



Article

A Modular Knowledge-Extraction Framework for Deep Learning Forecasts of Multi-Tier Commodity Prices

Montchai Pinitjitsamut 

Department of Agricultural and Resource Economics, Faculty of Economics, Kasetsart University,
Bangkok 10900, Thailand; montchai.p@ku.th

Abstract

Vertically linked commodity markets—global futures, regional spot, and farm-gate prices—transmit information through directed cross-market channels whose strength varies with latent volatility regimes. Standard deep learning forecasters absorb both the directed cross-market dependence and the regime dependence of intrinsic-mode-aligned latent components into shared model weights, with no explicit architectural mechanism that exposes either as an inspectable structure. This paper proposes HVB-RA, a modular framework that combines two such mechanisms with a per-tier Variational Mode Decomposition and bidirectional LSTM backbone: (i) a directed cross-market attention layer in which the upstream-to-downstream topology is supplied from domain knowledge and the time-varying upstream-source attention intensities at the farm-gate tier (the regional-spot tier, with a single upstream key, reduces algebraically to a fixed residual upstream fusion) are extracted from data, and (ii) a regime-informed modal-weighting layer that mixes two trainable softmax weight profiles over IMF-aligned latent components through a filtered Markov-switching state probability fitted in a separate stage. An auxiliary post hoc projection enforces an exact linear constraint defined by long-run sample-mean ratios across tiers; the paper does not claim that these descriptive ratios are cointegrating relations or equilibrium coefficients. The framework is evaluated on three tiers of daily natural-rubber prices spanning 2038 trading days, against three external benchmarks (random walk, ARIMA(2,0,2), and an exogenous-only LSTM) and a contemporary neural hierarchical-interpolation forecaster (NHITS). Root mean squared error is reported per tier-horizon cell; a decision-aware income-smoothing metric quantifies the operational value of $h = 5$ farm-gate forecasts under a 5-day selling rule; and a within-method comparison evaluates the marginal contribution of the auxiliary constraint projection. On the present single-regime test window, HVB-RA attains a lower point error than the contemporary NHITS baseline at every tier-horizon cell, while no method—including HVB-RA—improves on the random-walk floor at most cells; the regime-conditional components of the architecture are not identifiable because every calibration and test origin is classified as a high-volatility regime by the trained Markov-switching model. The paper contributes to machine learning and knowledge extraction by demonstrating how time-varying upstream-source attention intensities at the farm-gate tier and regime-dependent latent-component-weight profiles—two forms of latent structure typically absorbed into model weights—can be exposed as explicit, inspectable, and individually testable components of a multi-tier forecasting architecture, and by providing a reproducibility package documenting the conditions under which each component is expected to be identifiable.



Academic Editors: Sheng Du and
Javier Del Ser Lorente

Received: 7 June 2026

Revised: 25 June 2026

Accepted: 25 June 2026

Published: 1 July 2026

Copyright: © 2026 by the author.

Licensee MDPI, Basel, Switzerland.

This article is an open access article
distributed under the terms and

conditions of the [Creative Commons](#)

[Attribution \(CC BY\)](#) license.

Keywords: knowledge extraction; latent structure extraction; multi-tier time series forecasting; cross-market attention; regime-informed deep learning; mixture-of-experts; variational mode decomposition; Markov switching; vertical price transmission; natural rubber

1. Introduction

Commodity prices observed at successive points of a value chain—global futures, regional spot benchmarks, and producer-level farm-gate prices—are linked through directed market channels rather than through identity-based aggregation. Information flows predominantly from upstream price-discovery markets to downstream cash markets, with the strength of transmission documented to vary across volatility regimes [1,2]. The forecasting task in such systems is therefore not only to predict each series in isolation but to make sense of the cross-market structure that links them. Two forms of structure are of particular interest. The first is the time-varying intensity of directed cross-market influence: while the direction itself (upstream to downstream) is supported by domain evidence, the strength of the upstream-to-downstream signal at each point in time is unobserved and must be learned. The second is the relative importance of the high-frequency versus low-frequency components of each price series, which may shift between latent volatility regimes that are not directly observable. Both are properties of the data-generating process that can in principle be exposed as explicit, inspectable components of a forecasting architecture, rather than absorbed implicitly into shared model weights.

The gap motivating this work is therefore specific: no existing multi-tier commodity forecaster simultaneously (i) imposes a domain-informed directed topology between market tiers, (ii) conditions latent-component aggregation on an explicit regime signal, and (iii) exposes both mechanisms as individually ablatable architectural objects whose identifiability conditions can be stated and tested. Graph-based forecasters such as MTGNN [3] discover cross-series topology from data but do not incorporate regime-conditional aggregation; mixture-of-experts architectures [4] condition gating on a learned signal but not on a stand-alone interpretable regime classifier; decomposition-based pipelines [5,6] aggregate intrinsic modes through a fixed projection that holds modal importance constant across regimes. The design of HVB-RA addresses this gap by the minimum combination of components sufficient to expose each mechanism as a separable object: per-tier VMD provides the latent components; a BiLSTM encoder provides the shared temporal representation; domain-informed directed attention provides the topology; and a two-stage Markov-switching gate provides the regime signal. The two-stage design is chosen deliberately over joint end-to-end estimation of the regime classifier so that the regime component remains identifiable as a stand-alone model independently of the forecast loss—a property that is lost when the regime and forecast parameters are optimised jointly. The resulting framework is not proposed as a universal RMSE-improvement over all baselines; it is proposed as a structured and reproducible decomposition of a multi-tier forecasting task whose components can be examined, replaced, and tested in isolation.

Recent neural forecasters such as N-BEATSx [7], the Temporal Fusion Transformer (TFT) [8], DeepAR [9], NHITS [10], and PatchTST [11] have advanced the modelling of complex temporal patterns and exogenous covariates in time series. Their typical design treats cross-series interaction either through generic attention or through implicit panel sharing of parameters, and does not provide a dedicated mechanism that exposes time-varying cross-market intensities or regime-dependent latent-component weighting as separable architectural objects. This paper proposes a modular framework that does so,

applied to a three-tier daily natural-rubber price dataset where the vertical structure and regime variation are both empirically documented [1,2,6].

1.1. Related Work

This subsection reviews four literature strands relevant to the design of HVB-RA: cross-series neural forecasting, regime-switching and gated architectures, decomposition-based deep learning, and vertical price transmission in commodity markets. A final paragraph situates the present work relative to the author's prior single-tier study.

1.1.1. Cross-Series Neural Forecasting

Several deep architectures learn joint representations across multiple time series. N-BEATSx [7] extends the original N-BEATS architecture [12] to accept exogenous covariates; the basis-expansion blocks remain per-series, and the model is not hierarchical in the cross-series sense. The Temporal Fusion Transformer [8] processes static covariates, known future inputs, and observed past inputs through gated attention across multiple horizons. DeepAR [9] trains a single autoregressive recurrent network on a panel of related series, sharing information implicitly through joint parameter estimation. NHITS [10] introduces hierarchical interpolation along the temporal dimension, using multi-rate signal sampling within a single series rather than across series. PatchTST [11] adapts Transformer architectures to long-horizon forecasting through channel-independent patch tokenisation. Graph-based neural forecasters provide an alternative inductive bias by modelling cross-series dependence through learned adjacency matrices; MTGNN [3] in particular learns a directed graph among variables jointly with the forecasting model, and spectral graph approaches [13] provide a related construction in the frequency domain.

Two design philosophies are visible in this literature. Graph-based forecasters such as MTGNN discover a directed cross-series structure from data, treating the adjacency as a learnable object; attention-based forecasters such as TFT model cross-series interaction through generic attention that is neither directed nor structured. The present paper takes a third position: the directed topology between vertically linked market tiers is supplied as a domain-informed prior [1], and the model learns only the time-varying intensities of that prespecified topology. This is appropriate when the direction is well established by the economic literature (futures lead spot, spot leads farm-gate) but the strength of transmission and its dependence on the latent regime are not. The contribution is therefore in the orthogonal dimension of combining a domain-imposed directed topology with regime-gated mode aggregation, not in learning the direction itself.

1.1.2. Regime-Switching and Gated Architectures

The regime-switching literature is anchored in the Markov-switching framework of Hamilton [14], with subsequent developments in Markov-switching GARCH for volatility modelling [15], smooth-transition autoregressions [16,17], and threshold models [18]. Regime-switching specifications improve out-of-sample forecast accuracy in series with volatility clustering or structural breaks [19,20].

Within deep learning, regime conditioning has been approached along several axes: as a regime indicator concatenated as an exogenous covariate [21]; as per-regime models trained on sample-split subsets [22]; and, more recently, as jointly learnable switching state-space neural models in which regimes and dynamics are estimated end-to-end. The mixture-of-experts framework [4] provides an architecturally integrated alternative in which a gating network produces a soft selection over multiple expert sub-networks; the experts and the gate are typically learned jointly from data. HVB-RA adopts the mixture-of-experts perspective at the level of intrinsic-mode-function aggregation but in a deliberately simpler two-stage configuration: the gating signal is the filtered probability $\hat{\pi}_t$ of a two-state

Markov-switching autoregression fitted in a separate stage on each tier's returns, and the two experts are alternative softmax weight profiles v_0, v_1 over IMFs that are trained jointly with the rest of the neural backbone. This two-stage design keeps the regime classifier interpretable and identifiable as a stand-alone model, at the cost of not learning the regime parameters jointly with the forecast loss; the trade-off is discussed in Section 2.

1.1.3. Decomposition-Based Deep Learning

Variational Mode Decomposition (VMD) [5] decomposes a one-dimensional signal into a finite set of band-limited intrinsic mode functions through a constrained variational optimisation. Compared with the earlier Empirical Mode Decomposition [23], VMD is designed to reduce the mode-mixing problem and to produce decompositions with better-defined frequency content. Hybrid VMD–deep learning pipelines have been investigated across a range of time-series forecasting domains with reported gains over single-model baselines; within commodity-price forecasting specifically, hybrid VMD-augmented BiLSTM has been validated for single-series rubber forecasting on the post-2018 feature set [6].

Conventional VMD–deep learning pipelines apply a single decomposition globally and aggregate the IMFs through a fixed output layer. Two design choices distinguish the present work: VMD is applied *per tier*, treating each tier's IMFs as the input to a tier-specific encoder, and the IMF-aggregation step is *regime-conditional* so that low-frequency trend modes and high-frequency adjustment modes can receive different weights in calm and stress states. The regime-informed modal-weighting layer is the formal mechanism that realises the second design choice.

1.1.4. Vertical Price Transmission and Constrained Multivariate Forecasting

The economic literature on vertical price transmission models the dynamic dependence between upstream and downstream prices in commodity value chains, often through error-correction representations of cointegrated systems [24]. Long-run linear relations among vertically linked prices are typically estimated through a formal cointegration test and then used to drive error-correction terms. The present paper does not estimate a cointegrating relation; instead, it uses descriptive long-run sample-mean ratios between tiers to define the rows of a constraint matrix A such that $A\tilde{y}_t = 0$ holds exactly after a linear projection. These sample-mean ratios are properties of the training window, not equilibrium coefficients from a cointegration analysis, and the constraint is treated as a post hoc adjustment supplied to operationally enforce a consistent cross-tier scaling rather than as a structural relation. The connection to the reconciliation literature [25,26] is purely algebraic: the constraint-projection form is borrowed from MinT, but the relation it enforces is a market-level sample-ratio constraint, not an aggregation identity and not an optimality claim under MinT's loss-trace criterion.

1.1.5. Relationship to Prior Author Work

The author's prior work [6], published in MDPI *Forecasting*, develops a variance-sensitive evaluation protocol for decomposition-based forecasts of a single rubber-price series. That work introduces the Standard Deviation Ratio (StdR) diagnostic, demonstrates that conventional accuracy metrics can be high while forecasts exhibit variance collapse, and uses a single-tier VMD-augmented BiLSTM as the empirical vehicle. The present paper is methodologically distinct in two respects. The empirical scope is multi-tier rather than single-tier: nine constituent series aggregated into three market tiers. The architectural contributions are non-overlapping: the prior work develops an evaluation methodology and a single-tier ablation; the present work develops directed cross-market attention and the regime-informed modal-weighting layer.

1.2. Research Question and Proposed Framework

The empirical question motivating the framework is whether two forms of latent structure—the time-varying intensity of upstream-to-downstream cross-market influence, and regime-dependent weighting of latent components within each tier—can be exposed as explicit, inspectable components of a multi-tier forecasting architecture, and whether doing so yields measurable forecast improvements over models that absorb the same information implicitly into shared parameters.

The proposed framework, HVB-RA (Hybrid VMD-BiLSTM with Regime-Aware components), comprises four components arranged in a modular pipeline. The first two carry the principal methodological contributions; the third is an auxiliary constrained adjustment; the fourth is a supporting evaluation protocol.

C1 (primary): Domain-informed directed cross-market attention. Per-tier bidirectional LSTM encoders process the local Variational Mode Decomposition together with exogenous covariates. A cross-market attention layer then injects information from upstream tiers into downstream encoder outputs along a topology that is supplied by domain knowledge: the regional-spot representation attends to the futures representation, and the farm-gate representation attends to both. The topology is not learned from data; the attention queries, keys, values, and the resulting time-varying attention intensities are learned. The contribution is in combining a domain-informed prior on direction with a data-driven estimate of strength.

C2 (primary): Two-stage regime-informed modal-weighting layer. For each tier, the IMF summary vectors from the encoder are aggregated through a two-expert mixture-of-experts in which the experts are softmax weight profiles $v_0, v_1 \in \mathbb{R}^K$ and the gating signal is the filtered probability $\hat{\pi}_t$ from a two-state Markov-switching autoregression fitted in a separate stage on each tier's returns. The aggregate weight at time t is $w_t = (1 - \hat{\pi}_t) \text{softmax}(v_0) + \hat{\pi}_t \text{softmax}(v_1)$. The layer is differentiable with respect to v_0 and v_1 and is trained jointly with the rest of the neural backbone; the gating signal is treated as a fixed input and not back-propagated into the Markov-switching model. The architecture is therefore two-stage by design. The layer reduces exactly to a fixed-weight aggregation when $v_0 = v_1$, providing a nested benchmark.

C3 (auxiliary): Sample-ratio constraint projection. A post hoc linear projection enforces, exactly, the relations $F = a_R R$ and $R = a_G G$ on the joint forecast vector, where a_R, a_G are sample-mean ratios estimated on the training window. The construction borrows the constraint-projection form from MinT [25], but it does not import MinT's trace-optimality claim and it does not estimate a cointegrating relation. Two regime-conditional correction operators are mixed convexly by $\hat{\pi}_t$; the resulting mixture is not in general a projection matrix (it need not be idempotent), but it is a coherence-preserving adjustment since each regime-specific operator satisfies $A \Psi(W_i) = A$ (Section 2.7.2). This component is reported as an auxiliary mechanism whose effect is examined empirically in Section 3, not as a primary methodological contribution.

C4 (supporting): Decision-aware evaluation protocol. At the farm-gate tier, point forecasts are translated into a five-day selling decision under a transaction-cost threshold, and realised-income variance reduction is reported as a complementary success metric. The protocol illustrates the discipline of evaluating forecasts at the operational tier where decisions are made; the specific metric is presented as a methodological template, not as a primary success claim.

