

Article

Too Sharp to Be True? Illusory Gains in Regime-Weighted Conformal Prediction for Daily Rubber Price Changes

Montchai Pinitjitsamut 

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

Highlights

What are the main findings?

- Weighting past forecast errors by volatility regime made rubber-price intervals appear about 20% narrower. On one base forecaster, this advantage disappears once the interval is corrected for the small number of errors it is built from; across both base forecasters it disappears only when information from the day being predicted is withheld.
- The gain came from the timing of the signal, not from the weighting. The regime probability used the very day it was meant to predict. With no regime in the data at all, the same gain still appears.

What are the implications of the main findings?

- Use a regime probability computed from yesterday's information only. One computed from today's already contains the price move the interval is meant to cover.
- Even when corrected, the weighted interval can come out infinite. Plain split conformal, or a method that adapts to past misses, is the safer default.

Abstract

This study audits whether regime weighting can sharpen conformal forecast intervals without using target-period information or omitting the required finite-sample correction. Conformal prediction builds such intervals from past forecast errors. A natural refinement gives more weight to errors from days whose volatility resembles the forecast day. On daily natural-rubber prices, the refinement appears to work: intervals become about 20% narrower than plain split-conformal, with a significantly better Winkler score. This paper asks whether that gain is real. Three implementation choices are examined, one at a time. The first uses a regime signal that already sees the price move it is meant to predict. The second estimates the regime model on the same residuals the interval is calibrated on. The third omits a correction that the weighted quantile requires in finite samples. The forecast-feasible construction—predictive regime probabilities, a validation-fitted regime model, and the finite-sample correction—shows no detectable improvement over split-conformal, at a paired Winkler difference of +0.06 (95% CI −0.15 to +0.17). Applying the correction alone is not always enough: it removes the apparent advantage on the VMD-augmented ridge residuals, but the filtered comparison arm survives it on the AR(1) residuals at −0.54. Only withholding target-period information eliminates the artifact on both. Concentrated weights also leave some intervals unbounded, whereas adaptive conformal baselines remain finite throughout. A controlled simulation reproduces the same apparent gain where no regime information exists at all. Apparent sharpness must therefore be audited for information timing, calibration reuse, and the finite-sample correction.



Academic Editor: Alessia Paccagnini

Received: 19 July 2026

Revised: 3 September 2026

Accepted: 8 September 2026

Published: 10 September 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 \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: conformal prediction; prediction intervals; regime-switching; volatility clustering; distribution-free uncertainty quantification; Winkler score; time-series forecasting; natural rubber prices

1. Introduction

Prediction intervals for a non-stationary commodity price must satisfy two demands at once. They need a defensible coverage guarantee, and they need to be sharp when volatility shifts. Regime-weighted conformal intervals appear to deliver both: when an interval is formed, past forecast errors drawn from volatility states resembling the forecast origin receive more weight than the rest, so the interval widens in turbulent regimes and tightens in calm ones.

On daily Ribbed Smoked Sheet No. 3 (RSS3) natural-rubber price changes, such a construction seems to work exactly as advertised. Weighting the split-conformal calibration residuals by a two-state Markov-switching regime probability, with a per-regime asymmetric tail allocation, yields intervals about 20% narrower than split-conformal and a significantly better interval score. This paper asks whether that advantage is real, and finds that it is not. The audit examines three implementation features associated with it: an oracle filtered state that conditions on the target-period shock, calibration-dependent estimation of the regime model, and omission of the finite-sample weighted-quantile correction. Once the correction and fixed symmetric tails are applied, no construction that withholds target-period information retains a statistically detectable improvement over split-conformal.

1.1. What the Conformal Guarantee Rests on

Because the argument turns on the internal mechanics of the method, it is worth stating plainly what a conformal interval is. Conformal prediction builds intervals from the empirical distribution of realized forecast errors rather than from an assumed error distribution [1–3]; Stocker et al. [4] give an introduction aimed at time-series practitioners. In its split, or inductive, variant [5,6], the sample is divided into disjoint blocks: a model is fitted on a training block, and a separate calibration block supplies the nonconformity scores, here the residuals of that fitted model. The interval for a new observation is the point forecast plus the appropriate empirical quantile of those scores, a step that conformalized quantile regression [7] refines so that width tracks local variability.

The resulting interval carries a finite-sample marginal coverage guarantee of at least $1 - \alpha$, for any underlying model and any error distribution. That guarantee is what distinguishes conformal prediction from other distribution-free routes—resampling the standardized innovations of a state-space model [8], for instance, yields intervals with no such property. It is not, however, assumption-free. It requires the nonconformity scores to be exchangeable, meaning that their joint distribution is invariant to reordering, and daily commodity price changes violate this through volatility clustering and abrupt regime shifts. Everything that follows concerns how far that violation can be repaired by re-weighting the calibration scores.

Three quantities compare intervals throughout: empirical coverage, mean interval width, and the Winkler, or interval, score [9], which adds to the width a penalty of $2/\alpha$ times the distance by which a realization falls outside the interval. Lower is better, and a method is rewarded for narrowness only if it also contains the outcome. It is the metric on which the apparent 20% gain is claimed; Section 2.5 gives the formula and the limitations that shape how the results should be read.

1.2. What the Literature Does and Does Not Check

Two lines of work relax exchangeability, and the conditions each imposes on the calibration weights are central to this study. The first re-weights the calibration scores. Weighted conformal prediction [10] restores validity under covariate shift, and Barber et al. [11] bound the coverage gap for arbitrary sequences by a weighted sum of total-variation distances—for weights that are fixed, chosen in advance and not functions of the data.

A fixed weight is not the same as a merely predictable weight, one measurable at the forecast origin. Both differ categorically from a weight the calibration outcomes themselves produce: a state inferred using the target residual, or a regime model fitted on the very residuals that serve as nonconformity scores. Weights of that third kind are neither fixed nor past-measurable, so the bound does not reach them. The same construction also demands a finite-sample correction: point masses at $\pm\infty$ added to the weighted calibration distribution, without which the guarantee fails at any finite sample size. Practitioners drop them easily, and including them makes intervals unbounded when the weights concentrate. A separate route sidesteps weighting entirely, restoring exact validity for hidden-Markov data by blocking [12,13], though for classification sets rather than intervals.

The second line adapts online. Adaptive Conformal Inference [14] updates the mis-coverage level from realized coverage errors, its dynamically-tuned extension aggregates several learning rates [15], EnbPI [16] recalibrates from a sliding window, and Lu et al. [17] couple a deep switching state-space model with conformal inference. This family carries no explicit regime model yet survives drift on past miscoverage alone, so it sets the bar a regime-weighted construction must clear. Table 1 places all of these constructions, and the two arms examined here, on the four dimensions this paper audits.

Table 1. Where published regime-aware and locally weighted conformal constructions stand on the four audit dimensions.

