 6 Average loss values of pre- and post-adjusted risk forecasts over crisis periods.

	Panel A: CL						**Panel B: HO**					
	1%		**2.5%**		**5%**		**1%**		**2.5%**		**5%**	
	Before	**After**	**Before**	**After**	**Before**	**After**	**Before**	**After**	**Before**	**After**	**Before**	**After**
HS	3.39	**3.11**	2.86	**2.71**	2.49	**2.39**	2.85	**2.61**	2.40	**2.28**	2.10	**2.04**
WHS	2.65	**2.59**	2.27	**2.26**	2.06	**2.06**	2.31	**2.27**	2.00	**2.00**	1.80	1.81
CF	2.88	3.00	2.62	2.66	2.38	**2.38**	2.42	2.48	2.22	**2.23**	2.06	**2.04**
G-*s*	2.33	**2.28**	2.08	**2.06**	1.89	**1.87**	2.12	**2.11**	1.90	**1.89**	1.71	**1.70**
G-*skt*	2.30	**2.28**	2.07	**2.06**	1.88	**1.87**	2.09	2.11	1.89	**1.89**	1.71	**1.70**
EVT-POT	2.36	**2.28**	2.10	**2.06**	1.90	**1.87**	2.23	**2.11**	1.95	**1.89**	1.74	**1.70**
GAS-1F	2.43	**2.35**	2.15	**2.06**	1.88	**1.87**	2.19	**2.10**	1.89	**1.86**	1.70	**1.68**
FHS	2.32	**2.32**	2.08	2.08	1.88	1.88	2.20	2.21	1.90	1.90	1.71	**1.71**
CAViaR-SAV	2.39	**2.31**	2.12	**2.10**	1.89	**1.88**	2.17	**2.14**	1.90	**1.89**	1.70	**1.70**
CREA-SAV	2.61	**2.55**	2.26	**2.24**	1.92	**1.91**	2.57	**2.55**	2.32	**2.27**	1.89	**1.88**
# imp.	9		8		9		7		9		9	
	Panel C: NG						**Panel D: XB**					
	1%		**2.5%**		**5%**		**1%**		**2.5%**		**5%**	
	Before	**After**	**Before**	**After**	**Before**	**After**	**Before**	**After**	**Before**	**After**	**Before**	**After**
HS	2.16	2.18	2.03	2.04	1.88	1.90	3.19	**2.93**	2.70	**2.57**	2.38	**2.31**
WHS	2.19	**2.19**	2.02	2.02	1.86	1.86	2.57	**2.54**	2.26	2.26	2.05	2.06
CF	2.17	2.18	2.01	2.03	1.88	1.89	2.70	2.74	2.50	**2.47**	2.27	**2.25**
G-*s*	2.21	**2.13**	2.02	**2.00**	1.88	**1.86**	2.24	2.28	2.04	2.06	1.86	**1.87**
G-*skt*	2.13	2.13	2.01	**2.00**	1.87	**1.86**	2.34	**2.28**	2.11	**2.06**	1.90	**1.87**
EVT-POT	2.27	**2.13**	2.02	**2.00**	1.86	**1.86**	2.38	**2.27**	2.13	**2.06**	1.91	**1.87**
GAS-1F	2.24	**2.18**	2.16	**1.99**	1.86	**1.85**	2.46	**2.30**	2.14	**2.08**	1.95	**1.92**
FHS	2.10	**2.10**	1.96	1.96	1.84	**1.84**	2.31	2.33	2.09	**2.09**	1.89	1.90
CAViaR-SAV	2.12	**2.12**	2.00	**2.00**	1.85	**1.85**	2.46	**2.34**	2.15	**2.10**	1.91	**1.89**
CREA-SAV	2.12	2.12	1.95	1.96	1.85	**1.82**	2.88	**2.72**	2.40	**2.36**	2.07	**2.05**
# imp.	7		5		7		7		8		8	

Note: Each panel presents average *FZ*0 loss values of 10 candidate models before and after optimized adjustments are made to VaR and ES estimated at 1%, 2.5% and 5%, for a given asset. Results based on pre- and post-adjusted VaR and ES forecasts are labeled as column "before" and "after," respectively. The average loss values indicating the outperformance of post-adjusted forecasts than preadjusted forecasts are highlighted in bold. # imp. denotes the number of models that experiencing improvements after adjustment.

simple models, HS and CF, are less responsive to large market fluctuations and fail most of the backtests, whereas the more complex models such as G-*t*, G-*skt*, CAViaR-SAV, and CARE-SAV perform better over the crisis period. Generally, the risk models can generate better risk forecasts after the adjustment methodology is applied, as evidenced by the better backtesting performance in Figures 6, C3 and C4.

This adjustment methodology not only can improve on ex ante risk forecasts but also can give an indication of time-varying over- or underestimation of risks. Figure 7 presents a series of quantiles of optimal adjustments obtained in Equation (5), made to VaR (in the left column) and ES (in the right column) estimated at $\alpha = 2.5\%$ for several energy futures contracts and across various models, in which each line represents the quantile ranging from 5% to 95% with an increment of 10%. The value of the optimal adjustment being equal to 1 indicates that no correction will be made to VaR and ES forecasts. As shown in this figure, the underestimation or overestimation of risks is noticeable, thus calling for appropriate adjustments. Specifically, if the value of adjustment is above (below) 1, the current risk forecast is underestimated (overestimated), and the corresponding capital buffer should be increased (decreased). Among the

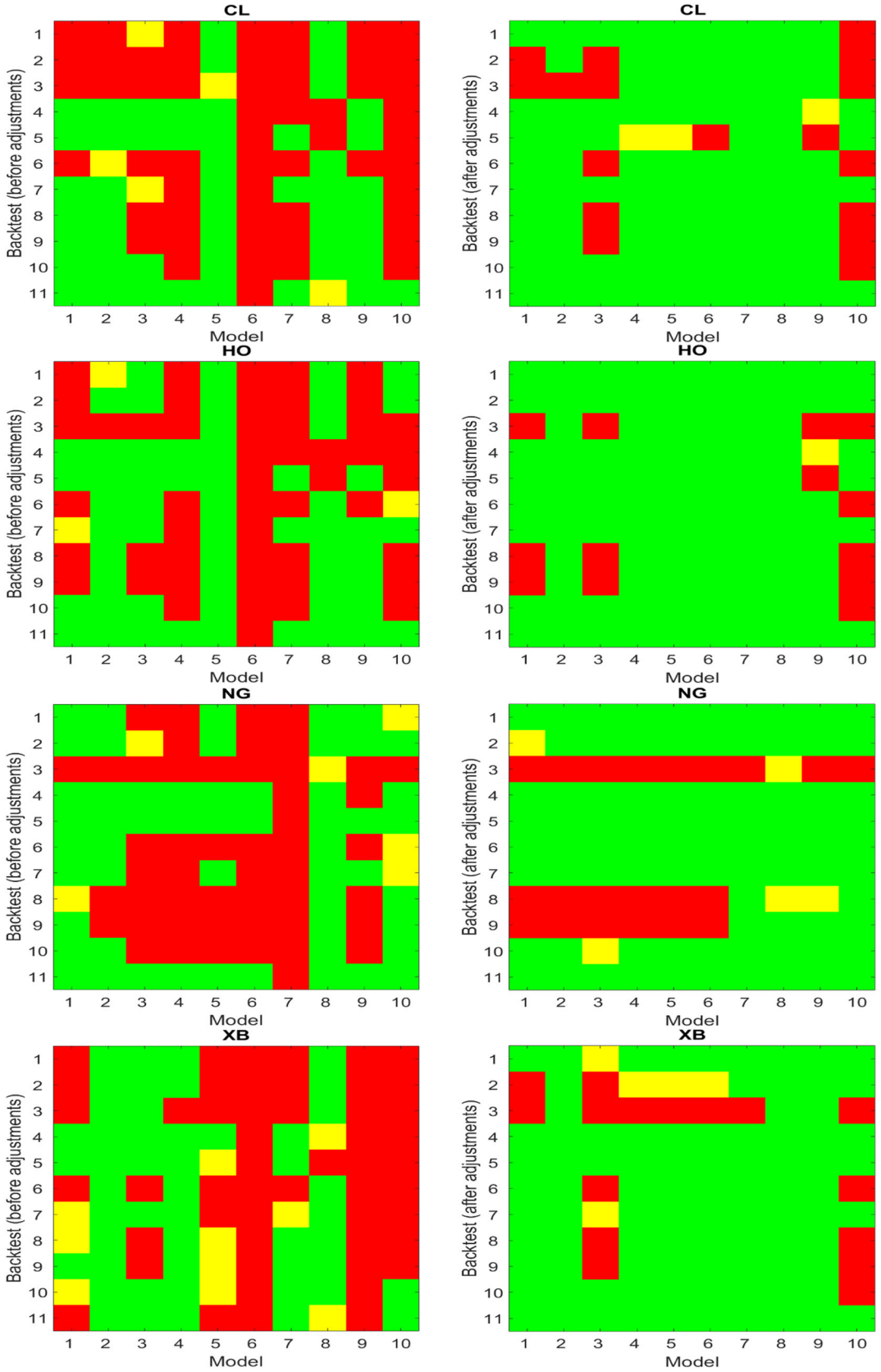

FIGURE 5 1% VaR and ES backtesting results with respect to *p*-values before and after adjustments, over the full period. A *p*-value smaller than 0.05 (between 0.05 and 0.1) is shaded with red (yellow); otherwise, it is green. Models 1–10 are HS, WHS, CF, G-*t*, G-*skt*, EVT-POT, GAS-1F, FHS, CAViaR-SAV, and CARE-SAV, accordingly. Backtests 1–11 are UC, CC, DQ, two-sided ER, one-sided ER, two-sided CCA, one-sided CCA, ESR Strict, ESR AUX, two-sided ESR Int, and one-sided ESR Int. [Color figure can be viewed at wileyonlinelibrary.com]

