

Forecasting the Volatility of Energy Transition Metals

Supplementary Information

Contents

S1 Estimation Details	S1
S1.1 Model specifications and prior distributions	S1
S1.2 Posterior Simulation	S2
S2 Additional results and robustness checks	S6
S2.1 Data sources, real prices, and realized volatility	S6
S2.2 Alternative volatility proxy	S8
S2.3 Forecast evaluation excluding extreme periods	S9
S2.4 In-sample comparison: GARCH versus stochastic volatility	S12
S2.5 Model confidence sets	S15
S2.6 Prior sensitivity	S18
S2.7 Rolling window estimation	S19
S2.8 Principal Component Analysis: further results	S22

S1 Estimation Details

We adopt the estimation steps and settings of Chan and Grant (2016) and present them below.

S1.1 Model specifications and prior distributions

Table S1: Model specifications and prior distributions.

Panel A: Model specifications

Model	Specification
GARCH(1, 1)	$y_t = \mu + \varepsilon_t$, $\varepsilon_t \sim \mathcal{N}(0, \sigma_t^2)$, $\sigma_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \beta_1 \sigma_{t-1}^2$
SV (baseline)	$y_t = \mu + \varepsilon_t^y$, $\varepsilon_t^y \sim \mathcal{N}(0, e^{h_t})$; $h_t = \mu_h + \phi_h(h_{t-1} - \mu_h) + \varepsilon_t^h$, $\varepsilon_t^h \sim \mathcal{N}(0, \omega_h^2)$
SV-2	Baseline with AR(2) log-volatility: $h_t = \mu_h + \phi_h(h_{t-1} - \mu_h) + \rho_h(h_{t-2} - \mu_h) + \varepsilon_t^h$
SV-J	Baseline with mean jumps: $y_t = \mu + k_t q_t + \varepsilon_t^y$, $q_t \sim \text{Bernoulli}(\kappa)$, $k_t \sim \mathcal{N}(\mu_k, \sigma_k^2)$
SV-t	Baseline with Student's t errors: $\varepsilon_t^y \sim t_\nu(0, e^{h_t})$
SV-L	Baseline with leverage: $\text{corr}(\varepsilon_t^y, \varepsilon_t^h) = \rho$

Panel B: Prior distributions and hyperparameter values

Parameter	Prior	Hyperparameters
<i>GARCH(1, 1)</i>		
μ	$\mathcal{N}(\mu_0, V_\mu)$	$\mu_0 = 0$, $V_\mu = 10$
$\gamma = (\alpha_0, \alpha_1, \beta_1)'$	$\log \gamma \sim \mathcal{N}(\gamma_0, V_\gamma) \mathbb{1}(\alpha_1 + \beta_1 < 1)$	$\gamma_0 = (1, \log 0.1, \log 0.8)'$, $V_\gamma = \text{diag}(10, 1, 1)$
<i>Common SV parameters (all SV models)</i>		
μ	$\mathcal{N}(\mu_0, V_\mu)$	$\mu_0 = 0$, $V_\mu = 10$
μ_h	$\mathcal{N}(\mu_{h0}, V_{\mu h})$	$\mu_{h0} = 1$, $V_{\mu h} = 10$
ϕ_h	$\mathcal{N}(\phi_{h0}, V_{\phi h}) \mathbb{1}(\phi_h < 1)$	$\phi_{h0} = 0.97$, $V_{\phi h} = 0.1^2$
ω_h^2	$\mathcal{IG}(\nu_h, S_h)$	$\nu_h = 5$, $S_h = 0.16$
<i>SV-2 (replaces the ϕ_h prior)</i>		
$\theta_h = (\phi_h, \rho_h)'$	$\mathcal{N}(\theta_{h0}, V_{\theta h}) \mathbb{1}(\theta_h \in A)$	$\theta_{h0} = (0.97, 0)'$, $V_{\theta h} = \text{diag}(0.1^2, 1)$
<i>SV-J (adds jump component)</i>		
κ	$\mathcal{U}(0, 0.1)$	—
$\delta = (\mu_k, \log \sigma_k^2)'$	$\mathcal{N}(\delta_0, V_\delta)$	$\delta_0 = (0, \log 10)'$, $V_\delta = \text{diag}(10, 1)$
<i>SV-t (adds t innovations)</i>		
ν	$\mathcal{U}(2, 100)$	—
<i>SV-L (adds leverage)</i>		
ρ	$\mathcal{U}(-1, 1)$	—

Notes: $\mathcal{IG}$ and $\mathcal{U}$ denote the inverse-gamma and uniform distributions. In Panel A, each SV variant modifies only the stated component of the baseline SV model. In Panel B, the SV-2 prior on θ_h replaces the common-SV prior on ϕ_h , with A the stationarity region of the AR(2) characteristic polynomial.

S1.2 Posterior Simulation

GARCH model. While the out-of-sample forecasting exercise relies on maximum-likelihood estimates of the GARCH(1,1) model, the in-sample comparison based on marginal likelihoods uses Bayesian estimation, since this provides a convenient framework for computing the marginal likelihood. For the GARCH(1,1) model we group the parameters into two blocks, μ and $\gamma = (\alpha_0, \alpha_1, \beta_1)'$, and sample sequentially from the full conditionals $p(\mu \mid \mathbf{y}, \gamma)$

and $p(\gamma \mid \mathbf{y}, \mu)$. Neither is of standard form—in particular μ is not Gaussian because it enters the conditional variance σ_t^2 recursively—so each block is updated with a Metropolis–Hastings (MH) step: a Gaussian proposal with mean $\bar{y}$ and variance s^2/T for μ , and a Gaussian proposal centred at the mode of $p(\gamma \mid \mathbf{y}, \mu)$ with covariance equal to the inverse negative Hessian at the mode for γ .

SV models: sampling scheme. The SV models are estimated with Markov chain Monte Carlo (MCMC). Collecting the parameters in $\boldsymbol{\theta} = (\mu, \mu_h, \phi_h, \omega_h^2)'$, posterior draws of $\boldsymbol{\theta}$ and the latent log-volatilities $\mathbf{h} = (h_1, \dots, h_T)'$ are obtained by sweeping through the full conditionals

1. $p(\mathbf{h} \mid \mathbf{y}, \mu, \mu_h, \phi_h, \omega_h^2)$;
2. $p(\mu \mid \mathbf{y}, \mathbf{h})$;
3. $p(\mu_h \mid \mathbf{h}, \phi_h, \omega_h^2)$;
4. $p(\omega_h^2 \mid \mathbf{h}, \mu_h, \phi_h)$;
5. $p(\phi_h \mid \mathbf{h}, \mu_h, \omega_h^2)$.

Step 1: the log-volatility vector $\mathbf{h}$. The conditional $p(\mathbf{h} \mid \mathbf{y}, \boldsymbol{\theta}) \propto p(\mathbf{y} \mid \mu, \mathbf{h}) p(\mathbf{h} \mid \mu_h, \phi_h, \omega_h^2)$ is high-dimensional and non-Gaussian, so $\mathbf{h}$ is sampled in one block with the acceptance–rejection Metropolis–Hastings (AR-MH) algorithm of Chan (2017). The proposal is the banded Gaussian

$$\mathbf{h} \sim \mathcal{N}(\hat{\mathbf{h}}, \mathbf{K}_h^{-1}), \quad (1)$$

characterised by its precision matrix $\mathbf{K}_h$ and mean $\hat{\mathbf{h}}$, defined as follows.