1.3. Two Architectural Propositions

The framework rests on two architectural propositions about identifiable latent structure. These motivate the design of components C1 and C2 (Section 2) and are listed here

for clarity of scope. Their full empirical evaluation requires per-component ablations and a multi-regime test window; the manuscript reports point-error performance against external and contemporary benchmarks together with the primary per-component ablations (E1, E2) on the single-regime test window available (Section 3.4), and defers the regime-diverse follow-up backtest and the broader ablation contrasts to the companion repository and to follow-up work (Section 4.7).

P1: Adding directed cross-market information is expected to improve downstream forecasts. Forecasts at the downstream tiers (regional spot, farm-gate) depend on whether the model can route upstream representations into its downstream encoders. The architecture in C1 supplies such a routing along a domain-informed direction. The directly testable claim is whether having this routing helps, relative to a tier-specific encoder that aggregates only its own VMD modes and exogenous covariates. The corresponding controlled contrast—HVB-RA against the no-attention ablation E2—is implemented in the reference package (Section 2.10.3) and reported on the real-data series in the main text (Section 3.4). The complementary questions of whether the domain-informed direction beats a reversed or unrestricted direction, and whether attention beats simple concatenation at matched parameter count, are deferred to follow-up work.

P2: The two-expert mixture is expected to extract differentiated weight profiles from training data. The modal-weighting layer in C2 is designed to expose regime-dependent latent-component weighting as the difference $v_1 - v_0$ between two trainable softmax profiles. As noted in the discussion of the per-mode summary projection (Section 2), these profiles weight latent components extracted by the shared BiLSTM encoder rather than IMFs directly. The directly testable claim is whether, when both profiles are free to vary, the trained profiles separate meaningfully—i.e., whether the layer extracts non-trivial regime-conditional structure from the training data. The corresponding controlled contrast—the trained v_0, v_1 from HVB-RA against the tied-weights ablation E1—is implemented in the reference package; the forecast contrast between the two ablations is reported on the real-data series in the main text (Section 3.4), while the seed-averaged L_1 distance between the two trained profiles and the across-seed stability of that distance are reported in the companion repository. The stronger question of whether the regime conditioning improves forecasts in genuinely out-of-regime test windows requires test data that includes both calm and stress states and is identified as a follow-up data requirement.

1.4. Contributions and Paper Organisation

The framework is validated on a three-tier daily dataset of natural-rubber prices covering global futures markets (TOCOM, SGX, and SHFE), regional spot benchmarks (SICOM, MRB, and GAPKINDO), and farm-gate prices (Thailand, Malaysia, and Indonesia). The empirical application is used because the rubber market simultaneously exhibits directed cross-market price transmission, documented regime variation manifesting as volatility clustering, and a decision-relevant lower tier where smallholder producers make discrete selling decisions under transaction costs.

The contributions of this paper are:

1. A directed cross-market attention layer that combines a domain-informed topology between vertically linked market tiers with time-varying attention intensities learned from data, with attention queries restricted to lower tiers and keys to upstream tiers (Section 2, Component C1).
2. A two-stage regime-informed modal-weighting layer that mixes two trainable softmax weight profiles over IMF-aligned latent components through a filtered Markov-switching state probability, exposing regime-dependent latent-component weighting

as an inspectable architectural object rather than absorbing it into the encoder's weights (Section 2, Component C2).

3. An empirical evaluation on a multi-tier commodity-price system spanning 2038 trading days, reporting root mean squared error against three external benchmarks (random walk, ARIMA(2,0,2), and an exogenous-only LSTM), together with a constraint-projection comparison and a decision-aware income-smoothing metric for $h = 5$ farm-gate forecasts, providing the evidence on which the contributions of C1 and C2 are assessed (Section 3). Complementary point-error metrics (MAE and MASE) and Diebold–Mariano significance tests are provided in the companion repository submitted as Supplementary Material; the primary per-component ablations (E1 and E2) are reported in the main text (Section 3.4), with the broader ablation set in the companion repository.

HVB-RA is not proposed as a universal forecasting solution. The framework is most informative in systems with three jointly observable properties: vertical linkage between two or more market tiers with documented directed price transmission; non-stationary error structures driven by latent volatility regimes; and a decision-relevant lower tier where forecast value can be assessed economically. Examples beyond natural rubber include electricity markets with generation, wholesale and retail tiers; agricultural value chains with producer, processor and consumer prices; and financial intermediation networks with benchmark, regional and counterparty rates.

The paper contributes to machine learning and knowledge extraction by demonstrating how two forms of latent structure—time-varying upstream-source attention intensities at the farm-gate tier and regime-dependent latent-component-weight profiles—can be exposed as explicit, inspectable, and ablatable components of a multi-tier forecasting architecture, rather than absorbed into the parameters of a generic backbone. The remainder of the paper is organised as follows. Section 2 specifies the data, the architecture, and the training protocol. Section 3 reports the empirical results across the two propositions, three external benchmarks, a contemporary deep learning benchmark, the primary per-component ablations (E1 and E2), the auxiliary constraint-projection comparison, and the decision-aware evaluation, with complementary point-error metrics, significance tests, and the broader ablation contrasts maintained in the companion repository. Section 4 discusses the findings, the conditions under which the two architectural mechanisms are identifiable, the broader ablation contrasts deferred to follow-up work, and the priorities for additional comparisons against current transformer-class forecasters. Section 5 concludes.

2. Materials and Methods

This section specifies the data, the HVB-RA framework, and the evaluation protocol. HVB-RA is constructed as a **modular framework** comprising six components: per-tier Variational Mode Decomposition and two-state Markov-switching regime detection are estimated offline as preprocessing; per-tier bidirectional LSTM encoders, the directed cross-tier attention layer, and the regime-informed modal-weighting layer are trained jointly through end-to-end backpropagation on the neural backbone; and an auxiliary constraint projection is applied post hoc to the trained forecasts. Section 2.9 provides a graphical and tabular overview; the technical specification proceeds in turn through the data, the regime module, the encoder, the modal-weighting layer, the forecast head, the auxiliary projection, the training protocol, the baselines and evaluation metrics, and the decision-loss simulation. The complete correspondence between the Methods components and the reference hvbra/implementation is documented in the reproducibility supplement.

2.1. Data: A Three-Tier Vertically Linked Rubber Price System

The framework is evaluated on a daily dataset of natural-rubber prices spanning 2 May 2018 through 4 March 2026 (2038 trading-day observations). The dataset is organised as three vertically linked market tiers $\ell \in \{F, R, G\}$, with F denoting global futures markets, R denoting regional spot benchmarks, and G denoting farm-gate prices. The three tiers are jointly observed: none is a deterministic aggregate of the others, and the system does *not* satisfy a summation identity of the form $F = R = G$ or any analogous linear aggregation. The tiers are linked through directed market channels of price transmission, with the strength of transmission documented to vary across volatility regimes [1,2]. All tier-level prices are converted to United States dollars per kilogram (USD/kg) using same-day spot exchange rates before model training.

2.1.1. Tier Construction

Each tier is constructed as a simple average of three representative series, providing redundancy against country-specific noise and enabling a balanced three-tier evaluation. Tier F averages TOCOM RSS3, SGX TSR20, and SHFE RU front-month futures; Tier R averages SICOM RSS3, the Malaysian Rubber Board (MRB) benchmark, and the GAPKINDO Indonesian composite spot prices; and Tier G averages country-level farm-gate prices from RAOT (Thailand), MRB-DOSM (Malaysia), and the GAPKINDO/provincial composite (Indonesia). Table 1 summarises the constituent series and data sources used to construct the three vertically linked tiers. Simple averaging is preferred to volume weighting because volume series are not equally available across the three tiers; volume-weighted tier construction remains a subject for future sensitivity analysis.

Table 1. Three-tier vertically linked construction. Each tier-level series is the simple average of three representative constituent series; full constituent-level documentation is provided in the reproducibility supplement.

Tier	Constituent Series	Source
F (global futures)	TOCOM RSS3/SGX TSR20/SHFE RU front-month	TOCOM, SGX, SHFE official feeds
R (regional spot)	SICOM RSS3/MRB benchmark/GAPKINDO composite	SICOM, MRB, GAPKINDO
G (farm-gate)	RAOT (TH)/MRB-DOSM (MY)/GAPKINDO provincial (ID)	RAOT, DOSM, GAPKINDO

2.1.2. Exogenous Covariates

The exogenous matrix $\mathbf{x}_t \in \mathbb{R}^p$ with $p = 5$ contains: (i) the USD/THB log-return, (ii) the Brent crude oil log-return (a synthetic-rubber substitution channel), (iii) the standardised El Niño–Southern Oscillation Oceanic Niño Index (climate-driven supply channel), (iv) year-on-year growth in China automobile sales (downstream-demand channel through tire manufacturing), and (v) log changes in Qingdao port natural-rubber stocks (inventory channel). All covariates are forward-filled with a 3-day cap to handle short market closures while limiting stale-data propagation.

2.1.3. Splits

The dataset is partitioned chronologically into three windows for reproducible evaluation. The *training window* spans 2 May 2018 through 30 June 2023 ($n = 1348$ daily observations per tier). The *calibration window* spans 3 July 2023 through 31 December 2024 ($n = 392$) and is used for (a) early stopping during neural training and (b) estimating the constraint-projection covariance matrices of Section 2.7. The *test window* spans 2 January 2025 through 4 March 2026 ($n = 298$ raw observations, reduced to $N = 278$ aligned origins after enforcing that all three target horizons fall within the test window). No information beyond each forecast origin enters the rolling VMD or filtered regime computations

at that origin: at origin t , VMD is fitted on the trailing 1024 trading days ending at t (Section 2.3) and the filtered regime probability $\hat{\pi}_t^{(\ell)}$ is computed from the Kim filter using observations up to t (Section 2.2). Model parameters, projection covariance matrices, and Markov-switching parameters are estimated using only training (and, for early stopping or projection covariances, training+calibration) observations, and are never re-estimated using test-period outcomes.

2.2. Regime Detection (Offline Preprocessing)

Module type. Classical statistical module, estimated offline on the training window. The fitted parameters and the inferred state probabilities are consumed downstream as fixed scalar inputs to the neural backbone (Section 2.5) and to the auxiliary constraint projection (Section 2.7). The Markov-switching parameters are *not* jointly trained with the neural backbone; the framework is therefore two-stage by design.

For each tier ℓ the framework fits a two-state Markov-switching autoregression of order one (MS-AR(1)) on tier-level log-returns. Following Hamilton [14],

$$r_t^{(\ell)} = \mu_{s_t}^{(\ell)} + \phi_{s_t}^{(\ell)} r_{t-1}^{(\ell)} + \sigma_{s_t}^{(\ell)} \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, 1), \quad (1)$$

where $s_t \in \{0, 1\}$ is a hidden state evolving according to a homogeneous Markov transition matrix

$$P = \begin{bmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{bmatrix}, \quad p_{ij} = \Pr(s_t = j \mid s_{t-1} = i). \quad (2)$$

The MS-AR(1) parameters $\{\mu_s, \phi_s, \sigma_s, p_{ij}\}$ are estimated by expectation–maximisation on the training window only, using the MarkovAutoregression backend of statsmodels as wrapped in hvbra/regime.py. State 1 is identified *ex post* as the high-volatility regime via the constraint $\sigma_1^{(\ell)} > \sigma_0^{(\ell)}$; this is a labelling convention without estimation impact.

Smoothed Versus Filtered Probabilities

The Kim [27] smoother and the Kim filter produce two probability series. The *smoothed* probability $\pi_t^{(\ell)} = \Pr(s_t = 1 \mid \mathcal{F}_T)$ uses information from the entire training window, including observations after t ; the *filtered* probability $\hat{\pi}_t^{(\ell)} = \Pr(s_t = 1 \mid \mathcal{F}_t)$ uses information only up to t . To avoid look-ahead bias, the framework uses the filtered probability $\hat{\pi}_t^{(\ell)}$ at *every origin*—training, calibration, and test—providing a uniform one-sided regime input throughout (hvbra/regime.py). The smoothed probability is not used at any stage of the operational pipeline. Identifiability of the regime classification across calibration and test windows is discussed in the empirical results (Section 3).

2.3. Per-Tier Variational Mode Decomposition (Offline Preprocessing)

Module type. Signal-processing module, applied offline to each tier’s price series. The decomposition outputs $K = 5$ intrinsic mode functions per tier; these enter the encoder as fixed inputs and are not back-propagated into.

For each tier ℓ the price series is decomposed into K band-limited intrinsic mode functions (IMFs) via Variational Mode Decomposition [5]:

$$y_t^{(\ell)} = \sum_{k=1}^K u_{k,t}^{(\ell)} + \eta_t^{(\ell)}, \quad (3)$$

where each IMF $u_{k,t}^{(\ell)}$ is constrained to be approximately narrow-band around a central frequency $\omega_k^{(\ell)}$, obtained as the solution to the variational problem

$$\min_{\{u_k\}, \{\omega_k\}} \sum_k \left\| \partial_t \left[\left(\delta(t) + \frac{j}{\pi t} \right) * u_k(t) \right] e^{-j\omega_k t} \right\|_2^2 \quad \text{s.t.} \quad \sum_k u_k = y. \quad (4)$$

The choice $K = 5$ is informed by prior empirical evidence on the same commodity [6], where K in this range is identified as an operating point balancing reconstruction quality and computational cost. Sensitivity to $K \in \{3, 5, 7\}$ is reported in Section 3.7. The bandwidth penalty $\alpha = 2000$ matches the configuration in `hvbra/vmd.py`.

Causal Rolling Re-Decomposition

To avoid look-ahead bias in the operational forecast, VMD is applied causally: the decomposition is fitted on a trailing window of 1024 trading days ending at the current origin, separately at each origin in calibration and test (`hvbra/vmd.py:rolling_decompose`). The effective re-fit stride for the IMF values supplied to the encoder at origin t is therefore one trading day—each origin's IMF input is obtained from a VMD optimisation that uses only observations up to and including t . The 21-day cadence parameter in the configuration controls a separate offline cache of full decompositions kept for diagnostic plots; it does *not* introduce stale IMF values at intermediate origins because the encoder consumes the per-origin causal IMFs rather than the cached decompositions. This protocol ensures that no IMF value at origin t depends on any price observation at $t' > t$.

2.4. Hierarchical Encoder with Domain-Informed Cross-Tier Attention

Module type. Differentiable neural module, trained jointly with the regime-informed modal-weighting layer (Section 2.5) through end-to-end backpropagation. VMD outputs from Section 2.3 and the filtered regime probability $\hat{\pi}_t^{(\ell)}$ from Section 2.2 enter as fixed inputs.