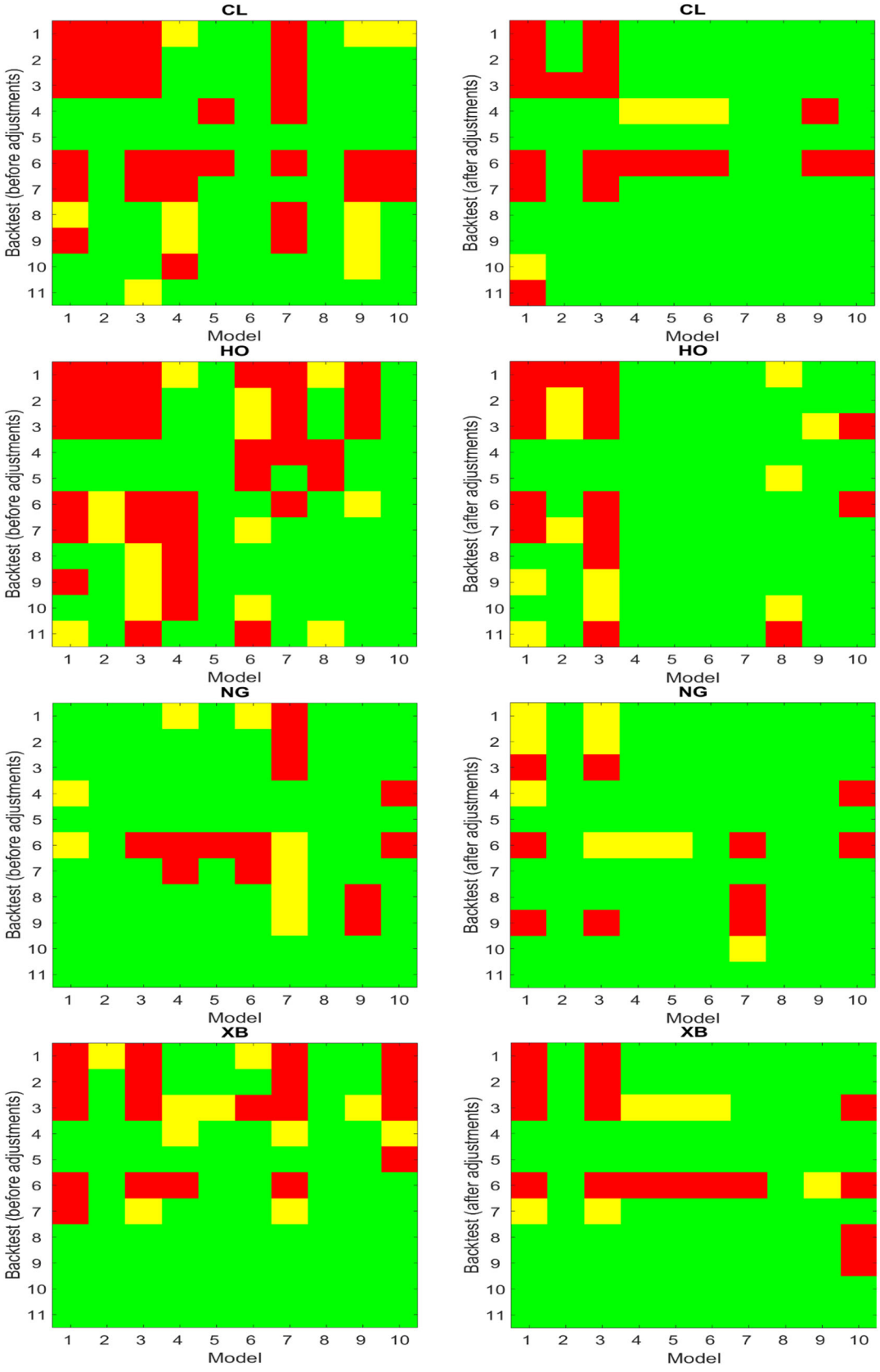

FIGURE 6 1% VaR and ES backtesting results with respect to *p*-values before and after adjustments, over the crisis period. A *p*-value smaller than 0.05 (between 0.05 and 0.1) is shaded with red (yellow); otherwise, it is green. Models 1–10 are HS, WHS, CF, G-*t*, G-*skt*, EVT-POT, GAS-1F, FHS, CAViaR-SAV, and CARE-SAV, accordingly. Backtests 1–11 are UC, CC, DQ, two-sided ER, one-sided ER, two-sided CCA, one-sided CCA, ESR Strict, ESR AUX, two-sided ESR Int, and one-sided ESR Int. [Color figure can be viewed at wileyonlinelibrary.com]

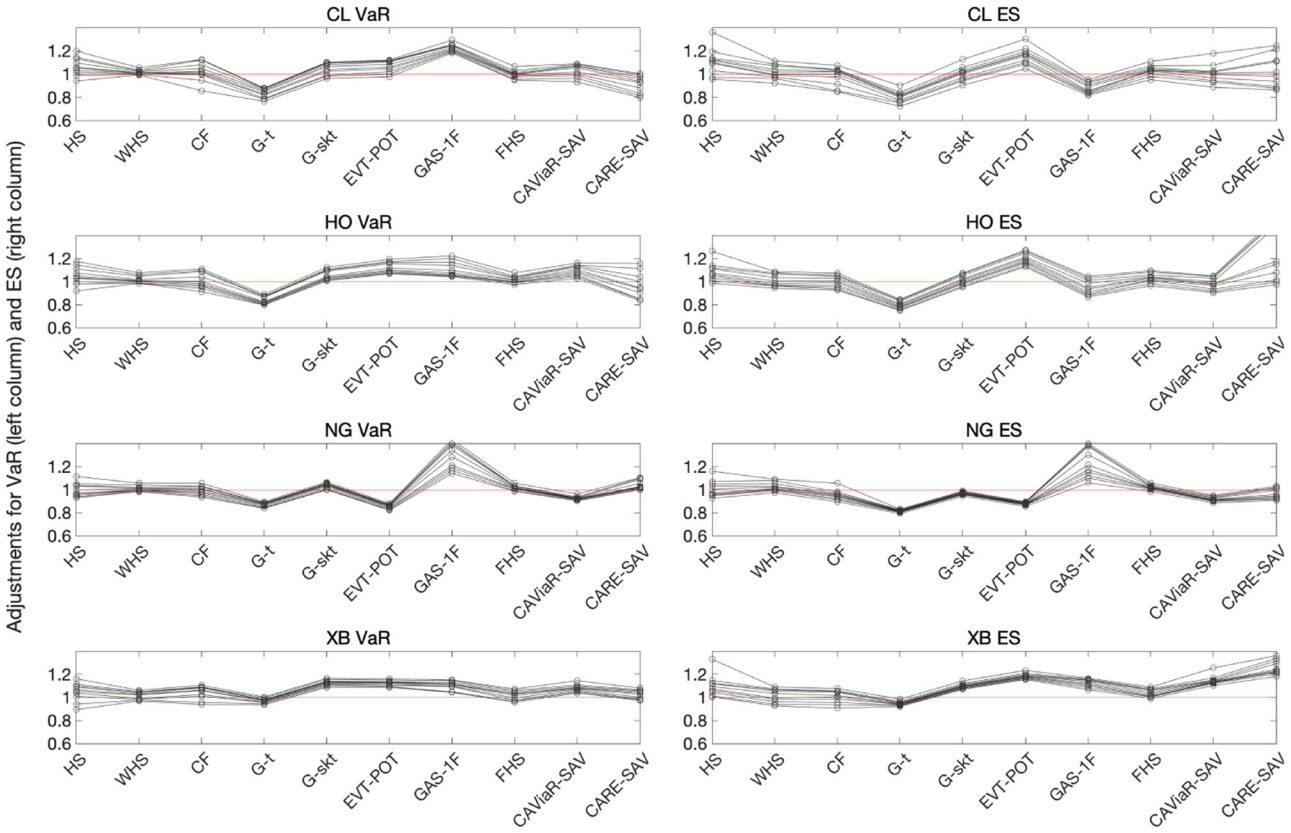

FIGURE 7 Adjustments made to risk forecasts, based on daily data of energy futures contracts. CL, HO, NG, and XB are abbreviations for WTI Crude Oil, Heating Oil, Natural Gas, and RBOB/Unleaded Gasoline, respectively. Each line represents the quantile, from 5% to 95% with an increment of 10%, of adjustments made to 2.5% VaR (in the left column) and ES (in the right column) forecasts across various risk models considered. [Color figure can be viewed at wileyonlinelibrary.com]

model set considered, the G–*t* applied to various energy futures tends to overestimate risks, whereas the HS, WHS, FHS, and GAS-1F methods tend to underestimate risks.

Complementary to traditional backtesting for model performance, we add comparative tests based on the *FZ*0 losses, namely, Diebold-Mariano (DM) (Diebold & Mariano, 2002) and model confidence set (MCS) tests (Hansen et al., 2011), to make model comparisons to check whether our adjustment methodology improves on existing risk forecasts.

In the DM test, a negative *t*-statistic indicates that the row forecast outperforms the column forecast with a significant loss difference. An absolute *t*-statistic greater than 2.575 (1.96 or 1.64) indicates that the average loss difference is significantly different from zero at the 1% (5% or 10%) significance level. Figures 8 and 9 present the results with the null hypothesis that the row forecast and the column forecast have equal values based on the loss function for 1% VaR and ES over the full sample and the crisis period, respectively. Blue blocks mean that the row forecast has a lower average loss than the column forecast at different significance levels (the darkest color means that we reject the null hypothesis at the 1% significance level and so on). White blocks mean that there is no significant difference between the row forecast and the column forecast. Red blocks mean that the row forecast has a higher average loss than the column forecast (the darkest shade means that we reject the null hypothesis at the 1% significance level and so on). To evaluate the efficiency of the adjusted forecasts compared with the original forecasts, we only focus on the blocks in the diagonal of the color maps. Blocks in the diagonal that are shaded blue indicate that our methodology significantly improves the forecasts; otherwise, there is no significant improvement after adjustment, or even worse. In Figures 8 and 9, it is obvious that most of the blocks in the diagonals are shaded blue, and it is rare to see red blocks in the diagonals. Overall, the DM test indicates the significant improvement of our methodology for 1% VaR and ES forecasts.[17]

[17]The backtesting results for VaR and ES at other significance levels are presented in Figures C5, C6, C7, C8 in Appendix C. The results are consistent.

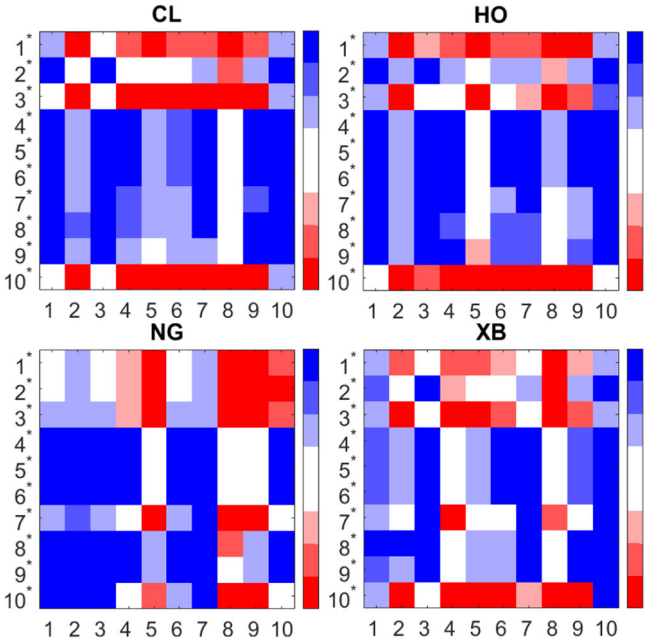

FIGURE 8 Color map based on the Diebold–Mariano (DM) test comparing the average losses using the *FZ*0 loss function for 1% VaR and ES over the full sample. Forecasts marked with 1* to 10* are the adjusted risk measures forecasts based on the original forecasts marked with 1–10. [Color figure can be viewed at wileyonlinelibrary.com]

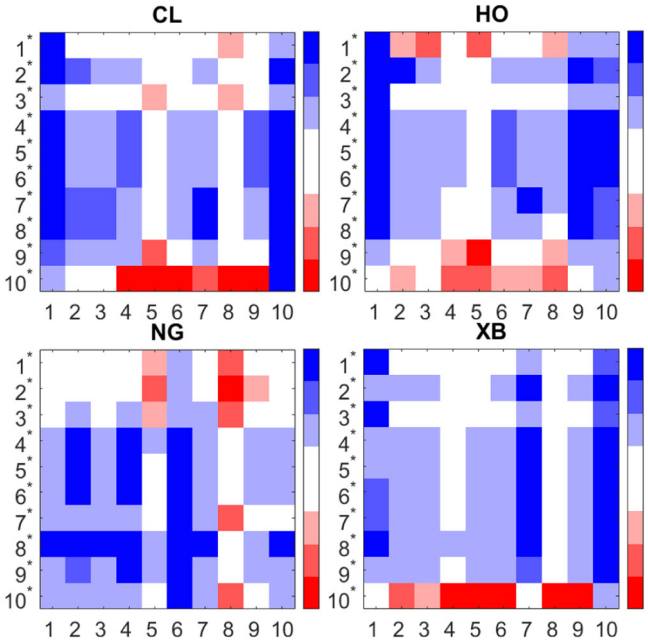

FIGURE 9 Color map based on the Diebold–Mariano (DM) test comparing the average losses using the *FZ*0 loss function for 1% VaR and ES over the crisis sample. Forecasts marked with 1* to 10* are the adjusted risk measures forecasts based on the original forecasts marked with 1–10. [Color figure can be viewed at wileyonlinelibrary.com]

TABLE 7 The Model Confidence Set (MCS) test with the R and SQ methods over crisis periods.