Construction	Localizing Signal	Timing	Fitted on Calib.	Atoms
Split-conformal [5,6]	none	—	no	n/a
Weighted conformal [10]	likelihood ratio	fixed in advance	no	required
Beyond exchangeability [11]	arbitrary fixed weights	fixed in advance	no	required
Block-permutation HMM [12]	block structure	—	no	n/a
ACI [14], DtACI [15]	realized miscoverage	past-measurable	n/a	n/a
EnbPI [16]	sliding window	past-measurable	no	n/a
Similarity clustering [18]	observed weather features	past-measurable	no	no
Regime-similarity risk control [19]	observed volatility feature	past-measurable	no	yes
RareCP [20]	learned expert mixture	past-measurable	off-calibration	yes
Deep switching state space [17]	latent state, filtered	conditions on period t	jointly fitted	not reported
This study, coupled arm	latent state, filtered	conditions on period t	yes	omitted, then applied
This study: forecast-feasible construction	latent state, predictive	past-measurable	no	applied

Entries record what each source states or implements, not an assessment of its merit. “Timing” is whether the localizing signal is measurable at the forecast origin, “Fitted on calib.” whether the model producing it is estimated on the calibration scores, and “Atoms” whether the $\pm\infty$ point masses the weighted conformal quantile requires are applied. Rows with weights fixed in advance, or restoring exchangeability by blocking, lie outside the diagnostic.

A third strand sharpens intervals by localizing calibration toward states resembling the forecast origin, and it is the one this paper audits. The idea predates conformal prediction: Zarnani et al. [18] draw a new case’s interval from the error distribution of its fuzzy-clustered weather regime, with no coverage guarantee. Schmitt [19] weights conformal risk-control scores by a time decay and a regime-similarity kernel over an

observed volatility feature, and proves a weighted-exchangeability guarantee; RareCP [20] retrieves similar calibration examples through a mixture of learned regime experts; and a parallel line refines the tail budget rather than the weighting [21]. The closest instance in this journal, DG-TFT-CQR [22], calibrates temporal-fusion-transformer quantiles with a single pooled, unweighted split-conformal step, and its authors note the absence of state-adaptive calibration as a limitation—the opening that regime-aware weighting seeks to exploit.

These constructions differ in how localization is obtained: an observed feature, cluster membership, or a latent volatility state inferred from a Markov-switching model, as studied here. None of them, however, examines systematically whether the localizing signal is available at the forecast origin, that is, filtered versus one-step predictive, nor whether the auxiliary model producing it is estimated on data disjoint from the calibration scores. These two questions are immaterial to the headline sharpness metrics yet decisive for admissibility. Two patterns in Table 1 fix the scope of what follows: constructions with weights fixed in advance, and those restoring exact exchangeability by blocking, lie outside the diagnostic entirely, while it binds wherever a fitted auxiliary model supplies the signal—and in that group the timing of the signal and the provenance of the model are rarely stated explicitly.

1.3. Why Natural Rubber

Within commodity forecasting, latent regimes have a long modeling tradition. Markov-switching models [23] formalize the alternation between calm and stress states, and regime dependence in commodity and energy price volatility is well documented [24,25]. GARCH and stochastic-volatility models describe the same heteroskedasticity through a continuous conditional variance; a discrete state is used here because the conformal weights need a similarity between the forecast origin and each calibration day, which a two-state probability vector supplies directly. The closest agricultural precedent is Kummaraka and Srisuradetchai [26], whose Monte-Carlo-dropout intervals for Thai durian exports improve empirical coverage over ARIMA but carry no finite-sample guarantee.

Rubber itself has been forecast with econometric models [27,28], machine-learning regressors, and decomposition-augmented deep models [29–31], while latent-regime forecasters [17] tie the regime variable to the conditional mean. All of that work targets the point or volatility forecast. This study relocates the regime structure from the point layer to the uncertainty layer, where the calibration gap resides.

1.4. This Paper

The present study is diagnostic rather than constructive: it audits the regime-aware conformal family rather than adding a method to it. The construction it examines—a two-state Markov-switching regime driving a similarity-weighted, per-regime-asymmetric conformal interval—is a representative of that family, not a proposal.

Three results follow. A two-factor empirical design separates the timing of the regime signal from the reuse of calibration outcomes, and under the corrected construction the naive 20% sharpening disappears: the filtered, calibration-fitted cell falls to a paired Winkler difference of -0.22 with an interval spanning zero, and the forecast-feasible construction shows no detectable gain. A controlled Markov-regime simulation with an exact no-regime null then shows why. Replacing predictive with filtered probabilities manufactures a spurious advantage even where no regime information exists, while calibration-fitting adds none that is separately detectable under a stationary null. The controlled null attributes the manufactured gain to target-state leakage; calibration-dependent fitting remains associated with the empirical pattern, but its contribution cannot be separated from regime-model recency and transfer and is explicitly not test-data leakage. From these results the paper

distills an audit protocol: information timing, calibration reuse, finite-sample correction and effective calibration size must be checked jointly.

Three limits belong at the outset. No finite-sample coverage guarantee is established for those weights; the fixed-weight correction is a benchmark, not a certificate. No published method is audited—the construction is assembled from components in the literature, and Table 1 records where the diagnostic binds. And an interval spanning zero is not equivalence, since the sample sizes here would not resolve small effects. Section 4.4 states each claim in full.

Section 2 describes the data, the base forecasters and the experimental design; Section 3 reports the decomposition, its robustness and the simulation; Section 4 interprets the two couplings and states the scope; and Section 5 concludes with guidance for practitioners. The conclusions are limited to the implementations and regime conditions examined.

2. Data and Empirical Methodology

Figure 1 sets out the pipeline and the point at which its information properties are decided: a base forecast yields residuals, a Markov-switching model supplies the regime signal, and a regime-weighted conformal interval follows. The predictive, validation-fitted branch is forecast-feasible and keeps regime-parameter estimation separate from conformal calibration; the filtered branch is an oracle diagnostic when its origin state conditions on the target-period residual, and calibration-fitted regime parameters introduce calibration-dependent auxiliary estimation. These distinctions concern information availability and score reuse; formal coverage for the resulting data-dependent latent-state weights is treated separately in Supplementary Section S2.