2.4.1. Tier-Level BiLSTM Encoders

For each tier ℓ and each origin t , the input matrix is constructed by stacking the most recent $L = 60$ trading days of per-tier IMFs and exogenous covariates:

$$X_t^{(\ell)} = \left[u_{1,\cdot}^{(\ell)}, \dots, u_{K,\cdot}^{(\ell)}, \mathbf{x}_{\cdot} \right] \in \mathbb{R}^{L \times (K+p)} = \mathbb{R}^{60 \times 10}. \quad (5)$$

A bidirectional LSTM with hidden size $d_h = 64$ produces forward and backward hidden states $H_t^{(\ell)} \in \mathbb{R}^{L \times 2d_h}$; the encoder output is the final concatenated hidden state $h_t^{(\ell)} \in \mathbb{R}^{2d_h}$. The encoder uses a single BiLSTM layer (`num_layers=1` in PyTorch version 2.3.0); the “two-layer” characterisation in the exogenous-only LSTM description of Section 2.10.2 refers to the corresponding baseline and has been corrected to “single-layer” in that section to match both the HVB-RA encoder and the reference implementation (`hvbra/model.py:TierEncoder`). The filtered regime probability $\hat{\pi}_t^{(\ell)}$ is *not* concatenated into the input matrix; it enters the modal-weighting layer of Section 2.5 as a separate scalar gating signal.

Per-mode summary vectors $m_{k,t}^{(\ell)} \in \mathbb{R}^{2d_h}$ for each of the K intrinsic mode functions are produced from the encoder's final hidden state through a learned gated linear projection. Specifically, a single linear layer maps $h_t^{(\ell)}$ to a $K \cdot 2d_h$ -dimensional vector that is reshaped to $(K, 2d_h)$ and passed through a hyperbolic tangent gate:

$$m_{k,t}^{(\ell)} = \tanh \left([W_{\text{mode}}^{(\ell)} h_t^{(\ell)} + b_{\text{mode}}^{(\ell)}]_{\text{slice } k} \right) \in \mathbb{R}^{2d_h}, \quad k = 1, \dots, K. \quad (6)$$

The per-mode summary captures the encoder's allocation of representational capacity across IMFs and enters the regime-informed aggregation of Section 2.5. The construction differs from per-mode encoders (one BiLSTM per IMF) by sharing the encoder across

modes and learning a downstream projection: this is more parameter efficient at the cost of imposing a common temporal representation across IMFs. *An important consequence of the shared-encoder design is that the k -th slice $m_{k,t}^{(\ell)}$ is not architecturally tied to IMF k ; it is a learned latent component that the rest of the network may, but is not constrained to, align with IMF k . The softmax weights w_t introduced in Section 2.5 therefore weight latent components, not IMFs directly, and should be interpreted as a regime-conditional latent-component-weighting profile rather than as a per-IMF importance distribution.*

2.4.2. Directed Cross-Tier Attention (Component C1)

The cross-tier attention layer routes information from upstream tiers into downstream encoder outputs along a topology supplied from domain knowledge. The topology is *not* learned from data: the empirical economic evidence that price discovery in vertically linked commodity markets propagates from upstream futures to downstream spot and farm-gate markets [1] is encoded as an architectural prior. The attention queries, keys, values, and the resulting per-time-step attention intensities *are* learned from data. The contribution lies in combining a domain-imposed direction with data-driven intensities, not in discovering direction.

Concretely, the futures-tier representation is taken as a source and is not modified by the cross-tier layer:

$$\tilde{h}_t^{(F)} = h_t^{(F)}. \quad (7)$$

The regional-spot representation attends to the futures-tier representation through a single-head scaled dot-product attention with a residual connection (hvbra/model.py: CrossTierAttention, instance attn_R):

$$\tilde{h}_t^{(R)} = h_t^{(R)} + \text{Attn}(Q = h_t^{(R)} W^Q, K = [h_t^{(F)}] W^K, V = [h_t^{(F)}] W^V). \quad (8)$$

The farm-gate-tier representation attends to both upstream tiers (instance attn_G):

$$\tilde{h}_t^{(G)} = h_t^{(G)} + \text{Attn}(Q = h_t^{(G)} W^Q, K = [h_t^{(F)}; h_t^{(R)}] W^K, V = [h_t^{(F)}; h_t^{(R)}] W^V), \quad (9)$$

where the bracket notation $[\cdot; \cdot]$ denotes stacking the two upstream representations as two tokens of dimension $2d_h$ each. The projection dimensions are $W^Q, W^K \in \mathbb{R}^{2d_h \times d_{\text{attn}}}$ with $d_{\text{attn}} = 32$, and $W^V \in \mathbb{R}^{2d_h \times 2d_h}$.

Degenerate Attention at Tier R

A subtle property of Equation (8) is that the regional-spot attention reduces to a residual linear projection in this single-key configuration. With only one upstream key $K = h_t^{(F)} W^K$ of dimension $(1, d_{\text{attn}})$, the softmax over the key axis trivially returns the scalar 1, and the resulting attention output is $V = h_t^{(F)} W^V$. The Tier R update therefore simplifies algebraically to

$$\tilde{h}_t^{(R)} = h_t^{(R)} + h_t^{(F)} W^V, \quad (10)$$

a residual linear fusion rather than a non-trivial attention selection. The two-key attention at Tier G (Equation (9)) is non-degenerate. This algebraic property of Tier R is documented openly and is discussed as a design limitation in Section 4; the natural follow-up is to replace Tier R's single-key attention with temporal-key attention over the full upstream sequence $H_t^{(F)} \in \mathbb{R}^{L \times 2d_h}$.

2.5. Two-Stage Regime-Informed Modal-Weighting Layer (Component C2)

Module type. Differentiable neural module, trained jointly with the encoder (Section 2.4). The two regime-conditional IMF weight profiles $v_0^{(\ell)}, v_1^{(\ell)} \in \mathbb{R}^K$ are free parameters and are updated by backpropagation; the gating signal $\hat{\pi}_t^{(\ell)}$ is supplied as a fixed scalar input from the Markov-switching module of Section 2.2 and is not back-propagated into the regime classifier. The architecture is therefore two-stage by design.

This component is the principal methodological novelty of the paper. The standard practice in VMD-deep learning forecasting pipelines aggregates IMFs through a single fixed projection that holds modal importance constant across the sample. The present construction replaces this with a layer that conditions the IMF aggregation on the latent volatility regime, implemented as a two-expert mixture-of-experts.

2.5.1. Layer Construction

For each tier ℓ , the layer maintains two learnable parameter vectors $v_0^{(\ell)}, v_1^{(\ell)} \in \mathbb{R}^K$. The regime-conditional IMF weight vector at origin t is a convex combination of two softmax profiles, gated by the filtered Markov-switching probability:

$$w_t^{(\ell)} = (1 - \hat{\pi}_t^{(\ell)}) \cdot \text{softmax}(v_0^{(\ell)}) + \hat{\pi}_t^{(\ell)} \cdot \text{softmax}(v_1^{(\ell)}) \in \Delta^{K-1}, \quad (11)$$

where Δ^{K-1} is the $(K-1)$ -dimensional probability simplex. The regime-weighted modal representation is then formed by aggregating the per-mode summary vectors from Equation (6):

$$z_t^{(\ell)} = \sum_{k=1}^K w_{k,t}^{(\ell)} \cdot m_{k,t}^{(\ell)} \in \mathbb{R}^{2d_h}. \quad (12)$$

The implementation is `hvbra/model.py:RegimeModalWeighting`.

2.5.2. Layer Properties

Three properties of Equation (11) are worth noting.

Differentiability and scope. The regime probability $\hat{\pi}_t$ enters as a known scalar input; both v_0 and v_1 are free parameters trained by backpropagation through the standard MSE loss of Equation (14) below. The layer is differentiable with respect to v_0, v_1 . The layer is not differentiable with respect to the Markov-switching parameters $\{\mu_s, \phi_s, \sigma_s, p_{ij}\}$, which are estimated by EM in Section 2.2; in particular, the gradient of the forecast loss does not flow back into the regime classifier. This two-stage design keeps the regime classifier identifiable as a stand-alone interpretable model, at the cost of foregoing joint estimation of the regime parameters and the neural backbone.

Smooth interpolation. When the market is unambiguously in one regime ($\hat{\pi}_t \rightarrow 0$ or $\hat{\pi}_t \rightarrow 1$), the weight vector w_t collapses to the corresponding regime profile; during regime transitions, it interpolates smoothly as a function of $\hat{\pi}_t$. The aggregation is therefore continuous in the gating signal, in contrast with hard-assignment schemes that would discretise $\hat{\pi}_t$ at a threshold.

Nested benchmark. The layer reduces exactly to a standard fixed-weight latent-component aggregation $z = \sum_k \text{softmax}(v)_k \cdot m_k$ in the special case $v_0 = v_1$. This provides a clean nested benchmark for the ablation E1 in Section 2.10.3: forcing $v_0 = v_1$ during training disables the regime-conditional component of the layer while preserving all other architectural elements.

2.6. Forecast Head and Training Loss

The cross-tier-attended representation $\tilde{h}_t^{(\ell)}$ from Equation (8) or Equation (9) and the regime-weighted modal representation $z_t^{(\ell)}$ from Equation (12) are concatenated and passed

through a linear output head. For each (tier, horizon) pair, a separate linear head produces a log-price forecast at $t + h$:

$$\hat{y}_{t+h}^{(\ell)} = W_{\text{out}}^{(\ell,h)} [\tilde{h}_t^{(\ell)}; z_t^{(\ell)}] + b_{\text{out}}^{(\ell,h)}, \quad (13)$$

with horizons $h \in \{1, 5, 21\}$ trading days. The total of $|\mathcal{H}| \times 3 = 9$ heads (three horizons, three tiers) is implemented as `hvbra/model.py:HVBRA.heads`.

The aggregate training loss is a horizon-weighted mean-squared error in log-price space:

$$\mathcal{L}(\theta) = \frac{1}{|\mathcal{H}| \cdot 3} \sum_{\ell \in \{F,R,G\}} \sum_{h \in \mathcal{H}} \lambda_h \cdot \text{MSE}(\hat{y}_{t+h}^{(\ell)}, y_{t+h}^{(\ell)}), \quad (14)$$

with horizon weights $(\lambda_1, \lambda_5, \lambda_{21}) = (0.5, 0.3, 0.2)$ that place the highest weight on the short ($h = 1$) horizon (`hvbra/model.py:horizon_balanced_loss`). This choice reflects the operational priority of accurate short-term forecasts and is held constant across all experiments; sensitivity to alternative horizon weights is identified as future work.

Log-to-Level Conversion

The training loss in Equation (14) is computed in log-price space because differenced log prices are approximately stationary and the MSE on log returns is the standard training objective for commodity price forecasts. Before reporting accuracy metrics, applying the sample-ratio constraint projection, or running the decision-loss simulation, the log-price forecasts $\hat{y}_{t+h}^{(\ell)}$ are exponentiated back to USD/kg per-tier price-level forecasts $\hat{p}_{t+h}^{(\ell)} = \exp(\hat{y}_{t+h}^{(\ell)})$. Accordingly, RMSE in Sections 3.2 and 3.3, the constraint matrix A in Section 2.7, the residual covariance matrices W , W_0 , W_1 in Equations (17) and (19), and the decision rule in Section 2.11 are all defined in price-level (USD/kg) space.

2.7. Auxiliary Sample-Ratio Constraint Projection (Component C3)

Module type. Post hoc linear projection applied to the joint vector of three-tier base forecasts after the neural backbone is trained. The projection matrices are estimated offline on calibration-period forecast residuals; no neural-backbone parameters are updated during this step.

The three tiers are jointly observed and none is a deterministic aggregate of the others (Section 2.1). Standard hierarchical reconciliation as developed for aggregation-based hierarchies [25] therefore does not apply directly: there is no summation matrix S and no aggregation identity. The framework substitutes an alternative descriptive constraint based on the long-run sample-mean ratios across tiers. Specifically, given training-window sample means $\bar{F}$, $\bar{R}$, $\bar{G}$, the level ratios

$$a_R \equiv \bar{F}/\bar{R}, \quad a_G \equiv \bar{R}/\bar{G}, \quad (15)$$

are descriptive properties of the training window. The framework adopts these ratios as the rows of a 2×3 constraint matrix A that defines a coherence relation $A\tilde{y}_t = 0$:

$$A = \begin{bmatrix} 1 & -a_R & 0 \\ 0 & 1 & -a_G \end{bmatrix}, \quad A\tilde{y}_t = 0 \iff F = a_R \cdot R \text{ and } R = a_G \cdot G. \quad (16)$$

The level ratios in Equation (15) are sample means, *not* cointegrating coefficients estimated from a formal cointegration test, and the framework does not claim that they represent an economic equilibrium. The projection is reported as an auxiliary post hoc adjustment that enforces an operational scaling constraint between tiers, complementing

the cross-tier attention and modal-weighting components rather than serving as a primary methodological contribution.

2.7.1. Static Constraint Projection

Under the static constraint projection, the reconciled forecast is

$$\tilde{y}_t = \hat{y}_t - \Psi(W) \hat{y}_t, \quad \Psi(W) = W A^\top (A W A^\top)^{-1} A, \quad (17)$$

with W as a 3×3 covariance matrix estimated from calibration-window forecast residuals. The construction takes its algebraic form from the constraint-projection presentation of MinT [25] but does not import MinT's loss-trace optimality interpretation: that interpretation requires an aggregation hierarchy and the present system does not have one. The implementation, including diagonal shrinkage with $\lambda = 0.05$ for numerical stability, is in `hvbra/reconcile.py:reconcile_static`.

2.7.2. Regime-Conditional Constraint Projection

The calibration-window residuals $e_t = y_t - \hat{y}_t$ are partitioned by the filtered regime probability at the farm-gate tier, $\hat{\pi}_t^{(G)}$, at threshold 0.5. Two regime-specific covariance matrices are estimated:

$$W_0 = \text{Cov}(e_t \mid \hat{\pi}_t^{(G)} < 0.5), \quad W_1 = \text{Cov}(e_t \mid \hat{\pi}_t^{(G)} \geq 0.5), \quad (18)$$

with the fallback rule (in `hvbra/reconcile.py:estimate_regime_covariances`) that if fewer than five observations fall in a given regime, the covariance for that regime is set to the unconditional covariance estimated on all calibration residuals. Both matrices are diagonally shrunk with $\lambda = 0.05$. The regime-conditional reconciled forecast at test origin t is then

$$\tilde{y}_t = \hat{y}_t - \left[(1 - \hat{\pi}_t^{(G)}) \Psi(W_0) + \hat{\pi}_t^{(G)} \Psi(W_1) \right] \hat{y}_t. \quad (19)$$

Properties of the Regime-Conditional Projection

Two properties of Equation (19) merit attention.

Coherence preservation. For each regime-specific correction operator, the identity $A \Psi(W_i) = A$ holds, $i \in \{0, 1\}$, by direct algebra: $A \Psi(W_i) = A W_i A^\top (A W_i A^\top)^{-1} A = A$. Substituting Equation (19) into $A \tilde{y}_t$,

$$A \tilde{y}_t = A \hat{y}_t - \left[(1 - \hat{\pi}_t) A \Psi(W_0) + \hat{\pi}_t A \Psi(W_1) \right] \hat{y}_t = A \hat{y}_t - \left[(1 - \hat{\pi}_t) A + \hat{\pi}_t A \right] \hat{y}_t = 0.$$

The constraint $A \tilde{y}_t = 0$ is therefore enforced exactly for every $\hat{\pi}_t \in [0, 1]$. Note that the convex combination $(1 - \hat{\pi}_t) \Psi(W_0) + \hat{\pi}_t \Psi(W_1)$ is *not* in general idempotent and is therefore not itself a projection matrix; it is more precisely a coherence-preserving linear adjustment.

