



# Forecasting the volatility of energy transition metals

Andrea Bastianin<sup>1,2</sup> · Xiao Li<sup>2,3</sup> · Luqman Shamsudin<sup>2</sup>

Received: 9 March 2026 / Accepted: 28 September 2026  
© The Author(s) 2026

## Abstract

The energy and digital transitions are increasing demand for metals used in renewable energy, electrification, storage, and digital infrastructure. Many of these metals are traded in thin and concentrated markets with limited substitutability, making volatility a key constraint on investment, hedging, and policy design. This paper analyzes the return and volatility dynamics of 16 Energy Transition Metals. We combine exploratory and dimension-reduction methods with in-sample estimation and out-of-sample point and density forecasts, comparing a baseline GARCH(1,1) model with several stochastic-volatility specifications. The results reveal substantial heterogeneity across metals. Commodities commonly grouped by data providers or linked through geological co-occurrence often exhibit markedly different return and volatility characteristics. Forecasting performance is similarly heterogeneous: neither GARCH nor any single stochastic-volatility specification consistently dominates across commodities. Although some associations emerge between empirical features and model performance, they are not robust across metals or evaluation criteria. Overall, the findings support a metal-specific approach to volatility modelling and risk management and suggest that market maturity, liquidity, price discovery, and hedging opportunities may help explain these differences.

**Keywords** Critical raw materials · Energy transition · Features · Volatility · Forecasting · Density forecasts

## Introduction

The adoption of a cleaner energy mix, which is a cornerstone of policies to achieve net-zero greenhouse gas emissions by 2050, together with the ongoing digital transition, will require massive efforts to secure the metals needed to deploy technologies at scale. Renewable energy generation, the electrification of entire sectors of the global economy, and the expansion of digital infrastructure are driving demand for minerals such as copper, lithium, nickel, cobalt,

and rare earth elements (World Bank 2020; International Energy Agency 2021a, b). Metal prices are expected to reach historic highs for an unprecedentedly long period in a net-zero scenario (Boer et al. 2024). While the total value of metals production would more than quadruple between 2021 and 2040 (European Commission 2020), current production and investment plans fall far short of what is needed to support the energy transition.

There is also widespread political recognition, at both national and supranational levels, of the growing strategic importance of minerals to economies around the world. Focusing on the EU, the European Commission (EC) regularly publishes a list of Critical Raw Materials (CRMs) and the political debate routinely addresses the issue of securing access to raw materials. CRMs are economically and strategically important materials characterized by a low degree of substitutability, high supply risk and production concentrated in a few countries see e.g., (Graedel et al., 2015). These characteristics, coupled with low price elasticity of supply and demand, exacerbate the price volatility of these commodities (Boer et al. 2024; Fally and Sayre 2018; Addison and Ghoshray 2023).

---

✉ Xiao Li  
xiao.li@unibs.it

Andrea Bastianin  
andrea.bastianin@unimi.it

Luqman Shamsudin  
a.luqman.shamsudin@gmail.com

<sup>1</sup> Department of Economics, Management, and Quantitative Methods, University of Milan, Milan, Italy

<sup>2</sup> Fondazione Eni Enrico Mattei, Milan, Italy

<sup>3</sup> Department of Economics and Management, University of Brescia, Brescia, Italy

This paper represents the first attempt to model the price volatility of a broad set of 16 Energy Transition Metals (ETMs), which are a subset of CRMs. The label “Energy Transition Metals” follows the International Monetary Fund (IMF), which groups these metals – selected for their importance to clean-energy technologies – and publishes a corresponding price index as part of its Primary Commodity Price System see also, (Boer et al., 2024). We first characterize the main empirical features of the volatility series using a mix of exploratory data analysis, data reduction and visualization methods (Kang et al. 2017). We then evaluate a number of volatility models, both in-sample and out-of-sample, and examine how differences in these features are associated with model performance across metals. Our main result is that ETMs exhibit complex and highly heterogeneous volatility dynamics. This heterogeneity extends to forecasting performance and does not align with the standard groupings used by data providers or with geological similarities between metals. Differences in market liquidity and maturity may contribute to these patterns, although these channels are not directly tested in our analysis.<sup>1</sup>

Because some metals are often classified into homogeneous groups by data providers (e.g., base, precious, or minor metals), and some are mined together due to their natural occurrence in the same geological deposits (e.g., copper, nickel, and cobalt found in sulfide-rich ores), one might expect them to exhibit similar volatility patterns. On the contrary, when we focus on a key set of features and summarize them using data reduction techniques, we uncover significant heterogeneity. This analysis highlights that distinct clusters emerge, which neither align with the standard groupings used by data providers nor reflect the geological properties of metals. Moreover, we show that neither Generalized Autoregressive Conditional Heteroskedasticity (GARCH) models, which we use as a benchmark specification, nor a single Stochastic Volatility (SV) specification systematically prevails in a forecasting horse race.

A broad spectrum of volatility models has been developed in the literature to capture the complex dynamics of financial and commodity time series. Two examples beyond GARCH and SV models include Generalized Autoregressive Score (GAS) models (Creal et al. 2013) and Heterogeneous Autoregressive (HAR) models (Corsi 2009), as well as other realized volatility models based on high-frequency data. While each of these specifications offers distinct advantages, their applicability is constrained by data frequency. Given the monthly nature of our dataset, we focus

on GARCH and SV models, which are widely used in macroeconomic forecasting and well-suited to lower-frequency data. For instance, as emphasized by Clark and Ravazzolo (2015), in macroeconomic applications Autoregressive (AR) and Vector AR specifications with conventional stochastic volatility dominate other volatility specifications in terms of point forecasting to some degree and density forecasting to a greater degree. Our empirical analysis confirms that neither GARCH nor any single SV specification systematically dominates in forecasting performance, reinforcing the importance of comparative model evaluation in this context.

Our paper sits at the intersection of two strands of literature. The first concerns the economics of CRMs and the energy transition. These works develop frameworks for assessing material criticality and supply risk (Graedel et al. 2015), examine the role of technology and substitution in shaping mineral demand (Diemer et al. 2022; Pommeret et al. 2022), and project the large increases in metal demand implied by decarbonization pathways, together with the associated price pressures and supply constraints (Boer et al. 2024; Saadaoui et al. 2025). Despite these advances, the literature remains largely scenario- and price-based, with limited attention to the volatility behavior of the metals central to the energy transition. We contribute by shifting the focus from demand–supply balances and price levels to volatility and, in particular, to out-of-sample forecasting.

The second strand examines volatility in commodity and financial markets, documenting its empirical properties, evaluating alternative models – including GARCH, stochastic-volatility, and realized-volatility specifications – and analyzing market spillovers, tail risk, and macroeconomic linkages see, e.g., (Bakas and Triantafyllou, 2019; Bollerslev et al., 2018; Chan and Grant, 2016; Ciner et al., 2020; Dinh et al., 2022; Han et al., 2022; Naeem et al., 2024; Pham and Hsu, 2024). Recent studies extend this literature to energy-transition and geopolitical risks, showing that renewable-energy penetration is associated with extreme volatility in wholesale electricity prices and that war-induced energy shocks can alter the roles of renewable and fossil energy (Wang et al. 2024b, 2026). ETMs, however, have received little systematic attention as a group. Existing studies typically focus on individual metals, such as rare earths, lithium, or nickel, or on broad base- and precious-metal aggregates, while the few contributions that explicitly consider ETMs emphasize prices and returns rather than volatility.<sup>2</sup> We complement this literature by providing a comparative analysis and out-of-sample evaluation of volatility models across the metals underpinning the energy transition.

<sup>1</sup> Our focus is on characterizing and forecasting volatility in these markets. We do not model future demand–supply balances under deployment scenarios, nor the channels through which volatility may feed back into investment or technology adoption; instead, we provide the volatility evidence on which such analyses can build.

<sup>2</sup> An exception is Bastianin et al. (2023), who study both ETM returns and volatilities.

Our specific contributions are twofold. First, we provide the first systematic comparative analysis of volatility across the full set of IMF ETMs, comparing the in- and out-of-sample performance of GARCH and SV models in forecasting returns and volatilities for 16 ETMs. Second, we characterize the time-series features of ETM volatility and examine their association with heterogeneity in forecasting accuracy across models and metals. While feature extraction has been used in the statistical forecasting literature to cluster time series and to select and combine forecasts (Kang et al. 2017; Montero-Manso et al. 2020; Talagala et al. 2023), its application to commodity volatility forecasting is, to our knowledge, new.

The remainder of the paper is structured as follows: Section 2 describes the data and methodology. Section 3 introduces the feature space, while Section 4 links features to the in-sample and out-of-sample evaluation of volatility models. Section 5 presents robustness checks and the potential implications of our results for correlation analysis. Section 6 concludes. Additional results and methodological details are presented in the [Supplementary Information](#) (SI).

## Data and methods

### Data and volatility proxies

We focus on the volatility of the real prices of the 16 constituents of the ETM index published by the IMF.<sup>3</sup> The ETM index consists of seven base metals (i.e., aluminium, cobalt, copper, lead, molybdenum, nickel, zinc), three precious metals (i.e., palladium, platinum, silver), and six minor metals (i.e., chromium, lithium, manganese, rare earth elements<sup>4</sup>, silicon, vanadium) that are in high demand for technologies related to the energy transition. In the following, we refer to the minor metals as “other ETMs”. The dataset in our analysis spans from July 2012 to December 2022 and consists of

<sup>3</sup> The price index is then calculated as a weighted average of the prices of the 16 ETMs, where the weight assigned to each metal is based on its share of imports in the total imports of the 16 ETMs. See the “Technical Documentation” accompanying the IMF Primary Commodity Price Index for further details (see: <https://www.imf.org/en/Research/commodity-prices>).

<sup>4</sup> The rare-earth series reports Shanghai Metal Market prices for mixed rare earth carbonate, specified under Chinese standards for ion-adsorption-type rare earth carbonate. The product contains 42–45% rare earth oxides, with minimum shares of cerium oxide of 48%, praseodymium oxide of 5%, neodymium oxide of 15%, and europium oxide of 0.18%. The index aggregates different rare earths into a single basket, reflecting joint mining and processing, but may mask element-level differences in end markets, price drivers, and volatility dynamics. Despite this limitation, it is one of the few high-frequency market-based price references for rare earths. For instance, USGS element-level measures are annual unit values rather than observed market transaction prices.

126 monthly observations. See Table 1 for some descriptive information, and Section S2.1 of the SI for further details about the data.

The daily nominal prices of the ETMs are sourced from LSEG (formerly Refinitiv Eikon) to exactly overlap with the price series tracked by the IMF. Monthly real prices,  $P_t$ , are calculated by taking the average daily nominal prices deflated by the interpolated US Consumer Price Index sourced from BLS. More precisely,  $P_t = D_t^{-1} \sum_{d=1}^{D_t} P_{d,t}$  where

$P_{d,t}$  is the daily real price and  $D_t$  represents the number of days in month  $t$ . Monthly percentage log-returns used to estimate the volatility models are  $r_t = 100 \times \log(P_t/P_{t-1})$ . This process also delivers daily percentage log real returns,  $r_{d,t} = 100 \times \log(P_{d,t}/P_{d-1,t})$ , which we use to construct the volatility proxies described in Section 2.3.

There are two reasons for using real prices at monthly sampling frequency rather than considering weekly or daily data. Firstly, the low liquidity in certain minerals markets makes the use of daily data impractical. Secondly, since true volatility is not directly observable, we require a volatility proxy to assess the predictive ability of models. In this paper, we construct such proxies by aggregating daily data to a monthly frequency (Schwert 1989).

### Volatility models

We compare two classes of volatility models: GARCH and SV.<sup>5</sup> The GARCH(1,1) model serves as a well-established benchmark, whereas SV models include a stochastic error component in the volatility equation, enabling them to capture more complex volatility dynamics. We use a range of SV models, each tailored to capture specific features such as volatility persistence, asymmetry, extreme movements, and infrequent jumps.

The GARCH(1,1) model of Bollerslev (1986) can be written as follows:

$$r_t = \mu + \epsilon_t, \quad \epsilon_t \sim N(0, \sigma_t^2) \quad (1)$$

$$\sigma_t^2 = \alpha_0 + \alpha_1 \epsilon_{t-1}^2 + \beta_1 \sigma_{t-1}^2 \quad (2)$$

where innovations  $\epsilon_t$  are normally distributed with zero mean and conditional variance  $\sigma_t^2$ . The parameters in the conditional variance equation should satisfy  $\alpha_0 > 0$ ,  $\alpha_1 > 0$ ,  $\beta_1 \geq 0$  to guarantee that  $\sigma_t^2 \geq 0$ . Moreover, the model is covariance-stationary if  $\alpha_1 + \beta_1 < 1$ . GARCH(1,1)

<sup>5</sup> For the in-sample comparison based on marginal likelihoods, both the GARCH and SV models are estimated within a Bayesian framework.