Panel A: The 95% model confidence set (MCS) test																
	Summed absolute values (R method)								**Summed squares (SQ method)**							
	1%		**2.5%**		**5%**		**Total**		**1%**		**2.5%**		**5%**		**Total**	
	Before	**After**	**Before**	**After**	**Before**	**After**	**Before**	**After**	**Before**	**After**	**Before**	**After**	**Before**	**After**	**Before**	**After**
HS1000	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
WHS	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CF	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G-s	0	0	1	1	0	0	1	1	0	0	1	1	0	1	1	2
G-skt	2	0	1	1	0	0	3	1	2	0	1	1	0	1	3	2
EVT-POT	0	0	0	1	0	0	0	1	0	1	0	1	0	1	0	3
GAS-1F	0	1	0	2	0	1	0	4	0	1	0	2	0	2	0	5
FHS	3	0	3	1	0	1	6	2	3	1	3	1	1	1	7	3
CAViaR-SAV	0	0	0	1	0	1	0	2	0	0	0	1	1	2	1	3
CARE-SAV	0	0	0	0	0	1	0	1	0	0	0	0	0	1	0	1
Total	5	1	5	7	0	4	10	12	5	3	5	7	2	9	12	19
Panel B: The 75% model confidence set (MCS) test																
	Summed absolute values (R method)								**Summed squares (SQ method)**							
	1%		**2.5%**		**5%**		**Total**		**1%**		**2.5%**		**5%**		**Total**	
	Before	**After**	**Before**	**After**	**Before**	**After**	**Before**	**After**	**Before**	**After**	**Before**	**After**	**Before**	**After**	**Before**	**After**
HS1000	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
WHS	0	0	0	0	0	0	0	0	0	0	1	1	1	0	2	1
CF	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G-s	0	0	1	1	1	1	2	2	1	0	2	1	1	1	4	2
G-skt	2	0	1	1	1	1	4	2	2	1	1	1	1	1	4	3
EVT-POT	0	1	0	2	0	1	0	4	0	2	0	2	0	2	0	6
GAS-1F	0	1	0	2	0	2	0	5	0	2	1	2	1	2	2	6
FHS	3	1	3	2	3	2	9	5	3	2	3	2	3	3	9	7
CAViaR-SAV	0	0	0	1	2	2	2	3	0	0	1	2	2	3	3	5
CARE-SAV	0	0	0	0	0	1	0	1	0	0	0	0	0	1	0	1
Total	5	3	5	9	7	10	17	22	6	7	9	11	9	13	24	31

Note: The number of commodity futures for which each method is within the model confidence set at the 75% and 95% confidence levels based on the *FZ0* loss function. Results based on pre- and post-adjusted VaR and ES forecasts are labeled as column "before" and "after," respectively.

Alternatively, we exploit the MCS test to compare the forecasts based on the losses generated from the *FZ*0 loss function. In this paper, we adopt two methods: (1) the R method using sums of absolute values for calculating the test statistics for MCS and (2) the SQ method using the summed squares.[18] Table 7 shows the backtesting results via the MCS test, in which the block bootstrap is used with a block length of 12 and 10,000 replications over the financial crisis period at 75% and 95% confidence levels, respectively. In general, the total number of the post-adjustment models included in the best model set is obviously larger than the one of preadjustment models, especially when we apply the SQ method. Specifically, after adjustment, the EVT-POT, GAS-1F, and CAViaR-SAV models are contained in the best model set more often than before the adjustment.

6 | CONCLUSION

To facilitate efficient financial risk management, this paper develops a generic adjustment methodology to improve on VaR and ES forecasts via the minimization of the average loss of the *FZ* loss function. This methodology is advantageous in explicitly indicating the degree of under- and overestimation of tail risks, a topic of concern to regulators and investors and applicable to any risk model that produces VaR and ES forecasts. Moreover, this adjustment methodology is built on the objective of minimizing a loss function, and as a result, we expect the risk disagreement among post-adjusted risk forecasts to be lessened.

In the empirical analysis, we apply the proposed methodology to a battery of risk forecasting models for several futures contracts in the energy commodity market in which tail risk management plays a significant role. After making adjustments to VaR and ES forecasts built using the risk models considered for four energy futures, the risk ratios decline over the full out-of-sample period and during seven energy crisis periods, indicating the abatement of risk disagreement within the model set. In addition, with the margin level change date signaling high volatility faced by market participants, around these dates, the shrinking range in values of post-adjusted VaR and ES forecasts given by various models shows the reduction in model disagreement. Taken together, this adjustment methodology helps alleviate the model disagreement among diverse risk models. Further, the forecasting accuracy of VaR and ES estimates is generally improved, as verified via various backtesting tests for VaR and ES. Notably, the UC test with respect to the frequency of VaR exceedances and the ER test focused on the magnitude of ES forecasts benefit most from adjustments made to VaR and ES forecasts.

ACKNOWLEDGMENTS

The authors would like to thank Andrew Patton (discussant) as well as participants at 11th International Conference on Futures and Other Derivatives and 2022 Australasian Finance and Banking Conference for helpful comments and suggestions. Any errors are our own. Yujing Gong gratefully acknowledges the support of the Economic and Social Research Council (ESRC) in funding the Systemic Risk Centre (grant numbers ES/K002309/1 and ES/R009724/1).

DATA AVAILABILITY STATEMENT

The energy futures pricing data that supports the findings of this study are available from Bloomberg. Restrictions apply to the availability of these data, which were used under license for this study. Data are available from the authors with the permission of Bloomberg, upon reasonable request.

ORCID

Ning Zhang http://orcid.org/0000-0002-1896-6888

REFERENCES

Bakshi, G., Gao, X., & Rossi, A. G. (2019). Understanding the sources of risk underlying the cross section of commodity returns. *Management Science*, *65*, 619–641. https://doi.org/10.1287/mnsc.2017.2840

Barone-Adesi, G., Giannopoulos, K., & Vosper, L. (1999). VaR without correlations for portfolios of derivative securities. *Journal of Futures Markets*, *19*, 583–602. https://doi.org/10.1002/(SICI)1096-9934(199908)19:5<583::AID-FUT5>3.0.CO;2-S

[18]Details can be found on page 465 of Hansen et al. (2011), and the Matlab code for MCS testing can be downloaded from www.kevinsheppard.com/MFEToolbox.

Barrieu, P., & Scandolo, G. (2015). Assessing financial model risk. *European Journal of Operational Research*, *242*, 546–556. https://doi.org/10.1016/j.ejor.2014.10.032

Basel Committee on Banking Supervision. (2019). *Minimum capital requirements for market risk*. https://www.bis.org/bcbs/publ/d457.pdf

Basel Committee on Banking Supervision. (2021). *Revisions to market risk disclosure requirements*. https://www.bis.org/bcbs/publ/d529.pdf

Bayer, S., & Dimitriadis, T. (2022). Regression-based expected shortfall backtesting. *Journal of Financial Econometrics*, *20*, 437–471. https://doi.org/10.1093/jjfinec/nbaa013

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, *31*, 307–327. https://doi.org/10.1016/0304-4076(86)90063-1

Boons, M., & Prado, M. P. (2019). Basis-momentum. *Journal of Finance*, *74*, 239–279. https://doi.org/10.1111/jofi.12738

Boudoukh, J., Richardson, M., & Whitelaw, R. (1998). The best of both worlds: a hybrid approach to calculating value at risk. *Risk*, *11*, 64–67.

Boudt, K., Peterson, B. G., & Croux, C. (2008). Estimation and decomposition of downside risk for portfolios with non-normal returns. *Journal of Risk*, *11*, 79–103. https://doi.org/10.2139/ssrn.1024151

Čech, F., & Baruník, J. (2019). Panel quantile regressions for estimating and predicting the value-at-risk of commodities. *Journal of Futures Markets*, *39*, 1167–1189. https://doi.org/10.1002/fut.22017

Chen, S. X. (2007). Nonparametric estimation of expected shortfall. *Journal of Financial Econometrics*, *6*, 87–107. https://doi.org/10.1093/jjfinec/nbm019

Christoffersen, P. F. (1998). Evaluating interval forecasts. *International Economic Review*, *39*, 841–862. https://doi.org/10.2307/2527341

Christoffersen, P. F. (2012). *Elements of financial risk management*. Academic Press.

Danielsson, J., James, K. R., Valenzuela, M., & Zer, I. (2016). Model risk of risk models. *Journal of Financial Stability*, *23*, 79–91. https://doi.org/10.1016/j.jfs.2016.02.002

Diebold, F. X., & Mariano, R. S. (2002). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, *20*, 134–144. https://doi.org/10.1198/073500102753410444

Ding, S., Cui, T., Zhang, Y., & Li, J. (2021). Liquidity effects on oil volatility forecasting: From fintech perspective. *Plos one*, *16*, e0260289. https://doi.org/10.1371/journal.pone.0260289

Ding, S., Zhang, Y., & Duygun, M. (2019). Modeling price volatility based on a genetic programming approach. *British Journal of Management*, *30*, 328–340. https://doi.org/10.1111/1467-8551.12359

Efron, B. (1991). Regression percentiles using asymmetric squared error loss. *Statistica Sinica*, *1*, 93–125. https://www.jstor.org/stable/24303995

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC Press.

Emmer, S., Kratz, M., & Tasche, D. (2015). What is the best risk measure in practice? A comparison of standard measures. *Journal of Risk*, *18*, 31–60. https://doi.org/10.48550/arXiv.1312.1645

Energy Futures Pricing Data (2021). Not publicly available; data set can be acquired by paying a data license fee at Bloomberg.

Engle, R. F., & Manganelli, S. (2004). CAViaR: Conditional autoregressive value at risk by regression quantiles. *Journal of Business & Economic Statistics*, *22*, 367–381. https://doi.org/10.1198/073500104000000370

Escanciano, J. C., & Olmo, J. (2010). Backtesting parametric value-at-risk with estimation risk. *Journal of Business & Economic Statistics*, *28*, 36–51. https://doi.org/10.1198/jbes.2009.07063

Farkas, W., Fringuellotti, F., & Tunaru, R. (2020). A cost-benefit analysis of capital requirements adjusted for model risk. *Journal of Corporate Finance*, *65*, 101753. https://doi.org/10.1016/j.jcorpfin.2020.101753

Fissler, T., & Ziegel, J. F. (2016). Higher order elicitability and osband's principle. *The Annals of Statistics*, *44*, 1680–1707. https://doi.org/10.1214/16-AOS1439

Gorton, G. B., Hayashi, F., & Rouwenhorst, K. G. (2013). The fundamentals of commodity futures returns. *Review of Finance*, *17*, 35–105. https://doi.org/10.1093/rof/rfs019

Guo, Z.-Y. (2020). Stochastic multifactor models in risk management of energy futures. *Journal of Futures Markets*, *40*, 1918–1934. https://doi.org/10.1002/fut.22154

Hansen, P. R., Lunde, A., & Nason, J. M. (2011). The model confidence set. *Econometrica*, *79*, 453–497. https://doi.org/10.3982/ECTA5771

Hedegaard, E. (2014). Causes and consequences of margin levels in futures markets. Arizona State University.