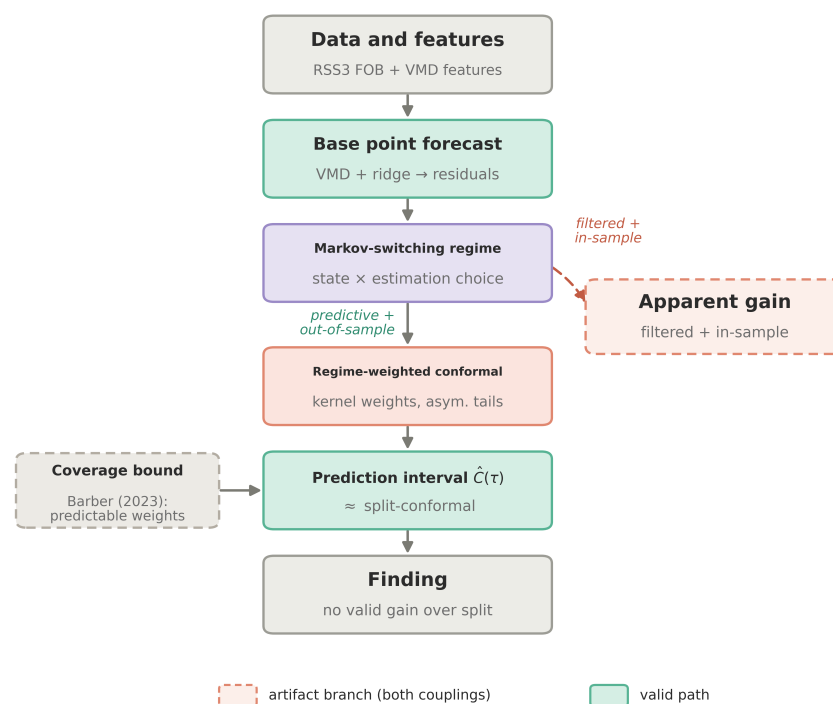


Figure 1. The regime-weighted conformal pipeline and its information fork. The predictive, out-of-sample branch is the forecast-feasible construction, which removes the two most direct couplings; the filtered, in-sample branch is an oracle comparison that yields the apparent gain identified in Section 3 as an information-coupling artifact. The coverage-bound element refers only to the fixed-weight benchmark in Supplementary Section S2; it does not provide a guarantee for the data-dependent latent-state weights implemented here [11]. The construction is labeled forecast-feasible because every input is measurable at the forecast origin, not because finite-sample validity is established.

2.1. Data and Feature Construction

The target variable is the first-differenced price of Ribbed Smoked Sheet grade 3 (RSS3) rubber quoted free on board (FOB) in Bangkok, $\Delta p_t = p_t - p_{t-1}$ (Baht/kg/day), where p_t is the near-month, first-position quote. The study window is 7 May 2018 to 13 July 2026, daily, and yields 1906 quoted trading days. Differencing on consecutive real observations leaves 1430 usable target observations. Figure 2 plots both series against the partitions.

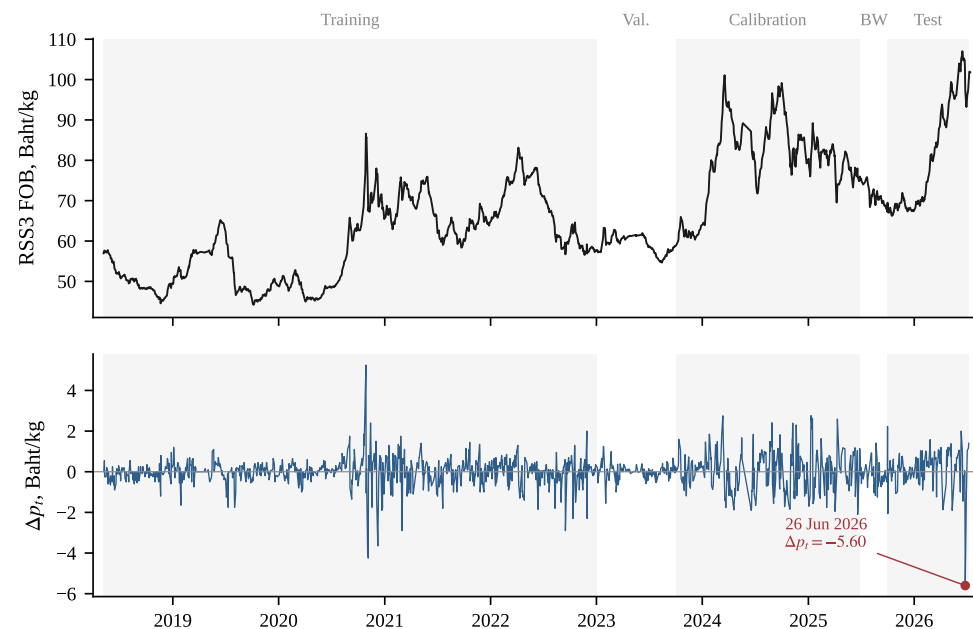


Figure 2. The RSS3 FOB near-month price (**upper panel**) and its first difference (**lower panel**), 7 May 2018–13 July 2026, with the five chronological partitions of Table 2 shaded alternately; “Val.” is the validation block and “BW” the bandwidth block. The level series drifts and shifts scale; the increments are centered but visibly heteroskedastic. The single largest decline of the sample, -5.60 Baht/kg on 26 June 2026, falls inside the test window and accounts for most of that window’s skew and excess kurtosis.

Table 2. Chronological partitions of the sample, their spans and their roles.

Partition	Span	Role	<i>n</i>
Training	7 May 2018–3 Jan 2023	base-model estimation	841
Validation	4 Jan–29 Sep 2023	ridge penalty; regime fit in the forecast-feasible arm	129
Calibration	3 Oct 2023–27 Jun 2025	conformal residuals	287
Bandwidth (BW)	1 Jul–30 Sep 2025	selection of γ	42
Test	1 Oct 2025–8 Jul 2026	final evaluation, never used for any choice	131

Counts are usable differenced observations. The partitions are contiguous and are never shuffled. “BW” abbreviates the bandwidth partition and is used as such in Figure 2 and in Figure S1 of the Supplementary Material. The roles are developed in Section 2.5; the table is placed here because every later table and figure refers to these blocks.

Three features of the data motivate the construction studied here. First, the level series is non-stationary and the differenced series is not. Augmented Dickey–Fuller and KPSS tests agree on both counts and at every conventional level, with the ADF statistic on Δp_t reaching -17.153 ($p < 0.001$); Panel C of Table 3 reports each test. The increment, rather than the level, is also the quantity whose uncertainty a one-step interval must express.

Table 3. Descriptive statistics and time-series properties of the RSS3 FOB near-month price and its first difference (7 May 2018–13 July 2026).

Panel A. Distribution								
	n	Mean	SD	Min	Max	Skew	Ex. kurt.	Jarque–Bera
Price level p_t (Baht/kg)	1906	65.70	13.52	44.20	107.03	0.55	−0.22	100.3 ***
First difference Δp_t	1430	0.030	0.754	−5.60	5.25	−0.35	7.12	3051.8 ***
Panel B. Δp_t by chronological partition								
	n	Mean	SD	Min	Max	Skew	Ex. kurt.	Mean p_t
Training (7 May 2018–3 Jan 2023)	841	−0.008	0.673	−4.25	5.25	−0.38	12.53	59.20
Validation (4 Jan–29 Sep 2023)	129	0.000	0.377	−1.55	1.25	−0.42	3.97	59.22
Calibration (3 Oct 2023–27 Jun 2025)	287	0.116	0.996	−2.10	2.76	0.10	−0.30	79.59
Bandwidth (1 Jul–30 Sep 2025)	42	−0.137	0.551	−1.25	1.41	0.46	0.49	71.59
Test (1 Oct 2025–8 Jul 2026)	131	0.171	0.918	−5.60	2.23	−2.11	10.83	80.31
Panel C. Unit-root and dependence tests								
Test	Series		Statistic	p	Lags	Conclusion		
ADF (intercept)	p_t		−1.099	0.716	9	unit root not rejected		
ADF (intercept + trend)	p_t		−2.675	0.247	9	unit root not rejected		
ADF (intercept)	Δp_t		−17.153	<0.001	4	unit root rejected		
KPSS (intercept)	p_t		3.959	<0.010	27	stationarity rejected		
KPSS (intercept)	Δp_t		0.380	0.086	1	stationarity not rejected		
Ljung–Box $Q(10)$	Δp_t		284.1	<0.001	—	serial dependence		
Ljung–Box $Q(10)$	Δp_t^2		497.9	<0.001	—	volatility clustering		
ARCH-LM(10)	Δp_t		275.4	<0.001	—	conditional heteroskedasticity		
Panel D. Autocorrelation at lag k , rank and linear								
	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$			
Spearman rank, Δp_t	0.388 ***	0.194 ***	0.088 ***	0.031	0.027			
Pearson, Δp_t	0.389	0.169	0.017	−0.073	−0.094			
Spearman rank, $ \Delta p_t $	0.364 ***	0.309 ***	0.278 ***	0.292 ***	0.239 ***			
Pearson, $ \Delta p_t $	0.463	0.304	0.278	0.279	0.259			