**Table 1** Technologies, expected demand increases and volatility of ETMs

| Metal      | Category | Technology |      |    |         |          |       | Projected                              |                                       |
|------------|----------|------------|------|----|---------|----------|-------|----------------------------------------|---------------------------------------|
|            |          | Solar      | Wind | EV | Battery | Hydrogen | Other | Demand <sup>2050</sup> <sub>2022</sub> | <sup>2016–22</sup> <sub>2012–15</sub> |
| Aluminium  | Base     |            |      |    |         |          | ✓     | 2.0                                    | 1.82                                  |
| Cobalt     | Base     |            |      | ✓  | ✓       | ✓        | ✓     | 4.3                                    | 4.76                                  |
| Copper     | Base     | ✓          | ✓    | ✓  | ✓       | ✓        | ✓     | 3.0                                    | 1.30                                  |
| Lead       | Base     | ✓          |      |    |         |          | ✓     | 1.3                                    | 1.28                                  |
| Molybdenum | Base     | ✓          | ✓    |    |         |          | ✓     | 2.5                                    | 1.48                                  |
| Nickel     | Base     | ✓          | ✓    | ✓  | ✓       | ✓        | ✓     | 8.2                                    | 1.72                                  |
| Zinc       | Base     | ✓          | ✓    |    |         |          | ✓     | 3.0                                    | 1.82                                  |
| Chromium   | Other    |            | ✓    |    |         |          | ✓     | 2.5                                    | 1.59                                  |
| Lithium    | Other    |            |      |    | ✓       |          | ✓     | 16.1                                   | 20.13                                 |
| Manganese  | Other    |            | ✓    | ✓  | ✓       |          | ✓     | 12.9                                   | 3.83                                  |
| REE        | Other    |            | ✓    | ✓  |         | ✓        |       | 175.4                                  | 1.43                                  |
| Silicon    | Other    | ✓          |      | ✓  | ✓       |          |       | 0.1                                    | 5.13                                  |
| Vanadium   | Other    |            |      | ✓  | ✓       |          |       | -                                      | 3.45                                  |
| PGMs       | Precious |            |      |    |         | ✓        |       | 119.5                                  | 0.65                                  |
| Silver     | Precious |            |      |    |         |          | ✓     | 0.4                                    | 3.83                                  |

*Notes:* in column 1, palladium and platinum are included in the Platinum Group Metals (PGMs). The technologies in columns 3–8 and the projected increase in demand (i.e. projected demand in 2050 divided by demand in 2022) in column 9 are taken from IEA (2023) and are based on the “Net Zero Emission Scenario by 2050”. For vanadium we do not show the projected increase in demand as there is no demand in 2022. The last column shows the ratio of the coefficient of variation of prices in 2016–22 to that in 2012–15. For PGMs, the average CV ratio for palladium and platinum is shown. The category “other” includes low-emission power generation technologies (e.g. hydropower, geothermal, bioenergy for electricity and nuclear power) and electricity networks

forecasts serve as a benchmark against which we compare five different SV specifications aimed at capturing different features of returns and volatilities of the ETMs.

The simplest SV model we analyze, denoted as SV(1), assumes that the log-volatility  $\log \sigma_t^2$  evolves as a stationary autoregressive process of order 1,  $AR(1)$ :

$$r_t = \mu + \epsilon_t, \quad \epsilon_t \sim N(0, e^{\log \sigma_t^2}) \quad (3)$$

$$\log \sigma_t^2 = \phi_0 + \phi_1(\log \sigma_{t-1}^2 - \phi_0) + \eta_t, \quad \eta_t \sim N(0, \omega^2) \quad (4)$$

with  $|\phi_1| < 1$  to ensure covariance-stationarity of the  $AR(1)$  process. The volatility equation incorporates a stochastic error term  $\eta_t$ , which is normally distributed with zero mean and a variance of  $\omega^2$ .

A simple way to extend the SV(1) model is to allow the log-volatility to follow an  $AR(2)$ , so that we can capture richer variance dynamics:

$$\log \sigma_t^2 = \phi_0 + \phi_1(\log \sigma_{t-1}^2 - \phi_0) + \phi_2(\log \sigma_{t-2}^2 - \phi_0) + \eta_t, \quad \eta_t \sim N(0, \omega^2) \quad (5)$$

in the SV(2) model, we assume that the roots of the characteristic polynomial associated with  $(\phi_1, \phi_2)$  lie outside the unit circle.

We also examine the SV model with  $t$ -distributed innovations, SV(1)- $t$ , which allows for more extreme observations compared to SV(1). In the SV(1)- $t$ , the log-volatility dynamics are again approximated by an  $AR(1)$  process, but now

we assume that innovations in the return distribution are  $t$ -distributed with  $\nu$  degrees of freedom:  $\epsilon_t \sim t_\nu(0, e^{\log \sigma_t^2})$ .

Next we consider the SV model with leverage, SV(1)-L, that allows for asymmetric effects of positive and negative returns on volatility. Log-volatility is approximated by an  $AR(1)$  process as in the SV(1) model, but now we assume that return innovations,  $\epsilon_t$ , and volatility innovations,  $\eta_t$ , follow a bivariate normal distribution:

$$\begin{bmatrix} \epsilon_t \\ \eta_t \end{bmatrix} \sim N \left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_t^2 & \rho \sigma_t \omega \\ \rho \sigma_t \omega & \omega^2 \end{bmatrix} \right) \quad (6)$$

where  $\rho$  is a correlation coefficient. When  $\rho < 0$  and a negative shock hits returns at time  $t$ , volatility at time  $t + 1$  will increase; when  $\rho = 0$  the SV(1)-L boils down to the standard SV(1) model.

Lastly, we consider a model with jumps, denoted as SV(1)-J, to capture infrequent, but significant price movements. In this case, log-volatility continues to be approximated by an  $AR(1)$  process, but the equation for returns becomes:

$$r_t = \mu + k_t q_t + \epsilon_t \quad (7)$$

where  $q_t \in \{0, 1\}$  is a jump variable with success probability  $\mathbb{P}(q_t = 1) = \kappa$  and  $k_t \sim N(\mu_k, \sigma_k^2)$  determines the size of the jump if it occurs at time  $t$ . As for the innovations, we continue assuming that:  $\epsilon_t \sim N(0, e^{\log \sigma_t^2})$ .

**Estimation and convergence diagnostics.** Priors, posterior simulation algorithms, and details on the computation of the marginal likelihoods are provided in Section S1 of the SI. Given the short sample and the flexibility of some of the SV specifications, we assess MCMC convergence using three parallel chains initialized from overdispersed starting values. Across all metals and models, the Gelman–Rubin statistic is no larger than 1.006, while the effective sample sizes generally range from several hundred to several thousand. Complete convergence diagnostics are reported in Table S3 of the SI.

## The forecasting exercise

We forecast volatilities and returns one step ahead using an expanding-window approach. We first estimate the models using data from July 2012 to December 2017 and forecast returns and volatilities for January 2018. We then add one observation to the estimation sample and forecast returns and volatilities for February 2018. This process is repeated until the end of the sample, resulting in a total of 60 one-step-ahead forecasts from January 2018 to December 2022. All model parameters are re-estimated at each forecast origin using only the observations available at that date.<sup>6</sup>

We evaluate models for their ability to produce accurate one-step-ahead point forecasts of volatilities, and density forecasts of returns. For the GARCH(1,1) model, point forecasts of the conditional variance,  $\hat{\sigma}_{t+1|t}^2$ , are obtained by iterating Equation 2 forward by one period. In the case of SV models, the one-step-ahead point forecast  $\hat{\sigma}_{t+1|t}^2$  is the mean of the predictive distribution of the conditional variance. Because the models are estimated using monthly returns,  $\hat{\sigma}_{t+1|t}^2$  measures the variance of the monthly percentage return. For comparison with the annualized volatility proxies, we annualize each model-based forecast as  $\hat{h}_{t+1|t}^{(m)} = 12\hat{\sigma}_{t+1|t}^{2(m)}$ , where the superscript  $m$  denotes the volatility model. Thus, the model forecasts and the volatility proxies used in the loss functions measure the same annualized conditional variance and are expressed in squared percentage points.

The predictive densities and the corresponding observed outcomes are expressed as monthly percentage log returns.<sup>7</sup> For SV models, the one-step-ahead predictive distribution of returns is obtained by simulating from the observation

equation across retained posterior draws of the parameters and latent states. The resulting predictive densities therefore integrate both parameter and latent-state uncertainty. For the GARCH(1,1) model, the one-step-ahead predictive distribution is generated from the conditional distribution of returns using the estimated conditional mean and the forecast conditional variance, while treating the maximum-likelihood parameter estimates as fixed.<sup>8</sup>

**Volatility proxies.** The models produce forecasts of the variance of the monthly return, while the volatility proxies are constructed by aggregating daily observations within each month. The use of daily data provides more informative proxies for the latent monthly variance without changing the monthly frequency of the forecasting exercise. Both the model forecasts and the proxies are annualized and expressed in squared percentage points, so they measure the same underlying quantity. Following Hansen and Lunde (2005) and Patton (2011), we use realized volatility and the adjusted squared range as alternative noisy proxies for latent conditional variance. Annualized realized volatility is defined as

$$RV_t = \frac{252}{D_t} \sum_{d=1}^{D_t} r_{d,t}^2, \quad (8)$$

where  $D_t$  is the number of daily observations in month  $t$ . The sum of squared daily log real returns estimates the variance accumulated during the month, while multiplication by  $252/D_t$  converts it into an annualized variance.

As an alternative proxy, we use the adjusted squared range. Let  $p_{d,t} = 100 \log(P_{d,t})$  denote the percentage log real price and define the within-month range as  $RG_t = \max_{d \in t}(p_{d,t}) - \min_{d \in t}(p_{d,t})$ . The adjusted squared range is then

$$RG2_t^* = 12 \left( \frac{RG_t}{2\sqrt{\log(2)}} \right)^2. \quad (9)$$

The divisor  $4\log(2)$  converts the squared high-low range into an estimator of the monthly variance, while multiplication by 12 annualizes the resulting measure.

**Loss functions.** We evaluate point forecasts of the conditional variance with two alternative loss functions, the Root

<sup>6</sup> In line with much of the literature, in the out-of-sample exercise the GARCH(1,1) model is re-estimated by maximum likelihood at each forecast origin, while the SV models are re-estimated using Bayesian methods. By contrast, the in-sample model comparison relies on Bayesian estimation for both classes of models, allowing us to compute the marginal likelihood consistently.

<sup>7</sup> No annualization is required because both refer to the return realized over the following month.

<sup>8</sup> The density-forecast comparison should therefore be interpreted as an evaluation of the forecasting procedures as implemented, rather than as a fully like-for-like comparison of the underlying model classes under a common treatment of parameter uncertainty (see Bassetti et al. 2020 for further details).

Mean Squared Forecast Error (RMSFE) and the QLIKE. These are defined as follows:

$$RMSFE^{(m)} = \sqrt{\frac{1}{Q} \sum_{t=1}^Q (h_t - \hat{h}_{t|t-1}^{(m)})^2}, \quad (10)$$

$$QLIKE^{(m)} = \frac{1}{Q} \sum_{t=1}^Q \left[ \frac{h_t}{\hat{h}_{t|t-1}^{(m)}} - \log \left( \frac{h_t}{\hat{h}_{t|t-1}^{(m)}} \right) - 1 \right]. \quad (11)$$

where  $Q$  denotes the size of the evaluation sample,  $h_t$  is either  $RV_t$  or  $RG2_t^*$ , and  $\hat{h}_{t|t-1}^{(m)} = 12\hat{\sigma}_{t|t-1}^{2(m)}$  is the annualized one-step-ahead conditional-variance forecast from model  $m$ . Thus, both the forecast and the proxy entering the loss function are measured in annualized squared percentage points. As shown by Patton (2011), the MSE and QLIKE belong to the class of loss functions that provide rankings robust to the use of conditionally unbiased but noisy volatility proxies. As we can see from Fig. 1, contrary to the MSFE, which is a symmetric loss function, the QLIKE criterion penalizes underforecasting more heavily than overforecasting.

Moving on to density forecasts of returns, we consider the quantile-based continuous ranked probability score (qCRPS) of Gneiting and Ranjan (2011). The interpretation of scoring rules is the same as that given to loss functions: in a pairwise comparison, the model with the lowest score is ranked as the most accurate. The qCRPS is defined as (we drop the model superscript,  $m$ , for ease of notation):

$$\widehat{QS}_t = \frac{1}{J-1} \sum_{j=1}^{J-1} \widehat{QS}_t^{\alpha_j} = \frac{1}{J-1} \sum_{j=1}^{J-1} 2 \left[ \mathbb{I}(r_{t+1} \leq \hat{q}_{t+1|t}^{\alpha_j}) - \alpha_j \right] \times (\hat{q}_{t+1|t}^{\alpha_j} - r_{t+1}), \quad (12)$$

where  $\mathbb{I}(\cdot)$  denotes the indicator function and  $\hat{q}_{t+1|t}^{\alpha_j}$  is the one-step ahead quantile forecast for  $r_{t+1}$  at level  $\alpha_j = j/J$  with  $J = 20$ , which corresponds to  $\alpha_j = 0.05, 0.10, \dots, 0.95$ . We focus on weighted versions of the qCRPS:

$$\widehat{wQS}_t = \frac{1}{J-1} \sum_{j=1}^{J-1} \pi(\alpha_j) \widehat{QS}_t^{\alpha_j} \quad (13)$$

where we select weights,  $\pi(\alpha_j)$ , to emphasize the center  $\pi(\alpha_j) = \alpha_j(1 - \alpha_j)$  or the tails of the distribution  $\pi(\alpha_j) = (2\alpha_j - 1)^2$ .