The precision is

$$\mathbf{K}_h = \mathbf{H}'_{\phi_h} \boldsymbol{\Sigma}_h^{-1} \mathbf{H}_{\phi_h} + \mathbf{G}.$$

The first term is the precision of the AR(1) prior on $\mathbf{h}$, in which $\mathbf{H}_{\phi_h}$ is the lower-bidiagonal matrix with ones on the diagonal and $-\phi_h$ on the first subdiagonal, and $\boldsymbol{\Sigma}_h = \text{diag}(\omega_h^2/(1 - \phi_h^2), \omega_h^2, \dots, \omega_h^2)$ holds the state-innovation variances. The second term $\mathbf{G} = \text{diag}(G_1, \dots, G_T)$

carries the curvature of the observation density, evaluated at an expansion point $\tilde{\mathbf{h}}$, with $G_t = \frac{1}{2}e^{-\tilde{h}_t}(y_t - \mu)^2$.

The mean is $\hat{\mathbf{h}} = \mathbf{K}_h^{-1}\mathbf{k}_h$, where

$$\mathbf{k}_h = \mathbf{f} + \mathbf{G}\tilde{\mathbf{h}} + \mathbf{H}'_{\phi_h}\boldsymbol{\Sigma}_h^{-1}\mathbf{H}_{\phi_h}\boldsymbol{\delta}_h,$$

the vector $\mathbf{f} = (f_1, \dots, f_T)'$ collects the observation scores $f_t = -\frac{1}{2} + \frac{1}{2}e^{-\tilde{h}_t}(y_t - \mu)^2$, and $\boldsymbol{\delta}_h = \mu_h\mathbf{1}_T$ is the AR(1) prior mean. The expansion point $\tilde{\mathbf{h}}$ is set to the mode by Newton–Raphson, iterating $\tilde{\mathbf{h}} \leftarrow \mathbf{K}_h^{-1}\mathbf{k}_h$ until convergence, so that at the fixed point $\tilde{\mathbf{h}} = \hat{\mathbf{h}}$. The vectors $\mathbf{f}$ and $\mathbf{G}$ are the only quantities that differ across the SV variants.

Because $\mathbf{K}_h$ is banded, candidate draws are obtained efficiently with the precision sampler of Chan and Jeliazkov (2009), and the AR-MH accept/reject step corrects the proposal so that the draws target the exact conditional.

Steps 2–5: the remaining parameters. Conditional on $\mathbf{h}$, each parameter is drawn from its own full conditional, summarised in Table S2; all are standard Gibbs draws except ϕ_h , which uses a Metropolis–Hastings step with a correction for the stationarity truncation.

Table S2: Full-conditional updates for the SV parameters, conditional on $\mathbf{h}$.

Parameter	Full conditional	Update
μ	$p(\mu \mid \mathbf{y}, \mathbf{h})$	Gibbs: $\mathcal{N}(\hat{\mu}, D_\mu)$
μ_h	$p(\mu_h \mid \mathbf{h}, \phi_h, \omega_h^2)$	Gibbs: $\mathcal{N}(\hat{\mu}_h, D_{\mu_h})$
ω_h^2	$p(\omega_h^2 \mid \mathbf{h}, \mu_h, \phi_h)$	Gibbs: $IG(\nu_h + \frac{T}{2}, \tilde{S}_h)$
ϕ_h	$p(\phi_h \mid \mathbf{h}, \mu_h, \omega_h^2)$	MH: $\mathcal{N}(\hat{\phi}_h, D_{\phi_h})\mathbb{I}(\phi_h < 1)$

The defining quantities are

$$\begin{aligned}
D_\mu^{-1} &= V_\mu^{-1} + \sum_{t=1}^T e^{-h_t}, & \hat{\mu} &= D_\mu \left(V_\mu^{-1} \mu_0 + \sum_{t=1}^T e^{-h_t} y_t \right), \\
D_{\mu_h}^{-1} &= V_{\mu_h}^{-1} + \mathbf{X}'_{\mu_h} \boldsymbol{\Sigma}_h^{-1} \mathbf{X}_{\mu_h}, & \hat{\mu}_h &= D_{\mu_h} \left(V_{\mu_h}^{-1} \mu_{h0} + \mathbf{X}'_{\mu_h} \boldsymbol{\Sigma}_h^{-1} \mathbf{z}_{\mu_h} \right), \\
D_{\phi_h}^{-1} &= V_{\phi_h}^{-1} + \mathbf{X}'_{\phi_h} \mathbf{X}_{\phi_h} / \omega_h^2, & \hat{\phi}_h &= D_{\phi_h} \left(V_{\phi_h}^{-1} \phi_{h0} + \mathbf{X}'_{\phi_h} \mathbf{z}_{\phi_h} / \omega_h^2 \right), \\
\tilde{S}_h &= S_h + \frac{1}{2} \sum_{t=1}^T \hat{\eta}_t^2,
\end{aligned}$$

with auxiliary vectors $\mathbf{X}_{\mu_h} = (1, 1 - \phi_h, \dots, 1 - \phi_h)'$, $\mathbf{z}_{\mu_h} = (h_1, h_2 - \phi_h h_1, \dots, h_T - \phi_h h_{T-1})'$, $\mathbf{X}_{\phi_h} = (h_1 - \mu_h, \dots, h_{T-1} - \mu_h)'$, $\mathbf{z}_{\phi_h} = (h_2 - \mu_h, \dots, h_T - \mu_h)'$, and AR(1) innovations $\hat{\eta}_1 = \sqrt{1 - \phi_h^2}(h_1 - \mu_h)$, $\hat{\eta}_t = h_t - \mu_h - \phi_h(h_{t-1} - \mu_h)$ for $t \geq 2$. The ϕ_h acceptance probability corrects for the stationary initialisation of h_1 .

Model-specific blocks. The variants modify Step 1 only through the gradient $\mathbf{f}$ and Hessian $\mathbf{G}$, and add their own parameter blocks: for SV-2 the pair (ϕ_h, ρ_h) is updated jointly subject to the stationarity region; for SV-J the jump indicators q_t , sizes k_t and parameters (κ, δ) are sampled in turn (after subtracting $q_t k_t$ from the observation); for SV-t the scale-mixing variables λ_t are sampled together with the degrees-of-freedom ν ; and for SV-L the leverage correlation ρ is updated with the observation and state innovations treated jointly, so that the banded structure of $\mathbf{K}_h$ is preserved. Full conditional distributions and sampling details are given in Chan and Grant (2014).

Marginal-likelihood computation. The marginal likelihoods of both the GARCH and SV models are computed using the adaptive importance-sampling approach adopted by Chan and Grant (2016), which is based on the improved cross-entropy method of Chan and Eisenstat (2015). For each model, the marginal likelihood integrates the likelihood over the complete parameter space with respect to the corresponding prior distribution. Thus, the GARCH entries reported in the model-comparison tables are Bayesian marginal likelihoods rather than maximized likelihood values. The same computational framework is applied to both model classes, and Monte Carlo standard errors are reported in parentheses.

Numerical settings and convergence. For each model we run 3 parallel chains from overdispersed starting values, each consisting of 20,000 draws with the first 2,000 discarded as burn-in. Convergence is monitored with the Gelman–Rubin potential scale reduction factor $\hat{R}$ (Gelman and Rubin, 1992) and the multi-chain effective sample size (ESS) (Geyer, 1992; Gelman et al., 2013), both computed across chains. The sampler mixes well (Table S3): across all metals and models the Gelman–Rubin statistic stays at or below 1.006, comfortably within our 1.01 threshold, while the multi-chain ESS runs from the hundreds into the thousands, with the lowest values close to 400 in only a handful of cases (typically

the leverage specification SV(1)-L, the least efficient of the models). This good behaviour is by design: because the banded Gaussian proposal in (1) is a second-order approximation of the conditional density about its mode, it closely matches the target, so candidate draws are accepted with high probability and the entire vector $\mathbf{h}$ is updated in a single block rather than element by element, which avoids the slow mixing of single-move samplers.