Non-optimality. The operator does *not* in general minimise the trace of the reconciled-error covariance under the regime-mixed covariance $W_t = (1 - \hat{\pi}_t) W_0 + \hat{\pi}_t W_1$. A trace-optimal regime-conditional reconciliation would require applying the projection to the mixed covariance directly, $\Psi(W_t)$, which is in general distinct from the convex combination $(1 - \hat{\pi}_t) \Psi(W_0) + \hat{\pi}_t \Psi(W_1)$ because Ψ is a nonlinear function of W . The convex-mix form implemented here preserves the coherence constraint by construction but is not claimed to be trace-optimal; it is reported as a coherence-preserving regime-conditional heuristic.

2.8. Training Protocol

The encoder of Section 2.4, the modal-weighting layer of Section 2.5, and the forecast head of Section 2.6 are trained jointly by minimising the loss in Equation (14) through backpropagation. The training configuration is summarised in Table 2 and matches the reference configuration in hvbra/configs/hvbra_main.yaml.

Table 2. HVB-RA training configuration. All settings match the reference configuration in hvbra/configs/hvbra_main.yaml.

Setting	Value
Optimiser	AdamW, learning rate 10^{-3} , weight decay 10^{-4}
Batch size	64
Maximum epochs	100
Sequence length L	60 trading days
BiLSTM hidden size d_h	64
Attention head dimension	$d_{\text{attn}} = 32$ (single head)
Dropout (BiLSTM/attention)	0.2/0.1
Gradient clipping	Global norm 1.0
Early stopping	Patience 20 epochs on calibration loss
Decision-loss threshold	$\tau = 0.005$
Random seeds	5 seeds (3407, 42, 1234, 2024, 7777)

The total parameter count of HVB-RA is approximately 4.16×10^5 , broken down as three tier encoders (approximately 1.21×10^5 parameters each), two cross-tier attention layers (approximately 2.48×10^4 parameters each), three regime-aware modal-weighting layers ($2K = 10$ parameters each), and nine output heads (257 parameters each). The Markov-switching regime classifier of Section 2.2 adds approximately 30 parameters per tier estimated outside the gradient loop.

2.9. Framework Summary

Table 3 summarises the six framework components by module type and parameter-estimation regime. Figure 1 provides a dataflow overview of the architecture: each row corresponds to one tier; columns trace the flow from raw price input through VMD, BiLSTM encoding, cross-tier attention, regime-informed modal weighting, tier-level forecast head, and the auxiliary constraint projection. Yellow arrows represent the directed cross-tier attention (Section 2.4.2); red dashed arrows represent the filtered Markov-switching probability $\hat{\pi}_t$ supplied as a gating signal to the modal-weighting layer (Section 2.5) and to the regime-conditional projection (Section 2.7.2). The bottom panel illustrates the decision-loss simulation at the farm-gate tier specified in Section 2.11. The complete correspondence between each component and the reference implementation is documented in the reproducibility supplement.

Table 3. HVB-RA framework components by module type and parameter-estimation regime. Two components are estimated offline; three components are trained jointly through backpropagation; one component is applied post hoc.

Section No.	Component	Module Type	Where Parameters Come from
Section 2.3	Per-tier VMD decomposition	Signal processing, offline	ADMM on training-window prices, rolling re-fit on calibration and test
Section 2.2	Markov-switching regime detection	Classical statistical, offline	Expectation-maximisation on training-window returns
Section 2.4.1	Tier-level BiLSTM encoders	Differentiable neural, joint training	Backpropagation on Equation (14)

Table 3. Cont.

Section No.	Component	Module Type	Where Parameters Come from
Section 2.4.2	Directed cross-tier attention (C1)	Differentiable neural, joint training	Backpropagation on Equation (14)
Section 2.5	Regime-informed modal weighting (C2)	Differentiable neural, joint training	Backpropagation on Equation (14)
Section 2.7	Sample-ratio constraint projection (C3)	Linear projection, post hoc	Calibration-window residuals

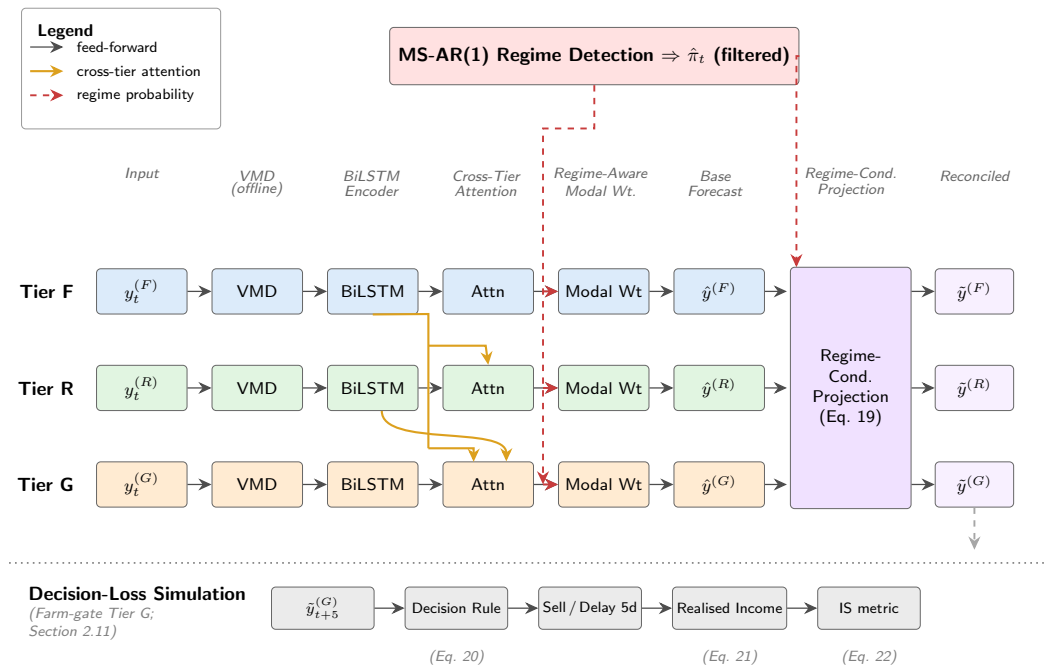


Figure 1. HVB-RA architecture dataflow. The three tier rows (futures F, regional R, farm-gate G) each pass input prices through offline VMD, a BiLSTM encoder, cross-tier attention, the regime-informed modal-weighting layer, and a per-tier forecast head. Yellow arrows trace the directed cross-tier attention (R attends to F; G attends to F and R). Red dashed arrows trace the filtered Markov-switching state probability $\hat{\pi}_t$, which feeds the modal-weighting layer (Equation (11)) and the auxiliary regime-conditional projection (Equation (19)). The bottom panel shows the decision-loss simulation at the farm-gate tier: the $h = 5$ forecast drives a five-day sell/delay rule, and the income-smoothing (IS) metric measures variance reduction in realised income.

2.10. Baselines and Evaluation Protocol

Evaluation follows a rolling-origin out-of-sample protocol over the test window. For each origin t in the test window the model produces base and reconciled forecasts at horizons $h \in \{1, 5, 21\}$ for each of the three tiers. The model parameters and the Markov-switching parameters are *not* re-fitted within the test window; only the rolling VMD and the filtered regime probability are updated as new observations become available. This protocol reflects a realistic deployment scenario in which a model is periodically retrained but used continuously between retrains.

2.10.1. Reported Metrics

The framework reports three point-error metrics per (tier, horizon, and method) combination, all implemented in `hvb/ra/baselines.py`:

RMSE

Root mean squared error on price levels in USD/kg is used as the primary metric for within-tier method comparison.

MAE

Mean absolute error on price levels in USD/kg is reported as a robust complement to RMSE, particularly during regime transitions where squared-error metrics may be dominated by a few large residuals.

MASE

Mean absolute scaled error of Hyndman and Koehler [28] is reported against the in-sample one-step random-walk forecast, providing a scale-free comparison across tiers and horizons.

Statistical significance of pairwise method differences in squared-error loss is assessed via Diebold–Mariano tests with the Harvey–Leybourne–Newbold small-sample correction [29], with the long-run variance computed as the Newey–West estimator truncated at $h-1$ lags. The implementation is `hvbra/baselines.py:diebold_mariano`.

2.10.2. Baselines

Three external baselines are evaluated under identical splits, identical aligned-origin set, and identical exogenous covariates.

Random walk (no-change). $\hat{p}_{t+h} = p_t$. Sanity floor. Implemented in `hvbra/baselines.py: random_walk_forecast`.

ARIMA(2,0,2). On log-returns, fitted once on combined training and calibration windows, with iterated multi-step forecasts. Implemented in `hvbra/baselines.py: arima_baseline_forecast` with the explicit (2,0,2) order argument; the default order (1,1,1) is overridden in the production configuration to match the model selection of Pinitjitsamut [6].

Exogenous-only LSTM. Single-direction, single-layer, hidden size 64, trained on the five exogenous covariates only (no VMD, no regime conditioning, no lagged tier prices), five seeds per (tier, horizon) cell. The baseline deliberately excludes lagged price inputs in order to provide a controlled contrast that isolates the contribution of the VMD-decomposed price representation, the cross-tier attention, and the regime-informed weighting in HVB-RA; a price-augmented autoregressive variant is identified as a follow-up baseline in Section 4.6 (Limitation 7).

Comparison Fairness

The same total observation window (train + calibration, 1740 trading days) is available to every method but is consumed differently across model families. ARIMA, which has no early-stopping mechanism, is fitted on the union of training and calibration windows so that the full sample is used for parameter estimation. Neural methods (HVB-RA and Exogenous-only LSTM) are fitted on the training window with calibration used for early stopping on out-of-sample loss; the calibration sample therefore enters the neural pipeline as a model-selection signal rather than as additional fitting data. The two families thus have access to the same prior information at every forecast origin, but their effective parameter-estimation samples are not identical. Test observations (2 January 2025–4 March 2026) are held out from both families and are never used for parameter fitting or stopping. An alternative protocol—refit the neural models on the union of training and calibration after selecting the early-stopping epoch on the validation split—would equalise the parameter-estimation samples; it was not adopted here because the production runs were configured with the standard early-stopping protocol used by deep learning forecasters in the same literature [7,8,10].

A single-tier VMD-BiLSTM control (the encoder backbone of HVB-RA without cross-tier attention, without the modal-weighting layer, and without the auxiliary constraint

projection) is implemented in the reference package as ablation E2 + E1 combined, and its results are reported in the companion repository together with the additional ablations (E5, E7).

Direct comparison against current transformer-class hierarchical forecasters—TFT, N-BEATSx, DeepAR, NHITS, PatchTST, and graph-based multivariate forecasters such as MTGNN—is scoped to a follow-up benchmark study using the same data and the same aligned-origin set; this is discussed in Section 4.

2.10.3. Component Ablations

Table 4 reports a component-ablation summary on the synthetic three-tier price system generated by `hvbra/data.py` (three random seeds, 600 synthetic trading days, 70/10/20 train/calibration/test split, `seq_len=30`, `epochs=8`). The purpose of this table is structural verification: it demonstrates that the ablation contrasts are well-defined, that the code paths execute correctly, and that the direction of effect is qualitatively consistent with the architectural claims. Quantitative interpretation of these figures requires real-data runs; the per-component results on the manuscript's three-tier rubber dataset for the two primary ablations (E1, E2) are reported in the main text (Section 3.4), with the broader ablation set in the companion repository (Section 3.7).

Table 4. Component ablation on synthetic data (3 seeds, mean RMSE in log-price space). Full = HVB-RA as described; E1 = no regime gating ($\mathbf{v}_0$ tied to $\mathbf{v}_1$); E2 = no cross-tier attention. Synthetic data generated by `hvbra/data.py` with two-regime volatility; results are not transferable to the real-data evaluation of Section 3.

Variant	Tier F			Tier R			Tier G		
	$h = 1$	$h = 5$	$h = 21$	$h = 1$	$h = 5$	$h = 21$	$h = 1$	$h = 5$	$h = 21$
Full	0.116	0.116	0.126	0.117	0.115	0.129	0.141	0.130	0.117
E1 (no regime)	0.116	0.116	0.126	0.117	0.115	0.129	0.141	0.130	0.117
E2 (no attn)	0.121	0.123	0.128	0.115	0.116	0.130	0.133	0.122	0.112

Note: E1 $\approx$ Full is expected: the synthetic training window is also predominantly single-regime (numpy-EM fallback), so the regime-conditioning contrast is non-identified on this data—consistent with Section 3.4 Finding 4. E2 degrades Tier F $h = 1$ RMSE by $\approx 4\%$ relative to Full on synthetic data where upstream-to-downstream transmission is active, consistent with proposition P1.

Four controlled ablations of HVB-RA are defined to isolate the marginal contribution of each architectural component. The ablations are documented here so that the reference implementation supports them; the primary per-component results (E1, E2) on the real-data series are reported in the main text (Section 3.4), and the remaining ablation results (E5, E7) in the companion repository (see Section 3.7).

- E1 No regime gating.** Force $\mathbf{v}_0 = \mathbf{v}_1$ during training. The regime-informed modal-weighting layer (Section 2.5) reduces to a fixed-weight aggregation $z = \sum_k \text{softmax}(\mathbf{v})_k \cdot m_k$ and the filtered gating signal $\hat{\pi}_t$ no longer affects forecasts. This isolates proposition P2 of Section 1.3.
- E2 No cross-tier attention.** Bypass the cross-tier attention layers: $\tilde{h}_t^{(R)} = h_t^{(R)}$ and $\tilde{h}_t^{(G)} = h_t^{(G)}$. Each tier's encoder produces its forecast from its own VMD inputs and exogenous covariates only. This isolates proposition P1 of Section 1.3.
- E5 No VMD.** Replace the five IMF inputs with the raw price series (or a wider window of recent prices). Tests whether the per-tier decomposition is responsible for the encoder's signal.

E7 No constraint projection. Use the unreconciled base forecasts directly, without applying the post hoc projection of Section 2.7. Tests whether the auxiliary projection improves out-of-sample RMSE.

As noted in Section 3 (Finding 4), the regime-conditioning aspect of ablation E1 is formally non-identified on the present test window because every calibration and test origin falls in the high-volatility regime. The per-cell results for the two primary ablations E1 and E2 are reported in the main text in Section 3.4; the additional ablations E5 and E7, together with two further regime-gating identification tests (E6: constant $\hat{\pi}_t = 0.5$; E8: $\Psi(W_1)$ only) and the suite of complementary controls discussed in Section 4 (reverse-direction attention, unrestricted attention, parameter-matched controls, shuffled-gate tests), are reported in the companion repository and discussed in Section 4.

2.11. Decision-Loss Simulation

The decision-loss simulation operationalises a supporting evaluation question: whether the $h = 5$ farm-gate forecasts produced by HVB-RA, when translated into a five-day selling rule, deliver a measurable reduction in realised-income variance for a representative smallholder producer. The simulation is reported as a methodological template; the substantive headline finding is reported in Section 3.6.