## The features of ETM time series

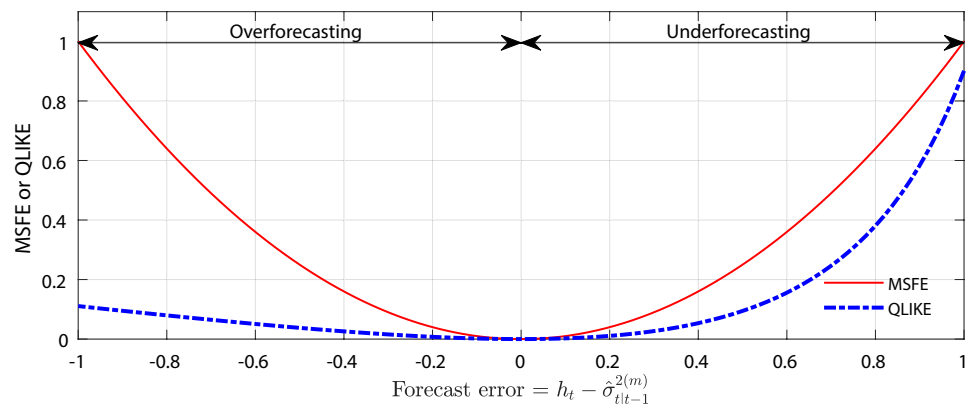
### Descriptive statistics

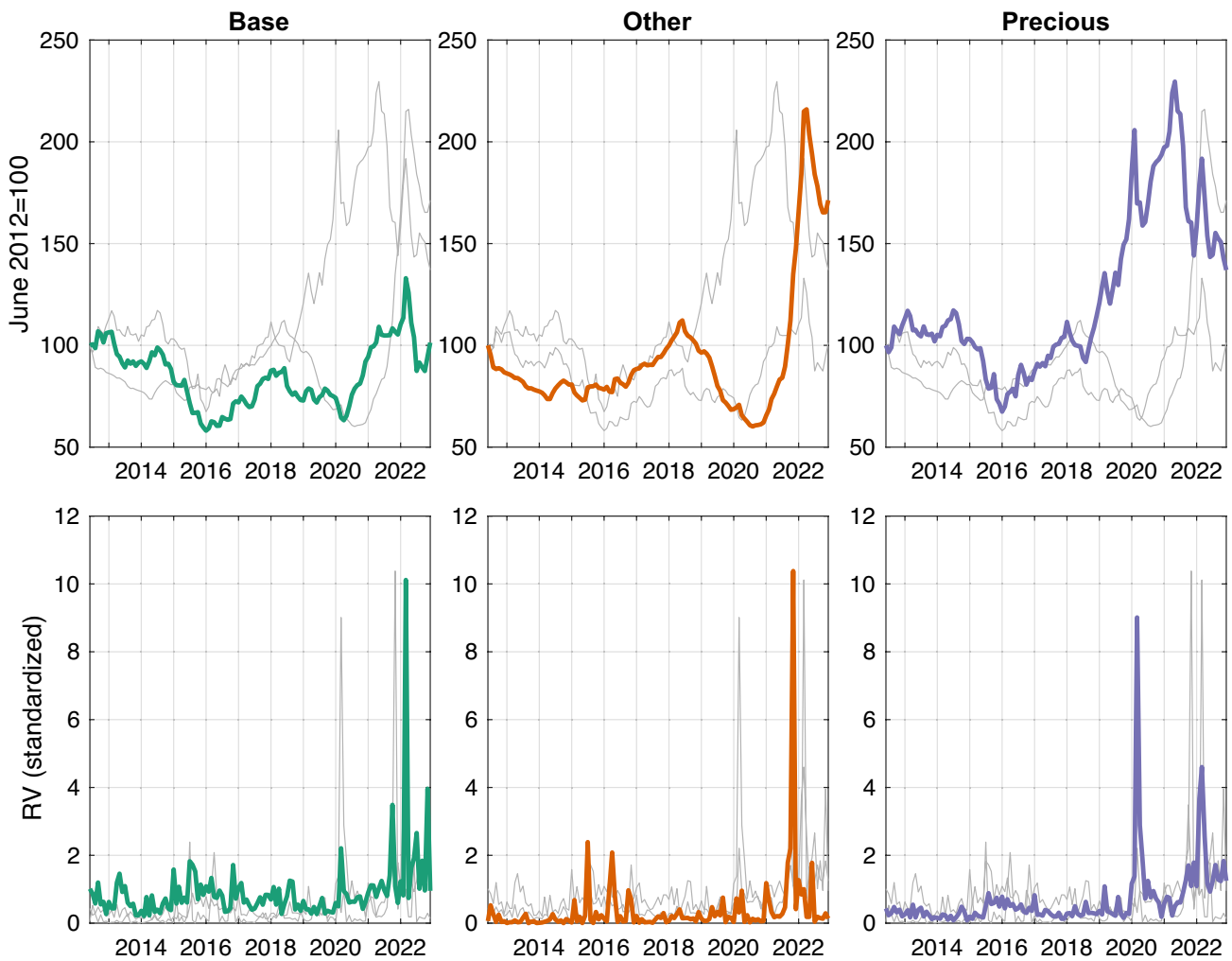
As can be seen from Table 1, we group ETMs into three broad categories: “base”, “other” and “precious”. Green energy technologies, hydrogen and electric mobility all use many ETMs. The projected increases in demand in 2050, as implied by the IEA’s net-zero emissions scenario, are in some cases orders of magnitude larger than 2022 levels. The last column shows the increase in price volatility, as measured by the coefficient of variation. This highlights that the period 2016–22 was a period of very high price volatility, particularly for base and precious metals. Lithium is a striking exception among the “other” metals, with by far the largest increase in volatility, as indicated by a CV ratio of about 20. This pattern coincides with substantial pressures on the battery supply chain, including rapidly expanding demand and slow-adjusting supply. Figure 2 shows the evolution of prices and RV for different categories of metals, while plots of the series for individual metals appear in Section S2.1 of the SI.

### Features of realized volatility for ETMs

To illustrate some stylized facts about the volatility of ETM prices, we focus on a few “features” of RV time series. While a “feature” refers to any measurable characteristic of a time series, their precise definition and selection is problem-specific (Talagala et al. 2023). In our analysis, we examine features

**Fig. 1** Mean squared forecast error and QLIKE loss. *Notes:* The figure plots the Mean Squared Forecast Error (MSFE) and QLIKE loss functions as functions of the forecast error





**Fig. 2** Evolution of prices and volatilities over time. *Notes:* The upper panels of the figure show the evolution of the prices of a group of metals, normalized to 100 in June 2012. The lower panels display the real-

ized volatility (RV) of each series, standardized to facilitate comparison. In all panels, thin gray lines represent the other series, while the thick colored line represents the series identified in the panel heading

of RV classified into three primary groups: autocorrelation (ACF) features, distributional features, and other features.<sup>9</sup> Each group of features maps onto a dimension of volatility behavior that the competing models are built to capture: ACF features summarize the persistence exploited by GARCH and by the autoregressive log-volatility of the SV models; distributional features (skewness, kurtosis, and the share of zero returns) capture the heavy tails, asymmetry, and illiquidity that motivate the SV(1)-*t*, SV(1)-L, and SV(1)-J specifications; and the remaining features (entropy and spikiness) gauge how forecastable each series is. Together they provide a compact description of the properties that may help account for differences in forecasting performance across models.

In the group of ACF features, we consider ( $F_1$ ) the first-order autocorrelation coefficient of RV  $\widehat{ACF}(1)$ ; ( $F_2$ ) the sum of squares of the first twelve autocorrelation coefficients of

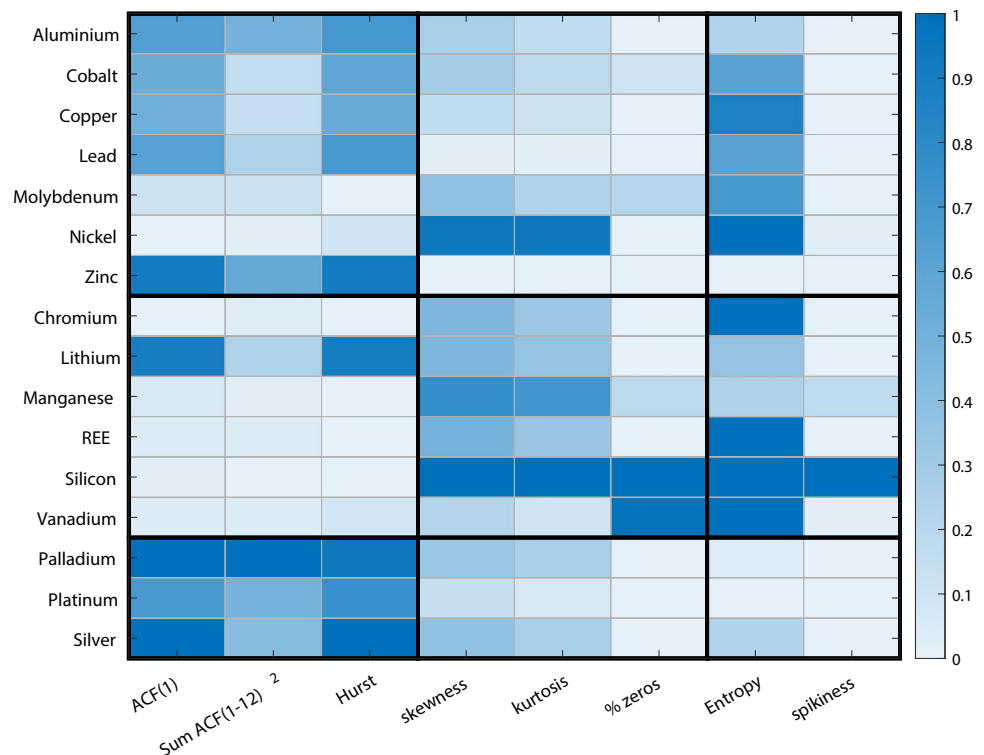
RV,  $\sum_{j=1}^{12} \widehat{ACF}(j)^2$ ; ( $F_3$ ) the Hurst coefficient, which serves

as a measure of “long memory” and thus, broadly speaking, signals the presence of significant autocorrelations over many lags.<sup>10</sup> As far as distributional characteristics are concerned, we analyze the sample skewness ( $F_4$ ) and kurtosis

<sup>10</sup> The Hurst coefficient or exponent  $H$  is estimated by fitting the equation  $E[\frac{R(n)}{S(n)}] = Cn^H$  to the data, where  $R$  represents the range of cumulative deviates of the data,  $S$  denotes the standard deviation of the data,  $n$  is the number of observations, and  $C$  is a constant. The estimation process involves taking the logarithm of both sides of the equation and subsequently fitting a straight line to the transformed data. The slope of this line corresponds to the value of  $H$ . The Hurst coefficient indicates the behavior of the trend in a time series, whether it is persistent (Hurst coefficient  $> 1/2$ ), antipersistent (Hurst coefficient  $< 1/2$ ), or exhibits short memory (Hurst coefficient  $= 1/2$ ). In the case of persistence, a high value is more likely to be followed by another high value. Conversely, in the case of antipersistence, a single high value is expected to be followed by a low value, with the subsequent value likely to be high.

<sup>9</sup> We use functions implemented in the tsfeatures R package by Hyndman et al. (2023) to compute some of these features.

**Fig. 3** Features of realized volatility. *Notes:* the heatmap shows the features of the RVs of the different metals. The thick black lines divide the metals into base, other and precious and the features into ACF, distributional and other. All features are normalized in the range (0,1). Darker shades indicate higher values of the feature, as shown in the colour bar



( $F_5$ ) of RV, as well as the share of zero monthly returns ( $F_6$ ), which we use as a simple proxy for infrequent price adjustment and market illiquidity. Specifically,  $F_6$  is calculated for each metal as the number of monthly observations for which the percentage log return  $r_t$  is equal to zero, divided by the total number of monthly return observations. Finally, in the group of other features, we analyze the Shannon spectral entropy measure ( $F_7$ ) and a proxy for the “spikiness” ( $F_8$ ) of RV see, e.g., (Hyndman and Athanasopoulos, 2018). The Shannon spectral entropy captures the noise level and the forecastability of the RV. A value close to 1 indicates high noise and makes forecasting difficult, while a value close to 0 indicates a strong trend (or seasonality), which makes forecasting easier. On the other hand, the “spikiness” of the RV measures the prevalence of spikes in the remainder component of the STL decomposition of the series.<sup>11</sup>

In general, the stronger the ACF features, the easier it should be for models to use historical patterns to accurately describe the stylized facts of ETM price volatility. On the other hand, the group of distributional features (i.e. skewness, kurtosis and percentage of zeros) and the group of other features (i.e. spectral entropy and spikiness) indicate the presence of outliers and noise in the time series.

<sup>11</sup> The STL (Seasonal and Trend decomposition using Loess) decomposition method decomposes a series, denoted as  $y_t$ , into three distinct components: the trend component, represented as  $T_t$ ; the seasonal component, denoted as  $S_t$ ; and the remainder component, represented as  $R_t$ . Mathematically, the decomposition can be expressed as follows:  $y_t = T_t + S_t + R_t$ . The spikiness is computed as the variance of the leave-one-out variances of  $R_t$ .

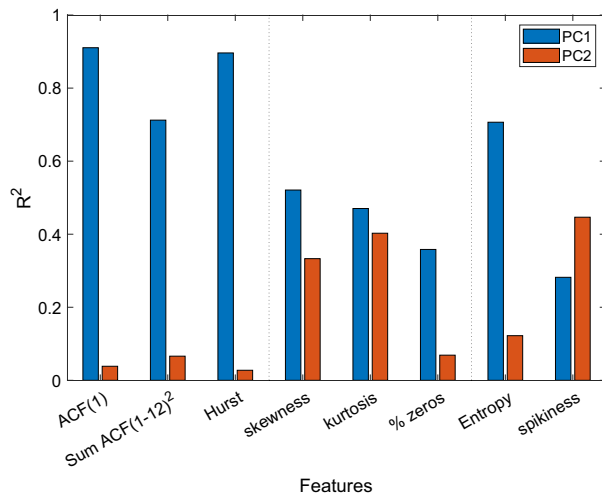
Consequently, a strong presence of these features indicates that it is more difficult for models to exploit past patterns and accurately model volatility.

After computing the eight features, we normalize them to the (0, 1) interval<sup>12</sup> and then collect them in the feature vector  $F = [F_1, F_2, \dots, F_8]'$ . In Fig. 3, we show a heatmap with the normalized feature values for different ETMs and the eight features. The bold black lines separate distinct clusters of ETMs (base, precious, other) and features (i.e., ACF, distributional, other). Base and precious metals tend to have high values for ACF features and low values for spikiness and zero returns<sup>13</sup>. On the contrary, most metals in the “other” category are characterized by low values for ACF features and high values concentrated on distributional features and entropy. Silicon is a case in point with high values for skewness, kurtosis, spectral entropy, spikiness and a high percentage of zero returns.

These feature patterns are associated with substantial differences across ETMs. Base and precious metals

<sup>12</sup> For ACF(1), we first take the absolute value.

<sup>13</sup> The prevalence of zero returns in the less liquid markets has direct consequences for both measurement and modeling. Infrequent trading mechanically dampens measured volatility between sporadic, large adjustments, producing series that combine long quiet stretches with occasional spikes. This is hard for volatility models to fit: it inflates apparent kurtosis and spikiness, weakens estimated autocorrelation, and pushes the data toward the jump- and tail-oriented behavior that SV(1)-J and SV(1)-t are designed to accommodate. It also means our proxies are noisier estimates of true volatility for the most illiquid metals.



**Fig. 4**  $R^2$  between principal components and features and the feature space of RVs. *Notes:* the left panel of Fig. 4 reports the  $R^2$  values from regressions of each of the eight features separately on  $PC_1$  and  $PC_2$ .