Jang, H. J., Lee, K., & Lee, K. (2020). Systemic risk in market microstructure of crude oil and gasoline futures prices: A Hawkes flocking model approach. *Journal of Futures Markets*, *40*, 247–275. https://doi.org/10.1002/fut.22048

Kang, W., Rouwenhorst, K. G., & Tang, K. (2020). A tale of two premiums: The role of hedgers and speculators in commodity futures markets. *Journal of Finance*, *75*, 377–417. https://doi.org/10.1111/jofi.12845

Koenker, R., & Machado, J. A. (1999). Goodness of fit and related inference processes for quantile regression. *Journal of the American Statistical Association*, *94*, 1296–1310. https://doi.org/10.1080/01621459.1999.10473882

Koijen, R. S., Moskowitz, T. J., Pedersen, L. H., & Vrugt, E. B. (2018). Carry. *Journal of Financial Economics*, *127*, 197–225. https://doi.org/10.1016/j.jfineco.2017.11.002

Kupiec, P. H. (1995). Techniques for verifying the accuracy of risk measurement models. *Journal of Derivatives*, *3*, 73–84. https://doi.org/10.3905/jod.1995.407942

Laporta, A. G., Merlo, L., & Petrella, L. (2018). Selection of value at risk models for energy commodities. *Energy Economics*, *74*, 628–643. https://doi.org/10.1016/j.eneco.2018.07.009

Lazar, E., & Zhang, N. (2019). Model risk of expected shortfall. *Journal of Banking & Finance*, *105*, 74–93. https://doi.org/10.1016/j.jbankfin.2019.05.017

Liu, F., & Stentoft, L. (2021). Regulatory capital and incentives for risk model choice under Basel 3. *Journal of Financial Econometrics*, *19*, 53–96. https://doi.org/10.1093/jjfinec/nbaa029

McNeil, A. J., & Frey, R. (2000). Estimation of tail-related risk measures for heteroscedastic financial time series: An extreme value approach. *Journal of Empirical Finance*, *7*, 271–300. https://doi.org/10.1016/S0927-5398(00)00012-8

Merlo, L., Petrella, L., & Raponi, V. (2021). Forecasting VaR and ES using a joint quantile regression and its implications in portfolio allocation. *Journal of Banking & Finance*, *133*, 106248. https://doi.org/10.1016/j.jbankfin.2021.106248

Miller, D. J., & Liu, W.-H. (2006). Improved estimation of portfolio value-at-risk under copula models with mixed marginals. *Journal of Futures Markets: Futures, Options, and Other Derivative Products*, *26*, 997–1018. https://doi.org/10.1002/fut.20224

Newey, W. K., & Powell, J. L. (1987). Asymmetric least squares estimation and testing. *Econometrica*, *55*, 819–847. https://doi.org/10.2307/1911031

Nieto, M. R., & Ruiz, E. (2016). Frontiers in var forecasting and backtesting. *International Journal of Forecasting*, *32*, 475–501. https://doi.org/10.1016/j.ijforecast.2015.08.003

Nolde, N., & Ziegel, J. F. (2017). Elicitability and backtesting: Perspectives for banking regulation. *Annals of Applied Statistics*, *11*, 1833–1874. https://doi.org/10.1214/17-AOAS1041

Park, Y.-H., & Abruzzo, N. (2016). An empirical analysis of futures margin changes: Determinants and policy implications. *Journal of Financial Services Research*, *49*, 65–100. https://doi.org/10.1007/s10693-014-0212-8

Patra, S. (2021). Revisiting value-at-risk and expected shortfall in oil markets under structural breaks: The role of fat-tailed distributions. *Energy Economics*, *101*, 105452. https://doi.org/10.1016/j.eneco.2021.105452

Patton, A. J., Ziegel, J. F., & Chen, R. (2019). Dynamic semiparametric models for expected shortfall (and value-at-risk). *Journal of Econometrics*, *211*, 388–413. https://doi.org/10.1016/j.jeconom.2018.10.008

Pitera, M., & Schmidt, T. (2018). Unbiased estimation of risk. *Journal of Banking & Finance*, *91*, 133–145. https://doi.org/10.1016/j.jbankfin.2018.04.016

Qiao, T., & Han, L. (2023). Covid-19 and tail risk contagion across commodity futures markets. *Journal of Futures Markets*, 242–272. https://doi.org/10.1002/fut.22388

Samuel, Y. M. Z.-t. (2008). Value at risk and conditional extreme value theory via Markov regime switching models. *Journal of Futures Markets: Futures, Options, and Other Derivative Products*, *28*, 155–181. https://doi.org/10.1002/fut.20293

Taylor, J. W. (2008). Estimating value at risk and expected shortfall using expectiles. *Journal of Financial Econometrics*, *6*, 231–252. https://doi.org/10.1093/jjfinec/nbn001

Taylor, J. W. (2019). Forecasting value at risk and expected shortfall using a semiparametric approach based on the asymmetric laplace distribution. *Journal of Business & Economic Statistics*, *37*, 121–133. https://doi.org/10.1080/07350015.2017.1281815

Taylor, J. W. (2022). Forecasting value at risk and expected shortfall using a model with a dynamic omega ratio. *Journal of Banking & Finance*, *140*, 106519. https://doi.org/10.1016/j.jbankfin.2022.106519

Xu, Y., & Lien, D. (2020). Optimal futures hedging for energy commodities: An application of the gas model. *Journal of Futures Markets*, *40*, 1090–1108. https://doi.org/10.1002/fut.22118

How to cite this article: Zhang, N., Gong, Y., & Xue, X. (2023). Less disagreement, better forecasts: Adjusted risk measures in the energy futures market. *The Journal of Futures Markets*, 1–41. https://doi.org/10.1002/fut.22412

APPENDIX A: BACKTESTING METHODS FOR VaR

In the following, we provide an overview of the backtesting methods employed in this paper.

Unconditional coverage (UC) test for VaR

The most popular procedures evaluating the performance of VaR forecasts are mainly based on VaR failures; that is,

$$I_t = \mathbf{1}\left\{r_t \leq VaR_t^{\alpha}\right\}.$$

The commonly used VaR backtesting method, known as the unconditional coverage (UC) test, is proposed by Kupiec (1995) and uses the proportion of failures as its main tool. In this test, the hit percentage is defined as the proportion of the returns below the estimated VaR. Then the difference between the hit percentage and its

theoretical value of α is examined. This performance test assesses the accuracy of VaR forecasts based on the following hypothesis:

$$H_0 : \mathbb{E}_{t-1}[I_t] = \alpha \quad \text{against} \quad H_1 : \mathbb{E}_{t-1}[I_t] \neq \alpha. \tag{A1}$$

A rejection of the null hypothesis implies that the risk model is not correctly specified.

Conditional coverage (CC) test for VaR

However, the UC test is statistically weak for a small sample size and is criticized by several studies (see Nieto and Ruiz, 2016) because it ignores the clustering of failures. To address these drawbacks, the conditional coverage (CC) test with respect to available information $\mathcal{F}_{t-1}$ is considered, and its hypothesis is given as

$$H_0 : \mathbb{E}_{t-1}[I_t|\mathcal{F}_{t-1}] = \alpha \quad \text{against} \quad H_1 : \mathbb{E}_{t-1}[I_t|\mathcal{F}_{t-1}] \neq \alpha.$$

The risk model fails the test if the null hypothesis is rejected, suggesting that the model is not correctly specified.

We also employ the dynamic quantile (DQ) test proposed by Engle and Manganelli (2004) to implement the CC test. The DQ test has power against the misspecification of ignoring conditionally correlated probabilities and can be extended to examine other explanatory variables. The DQ test examines whether the hit variable defined as $Hit_{v,t} = \mathbb{1}\{r_t \leq VaR_t\} - \alpha$, follows an i.i.d. Bernoulli distribution with probability level α and whether it is independent of the VaR estimator; the expected value of $Hit_{v,t}$ is 0. Furthermore, from the definition of the quantile function, the conditional expectation of VaR_t given any information known at $t-1$ must also be 0, which means that the hit function cannot be correlated with other lagged variables. Also, the $Hit_{v,t}$ must not be autocorrelated. If $Hit_{v,t}$ satisfies the conditions stated above, then there will be no autocorrelation in the hits and no measurement error. We include one lag of $Hit_{v,t}$ in the regression of the test. Consider the following DQ regression:

$$Hit_{v,t} = a_0 + a_1 Hit_{v,t-1} + a_2 VaR_{t-1} + u_{v,t}, \tag{A2}$$

where $\mathbf{a} = [a_0, a_1, a_2]$ is the set of parameters of the regression. Based on the null hypothesis, we test whether all parameters in the set $\mathbf{a}$ are zero. Performing this DQ test gives a test statistic, which is distributed $\mathcal{X}^2(3)$ asymptotically.

APPENDIX B: BACKTESTING METHODS FOR ES

Exceedance residual (ER) test for ES

The ER approach focuses on the magnitude of tail losses. Based on the ES-specified residuals, whenever there is a VaR exceedance, $ER_t = (r_t - ES_t)\mathbb{1}\{r_t \leq VaR_t\}$, this tests whether the expected value of the exceedance residual ER_t conditioning on the information $\mathcal{F}_{t-1}$, $\mu = \mathbb{E}[ER_t|\mathcal{F}_{t-1}]$, is zero using a bootstrap method (Efron & Tibshirani, 1994) without any assumption on the distribution of residuals. The two-sided and one-sided hypotheses are discussed in this paper and given as

$$\begin{aligned} &H_0^{2s} : \mu = 0 \quad \text{against} \quad H_1^{2s} : \mu \neq 0, \quad \text{and} \\ &H_0^{1s} : \mu \geq 0 \quad \text{against} \quad H_1^{1s} : \mu < 0. \end{aligned} \tag{B1}$$

The rejection of null hypotheses indicates the model misspecification. If the one-sided hypothesis is rejected, it suggests that ES is underestimated.

Conditional calibration test for ES

The conditional calibration test, introduced by Nolde and Ziegel (2017), uses identification functions V to test the model performance, in which V at time t is written as

$$V(r_t, VaR_t, ES_t; \alpha) = \begin{bmatrix} \alpha - \mathbb{1}\{r_t \leq VaR_t\} \\ VaR_t - ES_t - \mathbb{1}\{r_t \leq VaR_t\}(VaR_t - r_t)/\alpha \end{bmatrix} \tag{B2}$$

Based on the identification functions, this tests whether the pair of (VaR_t, ES_t) forecasts are the best possible predictions conditioning on the information set $\mathcal{F}$, with the two-sided and one-sided hypotheses shown as follows. For all t,

$$\begin{aligned} H_0^{2s} &: \mathbb{E}\left[V\left(r_t, VaR_t, ES_t; \alpha\right)|\mathcal{F}_{t-1}\right] = 0 \quad \text{against} \quad H_1^{2s} : \mathbb{E}\left[V\left(r_t, VaR_t, ES_t; \alpha\right)|\mathcal{F}_{t-1}\right] \neq 0, \\ & \qquad\qquad\qquad\qquad \textit{and} \\ H_0^{1s} &: \mathbb{E}\left[V\left(r_t, VaR_t, ES_t; \alpha\right)|\mathcal{F}_{t-1}\right] \geq 0 \quad \text{against} \quad H_1^{1s} : \mathbb{E}\left[V\left(r_t, VaR_t, ES_t; \alpha\right)|\mathcal{F}_{t-1}\right] < 0. \end{aligned} \tag{B3}$$

The economic rationale behind the one-sided null hypothesis is that the overestimation of risk may not be problematic in practice, whereas the underestimation of risk matters. The rejection of null hypotheses indicates the model misspecification. The rejection of the one-sided hypothesis suggests that ES is underestimated.