Table S3: MCMC convergence diagnostics: $\hat{R}$ and multi-chain ESS by model (3 chains $\times$ 20,000 draws)

	SV(1)		SV(2)		SV(1)-t		SV(1)-L		SV(1)-J		GARCH	
	$\hat{R}$	ESS	$\hat{R}$	ESS	$\hat{R}$	ESS	$\hat{R}$	ESS	$\hat{R}$	ESS	$\hat{R}$	ESS
Aluminium	1.001	6349	1.000	2332	1.001	6558	1.003	545	1.000	6328	1.002	924
Cobalt	1.001	1587	1.005	955	1.001	1337	1.006	411	1.002	1191	1.000	3859
Copper	1.001	4816	1.001	762	1.001	5298	1.003	549	1.001	5360	1.000	1529
Lead	1.000	6038	1.001	1371	1.000	5560	1.001	402	1.000	5392	1.003	366
Molybdenum	1.000	1521	1.000	1796	1.002	1111	1.005	603	1.001	430	1.002	666
Nickel	1.000	4209	1.003	1312	1.001	4567	1.006	357	1.001	4503	1.002	1200
Zinc	1.000	5541	1.000	2112	1.000	5133	1.002	654	1.000	4824	1.002	995
Silver	1.000	4928	1.000	4063	1.000	4165	1.002	721	1.001	4573	1.001	1455
Palladium	1.000	6291	1.000	394	1.000	5882	1.006	376	1.000	5352	1.003	546
Platinum	1.000	6272	1.000	697	1.001	6735	1.001	589	1.001	6442	1.002	1135
Chromium	1.001	1496	1.003	603	1.002	607	1.000	361	1.002	847	1.002	1661
Lithium	1.000	2572	1.001	2137	1.001	858	1.001	847	1.003	1484	1.001	1454
Manganese	1.001	1052	1.002	2436	1.002	1190	1.001	356	1.001	533	1.000	824
REE	1.000	2884	1.001	2212	1.001	1264	1.005	688	1.001	1348	1.000	3033
Silicon	1.000	2345	1.000	2568	1.001	2031	1.002	1257	1.003	539	1.002	1025
Vanadium	1.000	4529	1.001	4757	1.000	5062	1.001	1878	1.005	355	1.003	1254

Notes: $\hat{R}$ compares the between-chain with the within-chain variability and equals one when the chains have mixed to a common distribution. We treat $\hat{R} \leq 1.01$ as converged, a stricter standard than the traditional 1.1 (roughly a 1% versus 10% possible further shrinkage in the chains' spread). ESS is the number of independent draws the sample is effectively worth (we report the minimum across a model's parameters); at ESS = 400 the Monte Carlo error is about 5% of the posterior SD, so ESS < 400 reflects lower sampling efficiency rather than non-convergence.