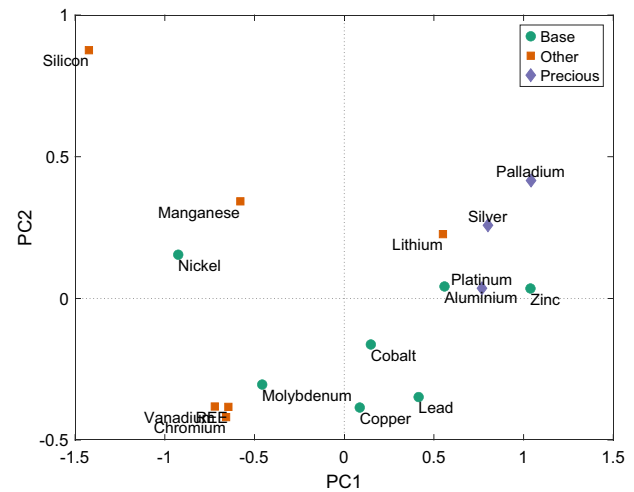
generally exhibit stronger autocorrelation and lower spikiness, whereas metals classified in the “other” group tend to display more irregular or abrupt price fluctuations.<sup>14</sup> The documented feature differences provide a useful basis for examining their association with forecasting performance in the following section.

### A principal component mapping of ETM features

Following Kang et al. (2017), we use Principal Component Analysis (PCA) to represent each time series of RV as a point in a two-dimensional “feature space”, also referred to as instance space.

While formal PCA diagnostics are reported in Section S2.8 of the SI, in what follows we retain two components because PCA is used for exploratory visualization rather than to estimate a latent factor structure or conduct formal inference. The first component explains 66.2% of the total variation and captures the main separation across metals. The second component explains an additional 15.1%, increasing the cumulative explained variation to 81.3%, and has a distinct interpretation, loading mainly on distributional features, entropy, and spikiness. It therefore distinguishes metals with similar persistence but different distributional properties, which would be concealed in a one-dimensional representation. Given that the analysis includes only 16 metals, the resulting instance space should be interpreted as descriptive, and none of our

<sup>14</sup> A related supply-side characteristic is by-product production. Several metals in our sample, most clearly cobalt, but also, to varying degrees, molybdenum and vanadium, are recovered largely as by-products of a primary metal. Their supply may therefore depend more strongly on the production conditions of the host metal than on their own market conditions.



The right-hand panel shows the instance space of RV time series.  $PC_1$  and  $PC_2$  are the first two principal components, projected from the eight-dimensional feature space

inferential results depends on retaining exactly two components. As shown in Equation (14),  $PC_1$  increases with the ACF features and decreases with the other features.

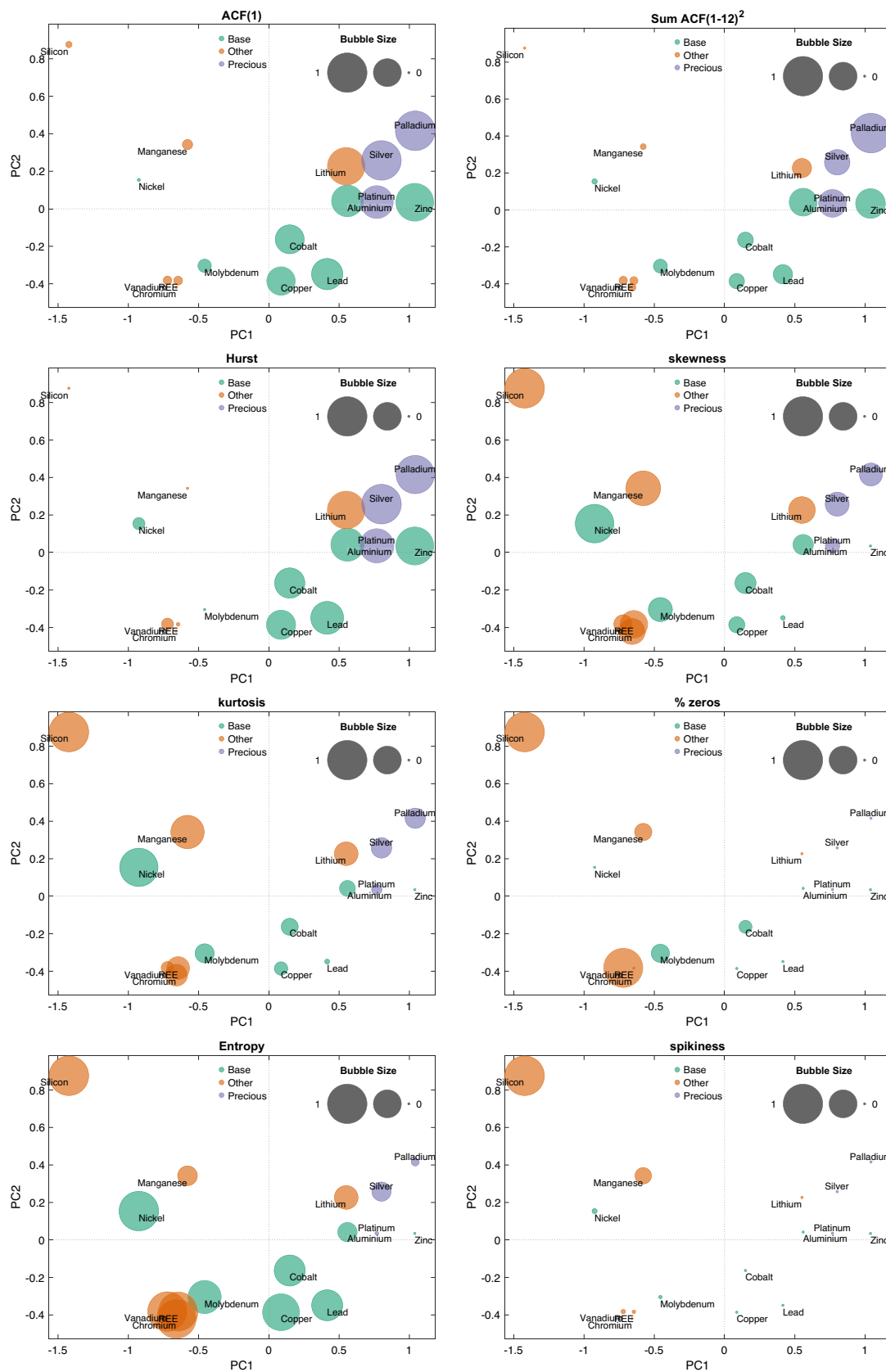
$$\begin{bmatrix} PC_1 \\ PC_2 \end{bmatrix} = \begin{bmatrix} 0.48 & 0.30 & 0.50 & -0.28 & -0.27 & -0.26 & -0.43 & -0.17 \\ 0.21 & 0.19 & 0.18 & 0.46 & 0.53 & 0.24 & -0.37 & 0.45 \end{bmatrix} F \quad (14)$$

The left panel of Fig. 4 reports the  $R^2$  values from regressions of each of the eight features separately on  $PC_1$  and  $PC_2$ . The resulting values show that  $PC_1$  is primarily associated with the ACF features, whereas  $PC_2$  is more strongly associated with distributional features, spectral entropy, and spikiness.

The scatter plot in the right-hand panel of Fig. 4 provides further interesting information. Focusing on the horizontal axis, we can see that  $PC_1$  clearly separates the series into two clusters, the first consisting mainly of precious and base metals and the second mainly of other metals.  $PC_2$  further separates the series and shows that the RV of silicon has distinctive features compared to other metals.

To better understand the relationship between the characteristics of the series and their division into clusters, Fig. 5 shows both the positioning of the ETMs and the distribution of the characteristic values in the instance space. The size of the bubbles in the graphs is proportional to the value of the features. The first three graphs show that all series associated with positive values of the first principal component, namely those in the right part of the instance space, have high values of the ACF features.

As far as skewness, kurtosis and entropy are concerned, the graphs in panels 4, 5 and 7 show that nickel is associated with values of these features that are much higher than those of most other base metals, which explains why it tends to be far from them in the instance space. Furthermore, we can see in Fig. 5 that nickel is associated with very low values



**Fig. 5** Instance space and feature values. *Notes:* each panel shows the instance space of RV time series.  $PC_1$  and  $PC_2$  are the first two principal components, projected from the eight-dimensional feature space.

In each panel, the size of the dots is proportional to the value of the feature in the title. Features are normalized in the (0,1) interval

for the ACF features. As for silicon, we can see that it is associated with low values of ACF features and high values of distributional and other features, especially spikiness, which explains why it tends to appear as an outlier in the top-left corner of the instance space.

## Evaluation of volatility models and forecasts

### In-sample model comparison

All models included in the in-sample comparison, including the GARCH(1,1) benchmark, are estimated within a common Bayesian framework. The marginal likelihood reported for GARCH is therefore the integrated likelihood obtained under the GARCH prior and is not the likelihood evaluated at the maximum-likelihood estimates. The in-sample performance of the  $m$ -th volatility model,  $\mathcal{M}_m$ , can be assessed focusing on its marginal likelihood (ML), defined as:

$$p(r|\mathcal{M}_m) = \int p(r|\theta_m, \mathcal{M}_m) p(\theta_m|\mathcal{M}_m) d\theta_m \quad (15)$$

where  $p(r|\theta_m, \mathcal{M}_m)$  is the likelihood function and the prior density  $p(\theta_m|\mathcal{M}_m)$ , both of which depend on a model-specific vector of parameters,  $\theta_m$ , of size  $p_m$ . The ML captures the fact that if data, collected in the vector of returns  $r$ , are likely under model  $m$ ,  $p(r|\mathcal{M}_m)$  should be large. In other terms, the ML is akin to a density forecast of the returns under the  $m$ -th model, evaluated at the observed data. Moreover, we note that since the ML depends on the vector of parameters  $\theta_m$ , it discounts model complexity from fit.

Denoting the GARCH(1,1) model as  $\mathcal{M}_0$  we can compare it against SV specifications using the Bayes factor defined as:  $BF_{m,0} = \frac{p(r|\mathcal{M}_m)}{p(r|\mathcal{M}_0)}$  for  $m = 1, \dots, 5$ . Therefore, observed returns are more likely under the  $m$ -th SV specification, compared to the GARCH(1,1) model,  $\mathcal{M}_0$ , if  $BF_{m,0} > 1$ . As shown in Kass and Raftery (1995), the log  $BF$  is asymptotically equivalent to the Schwarz information criterion; moreover, both are consistent model selection criteria (i.e. they can asymptotically recover the model with the correct structure with probability one).

Table 2 shows the log ML across metals and models.<sup>15</sup> As we can see from the first column of the table, silver is the only metal for which the log ML provides evidence in favor of the GARCH(1,1) against SV specifications. This is

also clear from the left panel of Fig. 6, which plots (twice the log of) the BF of the best model (i.e. the one associated with the largest or second-largest log ML) against the GARCH(1,1) model. Silver is associated with a negative value of log BF, shown as a red dot, indicating that there is no evidence to support any of the SV specifications against a simple GARCH(1,1) model.

Conversely, the dots representing the other metals are on a blue scale, indicating positive values of log BF. We can also see that the further we move away from the horizontal zero line, the greater the log BF (darker dots). Recall that in Section 3.2 we showed that  $PC_2$  captures distributional features, spikiness, and zero returns, while  $PC_1$  captures mostly ACF features. Therefore, as we move away from the horizontal zero line, the contribution of  $PC_2$  to the variance of the data increases. The relative performance of GARCH(1,1) tends to weaken as the contribution of  $PC_2$  increases, while the greater flexibility of the SV specifications is associated with higher in-sample fit for several of these metals.

The fourth column of Table 2 highlights that the SV(2) model provides the best fit for most metals. This is also clear from the right panel of Fig. 6, which plots (twice the log of) the BF of the best model (i.e. the one associated with the largest or second-largest log ML) against the SV(2) model. Note that most dots are red and hence indicate a negative log BF, which represent evidence in favor of the SV(2) model.

However, the plot also highlights metals for which there is no evidence in favor of the SV(2) model: aluminium, nickel, silver, lithium, and REE. For aluminium, nickel, silver, and lithium, the dots are in light shades of blue to indicate that  $\log BF \in (0, 2]$  and hence there is only weak evidence in favor of alternative specifications of the SV relative to the SV(2) model. Note that the dots associated with these metals lie near the horizontal zero line in the instance space, thus suggesting an increasing contribution of  $PC_2$  to explaining the variance of the data. Only for REE there is strong evidence that an alternative SV model provides gains over the SV(2). Interestingly, looking at Table 2, we can see that for nickel, lithium, and REE, the SV(1)- $t$  specification performs well. While these metals lie in different regions of the instance space, Fig. 3 shows that they exhibit strong values for entropy, skewness, and kurtosis.

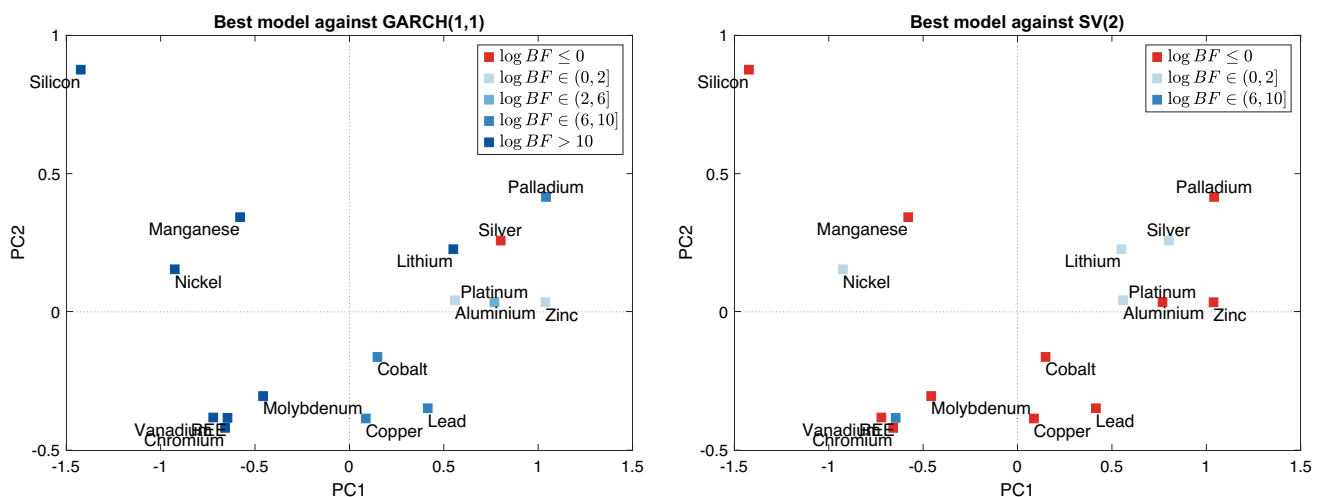
Overall, the results in this section point to the SV(2) as the model that best accommodates the heterogeneous features of ETM data. The presence of a second autoregressive log-volatility term seems, therefore, to capture the set of complex patterns that characterizes the ETM time series. The heterogeneous performance of different SV models can be interpreted in light of the underlying features of the metals, as shown in Fig. 3. Across most metals, the SV(2) model performs well, particularly for those characterized by a combination of autocorrelation (high ACF) and

<sup>15</sup> While the main analysis focuses on SV specifications, Section S2.4 of the SI extends the in-sample comparison beyond the baseline GARCH(1,1) by considering GARCH variants that mirror key features of the corresponding SV models, including heavy-tailed innovations, jumps and leverage effects.