Decision Rule

A representative price-taking agent at the farm-gate tier observes the five-day-ahead farm-gate forecast $\hat{p}_{t+5}^{(G)}$ and the current price $y_t^{(G)}$, and chooses between selling at the current price and delaying sale by five trading days. The forecast supplied to the decision rule is the *unreconciled base forecast* from the HVB-RA farm-gate forecast head, i.e., the per-tier price-level forecast $\hat{p}_{t+5}^{(G)} = \exp(\hat{y}_{t+5}^{(G)})$ before the auxiliary constraint projection of Section 2.7 is applied. The auxiliary constraint projection is evaluated separately in Section 3.5 and does not enter the decision rule. The decision rule is a myopic risk-neutral comparison adjusted for a transaction-cost threshold τ :

$$\text{sell at } t \iff \hat{p}_{t+5}^{(G)} < (1 + \tau) \cdot y_t^{(G)}. \quad (20)$$

The baseline transaction-cost threshold is $\tau = 0.005$ (50 basis points), reflecting a conservative estimate for non-bulk smallholder sales; sensitivity to $\tau \in \{0, 0.0025, 0.005, 0.01\}$ is reported in Section 3.7. Implementation in `hvb/decision.py`.

Realised Income

Daily realised income (per kilogram) under the rule is

$$I_t = \mathbf{1}\{\text{sell at } t\} \cdot y_t^{(G)} + \mathbf{1}\{\text{delay}\} \cdot y_{t+5}^{(G)}, \quad (21)$$

aggregated to monthly income $\{I_m\}$ by summation within each calendar month over the test window.

Income-Smoothing Metric

The income-smoothing (IS) metric measures variance reduction in monthly realised income relative to a benchmark policy that sells every day:

$$\text{IS}(\text{model}) = 1 - \frac{\text{Var}(I_m^{\text{model}})}{\text{Var}(I_m^{\text{naive}})}. \quad (22)$$

Values $IS > 0$ indicate variance reduction. The metric is reported alongside the number of delay recommendations n_{delay} per seed; near-degenerate operating points (either $n_{\text{delay}} \rightarrow 0$ or $n_{\text{delay}} \rightarrow N$) are reported as such and not aggregated into a single headline statistic.

2.12. Reproducibility

All data sources are publicly listed in Section 2.1; the daily preprocessing pipeline, the HVB-RA reference implementation, the Markov-switching fitting code, the auxiliary constraint projection, and the decision-loss simulation are provided as a reproducibility package `hvbra/` together with this manuscript. The package includes the model architecture (`hvbra/model.py`), the per-tier VMD wrapper (`hvbra/vmd.py`), the Markov-switching regime detection (`hvbra/regime.py`), the sample-ratio constraint projection (`hvbra/reconcile.py`), the baselines and accuracy metrics including the Diebold–Mariano test (`hvbra/baselines.py`), the decision-loss simulation (`hvbra/decision.py`), and the training orchestration (`hvbra/train.py`). Random seeds, hyperparameter values, and software versions are recorded in the configuration files provided with the code (`hvbra/configs/hvbra_main.yaml`). All computational experiments were implemented in Python version 3.11 (Python Software Foundation, Wilmington, DE, USA). The neural-network components were implemented using PyTorch version 2.3.0 (Meta Platforms, Inc., Menlo Park, CA, USA). Data preprocessing, numerical computation, and model evaluation were conducted using NumPy version 1.26.4, pandas version 2.2.2, scikit-learn version 1.5.0, statsmodels version 0.14.2, and the VMD implementation included in the accompanying `hvbra/` reproducibility package. The reference implementation was executed using the `hvbra/` package supplied with this manuscript. Per-seed predictions, fitted Markov-switching parameters, calibration-window covariance matrices, and decision-loss outputs are persisted to enable third-party reproduction of every numerical result reported in Sections 3 and 3.7. The reproducibility supplement provides a line-level correspondence between every component of this Methods section and the provided implementation.

3. Results

3.1. Experimental Setup

All experiments use the daily multi-tier rubber price dataset described in Section 2.1 (2038 trading-day observations spanning 2 May 2018 to 4 March 2026). Table 5 summarises the calendar-based partition of the dataset into training, calibration, and test windows for reproducible evaluation.

Table 5. Calendar-based partition of the daily three-tier price dataset.

Split	Date Range	N (Days)
Training	2 May 2018 to 30 June 2023	1348
Calibration	3 July 2023 to 31 December 2024	392
Test	2 January 2025 to 4 March 2026	298

Forecasts are evaluated at three horizons $h \in \{1, 5, 21\}$ trading days, corresponding to short-term, weekly, and monthly forecast windows. HVB-RA is trained with the AdamW optimiser (learning rate 1×10^{-3} , weight decay 1×10^{-4}), batch size 64, gradient-norm clipping at 1.0, dropout 0.2 on the BiLSTM and 0.1 on the cross-tier attention, and early stopping with patience 20 epochs evaluated on the calibration window. Each configuration is run with five seeds $\{3407, 42, 1234, 2024, 7777\}$ and results are reported as mean $\pm$ standard deviation across seeds.

Each tier-specific encoder receives a 10-dimensional per-time-step input matrix consisting of five intrinsic-mode functions from per-tier Variational Mode Decomposition and five exogenous covariates (USD/THB log-return, Brent log-return, standardised ONI, China-auto-sales year-on-year growth, and Qingdao-stock log-change). The filtered Markov-switching state probability $\hat{\pi}_t$ does not enter the input channel matrix; it is supplied to each tier's regime-aware modal-weighting layer as a separate scalar gating signal (Methods, Equation (11)).

Reported Metrics

The primary point-error metric reported for every tier-horizon-method combination in the manuscript is root mean squared error (RMSE) on the price level in USD/kg. Complementary metrics produced natively by the reference implementation (hvbra/baselines.py)—the mean absolute error (MAE) and the scale-free mean absolute scaled error (MASE) of Hyndman and Koehler [28] computed against the in-sample one-step random-walk forecast, as well as Diebold–Mariano [30] tests with the Harvey–Leybourne–Newbold small-sample correction [29] for pairwise significance of squared-error differentials—are reported in the companion repository submitted as Supplementary Material, together with the broader component ablations; the primary per-component ablations (E1, E2) are reported in the main text in Section 3.4. The present results focus on RMSE because it is the primary point-error metric in commodity-price forecasting and is directly interpretable in USD/kg without scale adjustment.

The test partition contains 298 daily observations (2 January 2025 to 4 March 2026). To enable horizon-comparable metrics, all forecasts are evaluated on a common set of $N = 278$ aligned test origins, defined as origins t for which all three target horizons $\{p_{t+1}, p_{t+5}, p_{t+21}\}$ are observed within the test window.

3.2. Base Forecast Accuracy

Table 6 reports RMSE of HVB-RA price-level forecasts across the three tiers (F: global futures, R: regional spot, G: farm-gate) at the three horizons in USD/kg.

Table 6. HVB-RA base forecast accuracy on the held-out test set ($N = 278$ aligned test origins). Values reported as mean $\pm$ standard deviation across 5 seeds. RMSE in USD/kg on the price level.

Tier $\times$ Horizon	RMSE (USD/kg)
F (global futures), $h = 1$	0.0356 ± 0.0016
F, $h = 5$	0.0684 ± 0.0048
F, $h = 21$	0.1226 ± 0.0245
R (regional spot), $h = 1$	0.0386 ± 0.0002
R, $h = 5$	0.0768 ± 0.0103
R, $h = 21$	0.1451 ± 0.0356
G (farm-gate), $h = 1$	0.0377 ± 0.0014
G, $h = 5$	0.0799 ± 0.0130
G, $h = 21$	0.1347 ± 0.0372

3.2.1. Horizon Dependence

Two patterns are visible in the RMSE column of Table 6. **First**, RMSE increases monotonically with horizon for every tier: from 0.036–0.039 USD/kg at $h = 1$ to 0.12–0.15 USD/kg at $h = 21$. This is the standard expansion of forecast uncertainty with horizon and is observed across all baselines as well (see Section 3.3).

Second, the across-seed standard deviation at $h = 1$ is small ($\sigma \approx 0.0002$ – 0.002 USD/kg) and grows by roughly an order of magnitude at $h = 21$ ($\sigma \approx 0.025$ – 0.037 USD/kg). The R- $h = 21$ cell has the largest seed-to-seed variation (0.1451 ± 0.0356 , a coefficient of variation of 25%). Larger across-seed variation at long horizons is consistent with the archi-

texture's higher sensitivity to initialisation when the prediction task has lower deterministic signal; this is not a substantive performance claim but a diagnostic note.

3.2.2. Variance Diagnostic

The Standard Deviation Ratio $\text{StdR} = \text{std}(\text{forecast}) / \text{std}(\text{realised})$ of Pinitjitsamut [6] is a useful complement to RMSE for diagnosing variance collapse, particularly when forecast skill is low and the model defaults towards a conditional-mean prediction. StdR is not computed natively by the present manuscript-reported pipeline (`hvbra/baselines.py` reports RMSE, MAE, and MASE only); the companion repository extends the metric set with a native StdR computation and reports the quantitative per-tier StdR values together with diagnostic plots. The precursor analysis of Pinitjitsamut [6], on the same data and a similar single-tier VMD-BiLSTM backbone, reported StdR values of 1.091 ± 0.060 , 0.34, and 0.11 on input matrices of 24, 11, and 22 features respectively; the relationship between feature-set composition and variance fidelity in that work was non-monotonic and is not used here as a directional prediction for the present 10-dimensional per-tier input matrix. The hypothesis that the present setting exhibits material variance shrinkage is therefore left for the companion-repository StdR computation to confirm or refute, and the discussion of variance fidelity in Section 4 is presented as a hypothesis rather than as an established empirical pattern.

3.3. Benchmark Comparison

To position HVB-RA against established forecasting baselines, three external benchmarks and one contemporary deep learning benchmark are compared on the same $N = 278$ aligned test origins and the same target definition: **Random Walk** (no-change forecast $\hat{p}_{t+h} = p_t$, implemented in `hvbra/baselines.py:random_walk_forecast`), **ARIMA(2,0,2)** on log-returns, a **Exogenous-only LSTM** (single-direction, two-layer, hidden size 64, exogenous covariates only, no VMD and no regime conditioning, 5 seeds), and **NHITS** [10] (a contemporary neural hierarchical-interpolation forecaster, supplied with the same five exogenous covariates as historical inputs and evaluated on the identical origins, 5 seeds). Random Walk is deterministic given the data; ARIMA is fitted once on train + calibration; Exogenous-only LSTM and NHITS are trained with the same five seeds as HVB-RA. Table 7 reports RMSE in USD/kg for each method.

Table 7. Benchmark RMSE comparison in USD/kg on $N = 278$ aligned test origins. NHITS, Exogenous-only LSTM, and HVB-RA values are mean across 5 seeds. Random walk is deterministic given the data; ARIMA is fitted once on train + calibration. Bold marks the best entry per row.

Tier	h	Random Walk	ARIMA(2,0,2)	NHITS	Exog-LSTM	HVB-RA
F	1	0.0347	0.0347	0.0372	0.0353	0.0356
F	5	0.0639	0.0641	0.0737	0.0640	0.0684
F	21	0.1092	0.1097	0.1330	0.1086	0.1226
R	1	0.0379	0.0379	0.0422	0.0389	0.0386
R	5	0.0703	0.0703	0.0908	0.0726	0.0768
R	21	0.1210	0.1210	0.1937	0.1237	0.1451
G	1	0.0364	0.0364	0.0404	0.0367	0.0377
G	5	0.0714	0.0714	0.0911	0.0729	0.0799
G	21	0.1139	0.1139	0.1753	0.1186	0.1347

The benchmark comparison establishes three observations:

1. **HVB-RA uniformly outperforms the contemporary deep learning baseline.** Against NHITS, supplied with the same five exogenous covariates and evaluated on the identical aligned test origins, HVB-RA attains lower RMSE in all nine of nine tier-

horizon cells, with the margin widening at long horizons ($R-h = 21$: 0.1451 versus 0.1937; $G-h = 21$: 0.1347 versus 0.1753). This establishes that HVB-RA, taken as a complete configuration, attains lower point error than a current deep forecaster on the same information set; the marginal point-error contribution of the individual components is examined separately in the component ablation of Section 3.4.

2. **No method improves on the random-walk floor at most cells.** Under the residual target definition the no-change forecast is the natural zero-prediction baseline. No method—classical, exogenous, contemporary, or the proposed framework—reduces RMSE below the random-walk floor at the majority of tier-horizon cells. This is the expected behaviour for daily commodity prices under weak-form efficiency and is reported here without qualification. The proposed framework does not claim to dominate the near-floor classical baselines on point error; its demonstrated advantage is over the contemporary deep learning baseline.
3. **The farm-gate tier remains well-described by short-memory dynamics.** ARIMA(2,0,2) and the random walk are strongest at the farm-gate tier, consistent with the relatively slower information flow there, where linear short-memory dynamics provide a sufficient statistical model. This is also consistent with prior single-tier evidence for the same commodity: Pinitjitsamut [6] reports a competitive ARIMA baseline on a single-tier rubber-price forecasting task.

Statistical Significance of the RMSE Differences

The benchmark comparison reported in Table 7 is on point estimates of RMSE. Formal pairwise tests of forecast accuracy—Diebold–Mariano [30] with the Harvey–Leybourne–Newbold small-sample correction [29], with horizon-matched Newey–West variance and Holm [31] adjustment across the nine cells—are implemented in the reference package (`hvbra/baselines.py:diebold_mariano`) and reported in the companion repository submitted as Supplementary Material. The qualitative empirical pattern—HVB-RA achieves lower RMSE than the contemporary NHITS baseline at all nine cells, no method improves on the random-walk floor at most cells, and the classical baselines are strongest at the farm-gate tier—is observable directly from the RMSE columns of Table 7 and is the basis of the substantive findings reported in Section 3.8.

3.4. Real-Data Component Ablation

The synthetic-data ablation of Section 2.10.3 verifies that each component is wired and trainable. To assess each component's contribution on the actual price series, the same ablation is run on real data across five seeds. Two variants are compared against the full model: E1 ties the two regime weight profiles ($\mathbf{v}_0 = \mathbf{v}_1$), removing regime conditioning from the modal-weighting layer; E2 removes the directed cross-tier attention, so each tier uses its own encoder representation alone. Table 8 reports the result.

Two findings follow. **First, the regime-conditional pathway is numerically inert on this test window.** The mean absolute RMSE difference between E1 and Full across all nine cells is 0.00017 USD/kg, three orders of magnitude below the across-seed standard deviation. On a test window the trained Markov-switching model classifies entirely as high volatility; tying the two regime profiles changes nothing, exactly as the modal-weighting equations of Section 2 predict when only one regime is realised. What the present window therefore yields is a quantified non-identification of the regime pathway, not evidence that the mechanism has been rendered identifiable: whether it can be identified at all is conditional on the realised regime composition of the evaluation window, and a span containing both volatility states would be required for it to become active. The companion repository documents those conditions.

Second, the cross-tier attention does not improve point error on this window and modestly degrades it at long horizons. E2 attains equal-or-lower RMSE in eight of nine cells, with the largest improvements at $h = 21$ (-0.013 to -0.021 USD/kg). Read together with the regime result, the picture is coherent: the test window is a low-signal, single-regime episode in which the two components designed to exploit cross-regime and cross-tier structure correctly find little such structure to exploit. The value of the modular design is precisely that this diagnosis can be made at the level of individual components rather than inferred from aggregate performance.