S2 Additional results and robustness checks

S2.1 Data sources, real prices, and realized volatility

Figure S1: Real prices of energy transition metals

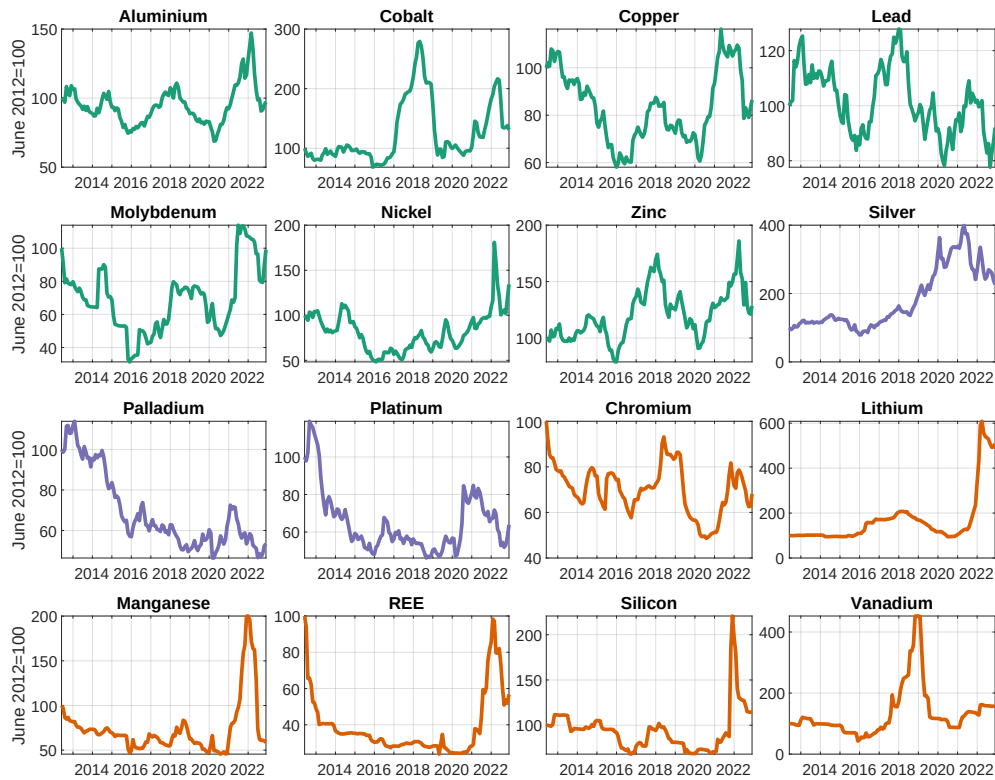


Figure S2: Realized volatility of energy transition metals

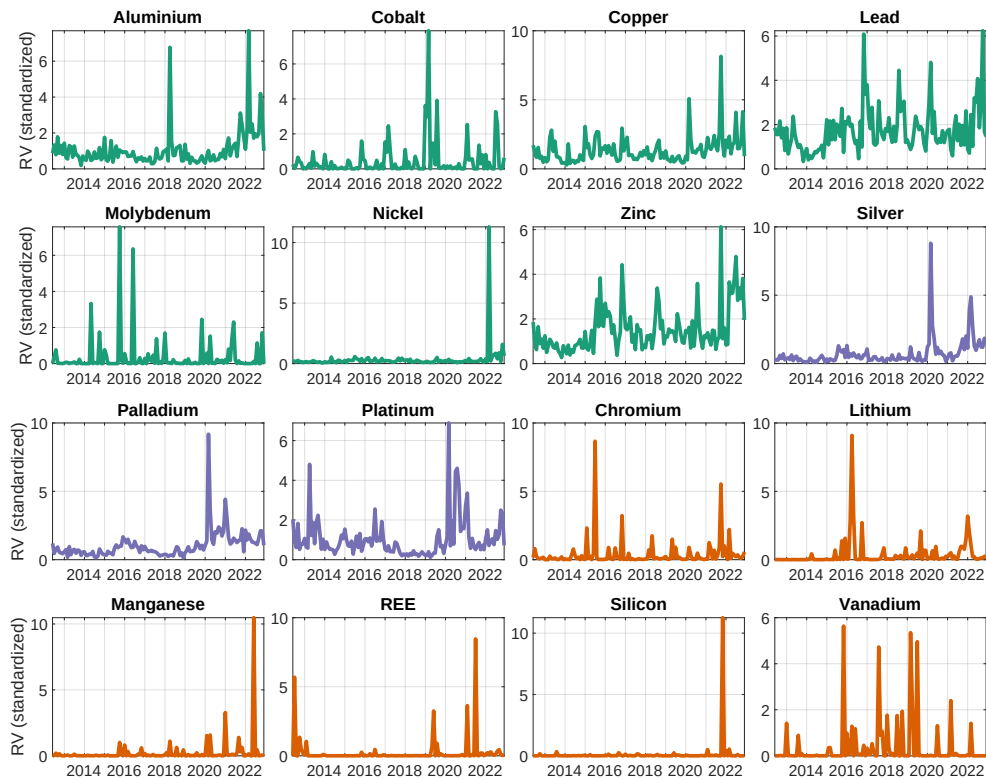


Table S4: Data Sources

Metal	Price specification	Unit	Source
Aluminium	Aluminium, 99.5% minimum purity, LME spot price, CIF UK ports	USD/mt	London Metal Exchange
Chromium	Chromium, #1 Chromium = 99.2%, 99A, Coarse Particle, Fine Particle	USD/mt	Shanghai Metal Exchange
Cobalt	Cobalt, minimum 99.80% purity, LME spot price	USD/mt	London Metal Exchange
Copper	Copper, grade A cathode, LME spot price, CIF European ports	USD/mt	London Metal Exchange
Lead	Lead, 99.97% pure, LME spot price, CIF European Ports	USD/mt	London Metal Exchange
Lithium	Lithium, 99% pure, industrial grade, battery grade	USD/mt	Shanghai Metal Exchange
Manganese	Manganese Electro CIF North West Europe	USD/mt	Refinitiv
Molybdenum	Molybdenum MO3, Insurance and Freight North West Europe	USD/mt	Refinitiv
Nickel	Nickel, melting grade, LME spot price, CIF European ports	USD/mt	London Metal Exchange
Palladium	Palladium, LME spot price	USD/troy ounce	London Metal Exchange
Platinum	Platinum, LME spot price	USD/troy ounce	London Metal Exchange
Rare Earth Elements	Rare earth carbonate REO 42-45% purity	USD/mt	Shanghai Metal Exchange
Silicon	Silicon lumps, CIF North West Europe	USD/mt	Refinitiv
Silver	Silver, London Bullion Market Association	USD/troy ounce	ICE Benchmark Administration
Vanadium	Vanadium pentoxide, CIF North West Europe	USD/mt	Refinitiv
Zinc	Zinc, minimum special high-grade zinc of 99.995% purity	USD/mt	London Metal Exchange

S2.2 Alternative volatility proxy

Table S5 reports the point-forecast evaluation using $RG2^*$, which is complementary to the RV-based results in the main paper. SV specifications reduce the loss relative to GARCH(1,1) in 74 of the 160 comparisons (about 46%, against 41.3% under RV), the largest reductions again come most often from SV(1)-L, notably for aluminium and lead under both loss functions, and SV(1)-J is rarely the most accurate. The cross-sectional ranking also carries over: gains are concentrated and statistically strongest for lead (significant under both losses) and, more modestly, for aluminium and palladium, whereas silver, copper, REE and vanadium continue to favour GARCH(1,1). The main difference relative to RV concerns significance rather than ordering: under $RG2^*$ the DM test rejects equal predictive ability in only 14 of 160 comparisons (about 9%, against 20% under RV), and the improvements for platinum lose significance. The choice of proxy therefore affects the strength of the evidence against GARCH(1,1) but not the substance of the model comparison.

Table S5: Evaluation of volatility point forecasts using RG2* as proxy

(a) Root Mean Squared Forecast Error loss					
	SV(1)	SV(2)	SV(1)-J	SV(1)-t	SV(1)-L
Aluminium	-0.0281	-0.0104	-0.0242	-0.0249	-0.0513**
Cobalt	-0.0802	-0.1756	-0.0878	-0.1450	-0.0801
Copper	0.0053	0.0124	0.0118	0.0157	0.0116
Lead	-0.0679**	-0.0322**	-0.0483**	-0.0469**	-0.0694**
Molybdenum	3.5362	0.4073	-0.1096	-0.1576*	0.0270
Nickel	-0.0043	0.0025	-0.0060	-0.0034	-0.0056
Zinc	-0.0210	0.0056	-0.0128	-0.0089	-0.0131
Silver	0.0173	0.0607	0.0191	0.0230	0.0219
Palladium	-0.0102*	0.0001	-0.0061	-0.0070	-0.0089
Platinum	-0.0190	0.0068	-0.0157	-0.0130	-0.0161
Chromium	0.1814	-0.0590*	-0.0535	-0.0894	0.2805
Lithium	0.3145	-0.0610	0.0250	-0.0091	0.1177
Manganese	-0.0066	-0.1294	-0.0347	-0.1355	-0.0396
REE	0.7648	0.5124	0.4491	0.4057	0.4397
Silicon	1.1209	0.1591	1.2934	2.7540	-0.0477
Vanadium	1.4447	0.7754	1.6972	1.9673	0.1086
(b) QLIKE loss					
	SV(1)	SV(2)	SV(1)-J	SV(1)-t	SV(1)-L
Aluminium	-0.2296	-0.1235	-0.1987	-0.2021	-0.2812
Cobalt	0.0595	-0.0020	0.0626	0.0530	0.0690
Copper	0.0230	0.0722	0.1050	0.1312	0.0502
Lead	-0.2691**	-0.1442**	-0.1976**	-0.1932**	-0.2752**
Molybdenum	0.1108	0.0386	0.1464	0.0711	-0.0122
Nickel	0.0088	0.0349	0.0249	0.0479	-0.1009
Zinc	0.0656	0.0438	0.1056	0.1174	0.0973
Silver	0.0084	0.3839	0.0267	0.0578	0.0718
Palladium	-0.1435	-0.0260	-0.0932	-0.1089	-0.1300
Platinum	-0.2071	-0.1520	-0.1884	-0.1826	-0.1669
Chromium	0.1194	-0.0304	0.0406	0.0957	0.1389
Lithium	0.4215	-0.0426	-0.0031	-0.0751	0.1075
Manganese	0.0663	0.2472	0.1248	0.1972	0.0857
REE	0.1500	0.0233	0.6243	0.6814	0.1215
Silicon	-0.0998	0.1572	-0.0543	-0.1611	0.3154
Vanadium	0.4617	0.1905	0.5087	0.5724	0.8414

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.

S2.3 Forecast evaluation excluding extreme periods

Tables S6 and S7 report the point- and density-forecast evaluations on the restricted sample that excludes the most volatile months of 2020 and 2022 (54 of the 60 forecasts). These results

underlie the discussion at the end of the point-forecast section in the main paper. For point forecasts there is again no single dominant specification, with the lowest loss most often delivered by SV(1) or SV(1)-L. For density forecasts, SV(2)'s relative standing strengthens once the turbulent months are removed: it is the best-performing SV specification for 10 of the 16 metals (center panel), compared with 7 in the full sample. Thus, removing these episodes improves the relative ranking of SV(2), but 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.