**Table 2** Log marginal likelihood of the volatility models

| Commodity  | GARCH(1,1)              | SV(1)            | SV(2)                   | SV(1)-J          | SV(1)-t                 | SV(1)-L                 |
|------------|-------------------------|------------------|-------------------------|------------------|-------------------------|-------------------------|
| Aluminium  | -366.2<br>(0.03)        | -366.5<br>(0.11) | -366.1<br>(0.03)        | -367.2<br>(0.11) | -366.3<br>(0.01)        | <b>-365.9</b><br>(0.02) |
| Cobalt     | -443.2<br>(0.02)        | -445.1<br>(0.31) | <b>-438.2</b><br>(0.02) | -445.8<br>(0.19) | -440.2<br>(0.03)        | -443.9<br>(0.02)        |
| Copper     | -374.8<br>(0.03)        | -372.9<br>(0.08) | <b>-370.3</b><br>(0.01) | -372.6<br>(0.11) | -371.4<br>(0.02)        | -373.2<br>(0.02)        |
| Lead       | -376.8<br>(0.11)        | -376.6<br>(0.10) | <b>-373.6</b><br>(0.01) | -377.0<br>(0.09) | -376.7<br>(0.01)        | -376.9<br>(0.02)        |
| Molybdenum | -464.7<br>(0.07)        | -444.8<br>(0.28) | <b>-427.3</b><br>(0.06) | -447.3<br>(0.42) | -427.5<br>(0.02)        | -439.1<br>(0.08)        |
| Nickel     | -438.4<br>(0.04)        | -434.0<br>(0.17) | -432.0<br>(0.02)        | -434.3<br>(0.26) | <b>-431.7</b><br>(0.01) | -432.0<br>(0.01)        |
| Zinc       | -400.7<br>(0.02)        | -401.3<br>(0.09) | <b>-399.8</b><br>(0.02) | -401.4<br>(0.14) | -400.3<br>(0.01)        | -401.2<br>(0.01)        |
| Silver     | <b>-416.8</b><br>(0.02) | -420.3<br>(0.10) | -418.6<br>(0.02)        | -420.5<br>(0.12) | -418.4<br>(0.01)        | -419.3<br>(0.01)        |
| Palladium  | -391.2<br>(0.08)        | -388.0<br>(0.11) | <b>-386.7</b><br>(0.02) | -388.5<br>(0.10) | -387.2<br>(0.01)        | -387.6<br>(0.01)        |
| Platinum   | -410.4<br>(0.03)        | -409.1<br>(0.12) | <b>-407.4</b><br>(0.03) | -409.6<br>(0.18) | -407.9<br>(0.01)        | -408.8<br>(0.01)        |
| Chromium   | -369.9<br>(0.08)        | -363.2<br>(0.22) | <b>-359.0</b><br>(0.02) | -363.0<br>(0.30) | -360.0<br>(0.06)        | -363.3<br>(0.02)        |
| Lithium    | -369.0<br>(0.12)        | -325.7<br>(0.13) | -323.9<br>(0.05)        | -325.8<br>(0.16) | <b>-323.6</b><br>(0.07) | -325.4<br>(0.03)        |
| Manganese  | -453.3<br>(0.04)        | -436.3<br>(0.28) | <b>-424.5</b><br>(0.04) | -434.4<br>(0.33) | -428.0<br>(0.03)        | -432.5<br>(0.02)        |
| REE        | -417.4<br>(0.02)        | -362.5<br>(0.20) | -362.2<br>(0.49)        | -363.7<br>(0.25) | <b>-358.6</b><br>(0.05) | -359.9<br>(0.08)        |
| Silicon    | -449.9<br>(0.08)        | -346.0<br>(0.03) | <b>-328.2</b><br>(0.22) | -349.2<br>(0.43) | -340.7<br>(0.16)        | -343.4<br>(0.17)        |
| Vanadium   | -482.7<br>(0.05)        | -397.7<br>(0.36) | <b>-374.1</b><br>(0.26) | -399.3<br>(0.36) | -387.0<br>(0.17)        | -390.3<br>(0.22)        |

Notes: values in bold represent the best model associated with the highest marginal likelihood. Numerical standard errors are given in brackets



**Fig. 6** Instance space and Bayes Factors. Notes: the colour of the dots is based on the scale given by Kass and Raftery (1995) for interpreting twice the natural logarithm of the Bayes factor. A negative value (red dot) provides no evidence in favour of the best model against the GARCH(1,1) or the SV(2) model in the left and right panels, respec-

tively. Dots on a blue scale indicate positive values and provide evidence (weak, positive, strong, very strong) in favour of the best model compared to one of the two benchmarks. See notes to Fig. 5 for further details

distributional complexity (e.g., skewness and kurtosis). Its additional autoregressive component enables it to better capture persistent volatility patterns. However, in some cases, alternative specifications yield better results. The SV(1)-t model ranks highly for nickel, lithium, and REEs, which are characterized by relatively high entropy, skewness, and kurtosis.

For aluminium, the SV(1)-L model records the highest marginal likelihood, consistent with the possibility that allowing for asymmetric volatility responses improves in-sample fit. In this sense, features such as kurtosis, skewness, and entropy provide useful descriptive context for the differences in model rankings observed across metals.

### Point forecasts

As outlined in Section 2.3, models are recursively estimated<sup>16</sup> to produce a total of 60 one-step-ahead forecasts from January 2018 to December 2022. Point volatility forecasts are evaluated against two volatility proxies – RV and  $RG2^*$  – using two alternative loss functions: the RMSFE and the QLIKE. Table 3 focuses on RV as a volatility proxy, while results for  $RG2^*$  appear in Section S2.2 of the SI. We implement the test for equal predictive ability of Diebold and Mariano (1995), as modified by Coroneo and Iacone (2020) to improve its performance in small samples (DM test, henceforth). The null hypothesis of the test is that the observed difference in forecasting performance of SV specifications against the GARCH(1,1) model is statistically indistinguishable from zero.

Each entry in Table 3 represents (if multiplied by 100) the percentage change in loss when moving from a GARCH(1,1) forecast to one based on an SV specification. For instance, in the case of aluminium, the SV(1) forecast results in a 1.76% reduction in RMSFE and a 28.43% decrease in QLIKE loss.

The large difference between these figures is explained by the fact that, while the RMSFE is a symmetric loss function, the QLIKE loss penalizes underforecasting more heavily than overforecasting. In both cases, the table displays an asterisk indicating that, according to the DM test, we can reject the null hypothesis that the observed difference in forecasting performance between the SV(1) model and the GARCH(1,1) model is statistically indistinguishable from zero.

Although there is considerable heterogeneity in the accuracy of SV forecasts across different metal categories, some common patterns emerge. Looking at Table 3,

<sup>16</sup> A rolling window approach, where forecasts are generated using a sample of 66 observations that is moved forward one period each time we generate a new forecast, yields qualitatively similar results. See Section S2.7 of the SI.

**Table 3** Evaluation of volatility point forecasts using RV as proxy

|                                           | SV(1)            | SV(2)           | SV(1)-J   | SV(1)-t        | SV(1)-L          |
|-------------------------------------------|------------------|-----------------|-----------|----------------|------------------|
| (a) Root Mean Squared Forecast Error loss |                  |                 |           |                |                  |
| Aluminium                                 | -0.0176*         | 0.0587          | -0.0101   | -0.0078        | <b>-0.0182</b>   |
| Cobalt                                    | 0.1293           | 0.1400          | 0.1393    | 0.1617         | 0.1322           |
| Copper                                    | <b>-0.0094</b>   | 0.0335          | 0.0204    | 0.0259         | -0.0053          |
| Lead                                      | -0.0719**        | -0.0345**       | -0.0515** | -0.0500**      | <b>-0.0777**</b> |
| Molybdenum                                | 1.8386           | 0.2244          | 0.0313    | 0.0466         | 0.0637           |
| Nickel                                    | -0.0026          | 0.0011          | -0.0023   | -0.0004        | <b>-0.0033*</b>  |
| Zinc                                      | 0.0020           | 0.0889          | 0.0233    | 0.0363         | 0.0127           |
| Silver                                    | 0.0421           | 0.1054          | 0.0451    | 0.0511         | 0.0463           |
| Palladium                                 | <b>-0.0416**</b> | 0.0215          | -0.0153** | -0.0122**      | -0.0355**        |
| Platinum                                  | -0.0930**        | -0.0173         | -0.0809*  | -0.0679**      | <b>-0.0967**</b> |
| Chromium                                  | 0.0885           | -               | -0.0291   | -0.0012        | 0.1446           |
|                                           |                  | <b>0.0370**</b> |           |                |                  |
| Lithium                                   | 0.4439           | -0.1019         | 0.1646    | <b>-0.1581</b> | 0.5451           |
| Manganese                                 | -0.0014          | -0.0126         | -0.0041   | <b>-0.0133</b> | -0.0070          |
| REE                                       | 0.3698           | 0.2324          | 0.2060    | 0.1928         | 0.2025           |
| Silicon                                   | 0.2381           | 0.0257          | 0.2818    | 0.7451         | <b>-0.0063</b>   |
| Vanadium                                  | 0.4387           | 0.2529          | 0.5309    | 0.5881         | 0.0398           |
| (b) QLIKE loss                            |                  |                 |           |                |                  |
| Aluminium                                 | -0.2843**        | -0.0112         | -0.2274*  | -0.2296*       | <b>-0.3181**</b> |
| Cobalt                                    | 0.0738           | 0.0546          | 0.1140    | 0.2164         | 0.0920           |
| Copper                                    | 0.0186           | 0.2493          | 0.2599    | 0.2843         | 0.0834           |
| Lead                                      | -0.3279**        | -0.1866**       | -0.2398** | -0.2358**      | <b>-0.3402**</b> |
| Molybdenum                                | 0.1134           | 0.2386          | 0.7838    | 0.7711         | 0.2323           |
| Nickel                                    | 0.0063           | 0.0159          | 0.0353    | 0.0716         | <b>-0.1338</b>   |
| Zinc                                      | 0.0998           | 0.2252          | 0.2082    | 0.2545         | 0.2248           |
| Silver                                    | <b>-0.0739</b>   | 0.5476          | -0.0409   | 0.0082         | 0.0181           |
| Palladium                                 | <b>-0.2430**</b> | 0.0856          | -0.1187** | -0.1151*       | -0.2076**        |
| Platinum                                  | <b>-0.3906**</b> | -0.2063         | -0.3477** | -0.3203**      | -0.3562**        |
| Chromium                                  | 0.1021           | <b>-0.0399</b>  | 0.0263    | 0.2800         | 0.1114           |
| Lithium                                   | 0.2764           | -0.0119         | 0.0279    | <b>-0.0585</b> | 0.1684           |
| Manganese                                 | 0.0799           | 0.5965          | 0.2557    | 0.5985         | 0.1164           |
| REE                                       | -0.1586          | <b>-0.1911</b>  | 0.2420    | 0.3806         | -0.1279          |
| Silicon                                   | -0.1849          | 0.2518          | -0.0963   | <b>-0.2449</b> | 0.4818           |
| Vanadium                                  | 1.5085           | 0.6045          | 1.4553    | 1.5907         | 3.4599           |

Notes: each entry in panel (a) of the table shows  $(RMSFE_j/RMSFE_0) - 1$ , where  $RMSFE_0$  is the root mean squared forecast error of the GARCH(1, 1) forecast, and  $RMSFE_j$  corresponds to one of the SV specifications:  $j = SV(1), SV(2), SV(1) - J, SV(1) - t, SV(1) - L$ . In panel (b) the loss function is the QLIKE. A negative value indicates that the GARCH(1, 1) forecast is less accurate than that of the corresponding SV specification. Asterisks denote the rejection of the null hypothesis of equal predictive ability in the Diebold-Mariano test, with \* indicating significance at the 10% level and \*\* at the 5% level. The values in bold represent the best model

we can see that in a total of 160 pairwise comparisons with GARCH(1,1) forecasts, SV specifications yield a reduction in the loss function in 41.3% of cases, and we can reject the null of equal predictive accuracy in 20% of cases.

The reductions in the loss function associated with the SV models vary across mineral classes and specifications. Focusing on base metals, we observe that SV forecasts are more accurate than GARCH(1,1) forecasts for aluminium, lead, copper, and nickel. However, for the latter two metals,

the reductions in forecast errors are often statistically indistinguishable from zero. For the precious metals, the SV specifications lead to an improvement in forecast accuracy for palladium and platinum, but not for silver. This is in line with the model comparison in Section 4.1 where we showed that silver was the only metal for which none of the SV specifications performed well against the GARCH(1,1) model. Turning to the “other ETMs”, we observe reductions in the loss function associated with the SV models in 20 out of 60 pairwise comparisons, but only in one case can we reject the null hypothesis of the DM test.

As shown in Table 3 and Fig. 7, which also incorporates results based on the RG2\* volatility proxy, the SV(1)-J forecasts are rarely more accurate than those of the GARCH(1,1) model. On the contrary, the specification that most often appears to offer the largest percentage change loss reduction compared with GARCH(1,1) forecasts is the SV(1)-L model.

A closer look at Fig. 7, where in each panel we show the instance space and highlight the number of times that the model yields the lowest loss for a given metal, reveals that often the most accurate forecasts are either those from SV(1) and SV(2) models or those from the SV(1)-L. This pattern is confirmed across different classes of metals and seems to be related both to ACF and distributional features.