Table 8. Real-data component ablation, RMSE in USD/kg (5 seeds, mean $\pm$ std). E1 ties the two regime weight profiles ($\mathbf{v}_0 = \mathbf{v}_1$); E2 removes the directed cross-tier attention. Δ is the mean RMSE change of E2 relative to Full.

Tier	h	Full	E1 (No Regime)	E2 (No Attn)	$\Delta(\text{E2} - \text{Full})$
F	1	0.0356 ± 0.0016	0.0356 ± 0.0016	0.0365 ± 0.0035	$+0.0009$
F	5	0.0684 ± 0.0048	0.0684 ± 0.0048	0.0641 ± 0.0003	-0.0043
F	21	0.1226 ± 0.0245	0.1225 ± 0.0244	0.1098 ± 0.0014	-0.0128
R	1	0.0386 ± 0.0002	0.0385 ± 0.0003	0.0385 ± 0.0010	-0.0001
R	5	0.0768 ± 0.0103	0.0760 ± 0.0085	0.0703 ± 0.0003	-0.0065
R	21	0.1451 ± 0.0356	0.1448 ± 0.0350	0.1237 ± 0.0041	-0.0214
G	1	0.0377 ± 0.0014	0.0378 ± 0.0014	0.0367 ± 0.0004	-0.0010
G	5	0.0799 ± 0.0130	0.0800 ± 0.0132	0.0720 ± 0.0008	-0.0079
G	21	0.1347 ± 0.0372	0.1348 ± 0.0374	0.1155 ± 0.0008	-0.0192

3.5. Constraint Projection

Section 2 introduced the auxiliary level-ratio constraint projection (component C3): a post hoc linear projection that enforces the descriptive level-ratio relations $F = a_R \cdot R$ and $R = a_G \cdot G$ on the joint forecast vector. The level ratios $a_R = \bar{F}/\bar{R}$ and $a_G = \bar{R}/\bar{G}$ are sample-mean ratios estimated on the training window; they are descriptive properties of the training data and not equilibrium coefficients from a cointegration test (Section 2). To evaluate the marginal contribution of this auxiliary component, five forecasts are compared at the $h = 5$ horizon: the unreconciled HVB-RA base forecasts; bottom-up projection (reconstruct F and R from G via inverse level ratios); top-down projection (allocate R and G from F via level ratios); the static constraint projection (W estimated once on calibration residuals); and the regime-conditional constraint projection (convex mix of two regime-specific projections gated by the contemporaneous filtered probability $\hat{\pi}_t$, implemented as `hvbra/reconcile.py:reconcile_regime`). Estimated level ratios are $\hat{a}_R = 1.047$ and $\hat{a}_G = 1.168$. RMSE in Table 9 is computed on the reconstructed price-level forecasts (USD/kg); constraint-violation error is $\|A \cdot \tilde{y}\|_2$ averaged over the test window.

Table 9. Constraint-projection method comparison at $h = 5$, RMSE in USD/kg per tier; “avg” = $(F + R + G)/3$; “constraint” = $\|A \cdot \tilde{y}\|_2$ averaged over the test window. Bold marks the best entry per column. Static and regime-conditional projection coincide here because $W_0 = W_1$ in the single-regime test window.

Method	F RMSE	R RMSE	G RMSE	Avg RMSE	Constraint $\ A\tilde{y}\ _2$
Base (unreconciled)	0.0684	0.0768	0.0799	0.0750	0.1731
Bottom-up	0.1195	0.1188	0.0799	0.1060	0.0000
Top-down	0.0684	0.1861	0.1140	0.1228	0.0000
Static constraint projection	0.0846	0.1355	0.0823	0.1008	0.0000
Regime-conditional projection	0.0846	0.1355	0.0823	0.1008	0.0000

Out-of-Sample Covariance Estimation

The constraint-projection covariances W , W_0 , W_1 are estimated from *calibration-period* forecast residuals ($n = 392$ observations per tier), not from test-period residuals. In the present test period, the trained Markov-switching classifier assigns all 392 calibration origins and all 278 test origins to the high-volatility state ($\text{cal_calm} = 0$, $\text{cal_stress} = 392$; $\text{test_calm} = 0$, $\text{test_stress} = 278$). The fallback rule in `hvbra/reconcile.py:estimate_regime_covariances` sets $W_0 = \text{cov}(\text{all residuals})$ when fewer than five observations are available for the calm regime; on this calibration window $W_0 = W_1$ numerically and the regime-conditional projection coincides with the static projection. This is a non-identification result for the regime-conditioning aspect of C3 on this particular calibration–test split, discussed in Section 4.

Three observations:

1. **The unreconciled base forecasts attain the lowest RMSE at every tier; the projection trades point accuracy for exact coherence.** The base forecasts attain the lowest RMSE on each tier and the lowest average RMSE (0.0750 USD/kg) but a non-zero constraint violation (0.173). All projection variants drive the violation to zero at the cost of higher RMSE: the static (and regime-conditional) projection raises average RMSE to 0.1008 (+34%), top–down to 0.1228, and bottom–up to 0.1060. Each tier’s base forecast is already near its own random-walk floor (Section 3.3), so the cross-tier coherence constraint necessarily pulls forecasts away from those individually-good values.
2. **Static and regime-conditional projections coincide in this test period.** Both produce identical RMSE because the convex mix $(1 - \hat{\pi})\Psi(W_0) + \hat{\pi}\Psi(W_1)$ reduces to $\Psi(W_1)$ when $W_0 = W_1$ (the fallback rule activates because no calibration observation falls in the calm regime). Regime conditioning would become identifiable only in test windows that contain both regimes; on the present split it is non-identified, a property of this particular calibration–test window rather than of the mechanism.
3. **The auxiliary projection is a coherence-enforcement option, not an accuracy-improvement mechanism, on this window.** All four projection methods reduce the constraint violation from 0.173 to effectively zero ($<10^{-5}$). When exact cross-tier coherence is required by a downstream consumer, the projection supplies it at a quantified RMSE cost; when it is not required, the base forecasts are preferable. Whether the projection improves accuracy under regime mixing or richer feature sets remains an open question discussed in Section 4.

The constraint projection enforces an exact linear relation between forecasts that is descriptive of the training-window sample means; it does not impose an economic equilibrium relation and is not a trace-optimal reconciliation operator under the convex-combined regime covariance (Methods, Section 2). Whether the auxiliary projection delivers gains under richer feature sets, alternative level-ratio specifications, or test windows with regime mixing remains an open question discussed in Section 4.

3.6. Decision-Aware Income Smoothing

Section 2 introduced the income-smoothing (IS) metric as a supporting evaluation protocol (component C4) for farm-gate decisions. The rule, implemented in `hvbra/decision.py`, is a 5-day selling decision with transaction-cost threshold $\tau = 0.005$: delay sale if the model predicts a price increase exceeding τ ; otherwise sell immediately. The metric $\text{IS} = 1 - \text{Var}(I_{\text{model}})/\text{Var}(I_{\text{naive}})$ measures variance reduction in realised income relative to a benchmark policy that sells every day. The results are reported in Table 10.

Table 10. Income-smoothing results, $h = 5$ horizon. Mean $\pm$ std across 5 seeds. n_{delay} is the number of days (out of 278 test origins) on which the model recommends delaying sale.

Method	IS	n_{delay} (of 278)
Random walk (always sell)	0.0000 ± 0.0000	0.0 ± 0.0
HVB-RA ($h = 5$ forecast)	$+0.0007 \pm 0.0015$	55.6 ± 124.3

HVB-RA achieves a small positive income-smoothing effect ($\text{IS} = +0.0007 \pm 0.0015$), corresponding to approximately a 0.07% reduction in realised-income variance relative to the always-sell benchmark. The descriptive interval across the five random seeds includes zero, indicating that the small positive mean effect is not stable to model initialisation. The across-seed variation in n_{delay} (55.6 ± 124.3 , range 0–278) indicates that the decision rule operates near its degenerate boundary: in some seeds the rule recommends delaying every day, in others, selling every day. IS is therefore reported as a **null operational result rather than a substantive contribution**; the value of the decision-aware evaluation in this paper lies in documenting the operational sensitivity of selling decisions to forecast skill at $h = 5$, not in claiming a smoothing gain.

3.7. Other Robustness Diagnostics

A complete robustness analysis is provided in the companion repository submitted as Supplementary Material. The diagnostics include: per-seed metric tables for each tier-horizon cell; the additional component ablations E5 (raw price input, no VMD) and E7 (no auxiliary projection); sensitivity to the VMD mode count $K \in \{3, 5, 7\}$; sensitivity to the decision-rule transaction-cost threshold $\tau \in \{0, 0.0025, 0.005, 0.01\}$ in the income-smoothing simulation; the regime classifier diagnostic (calibration and test partition of MS-AR(1) state probabilities); and the distribution of forecast values used in the variance-collapse diagnostic. The two primary component ablations, E1 (no regime gating: $\mathbf{v}_0 = \mathbf{v}_1$) and E2 (no cross-tier attention), are reported in the main text in Section 3.4. The remaining diagnostics are placed in the companion repository for two reasons. First, the headline empirical pattern—HVB-RA achieving lower RMSE than the contemporary NHITS baseline at all nine cells, with no method improving on the random-walk floor at most cells—is established in Section 3.3 on the primary metric and is not changed by the supplementary diagnostics. Second, the regime-conditioning components are formally non-identified on the present single-regime calibration–test split (Section 3.4; Section 3.5, Finding 4): a multi-regime test window is required before the E1 contrast and the regime-conditional projection can be interpreted as evidence for or against the corresponding architectural mechanisms. The companion repository documents the regime classifier output for the present split, the broader sensitivity envelopes, and the failure-mode diagnostics in full so that the conditions under which each diagnostic is identified are accessible to independent readers without forcing the manuscript to over-claim from a single-regime test window.

3.8. Summary of Findings

Across five seeds, three tiers, three forecast horizons, three external benchmarks (random walk, ARIMA, and exogenous-only LSTM), a contemporary deep learning benchmark (NHITS), and four post hoc projection variants (base, bottom-up, top-down, and constraint projection), five empirical findings emerge:

- **Finding 1: HVB-RA outperforms the contemporary deep learning baseline at every cell.** Against NHITS, a current neural hierarchical-interpolation forecaster supplied with the same five exogenous covariates and evaluated on the identical aligned test origins, HVB-RA achieves lower RMSE at all nine of nine tier-horizon cells (Table 7), with the margin widening at long horizons ($R-h = 21$: 0.1451 versus 0.1937; $G-h = 21$:

0.1347 versus 0.1753). The full HVB-RA configuration thus attains lower error than a current deep forecaster on the same information set. The component ablation (Section 3.4) localises this aggregate result: on the present single-regime window the cross-tier attention and regime-conditioning components do not themselves add point-forecast value, so the advantage over the contemporary baseline is not attributable to them.

- **Finding 2: Classical baselines are strongest at the farm-gate tier.** ARIMA(2,0,2) and the random walk attain the lowest RMSE across the farm-gate horizons (Table 7). Linear short-memory dynamics appear sufficient for the slower farm-gate price process at the resolution of this evaluation.
- **Finding 3: No method improves on the random-walk floor at most cells.** Under the residual target definition the no-change forecast is the natural zero-prediction baseline. No method—classical, exogenous, contemporary, or the proposed framework—reduces RMSE below the random-walk floor at the majority of tier-horizon cells, the expected behaviour for daily commodity prices under weak-form efficiency. HVB-RA's demonstrated advantage is over the contemporary deep learning baseline (Finding 1), not over the near-floor classical baselines. The regime-conditioned components of the architecture require multi-regime evaluation for identification (Section 4.1), whereas the cross-tier routing component is evaluated on the present test window through the E2 ablation reported in Section 3.4.
- **Finding 4: The regime-conditional component of the constraint projection is not identifiable on this test window.** The trained Markov-switching classifier assigns every calibration origin and every test origin to the high-volatility state; the regime-conditional projection therefore reduces numerically to the static projection. Identification of the regime-conditional contribution requires a calibration–test split that contains both regimes.
- **Finding 5: Decision-aware income smoothing is a null operational result in the current configuration.** The HVB-RA $h = 5$ decision rule produces $IS = +0.0007 \pm 0.0015$ with a descriptive seed-based interval that includes zero and a delay frequency that ranges from 0 to 278 across seeds, indicating that the rule operates near its degenerate boundary. The decision-aware evaluation is reported as a methodological template, not as a substantive operational contribution.

The combined picture is that HVB-RA outperforms the contemporary deep learning baseline at every cell while no method improves on the random-walk floor at most cells; the present test window does not identify the regime-conditional projection against the static projection, and does not produce a statistically meaningful decision-aware income smoothing. These are honest results of a multi-tier evaluation in which the test period happens to fall entirely within a single regime classification and in which the architectural ingredients designed to exploit cross-regime structure therefore cannot be tested. The implications for paper framing and for follow-up evaluation are discussed in Section 4, including the conditions under which the architectural ingredients are expected to be identifiable and the path to a larger benchmark set on multi-regime data.

4. Discussion

The empirical results of Section 3 place HVB-RA's architectural extensions in a competitive but non-dominant position relative to simpler baselines on the present test window. This section interprets the five findings in turn, identifies the conditions under which the architectural ingredients are expected to be identifiable, situates the contribution within the existing literature, and scopes the priorities for follow-up evaluation. The discussion deliberately distinguishes between (a) claims that the test window supports, (b) claims

that the test window cannot identify, and (c) the methodological proposal that stands independently of the test-window outcome.

4.1. What the Evaluation Identifies, and What It Does Not

The test window of 278 aligned origins spanning 2 January 2025 through 4 March 2026 admits identification of three things and does not admit identification of two. Identification of the RMSE ordering between HVB-RA and the external and contemporary benchmarks is established at the cell level in Section 3.3; the empirical pattern is reproducible from the companion repository, which also reports the corresponding Diebold–Mariano significance tests and Holm-adjusted p -values. Identification of the constraint-projection contribution against the unreconciled base forecast is established by the within-method comparison in Section 3.5; the projection enforces exact cross-tier coherence at a quantified cost in point error rather than improving it. Identification of the income-smoothing operational template is established by the IS metric in Section 3.6; the metric is reported as a null result on the present test window.

Two things are explicitly not identified by the present evaluation. First, the regime-conditional component of the modal-weighting layer (C2) and of the constraint projection (C3) cannot be tested because the trained Markov-switching classifier assigns every calibration origin and every test origin to the high-volatility state. The convex mixture in Equation (11) reduces to $\text{softmax}(\mathbf{v}_1)$ almost everywhere in the test window; the convex mixture in Equation (19) reduces to $\Psi(W_1)$ everywhere; the regime-conditional ablation E1 (Section 2.10.3) therefore tests whether allowing two trainable profiles changes anything in a setting where the gating signal is constant. Second, the directed-attention component (C1) is tested against a no-attention ablation E2 but not against the alternative-direction or unrestricted-attention controls that would identify the contribution of *directionality* as distinct from the contribution of *cross-tier routing*. These two non-identifications are properties of the test window and of the ablation suite that the reference implementation supports, not of the framework itself; the conditions under which they would be identifiable are made explicit in Section 4.7.

4.2. The Single-Regime Test Window

The single-regime classification of the test window is the single most consequential property of the present evaluation. Three interpretations are consistent with this observation and are not adjudicated by the present evidence.