Regression-based test for ES

Bayer and Dimitriadis (2022) estimate the parameter vectors ($\boldsymbol{\beta} = [\beta_1, \beta_2]^\top, \boldsymbol{\gamma} = [\gamma_1, \gamma_2]^\top$) in the regression equations (see below) to test for ES drawing on a joint loss function for VaR and ES from Fissler and Ziegel (2016). The forecasts VaR_t and ES_t are generated using the information set $\mathcal{F}_{t-1}$:

$$\begin{aligned} Y_t &= V_t\boldsymbol{\beta} + u_t^{VaR}, \quad \text{and} \\ Y_t &= W_t\boldsymbol{\gamma} + u_t^{ES}, \end{aligned} \tag{B4}$$

where V_t and W_t are two-dimensional, $\mathcal{F}_{t-1}$ measurable covariate vectors; and $VaR_\alpha\left(u_t^{VaR}|\mathcal{F}_{t-1}\right) = 0$ and $ES_\alpha\left(u_t^{ES}|\mathcal{F}_{t-1}\right) = 0$ almost surely. The three specifications of ESR tests are detailed as follows: (1) The strict ESR (ESR Strict) backtest using ES only as the explanatory variable: this specifies that $V_t = W_t = (1, ES_t)$ in (B4) and tests the hypothesis that

$$H_0 : (\gamma_1, \gamma_2) = (0, 1) \quad \text{against} \quad H_1 : (\gamma_1, \gamma_2) \neq (0, 1). \tag{B5}$$

The rejection of the null hypothesis suggests that the model is not correctly specified. (2) The auxiliary ESR (ESR AUX) backtest using VaR and ES as explanatory variables: this specifies that $V_t = (1, VaR_t)$ and $W_t = (1, ES_t)$ in (B4) and tests the hypothesis that

$$H_0 : (\gamma_1, \gamma_2) = (0, 1) \quad \text{against} \quad H_1 : (\gamma_1, \gamma_2) \neq (0, 1). \tag{B6}$$

The rejection of the null hypothesis suggests that the model is not correctly specified. (3) the intercept ESR (ESR INT) backtest allowing for the one-sided test that can give a clear indication of overestimated or underestimated risk: its specification is similar to the strict ESR test specification in (B5) but fixes the slope parameters (β_2 and γ_2) to one and only estimates the intercept terms. This tests the following two-sided and one-sided hypotheses:

$$\begin{aligned} H_0^{2s} &: \gamma_1 = 0 \quad \text{against} \quad H_1^{2s} : \gamma_1 \neq 0, \quad \text{and} \\ H_0^{1s} &: \gamma_1 \geq 0 \quad \text{against} \quad H_1^{2s} : \gamma_1 < 0. \end{aligned} \tag{B7}$$

The rejection of null hypotheses indicates the model misspecification. If the one-sided hypothesis is rejected, it suggests that ES is underestimated.

APPENDIX C: ADDITIONAL RESULTS

See Tables C1, C2 and Figures C1–C8.

TABLE C1 Risk ratio sensitivity over full sample—VaR.

	Panel A: CL						Panel B: HO					
	1%		2.5%		5%		1%		2.5%		5%	
	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After
None	3.06	2.54	2.37	2.18	1.77	1.71	2.28	2.12	2.12	2.00	1.92	1.83
HS1000	2.96	2.44	2.29	2.11	1.75	1.69	2.22	2.06	2.08	1.96	1.89	1.80
WHS	3.04	2.50	2.35	2.14	1.74	1.67	2.25	2.08	2.09	1.97	1.89	1.79
CF	2.52	2.14	2.10	1.92	1.74	1.68	2.18	2.07	2.08	1.96	1.89	1.80
G-s	3.05	2.54	2.36	2.18	1.75	1.71	2.27	2.12	2.11	2.00	1.90	1.83
G-skt	3.06	2.54	2.37	2.18	1.77	1.71	2.28	2.12	2.12	2.00	1.92	1.83
EVT-POT	3.06	2.54	2.37	2.18	1.77	1.71	2.26	2.12	2.10	2.00	1.91	1.83
GAS-1F	2.91	2.48	2.19	2.14	1.73	1.68	2.18	2.06	2.07	1.95	1.86	1.80
FHS	3.06	2.53	2.37	2.18	1.77	1.71	2.27	2.09	2.11	1.99	1.91	1.82
CAViaR-SAV	3.00	2.53	2.36	2.17	1.77	1.71	2.26	2.11	2.11	1.99	1.92	1.83
CARE-SAV	2.53	2.15	2.12	1.90	1.73	1.62	1.95	1.71	1.74	1.59	1.68	1.55
	Panel C: NG						Panel D: XB					
	1%		2.5%		5%		1%		2.5%		5%	
	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After
None	1.79	1.56	1.72	1.49	1.58	1.47	2.36	2.29	2.07	2.02	1.87	1.85
HS1000	1.78	1.54	1.71	1.47	1.57	1.46	2.31	2.23	2.01	1.96	1.84	1.82
WHS	1.77	1.51	1.70	1.44	1.57	1.46	2.34	2.24	2.03	1.97	1.82	1.79
CF	1.76	1.54	1.71	1.47	1.57	1.46	2.12	2.12	1.96	1.90	1.85	1.82
G-s	1.79	1.56	1.72	1.49	1.54	1.47	2.36	2.29	2.07	2.02	1.86	1.85
G-skt	1.78	1.56	1.69	1.49	1.55	1.47	2.36	2.29	2.07	2.02	1.86	1.85
EVT-POT	1.70	1.56	1.71	1.49	1.58	1.47	2.36	2.29	2.07	2.02	1.87	1.85
GAS-1F	1.72	1.53	1.57	1.47	1.56	1.45	2.14	2.05	1.99	1.94	1.81	1.78
FHS	1.77	1.55	1.72	1.48	1.58	1.47	2.36	2.28	2.07	2.01	1.87	1.84
CAViaR-SAV	1.78	1.54	1.72	1.46	1.58	1.46	2.35	2.28	2.06	2.01	1.87	1.84
CARE-SAV	1.79	1.55	1.72	1.49	1.56	1.47	2.18	2.07	1.81	1.74	1.67	1.63

Note: The mean of the ratio of the highest to the lowest daily pre- and post-adjusted VaR forecasts (risk ratio) across four commodity futures and three significance levels. Each panel presents average risk ratio of 10 candidate models before and after optimized adjustments are made to VaR estimated at 1%, 2.5% and 5%, for a given asset, including WTI Crude Oil (CL, Panel A), Heating Oil (HO, Panel B), Natural Gas (NG, Panel C), and Unleaded/RBOB gasoline (XB, Panel D). Results based on pre- and post-adjusted VaR forecasts are labeled as column "before" and "after," respectively. The sample period is from March 27, 2002 to December 31, 2021. Risk ratios presented in the "None" row are calculated from daily VaR forecasts by using all 10 candidate models. The rest of rows display risk ratios when one specific model is excluded to avoid the effect of outlying forecasts. The excluded model is indicated by the row name.

TABLE C2 Risk ratio sensitivity over full sample—ES.

	Panel A: CL						Panel B: HO					
	1%		2.5%		5%		1%		2.5%		5%	
	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After
None	3.17	1.80	2.66	1.71	2.23	1.58	2.30	2.59	2.17	2.29	1.95	2.08
HS1000	3.10	1.80	2.59	1.70	2.18	1.57	2.26	2.56	2.12	2.25	1.92	2.05
WHS	3.14	1.80	2.65	1.70	2.22	1.57	2.28	2.59	2.16	2.28	1.93	2.07
CF	2.32	1.72	2.04	1.66	1.83	1.56	2.16	2.14	2.06	1.98	1.89	1.89
G-s	3.16	1.80	2.65	1.66	2.20	1.53	2.29	2.59	2.16	2.28	1.93	2.07
G-skt	3.17	1.79	2.66	1.69	2.23	1.57	2.30	2.59	2.17	2.29	1.95	2.08
EVT-POT	2.94	1.79	2.53	1.71	2.13	1.58	2.14	2.55	2.07	2.26	1.89	2.06
GAS-1F	3.14	1.69	2.64	1.59	2.22	1.55	2.24	2.39	2.15	2.21	1.94	2.00
FHS	3.16	1.77	2.66	1.70	2.23	1.57	2.29	2.59	2.16	2.28	1.94	2.08
CAViaR-SAV	3.17	1.80	2.65	1.71	2.23	1.58	2.30	2.58	2.17	2.27	1.95	2.08
CARE-SAV	2.86	1.80	2.48	1.71	2.21	1.57	2.09	2.40	1.87	2.06	1.77	1.90
	Panel C: NG						Panel D: XB					
	1%		2.5%		5%		1%		2.5%		5%	
	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After
None	1.80	1.60	1.71	1.51	1.58	1.48	2.59	2.44	2.29	2.18	2.08	2.00
HS1000	1.80	1.58	1.70	1.50	1.57	1.46	2.56	2.42	2.25	2.15	2.05	1.98
WHS	1.80	1.58	1.70	1.49	1.57	1.47	2.59	2.41	2.28	2.16	2.07	1.98
CF	1.72	1.56	1.66	1.48	1.56	1.45	2.14	2.18	1.98	1.97	1.89	1.86
G-s	1.80	1.60	1.66	1.51	1.53	1.48	2.59	2.44	2.28	2.18	2.07	2.00
G-skt	1.79	1.60	1.69	1.51	1.57	1.48	2.59	2.44	2.29	2.18	2.08	2.00
EVT-POT	1.79	1.60	1.71	1.51	1.58	1.48	2.55	2.44	2.26	2.18	2.06	2.00
GAS-1F	1.69	1.56	1.59	1.49	1.55	1.45	2.39	2.23	2.21	2.08	2.00	1.91
FHS	1.77	1.58	1.70	1.50	1.57	1.47	2.59	2.42	2.28	2.17	2.08	1.99
CAViaR-SAV	1.80	1.58	1.71	1.49	1.58	1.46	2.58	2.43	2.27	2.17	2.08	1.99
CARE-SAV	1.80	1.59	1.71	1.51	1.57	1.47	2.40	2.18	2.06	1.86	1.90	1.77

Note: The mean of the ratio of the highest to the lowest daily pre- and post-adjusted ES forecasts (risk ratio) across four commodity futures and three significance levels. Each panel presents average risk ratios of 10 candidate models before and after optimized adjustments are made to ES estimated at 1%, 2.5%, and 5%, for a given asset, including WTI Crude Oil (CL, Panel A), Heating Oil (HO, Panel B), Natural Gas (NG, Panel C), and Unleaded/RBOB gasoline (XB, Panel D). Results based on pre- and post-adjusted ES forecasts are labeled as column "before" and "after," respectively. The sample period is from March 27, 2002 to December 31, 2021. Risk ratios presented in the "None" row are calculated from daily VaR forecasts by using all 10 candidate models. The rest of rows display risk ratios when one specific model is excluded to avoid the effect of outlying forecasts. The excluded model is indicated by the row name.