Three main conclusions can be drawn from the analysis in this section. First, contrary to the in-sample model comparison in Section 4.1, where the SV(2) model appeared to provide the best fit to the ETM data, there seems to be no clear winner in the case of point volatility forecasts. This, in turn, suggests that forecast combination might be a suitable approach for improving point volatility forecasts. Second, the SV(1)-L specification relatively often records the lowest point-forecast loss, despite rarely outperforming SV(2) in the in-sample comparison. This further illustrates that in-sample model rankings do not necessarily translate into superior out-of-sample forecasting performance. Third, there are no clear patterns linking point-forecast accuracy to the prevalence of specific features, and the SV(1)-J model is only rarely selected as the best-performing specification.

## Density forecasts

For density forecasts in Table 4, we focus on weighted versions of the qCRPS that emphasize either the center (Panel *a*) or both tails of the distribution of returns (Panel *b*). The structure and interpretation of the entries in the table are the same as those in Table 3. Inspection of Table 4 reveals that, in a total of 160 pairwise comparisons with GARCH(1,1) forecasts, SV specifications yield a reduction in the loss function in 43.8% of cases, and the null of equal predictive accuracy can be rejected in 23.1% of cases. Within this

implementation, SV(2) most frequently records the lowest qCRPS. This ranking should nevertheless be interpreted cautiously because parameter uncertainty is treated differently across the GARCH and SV forecasting procedures and the evaluation sample contains only 60 forecasts.

In Fig. 8, we report the forecasting procedure with the lowest qCRPS for each metal represented in the instance space. The location of metals in the instance space appears to be associated with the best-performing procedure: SV(2) and GARCH(1,1) are frequently selected for metals near the horizontal zero line, whereas SV(1)-L and SV(1)-t are more often selected for metals farther from it.

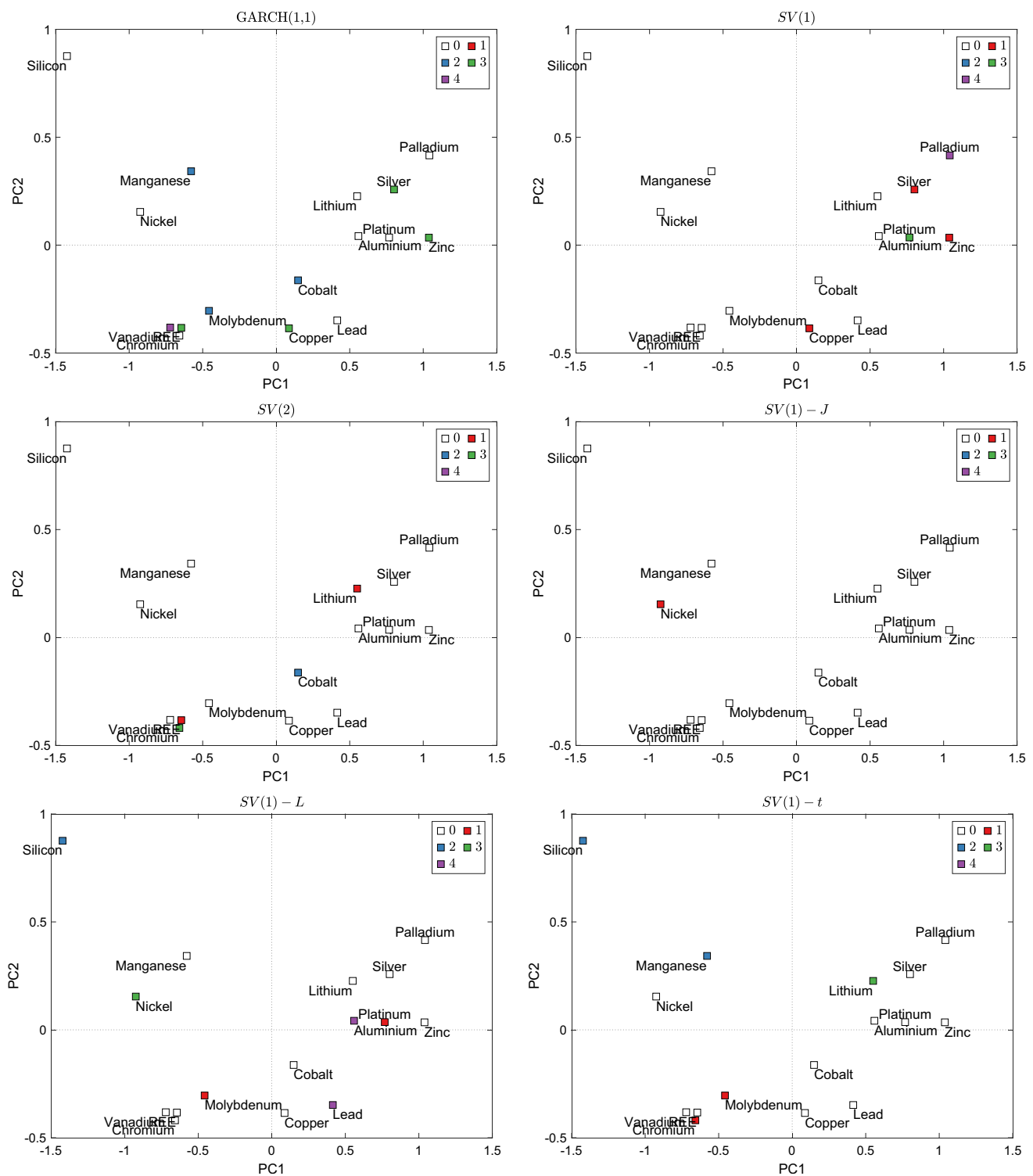
The contrast between the strong in-sample fit of SV(2) and its more limited out-of-sample advantage appears to be sensitive to a small number of turbulent months. As shown in Section S2.3 of the SI, when the point- and density-forecast evaluations exclude the most volatile periods of 2020 and 2022, the relative ranking of SV(2) improves. In the density-forecast comparison, it is the best-performing SV specification for 10 of the 16 metals, compared with 7 in the full sample. This result indicates that the rankings are sensitive to the composition of the short evaluation sample. Although the pattern is consistent with abrupt price movements being particularly difficult for specifications based primarily on smooth volatility persistence, it should not be interpreted as definitive evidence of a systematic relationship between particular metal features and model performance.

Taken together, our forecasting results suggest that combining forecasts from different models or adding extra features may be necessary to improve volatility forecasts in the metal market. One example would be to explicitly account for zero returns when estimating the models (Francq and Sucarrat 2022). We leave these topics for future research.

## Robustness checks and extensions

### Model confidence set

In Sections 4.2 and 4.3, we compared the predictive performance of each model against the benchmark GARCH(1,1) using pairwise Diebold–Mariano tests. As an alternative, we apply the Model Confidence Set (MCS) procedure of Hansen et al. (2011), which jointly evaluates the full set of competing models, accounts for multiple comparisons, and identifies the set of statistically superior forecasts. The idea is to sequentially eliminate the worst-performing models and retain a set of models that are statistically indistinguishable. Due to space constraints, technical details and full results are reported in Section S2.5 of the SI. The MCS results confirm that SV models consistently outperform GARCH(1,1) under both point-forecast and density-forecast



**Fig. 7** Instance space and volatility point forecasts. *Notes:* the colours of the dots indicate the number of times that the model in the title yields the lowest loss for a given metal. For each model we consider

2 loss functions and 2 volatility proxies, so a model can be selected a maximum of four times for each metal. See notes to Fig. 5 for further details

**Table 4** Quantile-based continuous ranked probability score (qCRPS)

|                                                 | SV(1)   | SV(2)            | SV(1)-J          | SV(1)-t         | SV(1)-L          |
|-------------------------------------------------|---------|------------------|------------------|-----------------|------------------|
| (a) Weighted qCRPS ( $\widehat{wQS}$ ) - center |         |                  |                  |                 |                  |
| Aluminium                                       | 0.0334  | <b>-0.0129</b>   | 0.0201           | 0.0243          | 0.0216           |
| Cobalt                                          | 0.0099  | <b>-0.0801**</b> | -0.0085          | -0.0738**       | 0.0138           |
| Copper                                          | 0.0062  | -0.0255**        | <b>-0.0305**</b> | -0.0274**       | -0.0030          |
| Lead                                            | 0.0600  | 0.0320           | 0.0428           | 0.0468          | 0.0635           |
| Molybdenum                                      | 0.4082  | -0.0961**        | -0.2229**        | -               | -0.1393**        |
| Nickel                                          | 0.0206  | <b>-0.0291</b>   | 0.0092           | -0.0030         | 0.0232           |
| Zinc                                            | -0.0044 | <b>-0.0405**</b> | -0.0153          | -0.0167         | -0.0049          |
| Silver                                          | -0.0139 | <b>-0.0786**</b> | -0.0203          | -0.0238         | -0.0227          |
| Palladium                                       | 0.0412  | <b>-0.0167*</b>  | 0.0123           | 0.0157          | 0.0331           |
| Platinum                                        | 0.0468  | 0.0102           | 0.0275           | 0.0222          | 0.0524           |
| Chromium                                        | 0.0389  | -0.0479**        | -0.0739**        | -               | 0.0511           |
|                                                 |         |                  |                  | <b>0.1342**</b> |                  |
| Lithium                                         | 0.1783  | -0.0032          | 0.0700           | 0.0008          | 0.1572           |
| Manganese                                       | 0.0211  | <b>-0.1966**</b> | -0.0294          | -0.1735**       | 0.0134           |
| REE                                             | 0.3375  | 0.2987           | 0.2216           | 0.2238          | 0.2579           |
| Silicon                                         | 0.4779  | 0.0042           | 0.3462           | 0.6023          | <b>-0.1098</b>   |
| Vanadium                                        | 0.0676  | -0.0402          | 0.1310           | 0.1865          | <b>-0.2583**</b> |
| (b) Weighted qCRPS ( $\widehat{wQS}$ ) - tails  |         |                  |                  |                 |                  |
| Aluminium                                       | 0.0527  | -0.0256          | 0.0313           | 0.0412          | 0.0368           |
| Cobalt                                          | 0.0124  | <b>-0.1192**</b> | -0.0174          | -0.1095*        | 0.0174           |
| Copper                                          | 0.0078  | -0.0358**        | <b>-0.0454**</b> | -0.0369**       | -0.0060          |
| Lead                                            | 0.0924  | 0.0502           | 0.0657           | 0.0757          | 0.0972           |
| Molybdenum                                      | 0.4914  | -0.1242**        | -0.2808**        | -               | -0.1797**        |
| Nickel                                          | 0.0344  | -0.0507          | 0.0166           | -0.0009         | 0.0366           |
| Zinc                                            | -0.0059 | <b>-0.0635**</b> | -0.0232          | -0.0223         | -0.0079          |
| Silver                                          | -0.0193 | <b>-0.1177**</b> | -0.0293          | -0.0313         | -0.0367          |
| Palladium                                       | 0.0595  | <b>-0.0255**</b> | 0.0170           | 0.0251          | 0.0474           |
| Platinum                                        | 0.0759  | 0.0214           | 0.0477           | 0.0435          | 0.0828           |
| Chromium                                        | 0.0501  | -0.0726**        | -0.1184**        | -               | 0.0638           |
|                                                 |         |                  |                  | <b>0.1869**</b> |                  |
| Lithium                                         | 0.2839  | -0.0087          | 0.1056           | 0.0024          | 0.2442           |
| Manganese                                       | 0.0327  | <b>-0.2738**</b> | -0.0480          | -0.2351**       | 0.0271           |
| REE                                             | 0.4621  | 0.4042           | 0.3061           | 0.3102          | 0.3572           |
| Silicon                                         | 0.5506  | 0.0155           | 0.4001           | 0.7161          | <b>-0.1096</b>   |
| Vanadium                                        | 0.0900  | -0.0544          | 0.1779           | 0.2543          | <b>-0.3246**</b> |

Notes: each entry of the table shows  $(\widehat{wQS}_j/\widehat{wQS}_0) - 1$ , where  $\widehat{wQS}_0$  is the weighted qCRPS the  $GARCH(1, 1)$  forecast, and  $\widehat{wQS}_j$  corresponds to one of the SV specifications:  $j = SV(1), SV(2), SV(1) - J, SV(1) - t, SV(1) - L$ . In panel (a) the weighted qCRPS emphasizes the center of the distribution, while in panel (b) the loss function emphasizes the tails of the distribution. A negative value indicates that the  $GARCH(1, 1)$  forecast is less accurate than that of the corresponding SV specification. Asterisks denote the rejection of the null hypothesis of equal predictive ability in the Diebold-Mariano test, with \* indicating significance at the 10% level and \*\* at the 5% level. The values in bold represent the best model

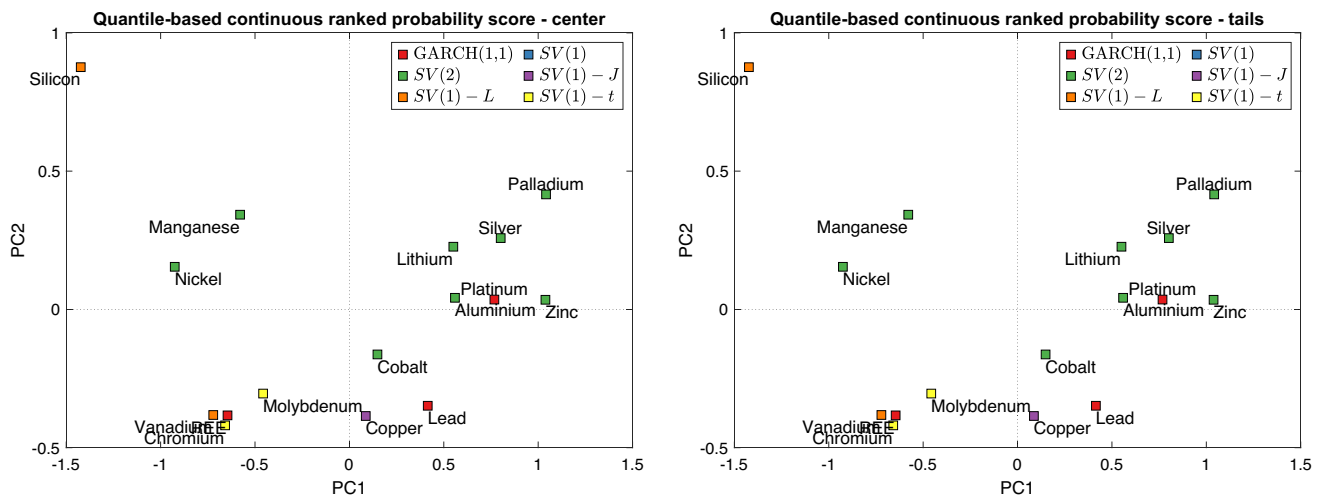
losses. For point forecasts, SV(1) and SV(1)-L are the most frequently retained specifications, while  $GARCH(1, 1)$  is excluded for 7 out of 16 commodities under squared-error loss and 6 under QLIKE loss. For density forecasts, SV(2) and SV(1)-t emerge as the most frequently retained models, with  $GARCH(1, 1)$  excluded in 11 out of 16 commodities under the centre-weighted specification.