Table S6: Evaluation of volatility point forecasts using RV as proxy (subset sample)

(a) Root Mean Squared Forecast Error loss					
	SV(1)	SV(2)	SV(1)-J	SV(1)-t	SV(1)-L
Aluminium	-0.0348**	0.0541	-0.0256**	-0.0236*	-0.0190
Cobalt	0.1257	0.1399	0.1359	0.1611	0.1284
Copper	-0.0177	0.0381**	0.0144	0.0217*	-0.0110
Lead	-0.0780**	-0.0402**	-0.0562**	-0.0547**	-0.0846**
Molybdenum	1.7203**	0.0800*	0.0388	0.0518	0.0713
Nickel	-0.1417	0.0375	-0.1289	-0.0590	-0.1087
Zinc	-0.0057	0.1022	0.0168	0.0310	0.0013
Silver	0.0612	0.2586*	0.0706	0.0924	0.0713
Palladium	-0.0897**	0.0565**	-0.0295	-0.0189	-0.0753*
Platinum	-0.1505**	-0.0087	-0.1320**	-0.1085**	-0.1608*
Chromium	0.0915*	-0.0392*	-0.0223	0.0152	0.1476*
Lithium	0.4519**	-0.2398	0.1684	-0.1648	0.5543**
Manganese	-0.0018	-0.0128	-0.0045	-0.0136	-0.0077
REE	0.3529*	0.2285**	0.1959**	0.1886**	0.1960**
Silicon	0.0794	0.0274	0.2293	0.6089	-0.0045
Vanadium	0.4524**	0.2598	0.5414**	0.5567**	0.0438
(b) QLIKE loss					
	SV(1)	SV(2)	SV(1)-J	SV(1)-t	SV(1)-L
Aluminium	-0.3424**	-0.1349	-0.2852**	-0.2915**	-0.3587**
Cobalt	0.0651	0.0535	0.1077*	0.2220	0.0842
Copper	-0.0373	0.2945**	0.2081**	0.2383**	0.0660
Lead	-0.3411**	-0.2109**	-0.2502**	-0.2471**	-0.3536**
Molybdenum	0.1275**	0.2383	0.8912	0.8449*	0.2756
Nickel	-0.3070	0.0903	-0.2611	-0.1482	-0.3249
Zinc	0.0515	0.2641*	0.1589	0.2031	0.1265
Silver	-0.3532	0.0887	-0.3200	-0.2749	-0.2626
Palladium	-0.2984**	0.2177**	-0.1233*	-0.0956*	-0.2410**
Platinum	-0.4809**	-0.0700	-0.4419**	-0.3927**	-0.4664**
Chromium	0.1199	-0.0558	0.0505	0.3403	0.1192
Lithium	0.2768	-0.0173	0.0102	-0.0695	0.1486
Manganese	0.0619	0.6096	0.2482*	0.5771**	0.0889
REE	-0.1705	-0.1981	0.2392	0.3861	-0.1332
Silicon	-0.2849	0.2455	-0.3792	-0.3338	0.5007
Vanadium	1.5741	0.6830	1.6625	1.8070	2.6586

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 forecast sample excludes Jan/Feb/Mar 2020 and Feb/Mar/Apr 2022 (54 of 60 observations).

Table S7: Quantile-based continuous ranked probability score (qCRPS) – subset sample

(a) Weighted qCRPS ($\widehat{wQS}$) - center					
	SV(1)	SV(2)	SV(1)-J	SV(1)-t	SV(1)-L
Aluminium	0.0379**	-0.0107	0.0242*	0.0285**	0.0228
Cobalt	-0.0084	-0.0922**	-0.0277	-0.0883**	-0.0046
Copper	0.0125	-0.0255**	-0.0255*	-0.0238*	0.0021
Lead	0.0612**	0.0324**	0.0439**	0.0479**	0.0644**
Molybdenum	0.4014**	-0.1235**	-0.2181**	-0.2924**	-0.1311**
Nickel	0.0135	-0.0367	0.0038	-0.0087	0.0160
Zinc	-0.0026	-0.0444**	-0.0140	-0.0156	-0.0030
Silver	-0.0088	-0.0769*	-0.0160	-0.0196	-0.0193
Palladium	0.0456**	-0.0176*	0.0140*	0.0171**	0.0370**
Platinum	0.0503	0.0084	0.0302	0.0242	0.0569
Chromium	0.0421	-0.0510**	-0.0776**	-0.1354**	0.0560
Lithium	0.2054**	0.0155	0.0779	-0.0008	0.1763**
Manganese	0.0196	-0.2016**	-0.0351	-0.1845**	0.0081
REE	0.3216**	0.3065**	0.1995**	0.2203**	0.2537**
Silicon	0.3700**	0.0272	0.3141**	0.4659	-0.1081
Vanadium	0.1028	-0.0529	0.1598	0.2111*	-0.2415**
(b) Weighted qCRPS ($\widehat{wQS}$) - tails					
	SV(1)	SV(2)	SV(1)-J	SV(1)-t	SV(1)-L
Aluminium	0.0600**	-0.0196	0.0381*	0.0482**	0.0387
Cobalt	-0.0109	-0.1373**	-0.0427	-0.1316**	-0.0058
Copper	0.0166	-0.0359**	-0.0382*	-0.0314*	0.0008
Lead	0.0958**	0.0524**	0.0685**	0.0787**	0.1003**
Molybdenum	0.4860**	-0.1533**	-0.2801**	-0.3314**	-0.1728**
Nickel	0.0213	-0.0632	0.0063	-0.0102	0.0253
Zinc	-0.0032	-0.0689**	-0.0207	-0.0200	-0.0050
Silver	-0.0082	-0.1035*	-0.0188	-0.0207	-0.0266
Palladium	0.0642**	-0.0256*	0.0188*	0.0268**	0.0520**
Platinum	0.0801*	0.0176	0.0510	0.0459	0.0881*
Chromium	0.0512	-0.0775**	-0.1291**	-0.1961**	0.0679
Lithium	0.3244**	0.0207	0.1167	0.0019	0.2724**
Manganese	0.0308	-0.2757**	-0.0537	-0.2453**	0.0200
REE	0.4270**	0.4060**	0.2628**	0.2917**	0.3343**
Silicon	0.3960**	0.0440	0.3387*	0.5314	-0.1160
Vanadium	0.1294	-0.0748	0.2077	0.2777*	-0.3069**

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 forecast sample excludes Jan/Feb/Mar 2020 and Feb/Mar/Apr 2022 (54 of 60 observations).

S2.4 In-sample comparison: GARCH versus stochastic volatility

Table S8: Log marginal likelihoods: GARCH families vs. stochastic volatility models