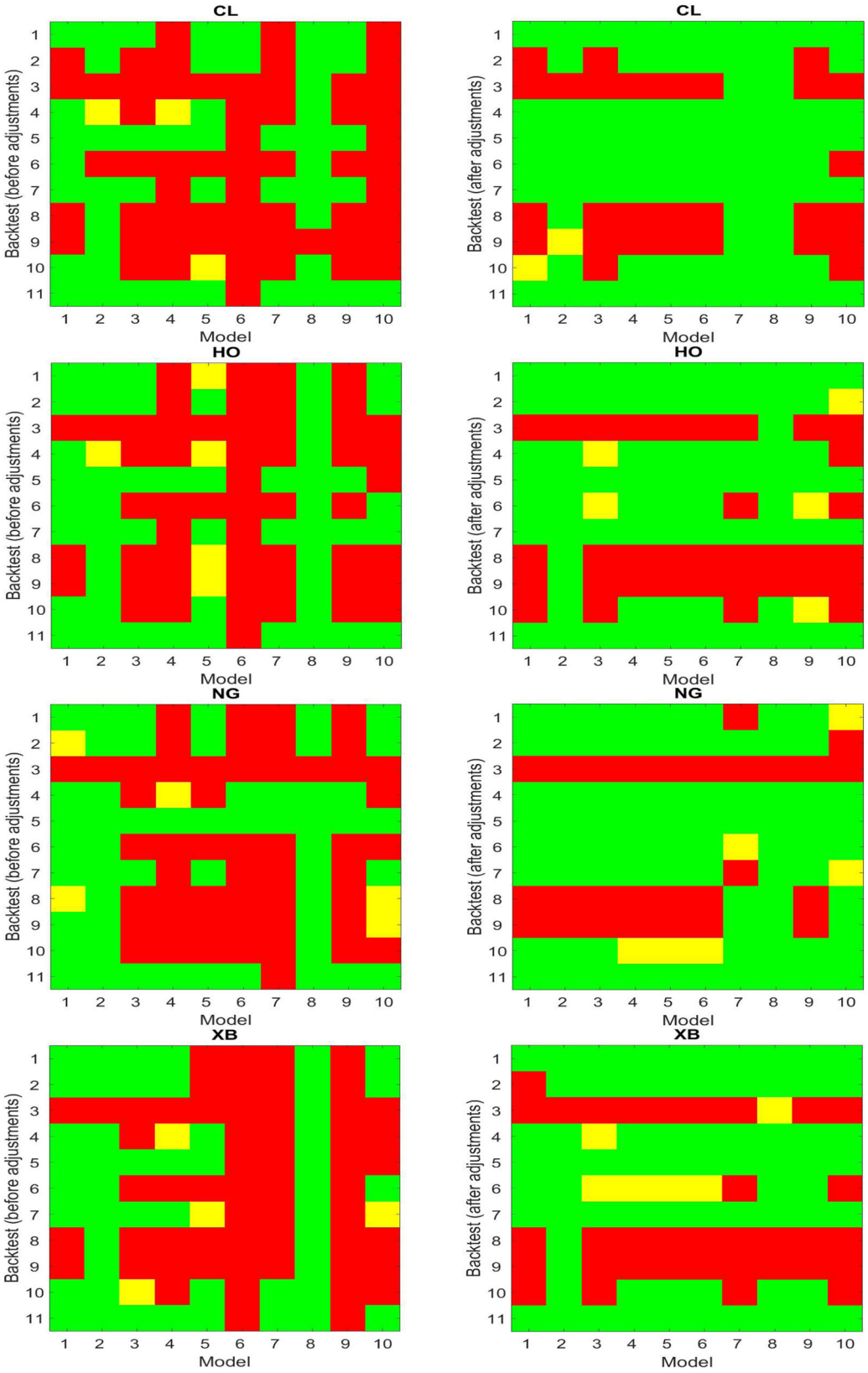

FIGURE C1 2.5% VaR and ES backtesting results with respect to *p*-values before and after adjustments, over the full period. A *p*-value smaller than 0.05 (between 0.05 and 0.1) is shaded with red (yellow); otherwise, it is green. CL, HO, NG, and XB are abbreviations for WTI Crude Oil, Heating Oil, Natural Gas, and RBOB/Unleaded Gasoline, respectively. Models 1–10 are HS, WHS, CF, G-*t*, G-*skt*, EVT-POT, GAS-1F, FHS, CAViaR-SAV, and CARE-SAV, accordingly. Backtests 1–11 are UC, CC, DQ, two-sided ER, one-sided ER, two-sided CCA, one-sided CCA, ESR Strict, ESR AUX, two-sided ESR Int, and one-sided ESR Int. [Color figure can be viewed at wileyonlinelibrary.com]

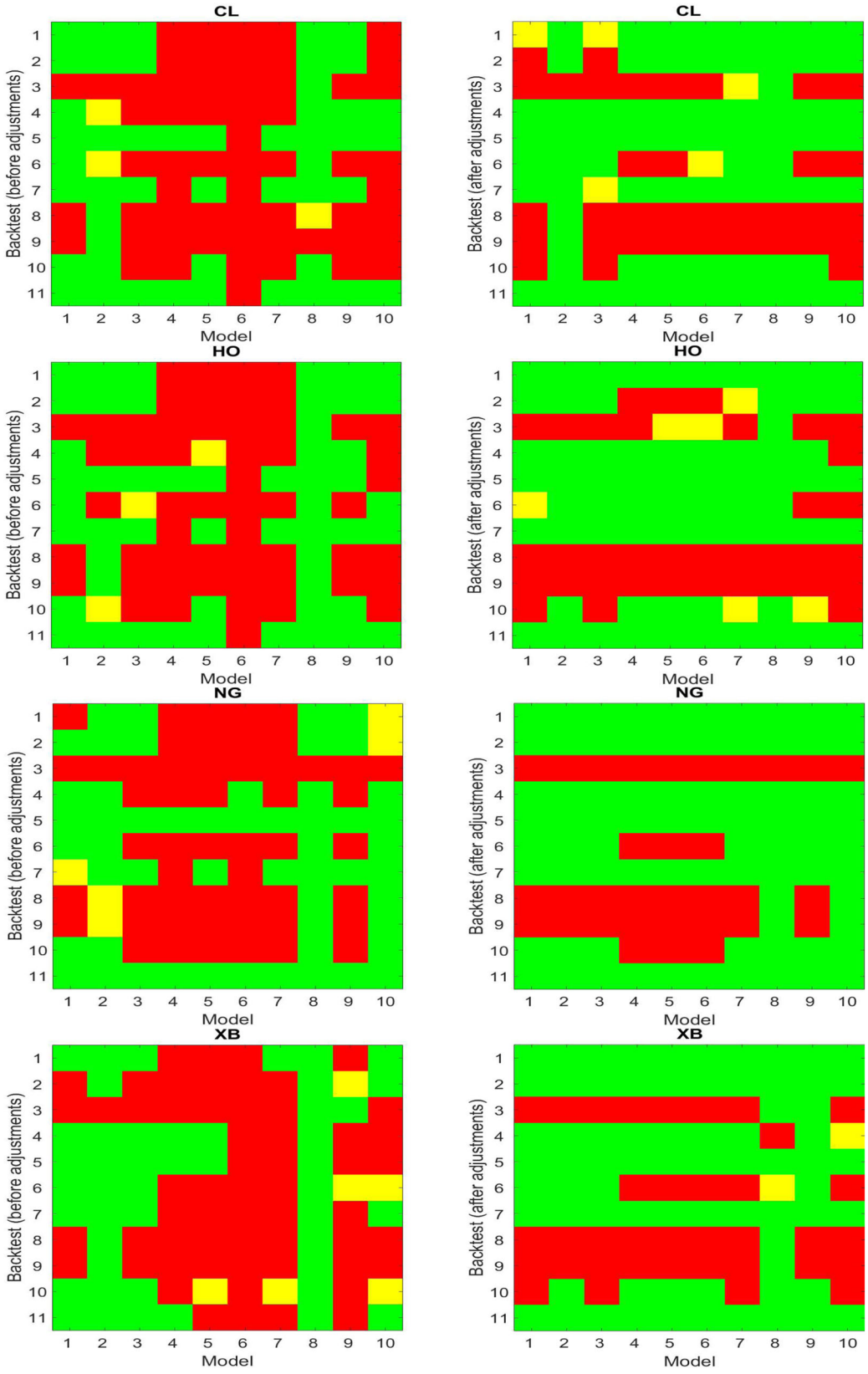

FIGURE C2 5% VaR and ES backtesting results with respect to *p*-values before and after adjustments, over the full period. A *p*-value smaller than 0.05 (between 0.05 and 0.1) is shaded with red (yellow); otherwise, it is green. CL, HO, NG, and XB are abbreviations for WTI Crude Oil, Heating Oil, Natural Gas, and RBOB/Unleaded Gasoline, respectively. Models 1–10 are HS, WHS, CF, G-*t*, G-*skt*, EVT-POT, GAS-1F, FHS, CAViaR-SAV, and CARE-SAV, accordingly. Backtests 1–11 are UC, CC, DQ, two-sided ER, one-sided ER, two-sided CCA, one-sided CCA, ESR Strict, ESR AUX, two-sided ESR Int, and one-sided ESR Int. [Color figure can be viewed at wileyonlinelibrary.com]