## Robustness checks

We conduct several robustness checks to assess the sensitivity of our main findings. The full results are reported in the SI.

**Comparing extended GARCH and SV specifications** Section S2.4 reports the comparison of log marginal likelihoods for a broader set of GARCH specifications ( $GARCH$ ,  $GARCH-2$ ,  $GARCH-J$ ,  $GARCH-t$  and  $GARCH-GJR$ ) and their SV counterparts. The results confirm that SV models, especially SV(2), provide the best fit for the majority of metals. A few exceptions (cobalt, chromium and silver) occur where  $GARCH$  augmented with heavy-tailed or jump features becomes competitive. This is consistent with Chan and Grant (2016): once a  $GARCH$  model is augmented with Student- $t$  innovations or a jump component, it becomes considerably more capable of accommodating heavy tails and isolated extreme observations. In such cases, these extensions provide less additional benefit to SV models, whose latent volatility process can already accommodate large shocks through endogenous increases in volatility. Consequently, the performance gap between the two classes of models narrows substantially.

**Sensitivity to the Volatility Prior.** Section S2.6 examines the sensitivity of the marginal likelihood comparisons to the prior scale of the volatility-of-volatility parameter,  $S_h$ , which directly controls the prior flexibility of the latent volatility process. We vary  $S_h$  over a grid from one quarter to four times the baseline value. The best model remains unchanged for most commodities across the entire grid. The only exceptions occur at the smallest prior scale, where  $GARCH$  occasionally emerges (Zinc) or SV(1)-t is favoured (Chromium, Lithium); at the largest scale, SV(2) can displace SV(1)-L (Aluminium). A small  $S_h$  restricts movements in the latent volatility, so outliers must be absorbed by heavy-tailed innovations, favouring SV(1)-t; a large  $S_h$  permits volatility to move freely, which benefits the richer



**Fig. 8** Instance space and density forecasts. *Notes:* The colours of the dots indicate which model delivers the lowest loss. The loss function is given in the title. See notes to Fig. 5 for further details

volatility dynamics of SV(2) and renders heavy tails partly redundant.

**Alternative volatility proxy.** Section S2.2 reports the point-forecast evaluation using  $RG2^*$  as an alternative volatility proxy. The ranking of models remains broadly unchanged, although the statistical evidence against GARCH(1,1) becomes weaker: the Diebold-Mariano test rejects equal predictive ability in only 14 of 160 comparisons (about 9%), compared with 20% under the baseline RV measure. The choice of proxy therefore affects the strength of the evidence against GARCH(1,1) but not the substance of the model comparison.

**Sensitivity to extreme volatility.** In Section S2.3, we re-estimate the forecast comparisons on a restricted sample that excludes the most volatile months of 2020 and 2022 (54 of the 60 forecast observations). While the point-forecast rankings are qualitatively similar, SV(2) strengthens its relative performance for density forecasts once these turbulent episodes are removed. Nevertheless, the exercise does not identify a uniformly dominant specification and should primarily be interpreted as evidence that the rankings are sensitive to the composition of the short evaluation sample.

**Rolling window estimation.** Finally, in Section S2.7 of the SI, we re-evaluate the point and density forecasts using a 66-month rolling window approach. Relative to the recursive scheme, the rolling window amplifies the predictive gains of SV models over GARCH(1,1): the share of cells where SV beats GARCH rises from 41% to 51% for point forecasts, and reaches 95 out of 160 cases for density forecasts. The improvements extend to lead and platinum, where GARCH was previously competitive. These results suggest

that the rolling window is more sensitive to time-varying volatility dynamics, thereby better capturing the flexibility of SV specifications.

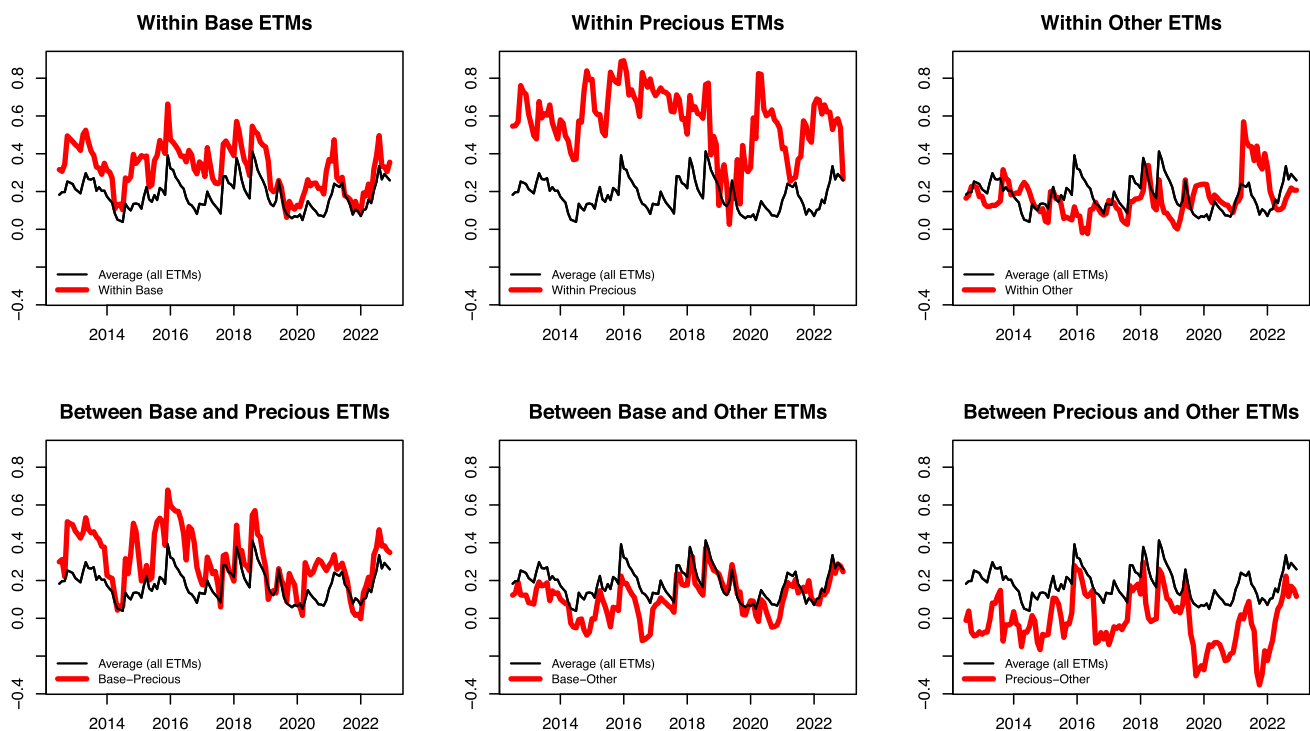
## A look at the dynamic correlation of ETMs

While the primary goal of this work is to analyze univariate models for volatility and examine how the features of each series relate to the models' in-sample and out-of-sample performance, our findings also carry implications for the analysis of correlations across different series. Such analysis is essential for a wide range of applications, including portfolio allocation and hedging strategies. A comprehensive investigation of these correlation dynamics, however, is beyond the scope of this study. As such, the discussion that follows is intentionally kept simple and aims only to offer an intuitive illustration of the potential implications of our results.<sup>17</sup> Accordingly, in this section we provide a brief examination of the dynamic correlations among the series using a simple multivariate model based on the RiskMetrics framework, which is widely used as a benchmark in the literature see e.g., (Engle, 2009).

To examine the correlation dynamics between ETMs, we compute the time-varying covariance matrix  $H_t$  as follows:

$$H_t = (1 - \lambda)H_{t-1} + \lambda r_{t-1}r_{t-1}^\top \quad (16)$$

<sup>17</sup> For example, one could adopt a two-stage approach, as is common in models of dynamic conditional correlation or copula-based frameworks see e.g., (Engle, 2009; Oh and Patton, 2017), where the variance of each series is first modeled individually using the most appropriate specification, and the correlation structure is then estimated using the chosen multivariate method.



**Fig. 9** Dynamics of average correlations. *Notes:* The black thin lines depict the dynamic average correlations across all ETMs, while the colored thick lines illustrate the average correlations within and between predefined ETM groups over time

where  $H_t$  represents the covariance matrix at time  $t$ ,  $\lambda$  is the decay rate which controls how much weight is placed on more recent observations, and  $r_t$  denotes the vector of real returns on ETMs at time  $t$ . The dynamic conditional correlation matrix  $R_t$  is given by<sup>18</sup>:

$$R_t = D_t^{-1} H_t D_t^{-1} \quad \text{where} \quad D_t^2 = \text{diag}[H_t]$$

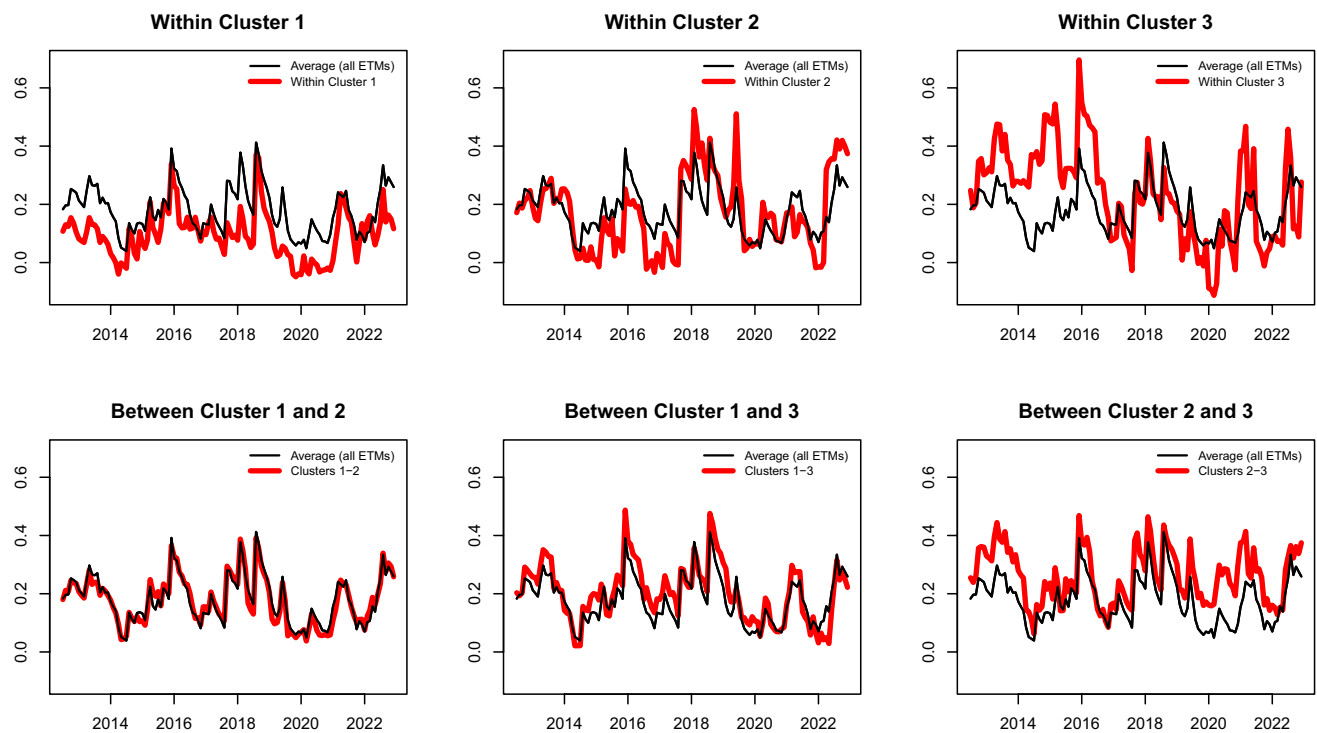
Figure 9 displays the dynamics of average correlations among ETMs, both within and between predefined groups. Precious ETMs exhibit persistently high within-group correlations (top center panel), consistently exceeding the overall average (black line), reflecting their shared features—such as strong autocorrelation and low spikiness (see Fig. 3)—and their dual role as industrial inputs and investment assets. By contrast, base (top left) and other ETMs (top right) show lower and more volatile within-group correlations, suggesting more heterogeneous price behaviors and weaker common dynamics.

The bottom panels highlight correlations between groups. Base and precious ETMs (bottom left) maintain

relatively strong positive correlations, especially before 2019. In contrast, correlations between precious and other ETMs (bottom right) are notably weaker and occasionally negative, indicating divergent price dynamics. The base–other ETM correlation (bottom center) lies between these extremes, remaining relatively stable but below the overall average. Across all panels, a marked decline in correlations is observed in 2020, particularly within and between base and precious ETMs, suggesting that COVID-19 disrupted typical co-movement patterns. This breakdown suggests that a strong systemic shock, such as the pandemic, can temporarily break the usual co-movements among ETMs: a system-wide disruption hits all these markets at once but through heterogeneous channels, so the correlations that normally link the metals weaken. In other words, the historical correlations that diversification and hedging rely on can break down just when they are needed most, so systemic risk should be considered as a common factor that cannot be diversified away.

To move beyond the limits of application-based groupings, we reclassify the ETMs using feature-based hierarchical clustering see, e.g., (Fulcher, 2018; Wang et al., 2006). The clustering is performed on the same eight normalized features used in the PCA, so each metal is represented by its vector of feature values scaled to the (0, 1) interval. We compute pairwise distances between metals using Euclidean distances and construct the hierarchical tree using Ward's

<sup>18</sup> The initial covariance matrix  $H_0$  is calculated as the sample covariance based on all the observations. This pre-sample estimate serves as the starting value for the exponential smoothing recursion. For our monthly data, we choose a decay factor of  $\lambda = 0.2$ , to ensure that it is responsive enough to capture meaningful shifts in ETM correlations over time, while still providing sufficient smoothing to avoid excessive noise from month-to-month variation.