Commodity	GARCH	GARCH-2	GARCH-J	GARCH-t	GARCH-GJR	SV(1)	SV(2)	SV(1)-J	SV(1)-t	SV(1)-L
Aluminium	-366.2 (0.03)	-366.2 (0.12)	-366.3 (0.03)	-366.2 (0.03)	-367.9 (0.19)	-366.5 (0.11)	-366.1 (0.03)	-367.2 (0.11)	-366.3 (0.01)	-365.9 (0.02)
Cobalt	-443.2 (0.02)	-443.7 (0.01)	-442.0 (0.07)	-437.6 (0.03)	-442.9 (0.02)	-445.1 (0.31)	-438.2 (0.02)	-445.8 (0.19)	-440.2 (0.03)	-443.9 (0.02)
Copper	-374.8 (0.03)	-374.8 (0.04)	-373.2 (0.10)	-372.8 (0.07)	-376.4 (0.04)	-372.9 (0.08)	-370.3 (0.01)	-372.6 (0.11)	-371.4 (0.02)	-373.2 (0.02)
Lead	-376.8 (0.11)	-376.8 (0.09)	-376.7 (0.08)	-377.3 (0.15)	-378.4 (0.13)	-376.6 (0.10)	-373.6 (0.01)	-377.0 (0.09)	-376.7 (0.01)	-376.9 (0.02)
Molybdenum	-464.7 (0.07)	-465.1 (0.10)	-446.3 (0.16)	-430.6 (0.02)	-466.8 (0.06)	-444.8 (0.28)	-427.3 (0.06)	-447.3 (0.42)	-427.5 (0.02)	-439.1 (0.08)
Nickel	-438.4 (0.04)	-438.9 (0.05)	-433.3 (0.06)	-432.1 (0.04)	-436.7 (0.07)	-434.0 (0.17)	-432.0 (0.02)	-434.3 (0.26)	-431.7 (0.01)	-432.0 (0.01)
Zinc	-400.7 (0.02)	-400.8 (0.05)	-400.6 (0.04)	-400.5 (0.04)	-401.8 (0.07)	-401.3 (0.09)	-399.8 (0.02)	-401.4 (0.14)	-400.3 (0.01)	-401.2 (0.01)
Silver	-416.8 (0.02)	-416.7 (0.03)	-416.7 (0.12)	-416.8 (0.04)	-418.0 (0.03)	-420.3 (0.10)	-418.6 (0.02)	-420.5 (0.12)	-418.4 (0.01)	-419.3 (0.01)
Palladium	-391.2 (0.08)	-391.2 (0.27)	-389.7 (0.04)	-390.0 (0.13)	-392.0 (0.13)	-388.0 (0.11)	-386.7 (0.02)	-388.5 (0.10)	-387.2 (0.01)	-387.6 (0.01)
Platinum	-410.4 (0.03)	-410.5 (0.09)	-409.9 (0.05)	-409.4 (0.12)	-411.3 (0.05)	-409.1 (0.12)	-407.4 (0.03)	-409.6 (0.18)	-407.9 (0.01)	-408.8 (0.01)
Chromium	-369.9 (0.08)	-370.4 (0.02)	-359.0 (0.10)	-356.6 (0.02)	-371.3 (0.04)	-363.2 (0.22)	-359.0 (0.02)	-363.0 (0.30)	-360.0 (0.06)	-363.3 (0.02)
Lithium	-369.0 (0.12)	-375.7 (0.20)	-331.5 (0.11)	-326.5 (0.03)	-365.8 (0.49)	-325.7 (0.13)	-323.9 (0.05)	-325.8 (0.16)	-323.6 (0.07)	-325.4 (0.03)
Manganese	-453.3 (0.04)	-453.6 (0.06)	-435.3 (0.15)	-425.7 (0.03)	-454.5 (0.04)	-436.3 (0.28)	-424.5 (0.04)	-434.4 (0.33)	-428.0 (0.03)	-432.5 (0.02)
REE	-417.4 (0.02)	-417.2 (0.09)	-368.1 (0.17)	-419.5 (0.15)	-422.9 (0.20)	-362.5 (0.20)	-362.2 (0.49)	-363.7 (0.25)	-358.6 (0.05)	-359.9 (0.08)
Silicon	-449.9 (0.08)	-453.0 (0.23)	-368.8 (0.12)	-351.2 (0.03)	-480.1 (0.21)	-346.0 (0.03)	-328.2 (0.22)	-349.2 (0.43)	-340.7 (0.16)	-343.4 (0.17)
Vanadium	-482.7 (0.05)	-481.8 (0.03)	-478.4 (0.08)	-416.1 (0.02)	-483.3 (0.12)	-397.7 (0.36)	-374.1 (0.26)	-399.3 (0.36)	-387.0 (0.17)	-390.3 (0.22)

Notes: Entries are log marginal likelihoods; Monte Carlo standard errors are in parentheses. For each commodity the largest (best-fitting) value is in bold. GARCH-2, GARCH-J, GARCH-t and GARCH-GJR denote the GARCH family augmented with a second lag, jumps, Student- t innovations and a GJR leverage term; SV(1), SV(2), SV(1)-J, SV(1)-t and SV(1)-L are the corresponding stochastic-volatility specifications.

Table S8 reports the log marginal likelihoods for the GARCH family and the corresponding SV specifications. Only three of the sixteen commodities are best described by a GARCH specification: Cobalt and Chromium under GARCH-t, and Silver under GARCH-J. Importantly, these exceptions occur precisely in the settings where the traditional advantages of stochastic-volatility (SV) models are expected to be smallest. As emphasized by Chan and Grant (2016), SV models typically outperform GARCH because volatility evolves as a latent stochastic process rather than being fully determined by fixed parameters and past observations. This additional flexibility allows SV models to adapt more effectively to model misspecification, structural changes, and abrupt shifts in market conditions. However, 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, some of the flexibility that ordinarily distinguishes SV models is effectively

built into the GARCH specification. By contrast, 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. The few cases in which GARCH outperforms SV in our sample are fully consistent with this mechanism. GARCH-t dominates SV-t for Cobalt and Chromium, while GARCH-J dominates SV-J for Silver. Notably, these correspond exactly to the two model pairings, GARCH-t versus SV-t and GARCH-J versus SV-J, that Chan and Grant (2016) identify as the only situations in which GARCH can be expected to outperform its stochastic-volatility counterpart. Rather than contradicting the broader superiority of SV models, these results reinforce the view that GARCH becomes competitive only when enriched with features that mimic the flexibility naturally provided by stochastic volatility.

S2.5 Model confidence sets

We use the Model Confidence Set (MCS) approach of Hansen et al. (2011) to compare the predictive accuracy of competing volatility models. The idea is as follows. Let $L_{i,t}$ denote the loss of model i at time t , and define the relative performance variable $d_{ij,t} \equiv L_{i,t} - L_{j,t}$ with expectation $\mu_{ij} \equiv E(d_{ij,t})$. The set of superior models is defined as $\mathcal{M}^* \equiv \{i \in \mathcal{M}^0 : \mu_{ij} \leq 0 \text{ for all } j \in \mathcal{M}^0\}$. The MCS procedure determines $\mathcal{M}^*$ through a sequential testing framework for the null hypothesis $H_{0,\mathcal{M}} : \mu_{ij} = 0 \text{ for all } i, j \in \mathcal{M}$, where $\mathcal{M} \subset \mathcal{M}^0$. Starting from $\mathcal{M} = \mathcal{M}^0$, an equivalence test is applied at level α ; if $H_{0,\mathcal{M}}$ is rejected, an elimination rule removes one model, and the process repeats until the null is first accepted. For each pair of models $i, j \in \mathcal{M}$, the test is based on the sample mean loss differential $\bar{d}_{ij} = n^{-1} \sum_t d_{ij,t}$, standardized as $t_{ij} = \bar{d}_{ij} / \sqrt{\widehat{\text{Var}}(\bar{d}_{ij})}$, with variances estimated via a block bootstrap. We use the range statistic $T_{R,\mathcal{M}} = \max_{i,j \in \mathcal{M}} |t_{ij}|$ for the equivalence test. If the null is rejected, the elimination rule $e_{R,\mathcal{M}} = \arg \max_{i \in \mathcal{M}} \sup_{j \in \mathcal{M}} t_{ij}$ removes the model that performs worst relative to all others. The surviving set is the model confidence set $\widehat{\mathcal{M}}_{1-\alpha}^*$, which asymptotically contains the best model(s) with probability at least $1 - \alpha$. In Tables S9 and S10, we mark the models included in the 75% MCS with a checkmark ($\checkmark$), these are the models whose MCS p -value satisfies $\hat{p}_i \geq \alpha$, where $\alpha = 0.25$ is the threshold to yield a tighter set of superior models (see Hansen et al., 2011).

The point-forecast results in Table S9 show that GARCH(1,1) is excluded from the confidence set for 7 out of 16 commodities under squared-error loss and for 6 under QLIKE loss. The SV-class models generally outperform GARCH(1,1), with SV(1) and SV(1)-L being the most frequently retained. The density-forecast results in Table S10 reveal that, for the centre-weighted specification, GARCH(1,1) is excluded in 11 out of 16 commodities; SV(2) emerges as the most frequently retained model (9 of 16), followed by SV(1)-t (6 of 16). The tail-weighted specification yields similar patterns.