Panel A reports the quoted trading days for the level series and the usable consecutive-day increments for the difference. Ex. kurt. is excess kurtosis. *** denotes $p < 0.001$. Partition boundaries in Panel B are those of Table 2; Mean p_t is the mean quoted level over the same span. ADF lag orders are selected by AIC; 5% critical values are −2.863 (intercept) and −3.413 (intercept and trend). KPSS uses automatic bandwidth selection, and its p -values are bounded by the tabulated range. Panel D reports rank and linear autocorrelations over the 1430 usable increments; stars mark the significance of the rank statistic.

Second, the increments are far from Gaussian, and their departure is concentrated in the tails: excess kurtosis is 7.12 over the full sample and 10.83 over the test window (Table 3, Panels A and B). One day accounts for almost all of the latter—excluding the −5.60 Baht/kg decline of 26 June 2026 returns the window to a skew of −0.51 and an excess kurtosis of 0.68—and it matters in advance, because a single exceedance enters the Winkler mean with a weight of twenty at $\alpha = 0.10$.

Third, and most directly relevant to regime weighting, the volatility of the increments is strongly clustered and shifts across the sample. Ljung–Box and ARCH-LM tests on Δp_t^2 both reject at $p < 0.001$, and the standard deviation of Δp_t ranges from 0.377 in the validation block to 0.996 in the calibration block (Table 3, Panels B and C). An undifferentiated split-conformal interval cannot accommodate this heteroskedasticity, which is exactly what regime weighting is designed to exploit. The increments also carry first-order autocorrelation of 0.389 over the full sample and 0.325 in the test window, and 11.4% of days record no price change at all, reflecting the partly administered character of the FOB quote.

Because that one observation accounts for most of the higher moments, the dependence structure is reported in rank as well as linear form (Table 3, Panel D), and the two cut in opposite directions. The persistence of $|\Delta p_t|$ is real rather than an artefact: its rank autocorrelation stays between 0.24 and 0.36 out to five lags, significant at every one, so the volatility clustering that motivates regime weighting survives a statistic the outlier cannot dominate.

The linear statistic nonetheless overstates the first lag, 0.463 against a rank value of 0.364, and in the conditional mean the leverage runs the other way: the Pearson autocorrelation turns negative at lags four and five where the rank statistic stays near zero, so that apparent mean-reversion is a property of one day rather than of the series. Section 3.4 returns to what the base forecaster does and does not recover from this structure.

All series are assembled into a single daily panel from primary sources: Thai physical-rubber prices from the Rubber Authority of Thailand [32] and the Thai Rubber Association, exchange settlements from SHFE, TOCOM and SGX, Bank of Thailand reference rates, Brent and WTI crude, and four Chinese demand indicators. Table 4 lists every series with its source, native frequency and alignment.

Table 4. Variable-level provenance and forecast-origin alignment.

Series	Source	Native Frequency	Time Stamp	Alignment to the Forecast Origin
RSS3 FOB Bangkok, near month	RAOT/TRA via CEIC	daily	quote date, Bangkok close	target; differenced, no lag
Concentrated latex, cup lump, unsmoked sheet	RAOT	daily	quote date, Bangkok close	lagged one trading day
SICOM RSS3, TSR20 settlement	SGX daily quotation	daily	settlement date, Singapore close	lagged one trading day
SHFE rubber settlement	SHFE	daily	settlement date, Shanghai close	lagged one trading day
TOCOM rubber settlement	TOCOM	daily	settlement date, Tokyo close	lagged one trading day
THB/USD reference rate	Bank of Thailand	daily	rate date	lagged one trading day
Brent, WTI crude spot	CEIC	daily	quote date	lagged one trading day
China manufacturing PMI	CEIC	monthly	reference month	effective first Bangkok trading day of the following month
China industrial inventory	CEIC	monthly	reference month	effective first Bangkok trading day of the following month
China natural-rubber imports	CEIC	monthly	reference month	effective first Bangkok trading day of the following month
China automobile production	CEIC	monthly	reference month	effective first Bangkok trading day of the following month

All series are aligned to the Bangkok trading calendar; foreign settlement series are carried forward across foreign holidays that are Bangkok trading days. The daily series are lagged a full trading day, so a forecast for day t uses no quote later than the close of $t - 1$ in any market. Monthly indicators follow a uniform one-month rule rather than series-specific release dates, and are drawn as currently available rather than as first prints; Section 2.2 discusses both limitations.

Monthly indicators are release-lagged: a value stamped in month m becomes effective on the first trading day of month $m + 1$ and is held constant within it. Because several Chinese indicators are released mid-month and later revised, this uniform rule is a conservative approximation rather than an exact release calendar, and a series-specific audit of release dates and vintages remains a data-provenance limitation.

From the raw panel, 64 features are constructed under a single documented formula set: log returns of each price and exchange-rate series, trailing realized volatilities over 5, 10 and 22 trading days, cross-market spreads and futures–spot bases converted to Baht/kg, term-structure slopes, 252-day rolling z-scores, the Chinese demand signals, and calendar dummies for the wintering season and the COVID-19 period. Returns are computed only across consecutive real observations, so no artificial zero-return enters across market closures, and a machine-readable data dictionary records every formula.

Because the argument turns on what is knowable at a forecast origin, that provenance is part of the evidence rather than a technical appendix.

Table 5 collects the symbols used throughout, rather than introducing them piecemeal.

Table 5. Notation and abbreviations used throughout the paper.