**Fig. 10** Dynamics of the average correlation within clustered groups. *Notes:* The black thin lines show the dynamic average correlations across all ETMs, while the colored thick lines represent the dynamic average correlations within each ETM group and between ETM groups

minimum-variance linkage. The dendrogram is then partitioned into three clusters. Thus, the clustering is based entirely on similarities in the empirical features of the series and does not use the predefined base, precious, or other-metal classifications. The resulting clusters are as follows:

- Cluster 1: Chromium, manganese, molybdenum, nickel, REE, silicon, vanadium, characterized by weak autocorrelation and stronger distributional differences.
- Cluster 2: Aluminium, lithium, palladium, platinum, silver, zinc—exhibiting strong autocorrelation but smoother and more stable price behavior.
- Cluster 3: Cobalt, copper, lead—showing moderate autocorrelation and volatility features.

Figure 10 illustrates the dynamics of within-cluster average correlations. Cluster 1 shows consistently low within-group correlations, consistent with the highly heterogeneous features of the underlying series. Cluster 2's correlation pattern aligns closely with the total average, suggesting a more coherent group structure. Cluster 3 initially displayed high within-group correlations—likely due to shared external shocks—but this declined after 2017 as price trends diverged. Between-cluster correlations remain similar to the overall average, indicating relatively weak interdependence across the clusters.

These results highlight a key distinction between predefined groupings, which reflect functional uses (e.g., industrial vs. investment roles), and feature-based clusters, which capture intrinsic statistical similarities in price behavior. While predefined groups like precious ETMs show strong internal coherence, the feature-based clusters offer a more nuanced partitioning that aligns more closely with the empirical dynamics of volatility, autocorrelation, and distributional traits. This data-driven approach improves the interpretation of correlation structures and may better inform applications such as risk management, diversification strategies, and forecasting models.

## Conclusions

Some ETMs are traded in relatively liquid and established markets, whereas others have limited price histories and infrequent price changes. Modeling and forecasting their volatility is relevant for risk management in markets that play an important role in the energy and digital transitions.

Our analysis, which combines exploratory data analysis, data-reduction techniques, and visualization methods, reveals substantial heterogeneity in ETM return and volatility patterns. Although data providers often classify metals into broad homogeneous groups, and despite geological co-occurrence among some metals, the empirical features

of the individual series differ markedly. Our in-sample and out-of-sample evaluations also show that neither GARCH models nor any single SV specification consistently dominates across metals.

Some associations emerge between the characteristics of the series and forecasting performance, but they differ across evaluation criteria. Precious metals such as palladium, platinum, and silver exhibit relatively strong autocorrelation, and models emphasizing volatility persistence often rank favorably for these metals. For other metals, the rankings of the more flexible SV specifications vary substantially across loss functions and between point and density forecasts. In particular, the point-forecast results reveal no clear relationship between individual features and the best-performing specification, and the jump model is only rarely selected. The evidence therefore supports heterogeneity in model performance rather than a systematic mapping between particular data features and model classes.

The differences documented across metals may be consistent with variation in market maturity. Liquidity, the development of representative price benchmarks, trading frequency, transparency of price discovery, and the availability of derivatives and other hedging instruments may influence measured volatility and predictability. These mechanisms are not directly tested in our analysis, and this interpretation should therefore be regarded as suggestive rather than causal. In particular, our findings provide evidence of heterogeneity in volatility dynamics and forecast performance across metals relevant to the energy transition, rather than evidence of an effect of rising transition-related demand on volatility or predictability.

The results also show that the specification favored by the in-sample comparison does not necessarily produce the most accurate out-of-sample forecasts. For example, while the SV(1)-t model provides the best in-sample fit for rare earths, GARCH produces more accurate point forecasts. More generally, relative rankings vary across metals, loss functions, volatility proxies, and point and density forecasts. This reinforces the importance of evaluating alternative specifications rather than selecting a forecasting model exclusively on the basis of in-sample fit.

The heterogeneity we document implies that volatility modeling and risk management should be metal-specific rather than based exclusively on broad commodity groupings. Differences in liquidity, transparency, benchmark development, and hedging opportunities may also be relevant for understanding volatility and predictability across markets. Formally evaluating the effects of these institutional characteristics, however, lies beyond the scope of the present analysis.

This paper opens several directions for future research. One promising avenue is to combine forecasts from different

models and incorporate a broader set of volatility features. A larger panel of metal-price series would also make it possible to estimate formal feature-based model-selection or forecast-combination rules, in the spirit of the meta-learning approaches used in the forecasting literature. Another extension would be to further analyze the interconnectedness of ETMs using network-based approaches, building on recent evidence of state-dependent spillovers between ETMs and other commodity markets (Bastianin et al. 2026). This could reveal additional co-movement patterns and channels through which price shocks propagate across commodities, while connecting ETM volatility to the broader literature on sentiment, cross-market dependence, and hedging relationships across Bitcoin, crude oil, gold, and equity markets (Wang et al. 2024a).

Future work could also formally investigate whether market liquidity, benchmark development, and the availability of hedging instruments account for differences in volatility and predictability across metals. Finally, the relatively short sample remains an important limitation. The analysis is based on 126 monthly observations and 60 out-of-sample forecasts covering an unusually eventful period. Model rankings should therefore be interpreted cautiously. The robustness checks that exclude the most turbulent months and use a rolling estimation window indicate that the main qualitative findings are not driven solely by a few extreme observations. Nevertheless, these exercises cannot fully overcome the constraints that a short sample imposes on statistical power and on the stability of more demanding SV specifications.

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1007/s13563-026-00714-y>.

**Author Contributions** A.B. developed the core idea for the study. A.B., X.L., and L.S. all contributed to the coding, empirical analysis, and writing of the manuscript. All authors reviewed and approved the final version.

**Funding** Open access funding provided by Università degli Studi di Brescia within the CRUI-CARE. Agreement. Andrea Bastianin acknowledges financial support from the “Fund for Departments of Excellence” provided by the Ministero dell’Università e della Ricerca (MUR), established by the Stability Law (“Legge di Stabilità n. 232/2016, 2017”), within the projects of the Department of Economics, Management, and Quantitative Methods. Andrea Bastianin and Luqman Shamsudin acknowledge funding from the GREESCO (“Green Specialization and Circularity: Constraints and Opportunities”), funded by the European Union Next-GenerationEU, NATIONAL RECOVERY AND RESILIENCE PLAN (NRRP) – MISSION 4 COMPONENT 2 INVESTMENT 1.1, Fondo per il Programma Nazionale di Ricerca e Progetti di Rilevante Interesse Nazionale (PRIN)–CUP G53D23006810001.

**Data Availability** The data used in this study were obtained from Refinitiv Eikon under a subscription license and cannot be publicly shared.

## Declarations

**Competing interests** The authors declare no competing interests.

**Ethics approval and consent to participate** Not applicable

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Addison T, Ghoshray A (2023) Discerning trends in international metal prices in the presence of nonstationary volatility. *Resource Energy Econ* 71:101334
- Bakas D, Triantafyllou A (2019) Volatility forecasting in commodity markets using macro uncertainty. *Energ Econ* 81:79–94
- Bassetti F, Casarin R, Ravazzolo F (2020) Density forecasting. In: Fuleky P (ed) *Macroeconomic Forecasting in the Era of Big Data: Theory and Practice*. Springer International Publishing, pp 465–494
- Bastianin A, Casoli C, Galeotti MD (2023) The connectedness of energy transition metals. *Energ Econ* 128:107183
- Bastianin A, Casoli C, Kočenda E, Li X (2026) Extreme connectedness among energy transition metals and commodity markets. FEEM Working Paper 13-2026, Fondazione Eni Enrico Mattei
- Boer L, Pescatori A, Stuermer M (2024) Energy transition metals: bottleneck for net-zero emissions? *J Eur Econ Assoc* 22(1):200–229
- Bollerslev T (1986) Generalized autoregressive conditional heteroskedasticity. *J Economet* 31(3):307–327
- Bollerslev T, Hood B, Huss J, Pedersen LH (2018) Risk everywhere: Modeling and managing volatility. *Rev Financ Stud* 31(7):2729–2773
- Chan JC, Grant AL (2016) Modeling energy price dynamics: GARCH versus stochastic volatility. *Energy Econ* 54:182–189
- Ciner C, Lucey B, Yarovaya L (2020) Spillovers, integration and causality in lme non-ferrous metal markets. *J Commod Mark* 17:100079
- Clark TE, Ravazzolo F (2015) Macroeconomic forecasting performance under alternative specifications of time-varying volatility. *J Appl Economet* 30(4):551–575
- Coroneo L, Iacone F (2020) Comparing predictive accuracy in small samples using fixed-smoothing asymptotics. *J Appl Economet* 35(4):391–409
- Corsi F (2009) A simple approximate long-memory model of realized volatility. *J Financ Economet* 7(2):174–196
- Creal D, Koopman SJ, Lucas A (2013) Generalized autoregressive score models with applications. *J Appl Economet* 28(5):777–795
- Diebold FX, Mariano RS (1995) Comparing predictive accuracy. *J Bus Econ Stat* 253–263
- Diemer A, Iammarino S, Perkins R, Gros A (2022) Technology, resources and geography in a paradigm shift: the case of critical and conflict materials in icts. *Region Stud* 1–13
- Dinh T, Goutte S, Nguyen DK, Walther T (2022) Economic drivers of volatility and correlation in precious metal markets. *J Commod Mark* 28:100242
- Engle R (2009) *Anticipating correlations: a new paradigm for risk management*. Princeton University Press
- European Commission (2020). *Critical raw materials resilience: Charting a path towards greater security and sustainability*. COM(2020) 474
- Fally T, Sayre J (2018) *Commodity trade matters*. Technical report, National Bureau of Economic Research
- Francq C, Sucarrat G (2022) Volatility estimation when the zero-process is nonstationary. *J Bus Econ Stat* 41(1):53–66
- Fulcher BD (2018) Feature-based time-series analysis. In: *Feature engineering for machine learning and data analytics*. CRC press, pp 87–116
- Gneiting T, Ranjan R (2011) Comparing density forecasts using threshold-and quantile-weighted scoring rules. *J Bus Econ Stat* 29(3):411–422
- Graedel TE, Harper EM, Nassar NT, Nuss P, Reck BK (2015) Criticality of metals and metalloids. *Proc Natl Acad Sci* 112(14):4257–4262
- Han X, Liu Z, Wang S (2022) An R-vine copula analysis of non-ferrous metal futures with application in value-at-risk forecasting. *J Commod Mark* 25:100188
- Hansen PR, Lunde A (2005) A forecast comparison of volatility models: does anything beat a GARCH(1,1)? *J Appl Economet* 20(7):873–889
- Hansen PR, Lunde A, Nason JM (2011) The model confidence set. *Econometrica* 79(2):453–497
- Hyndman RJ, Athanasopoulos G (2018) *Forecasting: principles and practice*. OTexts Melbourne
- Hyndman R, Kang Y, Montero-Manso P, O'Hara-Wild M, Talagala T, Wang E, Yang Y (2023) tsfeatures: Time Series Feature Extraction. <https://pkg.robjhyndman.com/tsfeatures/>, <https://github.com/robjhyndman/tsfeatures>
- IEA (2023) *Critical minerals market review*. IEA, Paris. Available online at: <https://www.iea.org/reports/critical-minerals-market-review-2023>
- International Energy Agency (2021a) *Net Zero by 2050: A Roadmap for the Global Energy Sector*. International Energy Agency, Paris, France
- International Energy Agency (2021b) *The role of critical minerals in clean energy transitions*. World energy outlook special report, International Energy Agency, Paris, France
- Kang Y, Hyndman RJ, Smith-Miles K (2017) Visualising forecasting algorithm performance using time series instance spaces. *Int J Forecast* 33(2):345–358
- Kass RE, Raftery AE (1995) Bayes factors. *J Am Stat Assoc* 90(430):773–795
- Montero-Manso P, Athanasopoulos G, Hyndman RJ, Talagala TS (2020) FFORMA: Feature-based forecast model averaging. *Int J Forecast* 36(1):86–92
- Naeem MA, Hamouda F, Karim S (2024) Tail risk spillover effects in commodity markets: A comparative study of crisis periods. *J Commod Mark* 33:100370
- Oh DH, Patton AJ (2017) Modeling dependence in high dimensions with factor copulas. *J Bus Econ Stat* 35(1):139–154
- Patton AJ (2011) Volatility forecast comparison using imperfect volatility proxies. *J Economet* 160(1):246–256
- Pham L, Hsu K-C (2024) Metals of the future in a world in crisis: Geopolitical disruptions and the cleantech metal industry. *Energy Econ* 108004
- Pommeret A, Ricci F, Schubert K (2022) Critical raw materials for the energy transition. *Eur Econ Rev* 141:103991
- Saadaoui J, Smyth R, Vespignani J (2025) Ensuring the security of the clean energy transition: Examining the impact of geopolitical risk on the price of critical minerals. *Energy Econ* 142:108195

- Schwert GW (1989) Why does stock market volatility change over time? *J Financ* 44(5):1115–1153
- Talagala TS, Hyndman RJ, Athanasopoulos G (2023) Meta-learning how to forecast time series. *J Forecast* 42(6):1476–1501
- Wang X, Smith K, Hyndman R (2006) Characteristic-based clustering for time series data. *Data Min Knowl Disc* 13:335–364
- Wang G, Liu C, Sbai E, Sheng MS, Hu J, Tao M (2024a) Interrelations between Bitcoin market sentiment, crude oil, gold, and the stock market with Bitcoin prices: Vision from the hedging market. *Stud Econ Financ* 41(5):1166–1190
- Wang G, Sbai E, Wen L, Sheng MS (2024b) The impact of renewable energy on extreme volatility in wholesale electricity prices: Evidence from Organisation for Economic Co-operation and Development countries. *J Clean Prod* 484:144343
- Wang G, Do HX, Liu C, Sbai E, Silva E (2026) Shifting roles of renewable and fossil energy in the ENTSO-E countries: Evidence from a novel war-induced energy intensity index. *Energy Policy* 210:115083
- World Bank (2020) Minerals for Climate Action: The Mineral Intensity of the Clean Energy Transition. World Bank, Washington, D.C

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.