Table S9: Model confidence sets for point volatility forecasts

(a) MCS membership - squared error loss (RV proxy)						
	GARCH	SV(1)	SV(2)	SV(1)-J	SV(1)-t	SV(1)-L
Aluminium		✓				✓
Cobalt	✓	✓	✓	✓	✓	✓
Copper	✓	✓				✓
Lead						✓
Molybdenum	✓		✓	✓	✓	✓
Nickel		✓		✓		✓
Zinc	✓	✓				✓
Silver	✓					
Palladium		✓				✓
Platinum		✓				✓
Chromium			✓	✓	✓	
Lithium	✓		✓		✓	
Manganese	✓	✓	✓	✓	✓	✓
REE	✓					
Silicon						✓
Vanadium	✓					✓
(b) MCS membership - QLIKE loss (RV proxy)						
	GARCH	SV(1)	SV(2)	SV(1)-J	SV(1)-t	SV(1)-L
Aluminium		✓				✓
Cobalt	✓		✓			
Copper	✓	✓				✓
Lead		✓				✓
Molybdenum	✓					
Nickel		✓				✓
Zinc	✓	✓				✓
Silver		✓				
Palladium		✓				
Platinum		✓				
Chromium	✓	✓	✓	✓	✓	✓
Lithium	✓		✓		✓	
Manganese	✓					
REE	✓	✓	✓	✓	✓	✓
Silicon	✓	✓	✓	✓	✓	✓
Vanadium	✓	✓	✓	✓	✓	✓

Notes: each entry indicates whether a model belongs to the 75% Model Confidence Set (MCS) of Hansen et al. (2011), computed with the range statistic T_R and a circular block bootstrap (block length 3, 10,000 replications). A checkmark denotes inclusion in the 75% confidence set ($p \geq 0.25$). Losses are the period-by-period squared forecast errors (panel *a*) and QLIKE losses (panel *b*) of the annualized variance forecasts against realized variance.

Table S10: Model confidence sets for density forecasts (weighted qCRPS)

(a) MCS membership - weighted qCRPS, center						
	GARCH	SV(1)	SV(2)	SV(1)-J	SV(1)-t	SV(1)-L
Aluminium	✓		✓			
Cobalt			✓		✓	
Copper				✓		
Lead	✓					
Molybdenum					✓	
Nickel			✓			
Zinc			✓			
Silver			✓			
Palladium			✓			
Platinum	✓		✓	✓	✓	
Chromium					✓	
Lithium	✓		✓		✓	
Manganese			✓		✓	
REE	✓					
Silicon						✓
Vanadium						✓
(b) MCS membership - weighted qCRPS, tails						
	GARCH	SV(1)	SV(2)	SV(1)-J	SV(1)-t	SV(1)-L
Aluminium	✓		✓			
Cobalt			✓		✓	
Copper			✓	✓		
Lead	✓					
Molybdenum					✓	
Nickel			✓			
Zinc			✓			
Silver			✓			
Palladium			✓			
Platinum	✓		✓	✓	✓	
Chromium					✓	
Lithium	✓		✓		✓	
Manganese			✓		✓	
REE	✓					
Silicon	✓		✓			✓
Vanadium						✓

Notes: each entry indicates whether a model belongs to the 75% Model Confidence Set (MCS) of Hansen et al. (2011), computed with the range statistic T_R and a circular block bootstrap (block length 3, 10,000 replications). A checkmark denotes inclusion in the 75% confidence set ($p \geq 0.25$). Losses are the period-by-period weighted quantile scores emphasizing the center (panel a) and the tails (panel b) of the predictive distribution.

S2.6 Prior sensitivity

The marginal likelihood integrates the likelihood over the prior and therefore remains sensitive to the prior. We focus on the volatility-of-volatility prior scale S_h of the volatility-of-volatility prior $\omega_h^2 \sim \mathcal{IG}(\nu_h, S_h)$, as it directly controls the prior flexibility of the latent volatility process, which is the primary distinction between the SV and GARCH classes. 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 volatility dynamics of SV(2) and renders heavy tails partly redundant. Table S11 presents the prior sensitivity of S_h over a grid from one quarter to four times the baseline value. The baseline $S_h = 0.16$ follows Chan and Grant (2016), with $\nu_h = 5$, which centres the prior mean of the volatility-of-volatility at $\omega_h = 0.2$, the typical magnitude of posterior estimates. Most rankings are unchanged: SV(2) dominates for the majority of commodities across most of the grid. The only exceptions are 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).

Table S11: Prior sensitivity of the marginal likelihood comparison

Commodity	Best model at prior scale S_h				
	$S_h=0.04$	0.08	0.16	0.32	0.64
Aluminium	SV(1)-L	SV(1)-L	SV(1)-L	SV(1)-L	SV(2)
Cobalt	SV(1)-t	SV(2)	SV(2)	SV(2)	SV(2)
Copper	SV(1)-t	SV(2)	SV(2)	SV(2)	SV(2)
Lead	SV(2)	SV(2)	SV(2)	SV(2)	SV(2)
Molybdenum	SV(1)-t	SV(2)	SV(2)	SV(2)	SV(2)
Nickel	SV(1)-t	SV(1)-t	SV(1)-t	SV(1)-t	SV(1)-t
Zinc	GARCH	GARCH	SV(2)	SV(2)	SV(2)
Silver	GARCH	GARCH	GARCH	GARCH	GARCH
Palladium	SV(1)-t	SV(2)	SV(2)	SV(2)	SV(2)
Platinum	SV(1)-t	SV(1)-t	SV(2)	SV(2)	SV(2)
Chromium	SV(1)-t	SV(2)	SV(2)	SV(2)	SV(2)
Lithium	SV(1)-t	SV(1)-t	SV(1)-t	SV(1)-t	SV(1)-t
Manganese	SV(2)	SV(2)	SV(2)	SV(2)	SV(2)
REE	SV(1)-t	SV(1)-t	SV(1)-t	SV(1)-t	SV(1)-t
Silicon	SV(2)	SV(2)	SV(2)	SV(2)	SV(2)
Vanadium	SV(2)	SV(2)	SV(2)	SV(2)	SV(2)

Notes: The table reports the model with the highest log marginal likelihood for each value of the prior scale S_h of the volatility-of-volatility prior $\omega_h^2 \sim \mathcal{IG}(\nu_h, S_h)$. The baseline $S_h = 0.16$ follows Chan and Grant (2016); with $\nu_h = 5$, this centres the prior mean of the volatility-of-volatility at $\omega_h = 0.2$, and the grid spans one quarter to four times this value.

S2.7 Rolling window estimation

In Tables S12 and S13, we present the point- and density-forecast evaluations using a 66-month rolling window approach. For point forecasts (Table S12), under the RV proxy, the share of comparisons where SV models outperform GARCH(1,1) rises from 41% (66/160) under the recursive scheme to 51% (81/160) under the rolling scheme. SV(1) is most often the best-performing SV specification under RMSFE (4 metals), while SV(2) is most often the best under QLIKE (5 metals). For density forecasts (Table S13), SV models beat GARCH(1,1) in 95 out of 160 cases. Relative to the recursive scheme, the rolling window amplifies the predictive gains of SV models, with the improvements extending to lead and platinum, where GARCH was previously competitive. This is consistent with the view that the rolling window is more sensitive to time-varying volatility dynamics, thereby better capturing the flexibility of the SV specifications.