Symbol	Meaning	Range or Definition
Δp_t	first difference of the RSS3 FOB price	Baht/kg per day
r	regime index	$r \in \{1, 2\}$; 1 = calm, 2 = stress
i	calibration observation index	$i = 1, \dots, n$, with $n = 287$
τ	forecast origin; the interval is formed for day $\tau + 1$	—
$\hat{\pi}_t(r)$	filtered regime probability	conditions on information up to and including day t
$\rho_t(r)$	one-step predictive regime probability	$\rho_t = \hat{\pi}_{t-1}P$; conditions on information up to $t - 1$ only
P_{rs}	transition probability from regime r to regime s	P_{11}, P_{22} are the staying probabilities
g_i	regime vector attached to calibration residual i	$\hat{\pi}_i$ in the oracle arm, ρ_i in the forecast-feasible arm
γ	similarity-kernel bandwidth	selected from $\{0.5, 1, 2, 4, 8, 16, 32\}$
α	nominal miscoverage level	$\alpha = 0.10$ throughout
$\alpha_{L,r}, \alpha_{U,r}$	lower- and upper-tail miscoverage levels in regime r	$\alpha_{L,r} + \alpha_{U,r} = \alpha$
$w_i, \tilde{w}_i$	raw and normalized calibration weights	$\tilde{w}_i = w_i / (\sum_j w_j + 1)$
ϵ_0	floor added before the log transform of the squared residual	10^{-6}
M	number of VMD modes	$M = 6$
K	number of regime states	$K = 2$ unless stated otherwise
IMF	intrinsic mode function, the output of the decomposition	six per forecast origin
BW	the bandwidth partition, used to select γ	$n = 42$

The distinction between $\hat{\pi}_i$ and ρ_i carries the paper's argument and is developed in Section 2.3. "Calm" and "stress" name the two states of the fitted Markov-switching model and are used only in that sense; the variable supplied to the similarity kernel is called the regime signal. The interval bounds $[L, U]$ appearing in the Winkler score of Section 2.5 are distinct from the tail-level subscripts $\alpha_{L,r}$ and $\alpha_{U,r}$.

2.2. Base Point-Forecast Model

The conformal layer calibrates on whatever residuals a base forecaster leaves behind, so the diagnostic does not depend on which forecaster produces them. Two complementary forecasters of deliberately different complexity are used here, and the central decomposition is reported on both, so its conclusion does not depend on either. The simpler is a single-parameter autoregression,

$$\Delta p_t = c + \phi \Delta p_{t-1} + \varepsilon_t, \quad (1)$$

which uses no covariates at all and attains a test mean absolute scaled error of 0.921. The richer is a ridge-regularized linear regression on 64 lagged economic features augmented with Variational Mode Decomposition components, following Pinitjitsamut [31], which reaches 1.00 on the same window. Neither is offered as the recommended forecaster; reporting both is what makes the diagnostic separable from the choice of base model.

The regression stage that maps those features to a forecast is deliberately linear, chosen for transparency rather than point accuracy. Were the conditional mean the object of study, a more expressive model would be the tool of choice: the VMD-augmented bidirectional LSTM of Pinitjitsamut [31], on the same target, attains substantially higher point accuracy while preserving the series' dispersion. Here, a simple, well-understood regression keeps the regime-weighting question free of base-model tuning. It still produces residuals with

realistic dispersion, at a predicted-to-realized ratio of roughly 0.5–0.9 and a validation-partition correlation near 0.41. That this forecast sits close to a no-change benchmark on the test window (Section 3.4) is a property of the base model rather than of the conformal analysis, which turns on the regime structure of the residuals.

The ridge penalty is selected from $\{1, 3, 10, 30, 100, 300, 1000\}$ by validation mean-squared error, and the regime-similarity bandwidth γ from $\{0.5, 1, 2, 4, 8, 16, 32\}$ by the coverage-width criterion on the bandwidth block. Neither uses the test partition. The moving-block bootstrap is not a tuning choice but the inference tool used throughout; Section 2.5 gives its settings. Its penalty selection on a validation partition disjoint from the calibration set prevents model-selection leakage into the calibration residuals, though it does not restore the exchangeability that dependent residuals lack (Section 2.5, Supplementary Section S2).

Because VMD is a global operator, applying it to the whole sample would leak future information; the decomposition therefore uses an expanding window, recomputed at every origin and lagged one trading day together with all 64 economic features. Supplementary Section S1 sets out that treatment, and Table A1 the VMD settings and every other reproducibility parameter, together with a check in which the VMD components are removed entirely—which leaves the decomposition of Section 3.1 unchanged.

2.3. Regime Model and the Two Design Factors

This subsection defines the two experimental factors the paper varies. Both govern how much information the similarity weights may see: the first fixes *when* the regime signal is measured, the second *which data* the model producing it was estimated on. Every result in Section 3 is a cell of the grid they define.

Volatility regimes are inferred from the base-model residuals by a two-state Gaussian Markov-switching model fitted to the log-squared residual. Writing $z_t = \log(\hat{\epsilon}_t^2 + \epsilon_0)$ for that feature, with $\epsilon_0 = 10^{-6}$ guarding against a residual of exactly zero, and $S_t \in \{1, 2\}$ for the latent state,

$$z_t \mid S_t = r \sim \mathcal{N}(\mu_r, \sigma_r^2), \quad P(S_t = s \mid \mathcal{F}_{t-1}) = \sum_r P_{rs} P(S_{t-1} = r \mid \mathcal{F}_{t-1}), \quad (2)$$

so each state carries its own mean and variance in z_t , and the transition matrix P propagates the state one step ahead. Two design choices determine the information content of the weights this signal produces—analyzed in Supplementary Section S2—and the study varies both.

(i) State.

Two probabilities are available. The filtered probability, $\hat{\pi}_t(r) = P(\text{state} = r \mid \text{information up to } t)$, conditions on the target day and therefore encodes its shock. The one-step predictive probability, $\rho_t(r) = P(\text{state} = r \mid \text{information up to } t-1) = (\hat{\pi}_{t-1}P)(r)$, conditions only on the past. The distinction is one of availability: ρ_t is known before the period- t residual is observed, whereas $\hat{\pi}_t$ is not, so using $\hat{\pi}_{t+1}$ to build an interval for Δp_{t+1} is an oracle operation rather than a deployable forecast. Predictive timing does not, however, make the resulting similarity weights fixed, nor establish the coverage conditions of Barber et al. [11], because the weight vector remains a function of the historical score path (Supplementary Section S2.1). The filtered probability is retained only as an oracle comparison arm.

(ii) Estimation sample.

The Markov-switching parameters may be estimated on the calibration window, on a disjoint validation block, or on both. The first option is delicate because the calibration

window supplies the conformal scores themselves. Estimating on it is operationally feasible at the final origin, since those observations are historical, but it makes the auxiliary regime model depend on the same scores later used for calibration. This calibration-dependent estimation, or score reuse, falls outside the fixed-weight analysis, and it may improve apparent localization through calibration-period adaptation. Estimation on a disjoint validation block removes that particular dependence. The recursively generated state probabilities remain random. The calibration-informed variants are retained as comparison arms.

Crossing the two factors yields the six regime-weighted constructions of Section 3, which correspond to the forecast-feasible and coupled branches of Figure 1. One cell is distinguished: the predictive-state, validation-fitted construction removes the two most direct couplings and is referred to throughout as the forecast-feasible construction. The term is deliberately narrow.

It records only that every quantity entering the interval is measurable at the forecast origin and that the regime model is estimated on a block disjoint from the calibration scores; it asserts neither causal inference nor finite-sample validity, since the similarity weights remain data-dependent functions of the realized score path (Supplementary Section S2). The two states are named once and used consistently thereafter: the higher-mean state in z_t is the stress regime and the lower-mean state the calm regime. The naming is relative to each fitted model, and on this series the stress probability exceeds one-half on roughly 70% of sample days—a property of the fitted chain rather than a claim about turbulent days.