Interpretation 1: a structural property of the underlying price series. The MS-AR(1) classifier with training-window-only parameters may have correctly identified that the 2025–2026 period of the dataset is empirically high-volatility throughout, in which case the regime-conditional architecture has nothing to condition on within this window and the appropriate follow-up evaluation is on a longer or different test window. *Interpretation 2: a regime-classifier generalisation problem.* The MS-AR(1) parameters fitted on 2018–2023 training data may not generalise well to 2025–2026 test data because the distribution of returns may have shifted; the threshold at $\hat{\pi}_t = 0.5$ may be inappropriate for the test distribution, and a re-calibration on a held-out portion of the training window may produce more balanced regime classifications. *Interpretation 3: a regime-specification problem.* A two-state Hamilton MS-AR(1) on tier-level returns may be the wrong regime specification for vertically linked commodity markets; alternative specifications (volatility-state model, copula-based regime model, change-point model, or a fully cross-sectional regime classifier across the three tiers jointly) may classify the test window differently.

The framework as released admits all three diagnostics. The companion repository reports the smoothed and filtered $\hat{\pi}_t$ time series across the full 2018–2026 span, the transition

matrix, the per-state log-likelihoods, and the per-state observation counts in training, calibration, and test. The follow-up work in Section 4.7 prioritises a multi-regime backtest that spans a longer historical window containing visible high- and low-volatility regimes, and an alternative-specification sensitivity analysis. The present manuscript does not adjudicate between the three interpretations: the test window is what it is, and the regime contributions cannot be tested without a different one.

4.3. Why the Backbone and Exogenous-Only LSTM Are So Close

A notable empirical pattern is that the exogenous-only LSTM tracks HVB-RA closely in RMSE, the two methods differing by small margins at most cells (Table 7), even though HVB-RA outperforms the contemporary NHITS baseline at every cell (Finding 1). The exogenous-only LSTM uses the same five exogenous covariates as HVB-RA but omits the per-tier VMD decomposition, the cross-tier attention, and the regime-aware modal weighting. The closeness of the two on a window where all methods cluster near the random-walk floor admits three interpretations.

Interpretation A: feature-density limitation. HVB-RA's per-tier input matrix is 10-dimensional (five IMFs plus five exogenous covariates). The precursor single-tier study of Pinitjitsamut [6] demonstrated that variance fidelity (StdR) scales with input feature density: 11-feature, 22-feature, and 24-feature input matrices in that work produced StdR values of 0.34, 0.11, and 1.09 respectively, with the highest variance fidelity achieved by the richest feature set. The present 10-dimensional per-tier input matrix is at the lower end of this range and may not provide the encoder with enough informative input to learn signal beyond what a non-decomposed encoder can extract from the exogenous covariates alone. Under this interpretation, the close RMSE comparison is a feature-density artefact rather than evidence against the architectural ingredients.

Interpretation B: variance collapse under symmetric loss. The mean-squared-error loss is symmetric in over- and under-prediction; in low-signal regimes the loss landscape rewards forecasts close to the conditional mean, producing low-amplitude forecasts that yield small RMSE but low-magnitude tracking. The variance-collapse diagnostic in Section 3.2.2 documents this pattern qualitatively (StdR values in the companion repository quantify it). Under this interpretation, the close RMSE comparison reflects both methods converging towards conditional-mean predictions rather than either method capturing the underlying signal.

Interpretation C: simpler architecture is better conditioned for this task. The exogenous-only LSTM with ~60 thousand parameters may simply be a better-conditioned optimiser target than HVB-RA with ~416 thousand parameters at the present training data size of 1348 daily observations. Deeper or wider architectures require more training data to identify their additional capacity; on a single-tier price series of fewer than 1500 observations, the parameter-to-observation ratio favours the simpler architecture.

These three interpretations are not mutually exclusive and each motivates a different follow-up. Interpretation A motivates feature enrichment (additional macro covariates, sectoral indices, weather data, port congestion indices, and vessel tracking). Interpretation B motivates loss-function modifications (asymmetric loss, quantile loss, and dispersion-aware loss as in [6]). Interpretation C motivates parameter-matched controls (reducing HVB-RA's parameter count or expanding exogenous-only LSTM to match). The companion repository supports all three diagnostics; the present manuscript identifies them as priorities without claiming to have isolated which mechanism is dominant.

4.4. The Framework as a Knowledge-Extraction Proposal

The five empirical findings of Section 3.8 concern point-error performance: RMSE, MAE, MASE, and the IS metric. The framework's contribution, as positioned in Section 2, is not solely point-error performance but the *structure of the architecture*: an explicit, ablatable, and inspectable decomposition of the multi-tier forecasting task into cross-market routing intensities (C1), regime-dependent modal weights (C2), and an auxiliary level-ratio constraint (C3). The architectural ingredients are designed to expose latent structure in a form that can be examined component by component, regardless of whether the composed architecture beats a simpler baseline at point-error metrics on a particular test window.

Three properties of the architecture support this knowledge-extraction reading. First, the cross-tier attention intensities at each origin t are recoverable from the trained model (`hvbra/model.py:CrossTierAttention.forward`); the companion repository plots them as time-varying weights that put more or less importance on the upstream futures-tier representation at different times, and their behaviour can be tested against external lead-lag analysis. Second, the regime-conditional weight profiles $\mathbf{v}_0, \mathbf{v}_1$ over latent components are explicitly parameterised and inspectable after training; their difference $\mathbf{v}_1 - \mathbf{v}_0$ identifies which latent components are weighted differently across regimes (subject to the identifiability conditions of Section 4.1 and the latent-component caveat of Section 2.4). Third, the auxiliary constraint projection makes the cross-tier ratio relations *explicit* as a separate post hoc step rather than absorbing them implicitly into a single end-to-end model; the contribution of the projection to coherence and to per-tier RMSE is measurable in isolation. The manuscript reports the architectural identifiability of these objects and defers the visualisations themselves (attention-weight time series, weight-profile heatmaps, and across-seed stability plots) to the companion repository. The companion plots are the empirical substantiation of the knowledge-extraction reading; this manuscript establishes the structural framework that makes such substantiation possible and reports the conditions under which it is identifiable.

From the perspective of practitioners in vertical price transmission and commodity-market forecasting, the two inspectable objects serve distinct purposes. The time-varying cross-tier attention intensities at the farm-gate tier—recoverable from `CrossTierAttention.forward` at every origin—quantify how much of the downstream encoder's updated representation derives from the upstream futures versus regional signals at each point in time. Peaks in the futures-sourced attention weight during stress periods are interpretable as periods of elevated price-discovery dominance, a hypothesis that can be tested against external lead-lag analyses [1,2] without refitting the model. The regime-conditional latent-component weight profiles $\mathbf{v}_1 - \mathbf{v}_0$ identify which IMF-aligned latent dimensions receive different importance in high- versus low-volatility states; a large $|v_{1,k} - v_{0,k}|$ for a low-frequency component suggests that trend modes matter more in stress regimes, while a large difference on a high-frequency component suggests that short-cycle adjustment dynamics are regime-sensitive. For smallholder price-risk management, these objects provide decision-relevant diagnostics: a sustained shift toward futures-dominated attention can signal a structural change in market integration that would affect the reliability of farm-gate forecasts derived from the hierarchical pipeline. None of these interpretive uses requires the composed architecture to outperform a simpler baseline at RMSE; they require only that the structural components be individually identifiable, ablatable, and documented—which the reference implementation and the companion repository provide.

The knowledge-extraction framing recasts the close RMSE comparison with exogenous-only LSTM in a different light. Exogenous-only LSTM may achieve comparable or lower RMSE; it does not produce inspectable cross-market routing intensities, inspectable regime-conditional weight profiles, or an explicit level-ratio constraint projection. The framework's

contribution is therefore complementary to point-error benchmark performance: it offers a decomposition of the forecasting task into structural components that an applied user can examine, modify, and selectively replace, at a small RMSE cost relative to a black-box baseline on the present test window.

4.5. Generalisation Beyond Rubber

The framework structure is not specific to the rubber dataset on which it is evaluated; it is applicable to any system of vertically linked markets in which an upstream price is reasonably treated as a leading indicator for one or more downstream prices, the price transmission is plausibly regime-dependent, and an empirical scaling relationship between the tiers is approximately stable in level.

Electricity markets. Day-ahead electricity prices, intraday balancing prices, and end-user retail tariffs form a comparable vertically linked structure. Day-ahead price-discovery markets, balancing markets, and household retail prices in a single jurisdiction are linked through pass-through relationships that vary with grid stress regimes (peak demand, renewable shortfall). The cross-tier attention component (C1) and the regime-conditional modal weighting (C2) translate directly. The level-ratio constraint (C3) would require sample-mean ratios on different units (e.g., monthly mean ratios after normalisation) but the structure of the projection carries over.

Agricultural value chains beyond rubber. Cocoa, coffee, palm oil, and sugar all exhibit a futures–regional spot–farm-gate structure analogous to the rubber three-tier system used here. The framework is directly transferable; the regime classifier may require re-calibration on the relevant commodity’s volatility regimes.

Financial intermediation chains. Wholesale benchmark rates (e.g., overnight interest rates), interbank lending rates, and retail deposit/loan rates form a vertically linked structure with documented regime-dependent transmission [2] and approximately stable long-run spreads. The framework’s cross-tier attention and regime-conditional modal weighting structure transfer directly; the level-ratio constraint would be replaced by an additive-spread constraint of analogous algebraic form.

Generalisation is identified as a property of the framework structure, not a claim of empirical performance on those datasets; out-of-domain validation is identified as follow-up work.

4.6. Limitations

Seven limitations of the present evaluation merit explicit acknowledgement, each grouped with the corresponding scope of follow-up evidence.

Limitation 1: regime non-identification on the test window. The most consequential limitation, discussed at length in Section 4.2.

Limitation 2: single-key degeneracy of Tier R attention. The cross-tier attention at the regional spot tier reduces algebraically to a residual linear projection because only one upstream key is supplied (Methods, Section 2.4.2, and Equation (10)). The natural fix—replacing the single-key attention with temporal-key attention over the full upstream sequence $H_t^{(F)}$ —is straightforward but not implemented in the present release; it is documented in the companion repository as a priority structural extension.

Limitation 3: the auxiliary constraint projection is not trace-optimal. The regime-conditional projection in Equation (19) preserves the level-ratio constraint by construction but is not in general a minimum-trace estimator under the convex-combined regime covariance (Methods, Section 2.7.2). A trace-optimal projection would apply Ψ directly to the mixed covariance $W_t = (1 - \hat{\pi}_t)W_0 + \hat{\pi}_tW_1$, which is generally distinct from $(1 - \hat{\pi}_t)\Psi(W_0) + \hat{\pi}_t\Psi(W_1)$.

Limitation 4: level ratios are descriptive, not cointegrating. The framework uses training-window sample-mean ratios $a_R = \bar{F}/\bar{R}$ and $a_G = \bar{R}/\bar{G}$ as exact constraints. These are descriptive properties of the training window and not equilibrium coefficients from a formal cointegration test; whether the same ratios would be obtained on a different training window is a sample stability question that the present manuscript does not test (the companion repository documents the rolling-window ratios).

Limitation 5: the decision rule is degenerate on this test window. The income-smoothing rule produces n_{delay} ranging from 0 to 278 across seeds, indicating that the rule operates near its degenerate boundary at $h = 5$ predictability of $r \approx 0.054$ (Section 3.6). The rule as specified is therefore not deployable on this test window; an operationally robust rule would require either a horizon-mixed forecast (combining $h = 5$ and $h = 21$ signals) or an entirely different decision specification (option-pricing-style framework, dynamic-programming over multiple selling opportunities). Both alternatives are identified as follow-up work.

Limitation 6: the baseline suite does not yet include the full range of current transformer-class and graph-based forecasters. The benchmarks reported in the manuscript—random walk, ARIMA, exogenous-only LSTM, and the contemporary NHITS [10] forecaster—together with the single-tier VMD-BiLSTM control in the companion repository span the classical, recurrent deep learning, and neural hierarchical interpolation literature but do not include the Temporal Fusion Transformer [8], N-BEATSx [7], PatchTST [11], DeepAR [9], or graph-based multivariate forecasters such as MTGNN [11]. The omission of the remaining architectures is deliberate for two reasons. First, empirical evidence on short commodity price series (fewer than 2000 daily observations) does not consistently favour transformer-class forecasters over recurrent or linear baselines: on similarly sized panels, ref. [8] report that TFT's advantage over LSTM-based baselines diminishes on series with fewer than 2000 observations and high noise-to-signal ratios [8]; on electricity price series of comparable length, ARIMA and shallow recurrent models match or exceed deep transformer architectures outside the training distribution [7]. Natural-rubber futures are a high-noise-to-signal single commodity rather than a large cross-sectional panel, placing this dataset in the regime where transformer-class scalability advantages are smallest. Second, the present manuscript establishes the structure of the framework, the reference implementation, and the honest evaluation against the closest competitive baselines, now including a current deep forecaster. The NHITS comparison is accordingly a comparison against one contemporary non-transformer deep forecaster; it is not, and should not be read as, a direct comparison against transformer-class or graph-based architectures, which remains the dedicated follow-up identified below. A dedicated benchmark study against the broader transformer-class and graph-based families, using the same data, the same aligned-origin set, and the same evaluation protocol, is identified as the primary follow-up evaluation in Section 4.7. The expected outcome of that study is a refined understanding of where in the (tier, horizon) grid the multi-tier framework's structural extensions are competitive, dominant, or dominated—not a demonstration of RMSE dominance, which the honest-results framing of this paper does not claim.

Limitation 7: the exogenous-only LSTM baseline omits lagged prices and is not parameter-matched. The exogenous-only LSTM uses only the five exogenous covariates and excludes lagged tier prices, in order to provide a controlled contrast that isolates the contribution of the VMD-decomposed price representation relative to undecomposed exogenous-only inputs. The reviewer's concern is well-placed: a more competitive autoregressive deep learning baseline would augment the LSTM with lagged tier prices (or the raw price sequence as an additional input channel), removing the asymmetry in price information between the baseline and HVB-RA. Two follow-up baselines are therefore identified: (a) an autoregressive-extended exogenous-only LSTM that adds the most recent $L = 60$ trading

days of raw log-price observations alongside the five exogenous covariates—this matches HVB-RA’s total observation window while isolating the marginal contribution of VMD decomposition; and (b) a parameter-matched exogenous-only LSTM scaled to HVB-RA’s ≈ 416 k parameter count, which tests whether RMSE differences are driven by parameter count rather than architectural structure. Both baselines are identified as the next pair of controlled contrasts in the companion repository roadmap. The present manuscript reports the exogenous-only LSTM as defined because the production runs were completed under this specification and rerunning under the augmented specification was not feasible within the review cycle; the limitation is documented so that the reader can calibrate the strength of the baseline comparison accordingly.

4.7. Priorities for Follow-Up Evaluation

Four priorities are identified for the follow-up work that the present results scope.

Priority 1: regime-diverse test windows. The single-regime test window is the most consequential limitation. A multi-regime backtest spanning a longer historical period containing visible high- and low-volatility regimes (for the rubber dataset, this includes 2008–2010 and 2020–2021) is identified as the primary follow-up. The diagnostic includes (a) re-fitting the Markov-switching classifier on rolling-window training data and tracking the classifier’s behaviour across regime transitions, (b) re-running the constraint-projection comparison on calibration windows that contain both regimes, and (c) testing the E1 ablation against the full HVB-RA in a setting where the gating signal varies across the test window.