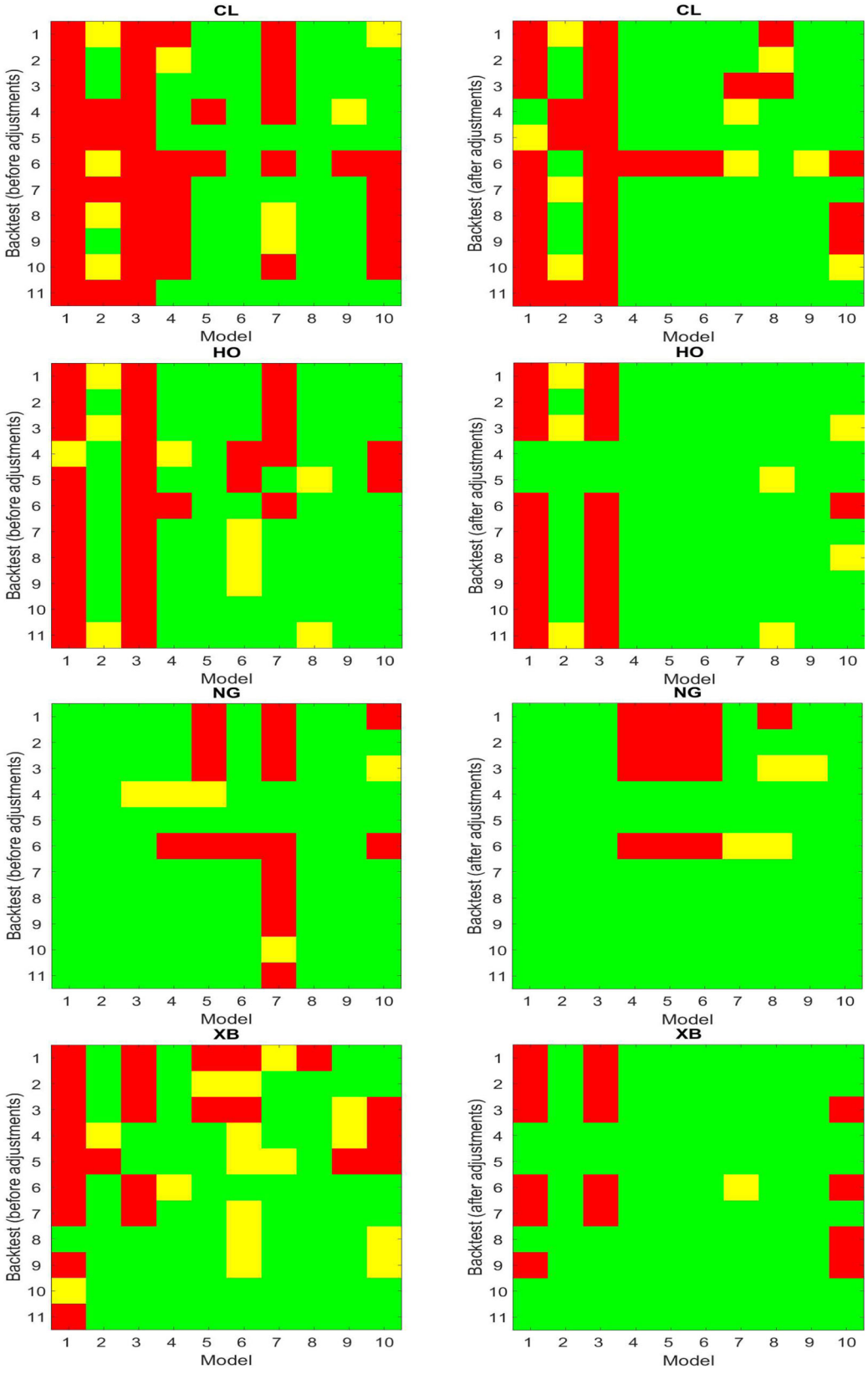

FIGURE C3 2.5% VaR and ES backtesting results with respect to *p*-values before and after adjustments, over the crisis period. A *p*-value smaller than 0.05 (between 0.05 and 0.1) is shaded with red (yellow); otherwise, it is green. CL, HO, NG, and XB are abbreviations for WTI Crude Oil, Heating Oil, Natural Gas, and RBOB/Unleaded Gasoline, respectively. Models 1–10 are HS, WHS, CF, G-*t*, G-*skt*, EVT-POT, GAS-1F, FHS, CAViaR-SAV, and CARE-SAV, accordingly. Backtests 1–11 are UC, CC, DQ, two-sided ER, one-sided ER, two-sided CCA, one-sided CCA, ESR Strict, ESR AUX, two-sided ESR Int, and one-sided ESR Int. [Color figure can be viewed at wileyonlinelibrary.com]

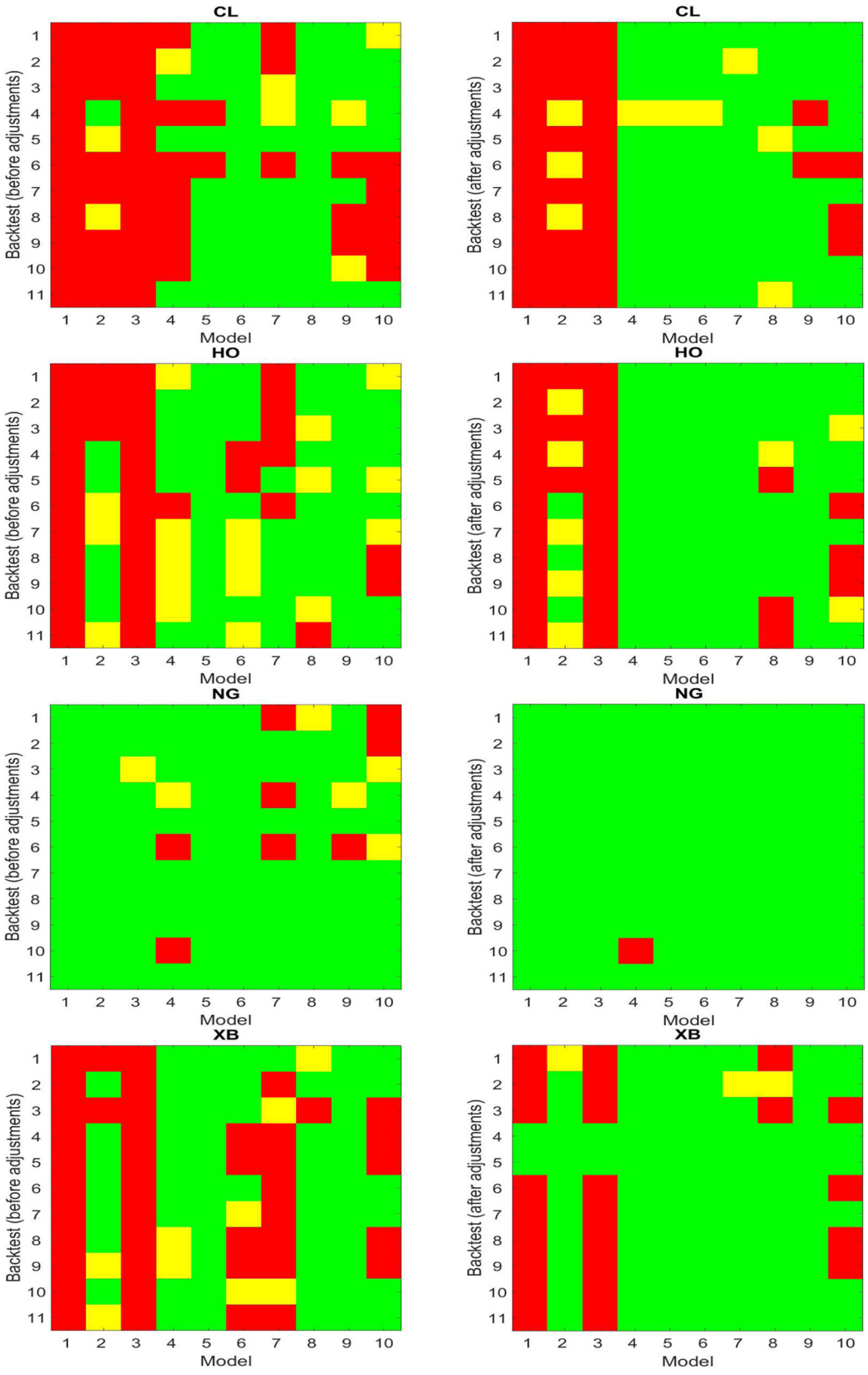

FIGURE C4 5% VaR and ES backtesting results with respect to *p*-values before and after adjustments, over the crisis period. A *p*-value smaller than 0.05 (between 0.05 and 0.1) is shaded with red (yellow); otherwise, it is green. CL, HO, NG, and XB are abbreviations for WTI Crude Oil, Heating Oil, Natural Gas, and RBOB/Unleaded Gasoline, respectively. Models 1–10 are HS, WHS, CF, G-*t*, G-*skt*, EVT-POT, GAS-1F, FHS, CAViaR-SAV, and CARE-SAV, accordingly. Backtests 1–11 are UC, CC, DQ, two-sided ER, one-sided ER, two-sided CCA, one-sided CCA, ESR Strict, ESR AUX, two-sided ESR Int, and one-sided ESR Int. [Color figure can be viewed at wileyonlinelibrary.com]

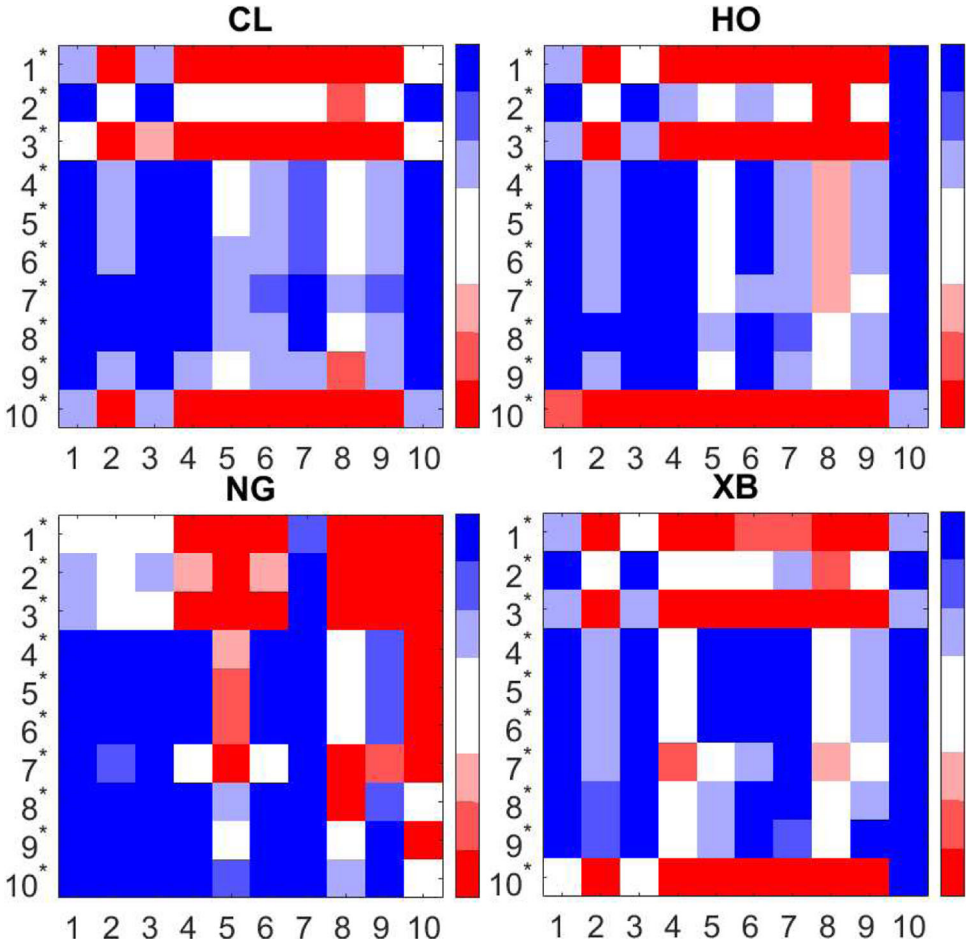

FIGURE C5 Color map based on the Diebold–Mariano (DM) test comparing the average losses using the *FZ*0 loss function for 2.5% VaR and ES over the full sample. Forecasts marked with 1* to 10* are the adjusted risk measures forecasts based on the original forecasts marked with 1–10. Blue blocks mean that the row forecast has a lower average loss than the column forecast at different significance levels (the darkest shade means that we reject the null hypothesis at the 1% significance level and so on). White blocks mean that there is no significant difference between the row forecast and the column forecast. Red blocks mean that the row forecast has a higher average loss than the column forecast (the darkest shade means that we reject the null hypothesis at the 1% significance level and so on). CL, HO, NG, and XB are abbreviations for WTI Crude Oil, Heating Oil, Natural Gas, and RBOB/Unleaded Gasoline, respectively. Forecasts 1–10 are generated from HS, WHS, CF, G-*t*, G-*skt*, EVT-POT, GAS-1F, FHS, CAViaR-SAV, and CARE-SAV, accordingly. [Color figure can be viewed at wileyonlinelibrary.com]