Table 6 reports the estimates, because their instability across blocks is itself part of the evidence. Expected stress durations range from 4.7 days when the model is fitted on validation to 20.0 days when fitted on validation and calibration together, and the mean stress probability over the test window ranges from 0.086 to 0.671. A regime model estimated on one block is therefore not the same object as one estimated on another—which is why the estimation sample enters as an experimental factor rather than an implementation detail, and why the transfer failure of Section 3.5 is unsurprising in retrospect.

Table 6. Estimated two-state Markov-switching models and the agreement of their regime signals with a volatility class the model cannot influence.

Panel A. Parameter estimates by fitting block									
Fitted on	P_{11}	P_{22}	μ_{calm}	μ_{stress}	σ_{calm}	σ_{stress}	Dur. calm	Dur. stress	$\bar{P}_{\text{stress}}$
Validation	0.992	0.785	−4.128	0.160	2.330	0.297	121.8	4.7	0.086
Calibration	0.560	0.844	−3.683	−0.563	2.565	1.247	2.3	6.4	0.671
Validation + calibration	0.939	0.950	−3.992	−0.762	2.536	1.429	16.5	20.0	0.621
Panel B. Agreement of each regime signal with realised volatility (test window, $n = 131$)									
Regime model	Regime signal		Target		AUC		95% CI		
Calibration-fitted	Filtered		same-day		0.633		[0.532, 0.727]		
Calibration-fitted	Filtered		next-day		0.643		[0.545, 0.740]		
Calibration-fitted	Predictive		next-day		0.632		[0.531, 0.728]		
Validation-fitted	Filtered		same-day		0.567		[0.465, 0.662]		
Validation-fitted	Predictive		same-day		0.547		[0.444, 0.648]		
Validation-fitted	Predictive		next-day		0.542		[0.442, 0.636]		

Panel A: models are fitted to the log-squared residual of the base forecaster. States are ordered by the mean of that feature, so “calm” is the lower-mean state and “stress” the higher-mean one. P_{11} and P_{22} are the probabilities of remaining in the calm and stress states; expected duration in days is $1/(1 - P_{ii})$, and $\bar{P}_{\text{stress}}$ is the mean stress probability over the test window. Panel B: the volatility class is the highest third of a trailing ten-day realized residual volatility, which no construction can influence. “Filtered” uses information up to and including the day itself, “predictive” up to the previous day only. AUC is the area under the ROC curve, 0.50 denoting chance; bootstrap intervals use 2000 resamples.

Panel B reports how well each signal tracks a volatility class the model cannot influence, and the two fitting blocks separate cleanly: the calibration-fitted signals reach an AUC of 0.63 to 0.64, the validation-fitted ones—those the forecast-feasible construction actually uses—only 0.54 to 0.57, with intervals including 0.50. In both cases, the predictive signal is no worse than the filtered one against next-day volatility, the quantity a one-step interval must anticipate. The signal is therefore weak where it is deployable, and its timing buys nothing for the target that matters.

2.4. Regime-Weighted Conformal Construction

At a forecast origin τ , each calibration residual i receives a weight reflecting how closely its regime vector g_i resembles that of the origin, $g_{\tau+1}$. The kernel is exponential in the ℓ_1 distance between the two,

$$w_i(\tau) = \exp(-\gamma \|g_i - g_{\tau+1}\|_1), \quad i = 1, \dots, n, \quad w_{n+1} = 1, \quad (3)$$

and the weights are normalized over the calibration set together with the test point,

$$\tilde{w}_i = \frac{w_i}{\sum_{j=1}^n w_j + 1}, \quad i = 1, \dots, n, \quad \tilde{w}_{n+1} = \frac{1}{\sum_{j=1}^n w_j + 1}. \quad (4)$$

Here, g is the chosen regime signal: the predictive ρ in the forecast-feasible construction, the filtered $\hat{\pi}$ in the oracle comparison arm. A larger γ concentrates the weight on calibration days whose regime vector is closest to the origin's; $\gamma \rightarrow 0$ returns uniform weights and hence ordinary split-conformal.

The interval bounds are the weighted lower- and upper-tail quantiles of the calibration residuals, added to the point forecast. Writing δ_x for a point mass at x , and augmenting the weighted calibration distribution with the $\pm\infty$ atoms that the weighted conformal quantile requires, the bounds are

$$q_L = Q_{\alpha_L}(\sum_i \tilde{w}_i \delta_{\hat{e}_i} + \tilde{w}_{n+1} \delta_{-\infty}), \quad q_U = Q_{1-\alpha_U}(\sum_i \tilde{w}_i \delta_{\hat{e}_i} + \tilde{w}_{n+1} \delta_{+\infty}), \quad (5)$$

$$\hat{C}(\tau) = [\Delta \hat{p}_{\tau+1} + q_L, \Delta \hat{p}_{\tau+1} + q_U], \quad (6)$$

where $\hat{e}_i$ is the i th calibration residual. The miscoverage budget is split between the tails by regime, $\alpha_{L,r} + \alpha_{U,r} = \alpha$, so that the heavier tail receives the smaller miscoverage level and hence the wider bound. Let s_r^- and s_r^+ denote the weighted lower- and upper-tail spreads of the regime- r residuals about their weighted median. The tail levels are set in inverse proportion to the spread each side carries,

$$\alpha_{L,r} = \alpha \frac{s_r^+}{s_r^- + s_r^+}, \quad \alpha_{U,r} = \alpha \frac{s_r^-}{s_r^- + s_r^+}, \quad (7)$$

which recovers symmetry when $s_r^- = s_r^+$.

Estimated from the calibration-tail spreads, these levels are data-dependent and fall outside the fixed-level coverage argument. The fixed symmetric allocation $\alpha_L = \alpha_U = \alpha/2$ is therefore the primary choice, and the asymmetric allocation is a sensitivity analysis. The bandwidth γ is selected on the held-out bandwidth partition by a coverage-width criterion, separately for each construction. Algorithm 1 summarizes the procedure. It is applied identically across all factor cells, with only the regime signal g and its estimation sample varying.

Algorithm 1 Regime-weighted interval at forecast origin τ

Input: calibration residuals $\{\hat{\epsilon}_i\}$ with regime vectors $\{g_i\}$; origin regime vector $g_{\tau+1}$; point forecast $\Delta\hat{p}_{\tau+1}$; level α ; bandwidth γ . (g = predictive ρ in the forecast-feasible construction, filtered $\hat{\pi}$ in the oracle comparison arm.)

1. Similarity weights. Form w_i and $\tilde{w}_i$ by Equations (3) and (4).
2. Tail budget. Identify the origin's dominant regime r from $g_{\tau+1}$; from the weighted tail spreads s_r^-, s_r^+ set $(\alpha_{L,r}, \alpha_{U,r})$ with $\alpha_{L,r} + \alpha_{U,r} = \alpha$ as above.
3. Weighted quantiles. Augment the calibration residuals with point masses at $-\infty$ and $+\infty$ carrying weight $\tilde{w}_{n+1}$ (the finite-sample correction of Barber et al. [11]). Compute q_L , the weighted α_L -quantile of the augmented set (with the $-\infty$ atom), and q_U , the weighted $(1 - \alpha_U)$ -quantile (with the $+\infty$ atom). The primary specification uses the fixed symmetric levels $\alpha_L = \alpha_U = \alpha/2$; the per-regime asymmetric levels of step 2 are a sensitivity variant.
4. Interval. Return $\hat{C}(\tau) = [\Delta\hat{p}_{\tau+1} + q_L, \Delta\hat{p}_{\tau+1} + q_U]$.