Table S12: Evaluation of volatility point forecasts using RV as proxy

(a) Root Mean Squared Forecast Error loss					
	SV(1)	SV(2)	SV(1)-J	SV(1)-t	SV(1)-L
Aluminium	-0.0051	0.0662	0.0054	0.0057	0.0134
Cobalt	-0.1142	-0.1071*	-0.0977	-0.0859	-0.1200
Copper	-0.1022	-0.0638	-0.0711	-0.0685	-0.0924
Lead	-0.0435**	-0.0132*	-0.0259**	-0.0253**	-0.0287*
Molybdenum	0.9666	0.0436	-0.0510	-0.0087	0.0182
Nickel	-0.0726	-0.0711	-0.0723	-0.0707	-0.0725
Zinc	0.0037	0.0647	0.0258	0.0375	0.0076
Silver	0.0452	0.0830	0.0502	0.0556	0.0663
Palladium	-0.0170	0.0435	0.0114	0.0163	-0.0300
Platinum	-0.0861**	-0.0338	-0.0745**	-0.0616**	-0.0920**
Chromium	-0.0629	-0.1096	-0.1049	-0.1118	-0.0371
Lithium	0.1788	0.0712	0.1305	0.2445	0.1316
Manganese	-0.0352	-0.0358	-0.0348	-0.0369	-0.0357
REE	0.0073	-0.0636*	0.0222	-0.0143	0.0140
Silicon	0.1282	0.0268	0.1082	0.0013	0.0009
Vanadium	0.3252	0.4773	0.3064	0.2228	0.0628
(b) QLIKE loss					
	SV(1)	SV(2)	SV(1)-J	SV(1)-t	SV(1)-L
Aluminium	-0.1821	0.0739	-0.1097	-0.1140	-0.0553
Cobalt	-0.1318	-0.1667	-0.0908	-0.0237	-0.1285
Copper	-0.1541*	0.0166	0.0754	0.0862	-0.0123
Lead	-0.2059**	-0.0910*	-0.1235**	-0.1183**	-0.1590*
Molybdenum	0.1121	-0.0259	0.1579	0.2581	0.3483
Nickel	0.0354	-0.0728	0.0856	0.1240	-0.0604
Zinc	0.1336	0.1104	0.2380	0.2840	0.1603
Silver	-0.0883	0.1955	-0.0401	0.0030	0.1658
Palladium	-0.1431	0.1628	-0.0064	0.0089	-0.1839
Platinum	-0.3473**	-0.2278	-0.2979**	-0.2761**	-0.3670**
Chromium	-0.0402	-0.1351	-0.0486	-0.0462	-0.0012
Lithium	0.1540	-0.0474	0.0282	-0.0053	0.1037
Manganese	0.0940	0.2453	0.1965	0.5779	0.1241
REE	-0.3102	-0.0958	1.5132	1.7541	-0.3592
Silicon	0.4171	0.5695	1.8358	1.7266	0.4654
Vanadium	0.8074	1.5809	1.9306	1.3282	1.1308

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.

Table S13: Quantile-based continuous ranked probability score (qCRPS)

(a) Weighted qCRPS ($\widehat{wQS}$) - center					
	SV(1)	SV(2)	SV(1)-J	SV(1)-t	SV(1)-L
Aluminium	-0.0030	-0.0486	-0.0158	-0.0112	-0.0294
Cobalt	-0.0053	-0.0607	-0.0246	-0.0780**	0.0039
Copper	0.0038	-0.0238	-0.0386	-0.0342	-0.0141
Lead	0.0360	0.0131	0.0231	0.0271	0.0368
Molybdenum	0.2306	-0.0128	-0.1611**	-0.2261**	-0.1015**
Nickel	-0.0310	-0.0692	-0.0429	-0.0564	-0.0313
Zinc	-0.0025	-0.0259	-0.0140	-0.0155	0.0069
Silver	-0.0196	-0.0517	-0.0270	-0.0299	-0.0439
Palladium	0.0112	-0.0362**	-0.0167*	-0.0147	0.0047
Platinum	0.0197	-0.0075	0.0033	-0.0057	0.0167
Chromium	-0.0271	-0.0818**	-0.1105**	-0.1296**	-0.0015
Lithium	0.1218	-0.0060	0.0567	0.0011	0.0976
Manganese	0.0341	-0.0886**	0.0093	-0.1122**	0.0379
REE	0.1581	-0.0090	0.1360	0.0642	0.1607
Silicon	0.3747	0.1521	0.2563	-0.1775**	-0.0486
Vanadium	0.0516	0.2260	0.0132	0.0224	-0.2149**
(b) Weighted qCRPS ($\widehat{wQS}$) - tails					
	SV(1)	SV(2)	SV(1)-J	SV(1)-t	SV(1)-L
Aluminium	-0.0030	-0.0789	-0.0233	-0.0134	-0.0412
Cobalt	0.0071	-0.0763	-0.0246	-0.1008**	0.0211
Copper	0.0074	-0.0293	-0.0531	-0.0423	-0.0207
Lead	0.0534	0.0201	0.0337	0.0429	0.0525
Molybdenum	0.2817	-0.0142	-0.2030**	-0.2569**	-0.1422**
Nickel	-0.0346	-0.0980	-0.0520	-0.0707	-0.0358
Zinc	-0.0021	-0.0393	-0.0195	-0.0187	0.0094
Silver	-0.0293	-0.0772	-0.0409	-0.0425	-0.0733
Palladium	0.0152	-0.0516**	-0.0254*	-0.0183	0.0063
Platinum	0.0365	-0.0036	0.0128	0.0035	0.0310
Chromium	-0.0384	-0.1106**	-0.1607**	-0.1724**	-0.0054
Lithium	0.1662	-0.0331	0.0637	-0.0090	0.1320
Manganese	0.0429	-0.1329**	0.0010	-0.1711**	0.0526
REE	0.2275	-0.0110	0.2059	0.0996	0.2299
Silicon	0.4851	0.2057	0.3510	-0.1957**	-0.0488
Vanadium	0.0748	0.2876	0.0348	0.0530	-0.2585**

Notes: each entry of the table shows $(\widehat{wQS}_j / \widehat{wQS}_0) - 1$, where $\widehat{wQS}_0$ is the weighted qCRPS of 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.

S2.8 Principal Component Analysis: further results

To assess the suitability of Principal Component Analysis (PCA), we first conducted Bartlett's test of sphericity (Bartlett, 1950). The test yielded a chi-square statistic of 144.50 with 28 degrees of freedom ($p < 10^{-16}$), indicating that the correlation matrix significantly departs from the identity matrix. This result supports the appropriateness of applying PCA to our data. In determining the number of components to retain, we considered both Parallel Analysis (Horn, 1965) and the Kaiser criterion (Kaiser, 1960). The observed eigenvalues were 4.96, 1.43, 0.93, 0.31, 0.18, 0.17, 0.01, and 0.004, while those obtained from randomly generated data were 2.24, 1.71, 1.32, 1.02, 0.75, 0.50, 0.31, and 0.15. Parallel Analysis supports the retention of one component, whereas the Kaiser criterion supports two.

References

- Bartlett, M. S. (1950). Tests of significance in factor analysis. *British journal of psychology*.
- Chan, J. C. (2017). The stochastic volatility in mean model with time-varying parameters: An application to inflation modeling. *Journal of Business & Economic Statistics*, 35(1):17–28.
- Chan, J. C. and Grant, A. L. (2014). Issues in Comparing Stochastic Volatility Models Using the Deviance Information Criterion. CAMA Working Papers 2014-51, Centre for Applied Macroeconomic Analysis, The Australian National University.
- Chan, J. C. and Grant, A. L. (2016). Modeling energy price dynamics: GARCH versus stochastic volatility. *Energy Economics*, 54:182–189.
- Chan, J. C. and Jeliazkov, I. (2009). Efficient simulation and integrated likelihood estimation in state space models. *International Journal of Mathematical Modelling and Numerical Optimisation*, 1(1-2):101–120.
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*. Chapman and Hall/CRC, Boca Raton, FL, third edition.
- Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, 7(4):457–472.
- Geyer, C. J. (1992). Practical markov chain monte carlo. *Statistical science*, pages 473–483.
- Hansen, P. R., Lunde, A., and Nason, J. M. (2011). The model confidence set. *Econometrica*, 79(2):453–497.
- Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30(2):179–185.
- Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and psychological measurement*, 20(1):141–151.