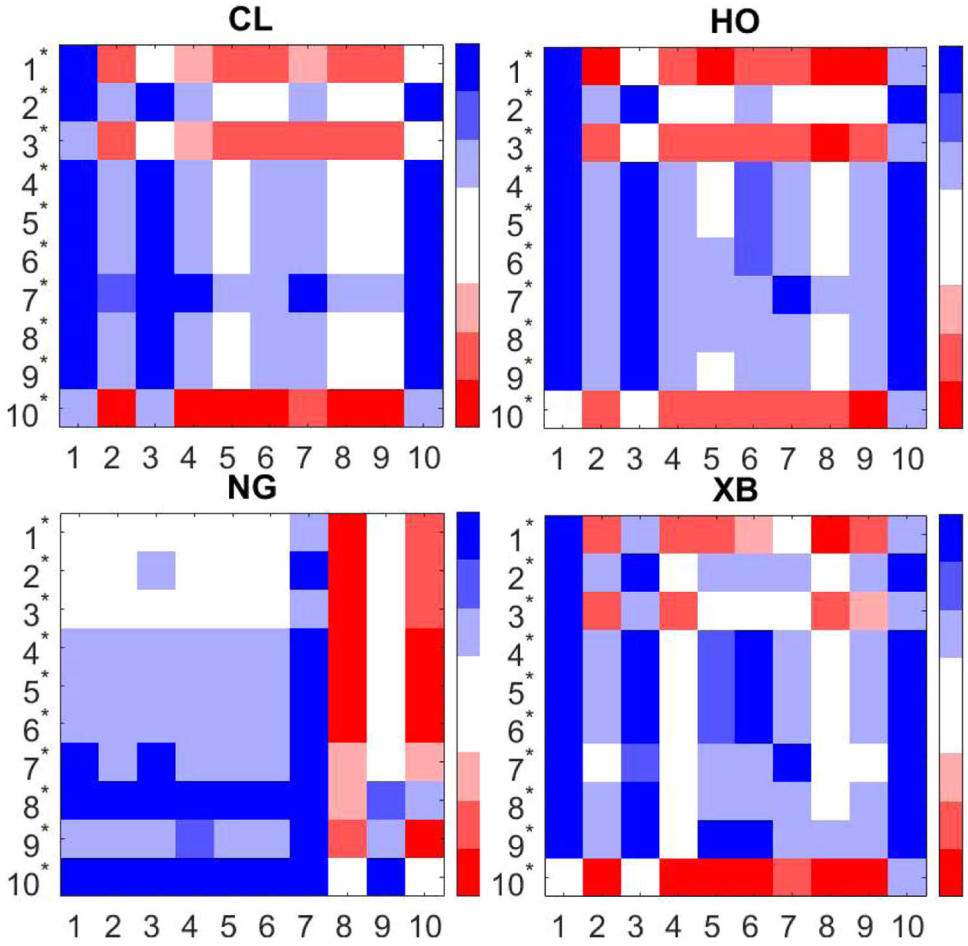

FIGURE C6 Color map based on the Diebold–Mariano (DM) test comparing the average losses using the *FZ*0 loss function for 2.5% VaR and ES over the crisis period. Forecasts marked with 1* to 10* are the adjusted risk measures forecasts based on the original forecasts marked with 1 to 10. Blue blocks mean that the row forecast has a lower average loss than the column forecast at different significance levels (the darkest shade means that we reject the null hypothesis at the 1% significance level and so on). White blocks mean that there is no significant difference between the row forecast and the column forecast. Red blocks mean that the row forecast has a higher average loss than the column forecast (the darkest shade means that we reject the null hypothesis at the 1% significance level and so on). CL, HO, NG, and XB are abbreviations for WTI Crude Oil, Heating Oil, Natural Gas, and RBOB/Unleaded Gasoline, respectively. Forecasts 1–10 are generated from HS, WHS, CF, G-*t*, G-*skt*, EVT-POT, GAS-1F, FHS, CAViaR-SAV, and CARE-SAV, accordingly. [Color figure can be viewed at wileyonlinelibrary.com]

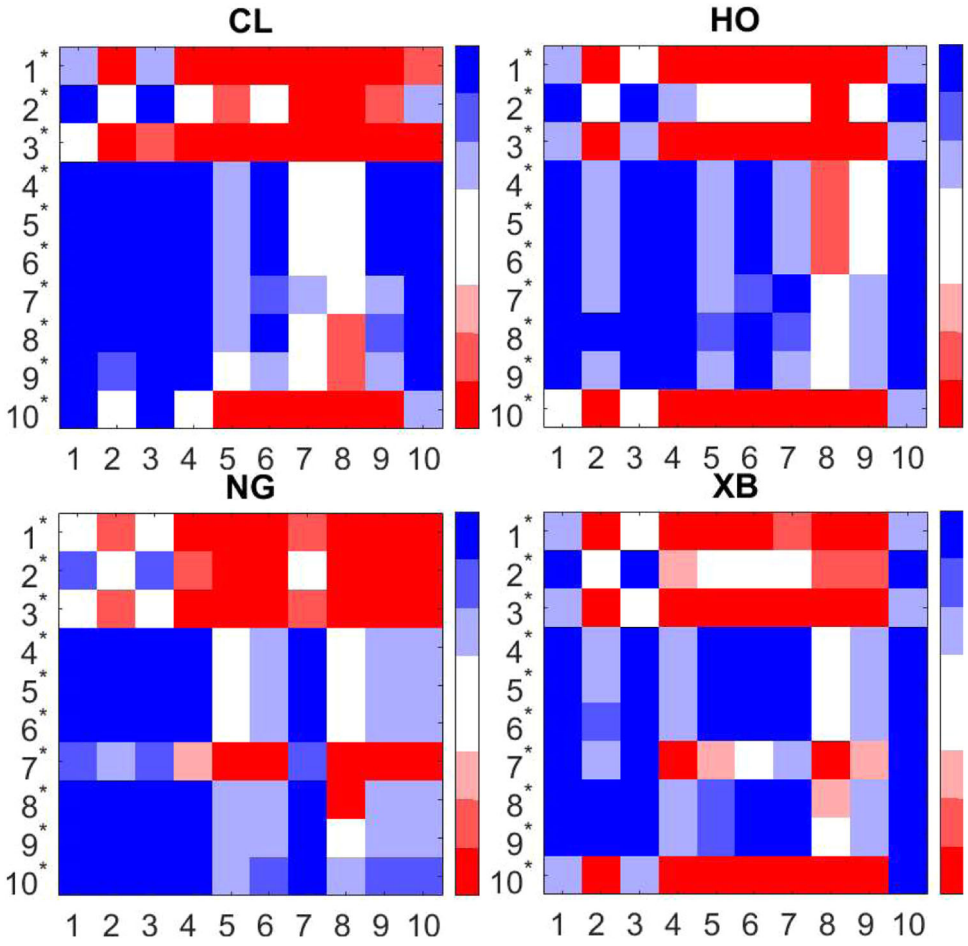

FIGURE C7 Color map based on the Diebold–Mariano (DM) test comparing the average losses using the *FZ*0 loss function for 5% VaR and ES over the full sample. Forecasts marked with 1* to 10* are the adjusted risk measures forecasts based on the original forecasts marked with 1–10. Blue blocks mean that the row forecast has a lower average loss than the column forecast at different significance levels (the darkest shade means that we reject the null hypothesis at the 1% significance level and so on). White blocks mean that there is no significant difference between the row forecast and the column forecast. Red blocks mean that the row forecast has a higher average loss than the column forecast (the darkest shade means that we reject the null hypothesis at the 1% significance level and so on). CL, HO, NG, and XB are abbreviations for WTI Crude Oil, Heating Oil, Natural Gas, and RBOB/Unleaded Gasoline, respectively. Forecasts 1-10 are generated from HS, WHS, CF, G-*t*, G-*skt*, EVT-POT, GAS-1F, FHS, CAViaR-SAV, and CARE-SAV, accordingly. [Color figure can be viewed at wileyonlinelibrary.com]

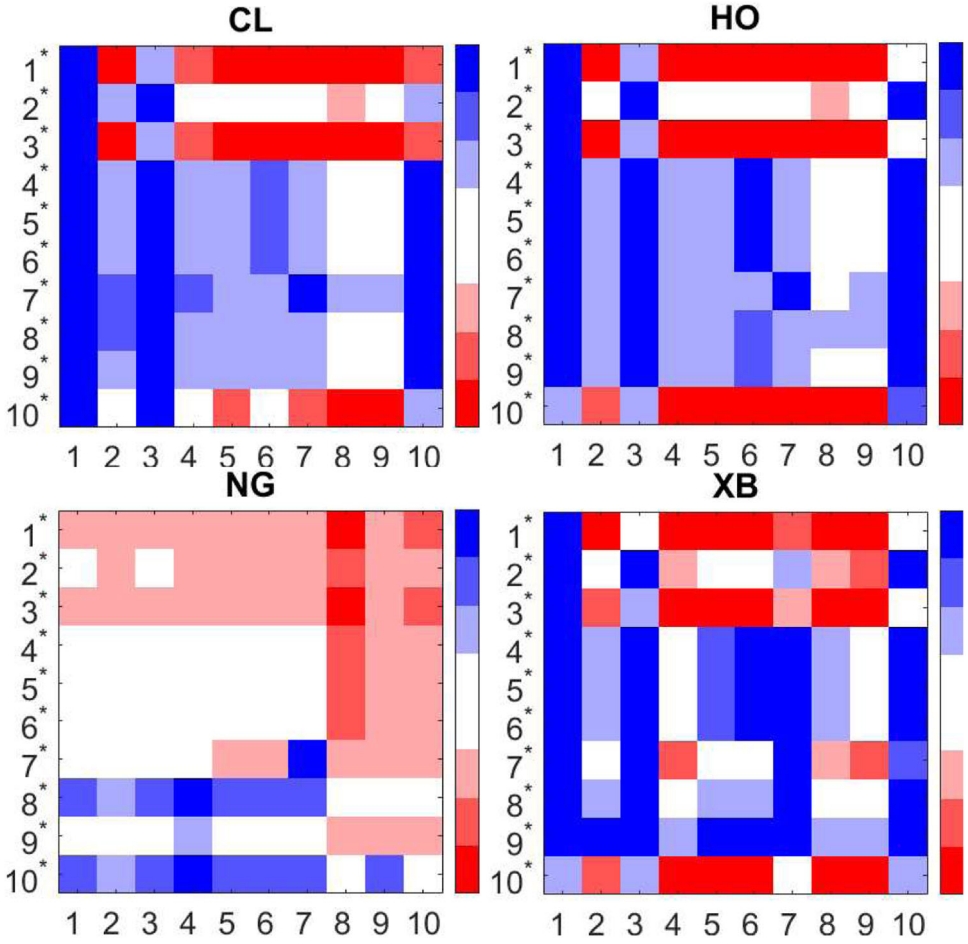

FIGURE C8 Color map based on the Diebold–Mariano (DM) test comparing the average losses using the *FZ*0 loss function for 5% VaR and ES over the crisis period. Forecasts marked with 1* to 10* are the adjusted risk measures forecasts based on the original forecasts marked with 1–10. Blue blocks mean that the row forecast has a lower average loss than the column forecast at different significance levels (the darkest shade means that we reject the null hypothesis at the 1% significance level and so on). White blocks mean that there is no significant difference between the row forecast and the column forecast. Red blocks mean that the row forecast has a higher average loss than the column forecast (the darkest shade means that we reject the null hypothesis at the 1% significance level and so on). CL, HO, NG, and XB are abbreviations for WTI Crude Oil, Heating Oil, Natural Gas, and RBOB/Unleaded Gasoline, respectively. Forecasts 1-10 are generated from HS, WHS, CF, G-*t*, G-*skt*, EVT-POT, GAS-1F, FHS, CAViaR-SAV, and CARE-SAV, accordingly. [Color figure can be viewed at wileyonlinelibrary.com]