2.5. Evaluation Protocol and Experimental Design

The five chronological blocks of Table 2 each serve one purpose, and the separation between them is what the diagnostic of this paper turns on. The base model is estimated on the training block; the validation block selects the ridge penalty and, in the forecast-feasible construction, estimates the regime parameters; the calibration block supplies the conformal residuals; the bandwidth block selects γ ; and the test block is held out.

Intervals target $\alpha = 0.10$ (90% coverage). Four measures assess performance. The first is marginal coverage, reported with moving-block-bootstrap confidence intervals [33] to account for serial dependence in the coverage indicators, the block resampling being what preserves the short-run dependence an i.i.d. bootstrap would destroy (block length $\lfloor \sqrt{n} \rfloor$, 2000 resamples, percentile intervals, seed 0). The second is regime-conditional coverage, evaluated on a high-volatility mask that no construction can influence—the highest third of test-window days ranked by realized volatility—so that no construction is scored on its own regime signal. The third is mean interval width. The fourth is the Winkler (interval) score, which for a $100(1 - \alpha)\%$ interval $[L, U]$ and realized y equals $(U - L) + \frac{2}{\alpha}(L - y)\mathbf{1}\{y < L\} + \frac{2}{\alpha}(y - U)\mathbf{1}\{y > U\}$, with lower values sharper and better calibrated. Paired differences in width and Winkler score against split-conformal are reported with moving-block-bootstrap intervals.

The Winkler score is proper and standard in the interval-forecasting literature. Three limitations shape how the results below should be read; Zhou et al. [34] survey them for load forecasting. First, the score implicitly favors unimodal predictive behavior and is ill-suited to heavy-tailed predictive distributions—a caution that matters for a target whose increments carry an excess kurtosis above seven (Table 3). Second, it emphasizes average performance, so a favorable mean score is compatible with poor behavior in exactly the rare, high-impact episodes where an interval matters most. Third, it is not comparable across datasets of differing predictability, which is why the comparisons here are always paired against split-conformal on the same origins rather than quoted as absolute levels.

Two further properties follow from the definition. The score compresses calibration and sharpness into one number, so two methods can attain the same value for opposite reasons—one narrow and under-covering, the other wide and well-covered. And the exceedance penalty is large: at $\alpha = 0.10$ it multiplies each miss by twenty, so a handful of exceedances dominates the mean, the sampling distribution is heavy-tailed, and paired comparisons on a short test window have limited power. A fifth point is specific to this study: the score is undefined whenever an interval is unbounded, a case the finite-sample correction makes real.

None of these is hypothetical here: each surfaces in Sections 3.1, 4.4 and 3.6, respectively. Coverage, width, and the Winkler score are therefore reported jointly throughout, each with moving-block-bootstrap intervals, and no conclusion rests on the interval score alone.

The experimental design has five parts. A decomposition grid crosses split-conformal with the six regime-weighted constructions of Section 2.3 on the test window, isolating each factor. The entire pipeline is then re-estimated at six rolling origins. The forecast-feasible construction is benchmarked against Adaptive Conformal Inference and its dynamically-tuned aggregate on a common rolling protocol. Point-forecast metrics locate the result in the forecastability of the residuals, and the grid is reported on both base forecasters, so that the null is not read off a single one. A controlled simulation (Sections 2.6 and 3.9) supplies ground-truth evidence.

2.6. Controlled Simulation Design

The application shows that the apparent gain disappears but cannot show what a genuine gain would have looked like, because the regime structure of the RSS3 residuals is whatever it is. A simulation supplies that counterfactual: the construction of Section 2.4 runs on synthetic scores whose predictability the design fixes in advance. Its parameters span that lever rather than replicate the rubber series, so the dispersion and shape of the simulated states deliberately differ from the empirical ones—a controlled experiment, not a resimulation.

Data-generating process. Scores are generated by a two-state Markov chain with symmetric persistence $p = P(\text{state unchanged})$. Conditional on the state, the residual is normal in the calm state (standard deviation $\sigma_{\text{calm}} = 1$) and skew-normal in the stress state (scale $\sigma_{\text{stress}} = \text{sep}$, shape -3). Each state is centered to a zero conditional mean, so the regimes differ only in dispersion and shape, and no conditional-mean shift is confounded with the volatility regime. The skew-normal scale is not its standard deviation. At $\text{sep} = 3.5$, the stress state has an actual standard deviation of about 2.3, so the state SD ratio is roughly 2.3 rather than 3.5. Its skewness is about -0.67 , a mild negative skew in keeping with commodity down-moves.

Construction and partitions. The construction builds intervals with the finite-sample $\pm\infty$ correction and fixed symmetric tails of Algorithm 1, identically to the empirical analysis. The series is partitioned chronologically into validation, calibration, bandwidth, and test blocks as in Table 2. One departure from the empirical specification must be recorded.

The Markov-switching model is fit to a causal local log-realized-variance, the trailing-window mean of the squared residual, because the one-point log-chi-square noise of a single residual prevents the filter from recovering even a strongly persistent state. That smoothing is a stated modeling choice, not a result, and it makes the latent state easier to recover than in the application, which fits the model to the log-squared residual (Appendix A). Section 3.9 quantifies the gap it creates.

Experimental lever. The design sweeps one lever, the persistence p , which governs how one-step-predictable the regime is. The separability $\text{sep} = \sigma_{\text{stress}}/\sigma_{\text{calm}}$, which governs how distinct the regimes are, is held fixed at 3.5. This is therefore a persistence experiment rather than a two-lever sweep, and a sweep of the separability lever is left to future work.

At each persistence, two constructions are compared with split-conformal by the paired Winkler difference: the forecast-feasible one, with predictive state and regime fit on validation, and the artifact one, with filtered state and regime fit on calibration. Differences are averaged over 250 Monte Carlo replications for the persistence experiment, 300 for the score-separation experiment, and 200 for the no-regime null, each with 95% Monte Carlo intervals, and no replication is discarded. Unbounded origins are retained for the

unbounded-rate calculation but excluded from the conditional score, and every paired comparison is taken over the constructions' common finite-origin set.

The no-regime null. To confirm that the pipeline does not itself manufacture an advantage, an exact no-regime null runs over 200 replications. Its two states share an identical $N(0, 1)$ emission, so the regime label carries no distributional information. The null runs as a full 2×2 factorial, crossing state timing (filtered versus one-step predictive) with the regime-estimation sample (calibration-fit versus disjoint validation-fit). This lets the state-timing and estimation contributions be read off as direct paired contrasts.