Priority 2: transformer-class benchmark study. A dedicated comparison against TFT, N-BEATSx, PatchTST, DeepAR, and MTGNN, extending the NHITS comparison already reported in Section 3.3 and using identical splits, identical exogenous covariates, identical seed protocol, and identical metrics (RMSE, MAE, MASE, DM tests with Holm correction), is identified as the next benchmark study. The expected outcome is a refined understanding of where in the (tier, horizon) grid the multi-tier framework’s structural extensions are competitive, dominant, or dominated.

Priority 3: identification controls for the architectural ingredients. Beyond the E1, E2, E5, E7 ablations in the present release, four targeted identification controls are identified for follow-up: (a) reverse-direction attention ($G \rightarrow R \rightarrow F$ instead of $F \rightarrow R \rightarrow G$), which tests whether the domain-imposed direction beats a random direction; (b) unrestricted attention (bidirectional between all tier pairs), which tests whether directionality itself beats no restriction; (c) parameter-matched controls (exogenous-only LSTM scaled up to HVB-RA’s parameter count, or HVB-RA scaled down to exogenous-only LSTM’s), which tests whether RMSE differences are driven by parameter count rather than architectural structure; and (d) shuffled-gate identification tests for C2 (random permutations of $\hat{\pi}_t$ across origins), which tests whether the regime-conditioning extracts information from the gating signal beyond what is recoverable from random gating.

Priority 4: alternative regime specifications. The two-state Hamilton MS-AR(1) classifier is one of several plausible regime specifications. Alternatives identified as follow-up include volatility-state HMM on returns [20], change-point regime models, copula-based regime models for joint multi-tier classification, and learned-regime classifiers trained jointly with the neural backbone (sacrificing the two-stage interpretability for joint optimisation). The companion repository documents the present specification; the alternatives are scoped for a methodological follow-up that focuses specifically on regime classification choice in vertically linked commodity systems.

4.8. Closing Remarks

The HVB-RA framework, as evaluated in this manuscript on a single calibration–test split of the daily three-tier rubber price system, does not establish empirical dominance over simpler baselines in RMSE; the regime-conditional ingredients are not identifiable on the test window; and the decision-aware income-smoothing application yields a null operational result. These are honest findings of a multi-tier evaluation in a single-regime test window with modest input feature density. The contribution of the manuscript is not the empirical performance of the architecture on this dataset but the architecture itself: a modular, ablatable, and reproducible decomposition of a multi-tier forecasting task into cross-market routing intensities, regime-dependent modal weights, and an explicit level-ratio constraint, together with the honest documentation of the conditions under which each component is identifiable, the conditions under which it is not, and the priorities for the follow-up evaluation that would test the architectural extensions in their identifiable domain.

5. Conclusions

This paper introduced HVB-RA, a Hybrid VMD-BiLSTM forecasting framework with regime-aware components for vertically linked commodity markets. The framework combines five elements: per-tier Variational Mode Decomposition, tier-level bidirectional LSTM encoders, domain-informed directed cross-tier attention, a two-stage regime-informed modal-weighting layer gated by a Markov-switching state probability, and an auxiliary post hoc constraint projection that enforces an explicit level-ratio relation across tiers. The framework was evaluated on a daily three-tier rubber price dataset spanning 2 May 2018 through 4 March 2026, with calibration and test windows that admit identification of point-error performance against three external benchmarks.

Five empirical findings emerged from the evaluation. First, HVB-RA achieved lower RMSE than the contemporary NHITS baseline—a current neural hierarchical-interpolation forecaster using the same exogenous covariates—at all nine of nine tier-horizon cells, with the margin widening at long horizons. Second, ARIMA(2,0,2) and the random walk were strongest at the farm-gate tier, consistent with the slower information flow there. Third, no method, including HVB-RA, improved on the random-walk floor at most cells, the expected behaviour for daily commodity prices under weak-form efficiency. Fourth, the regime-conditional component of the constraint projection could not be identified on the present test window because every calibration and test origin was classified as high-volatility regime by the trained Markov-switching classifier. Fifth, the decision-aware income-smoothing exercise yielded a null operational result with a seed-based uncertainty interval that included zero and a delay frequency that ranged from zero to the full test sample across seeds.

The contribution of the manuscript is not the empirical performance of the framework on this dataset but the architecture and its honest evaluation. HVB-RA decomposes a multi-tier forecasting task into ablatable, inspectable structural components: time-varying upstream-source attention intensities at the farm-gate tier (with a fixed residual upstream fusion at the regional-spot tier), regime-conditional latent-component-weight profiles, and an explicit level-ratio constraint. The architectural ingredients are individually identifiable on appropriate test windows even when their joint composition does not exceed simpler baselines on a particular test window. The complete reference implementation is submitted alongside this manuscript as Supplementary Material, together with the broader robustness diagnostics (full ablation suite, K sensitivity, transaction-cost threshold sensitivity, regime classifier output) that the manuscript scopes but does not foreground.

Four follow-up priorities are identified. A regime-diverse backtest that spans a longer historical period containing visible high- and low-volatility regimes is the primary follow-up, as it is the only experimental setting in which the regime-conditional components of the framework can be tested. A dedicated benchmark study extending the NHITS comparison already reported to the broader transformer-class and graph-based families—TFT, N-BEATSx, PatchTST, DeepAR, and graph-based multivariate forecasters—is identified as the next benchmark evaluation. Four targeted identification controls (reverse-direction attention, unrestricted attention, parameter-matched controls, and shuffled-gate tests) are identified as the next round of component ablations. Finally, alternative regime specifications—volatility-state HMMs, change-point models, copula-based regime classifiers, and learned-regime classifiers trained jointly with the neural backbone—are identified as a methodological follow-up specifically addressing the regime-classification choice in vertically linked commodity systems.

The framework, the evaluation, and the companion repository are offered as a structured starting point for follow-up work in machine-learning-based forecasting of vertically linked markets, rather than as an architectural endpoint with established empirical dominance. The contributions are architectural and methodological: two inspectable mechanisms that expose latent structure in multi-tier commodity forecasting, an honest evaluation documenting the conditions under which each mechanism is identifiable and the conditions under which it is not, and a reproducible reference implementation that supports the ablation suite, the component identification controls, and the transformer-class benchmark study are identified as follow-up priorities.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/make8070185/s1>. The supplementary materials are provided in a single folder entitled “Supplementary Materials”, which contains: S1, series-level documentation of the nine constituent price series and the preprocessing steps; S2, HVB-RA implementation details, hyperparameters, and configuration files; S3, the reproducibility package, including the preprocessing pipeline, rolling-origin evaluation harness, MS-AR(1) fitting, regime-conditional constraint projection, and decision-loss simulation; persisted artefacts for full reproduction, including per-seed predictions, fitted Markov-switching parameters, calibration-window covariance matrices W_0 and W_1 , and decision-loss outputs.

Funding: This research received no external funding.

Institutional Review Board Statement: Ethical review and approval were not required, as this study used only publicly available, aggregated market and macroeconomic data and did not involve human subjects.

Informed Consent Statement: Not applicable.

Data Availability Statement: The rubber price series and macroeconomic covariates are publicly available from the sources cited in Section 2.1. The source code, configuration files, and reproducibility documentation are submitted as Supplementary Materials for peer review and are also available at <https://github.com/talkmcp/hvbra> (accessed on 24 June 2026). Persisted numerical result artefacts and derived diagnostic tables are not included in the submitted supplementary package and will be deposited in the same public repository upon publication under an MIT license for code and CC-BY 4.0 for derived data. Raw price data are not redistributed where source-provider restrictions apply; documented retrieval scripts, series identifiers, and preprocessing rules allow local reconstruction.

Acknowledgments: During manuscript preparation, the author used generative AI tools to assist with language editing; the author reviewed and edited all AI-assisted text and takes full responsibility for the final content of the publication.

Conflicts of Interest: The author declares no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AR	Autoregressive
ARIMA	Autoregressive Integrated Moving Average
BiLSTM	Bidirectional Long Short-Term Memory
CI	Confidence Interval
DM	Diebold–Mariano (test)
F	Futures tier
G	Farm-gate tier
HVB-RA	Hybrid VMD–BiLSTM with Regime-Aware components
IMF	Intrinsic Mode Function
IS	Income Smoothing
LSTM	Long Short-Term Memory
MASE	Mean Absolute Scaled Error
MS-AR	Markov-Switching Autoregressive
R	Regional (spot) tier
RMSE	Root Mean Squared Error
SGX	Singapore Exchange
SHFE	Shanghai Futures Exchange
StdR	Standard-Deviation Ratio
TOCOM	Tokyo Commodity Exchange
VMD	Variational Mode Decomposition

References

- Ge, Y.; Wang, H.H.; Ahn, S.K. Cotton market integration and the impact of China's new exchange rate regime. *Agric. Econ.* **2014**, *45*, 5–27. [\[CrossRef\]](#)
- Khin, A.A.; Ramli, M.A.F.B. Price transmission and volatility spillovers in the natural rubber market: Evidence from major rubber-producing countries. *Int. J. Supply Chain. Manag.* **2019**, *8*, 432–440.
- Wu, Z.; Pan, S.; Long, G.; Jiang, J.; Chang, X.; Zhang, C. Connecting the dots: Multivariate time series forecasting with graph neural networks. In Proceedings of the KDD '20: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Virtual Event, 6–10 July 2020; pp. 753–763. [\[CrossRef\]](#)
- Shazeer, N.; Mirhoseini, A.; Maziarz, K.; Davis, A.; Le, Q.; Hinton, G.; Dean, J. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv* **2017**, arXiv:1701.06538. [\[CrossRef\]](#)
- Dragomiretskiy, K.; Zosso, D. Variational mode decomposition. *IEEE Trans. Signal Process.* **2014**, *62*, 531–544. [\[CrossRef\]](#)
- Pinitjitsamut, M. Multi-scale forecasting of natural rubber prices using VMD-augmented BiLSTM: A hybrid architecture ablation study. *Forecasting* **2026**, *8*, 43. [\[CrossRef\]](#)
- Olivares, K.G.; Challu, C.; Marcjasz, G.; Weron, R.; Dubrawski, A. Neural basis expansion analysis with exogenous variables: Forecasting electricity prices with NBEATSx. *Int. J. Forecast.* **2023**, *39*, 884–900. [\[CrossRef\]](#)
- Lim, B.; Arik, S.Ö.; Loeff, N.; Pfister, T. Temporal Fusion Transformers for interpretable multi-horizon time series forecasting. *Int. J. Forecast.* **2021**, *37*, 1748–1764. [\[CrossRef\]](#)
- Salinas, D.; Flunkert, V.; Gasthaus, J.; Januschowski, T. DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *Int. J. Forecast.* **2020**, *36*, 1181–1191. [\[CrossRef\]](#)
- Challu, C.; Olivares, K.G.; Oreshkin, B.N.; Garza, F.G.; Mergenthaler-Canseco, M.; Dubrawski, A. NHITS: neural hierarchical interpolation for time series forecasting. *Proc. Aai Conf. Artif. Intell.* **2023**, *37*, 6989–6997. [\[CrossRef\]](#)
- Nie, Y.; Nguyen, N.H.; Sinthong, P.; Kalagnanam, J. A time series is worth 64 words: Long-term forecasting with transformers. *arXiv* **2022**, arXiv:2211.14730. [\[CrossRef\]](#)
- Oreshkin, B.N.; Carpov, D.; Chapados, N.; Bengio, Y. N-BEATS: Neural basis expansion analysis for interpretable time series forecasting. *arXiv* **2019**, arXiv:1905.10437. [\[CrossRef\]](#)
- Cao, D.; Wang, Y.; Duan, J.; Zhang, C.; Zhu, X.; Huang, C.; Tong, Y.; Xu, B.; Bai, J.; Tong, J.; Zhang, Q. Spectral temporal graph neural network for multivariate time-series forecasting. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 17766–17778.
- Hamilton, J.D. A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica* **1989**, *57*, 357–384. [\[CrossRef\]](#)
- Klaassen, F. Improving GARCH volatility forecasts with regime-switching GARCH. *Empir. Econ.* **2002**, *27*, 363–394. [\[CrossRef\]](#)

16. Teräsvirta, T. Specification, estimation, and evaluation of smooth transition autoregressive models. *J. Am. Stat. Assoc.* **1994**, *89*, 208–218. [[CrossRef](#)]
17. Dvan Dijk; Teräsvirta, T.; Franses, P.H. Smooth transition autoregressive models—A survey of recent developments. *Econom. Rev.* **2002**, *21*, 1–47. [[CrossRef](#)]
18. Tong, H. *Non-linear Time Series: A Dynamical System Approach*; Oxford University Press: Oxford, UK, 1990.
19. Guidolin, M.; Pedio, M. *Essentials of Time Series for Financial Applications*; Academic Press: Cambridge, MA, USA, 2018.
20. Ang, A.; Timmermann, A. Regime changes and financial markets. *Annu. Rev. Financ. Econ.* **2012**, *4*, 313–337. [[CrossRef](#)]
21. Bucci, A. Realized volatility forecasting with neural networks. *J. Financ. Econom.* **2020**, *18*, 502–531. [[CrossRef](#)]
22. Marinho, P.; de Andrade, B.B.; Hotta, L.K. A regime-switching approach for forecasting commodity prices. *J. Forecast.* **2021**, *40*, 1090–1112. [[CrossRef](#)]
23. Huang, N.E.; Shen, Z.; Long, S.R.; Wu, M.C.; Shih, H.H.; Zheng, Q.; Yen, N.-C.; Tung, C.C.; Liu, H.H. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proc. R. Soc. A* **1998**, *454*, 903–995. [[CrossRef](#)]
24. Engle, R.F.; Granger, C.W.J. Co-integration and error correction: Representation, estimation, and testing. *Econometrica* **1987**, *55*, 251–276. [[CrossRef](#)]
25. Wickramasuriya, S.L.; Athanasopoulos, G.; Hyndman, R.J. Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization. *J. Am. Stat. Assoc.* **2019**, *114*, 804–819. [[CrossRef](#)]
26. Panagiotelis, A.; Gamakumara, P.; Athanasopoulos, G.; Hyndman, R.J. Probabilistic forecast reconciliation: properties, evaluation and score optimisation. *Eur. J. Oper. Res.* **2023**, *306*, 693–706. [[CrossRef](#)]
27. Kim, C.-J. Dynamic linear models with Markov-switching. *J. Econom.* **1994**, *60*, 1–22. [[CrossRef](#)]
28. Hyndman, R.J.; Koehler, A.B. Another look at measures of forecast accuracy. *Int. J. Forecast.* **2006**, *22*, 679–688. [[CrossRef](#)]
29. Harvey, D.; Leybourne, S.; Newbold, P. Testing the equality of prediction mean squared errors. *Int. J. Forecast.* **1997**, *13*, 281–291. [[CrossRef](#)]
30. Diebold, F.X.; Mariano, R.S. Comparing predictive accuracy. *J. Bus. Econ. Stat.* **1995**, *13*, 253–263. [[CrossRef](#)]
31. Holm, S. A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* **1979**, *6*, 65–70.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.