Under the null, predictable weighting carries no population-level regime information and should show no systematic expected advantage over split-conformal, though finite-sample differences may occur. Two limits follow. Stationarity leaves the estimation-sample coupling little to transfer, so the design isolates the state coupling and does not reproduce the non-stationarity that activates the other in the application (Section 3.1). And the score-separated arm, which fits the regime model on the first half of the calibration block and forms scores on the second, halves the effective calibration size to 150 of 300 scores, so any change it produces confounds score separation with a smaller calibration set.

3. Results

The evidence is presented in eight steps. Section 3.1 exhibits the apparent advantage of regime weighting and examines how it varies with the two information couplings. Section 3.2 shows the resulting null to be reproduced under a limited six-origin rolling-origin re-estimation, Section 3.3 places the forecast-feasible construction against adaptive-conformal baselines, and Section 3.4 reports the point-forecast context that locates the null in the residuals the conformal layer receives. Section 3.5 reports the same decomposition on the AR(1) residuals, and Section 3.6 evaluates two operational fallback rules for the unbounded intervals the finite-sample correction can produce.

Section 3.7 gathers the robustness evidence: the full bandwidth grid with weight-concentration diagnostics, a benchmark of the latent state against an observed volatility feature, and variation of the forecast origins, the nominal level, the benchmark family and the number of regime states. Section 3.9 corroborates the mechanism in a controlled simulation with a known regime. Throughout this section, a negative Δ Winkler denotes an improvement over split-conformal, and a bootstrap interval that spans zero denotes no statistically detectable difference from it. Results are reported on the ridge residuals of Section 2.2 unless the AR(1) residuals are named; Section 3.5 gives the same decomposition on the latter. All constructions target 90% coverage ($\alpha = 0.10$). Single-window results use the test partition ($n = 131$), and regime-conditional coverage is reported on the high-volatility mask defined in Section 2.5.

Two qualifications attach to that mask. It is an ex-post evaluation device defined on realized test-window volatility, not a deployable regime classifier available at the forecast origin. And with only 44 days it is severely underpowered: a 90% target implies roughly four to five expected misses, so coverage estimates carry sampling uncertainty of order ± 0.09 before serial dependence is considered. The high-volatility figures are therefore descriptive coverage on an ex-post subset, not evidence of conditional validity.

3.1. An Apparent Advantage and Its Decomposition

The first question is whether the headline gain belongs to regime weighting itself or to the information the weights are allowed to see. The two are separable only if the design choices that distinguish a regime-weighted interval from split-conformal are varied one at a time, rather than adopted together as a package. Table 7 crosses those two choices. The first is the regime signal: the filtered probability, which conditions on the target day, or

the one-step predictive probability, which conditions only on the past. The second is the estimation sample for the Markov-switching parameters: the calibration window, a disjoint validation block, or both.

Table 7. Decomposition of the regime-weighting effect (test partition, $\alpha = 0.10$, $n = 131$).

Construction	Coverage	Cov. (hi-vol)	Width	Winkler	Δ Winkler [95% CI]
Split-conformal	0.924	0.795	3.118	4.378	—
Filtered, calibration-fit	0.939	0.818	3.189	4.159	−0.22 [−0.36, +0.09]
Filtered, (val.+calib.)-fit	0.939	0.818	3.219	4.288	−0.09 [−0.27, +0.12]
Filtered, validation-fit	0.939	0.818	3.175	4.402	+0.02 [−0.04, +0.09]
Predictive, calibration-fit	0.931	0.818	3.170	4.404	+0.03 [+0.00, +0.06]
Predictive, (val.+calib.)-fit	0.939	0.818	3.195	4.409	+0.03 [−0.04, +0.08]
Predictive, validation-fit	0.916	0.773	2.990	4.433	+0.06 [−0.15, +0.17]

Cov. (hi-vol) is coverage on the highest third of test-window days by realized volatility (44 of 131 days), a ranking no construction can influence. Δ Winkler is the paired difference against split-conformal on the same origins; negative is an improvement, and intervals are 95% moving-block-bootstrap. All intervals use the finite-sample $\pm\infty$ correction and fixed symmetric tails. “val.” and “calib.” abbreviate the validation and calibration blocks; the predictive, validation-fitted cell is the forecast-feasible construction.

Consider first the naive implementation: the plain weighted quantile without the finite-sample correction, using the filtered state and a regime model estimated on the calibration window. At the bandwidth this criterion selects, $\gamma = 8$, its *mean interval width* is 2.434 Baht/kg against 3.047 for the corresponding split-conformal baseline, a reduction of 20.1%. The reduction is in width, not in coverage: coverage for every construction is reported separately in Table 7 and changes by at most 0.02. The naive arm also attains a paired Winkler difference of $-0.96 [-1.32, -0.48]$, excluding zero (Table S2, first row). Note that the naive arm uses the interpolating weighted quantile throughout, so its split-conformal baseline (3.047) differs from the order-statistic baseline of Table 7 (3.118); the two are never mixed within a comparison. Reported in isolation, this reads as a clear methodological success, and it is the number such implementations report.

The naive figure does not survive the corrected specification. The correction does not merely widen the interval at a given bandwidth; it removes the bandwidths at which the apparent gain lives. Without the correction, no interval is ever unbounded, so the coverage–width criterion is free to select a highly concentrated kernel; the naive figure above is obtained at $\gamma = 8$. Under the correction, that bandwidth is inadmissible: it leaves 31 of the 131 test origins unbounded, and $\gamma = 4$ already leaves 8. The selection is therefore forced down to $\gamma = 2$, where the paired difference is -0.22 rather than -0.96 . Holding the bandwidth fixed at $\gamma = 2$ separates the two effects: the uncorrected construction gains -0.43 there and the corrected one -0.22 , so roughly half of the headline advantage is the atom and the remainder is the concentration the atom forbids.

The intervals of Table 7 are accordingly built with the finite-sample $\pm\infty$ correction that the weighted-conformal quantile requires (Supplementary Section S2.1) and with fixed symmetric tails. Under this specification, the coupled cell falls to a paired Winkler difference of -0.22 , with a bootstrap interval of $[-0.36, +0.09]$ that spans zero. Its width, 3.189, is no longer distinguishable from split-conformal at 3.118. Supplementary Section S3 reports the full six-cell grid with and without the correction. The finite-sample correction widens most exactly the cells whose weights are most concentrated—the coupled cells—because their effective sample is smallest; the apparent narrowing was therefore in large part a small-sample artifact, not only an information coupling.

Under the finite-sample-corrected specification, no cell significantly improves on split-conformal, and the two couplings remain separately identifiable in what survives. Replacing the filtered state with the predictive state is the only choice that removes the

direct dependence on the target-day shock. It leaves a paired interval-score difference of $+0.03$ [$+0.00, +0.06$] for the calibration-fitted regime, and $+0.06$ [$-0.15, +0.17$] for the forecast-feasible construction. Both bootstrap intervals span zero. Independently, retaining the filtered state but estimating the regime out of sample, on the validation block alone, yields $+0.02$ [$-0.04, +0.09$], again with no statistically detectable difference from split-conformal. The forecast-feasible construction—predictive state, out-of-sample estimation—shows no statistically detectable difference from split-conformal in interval score; its mean width is 2.990 against 3.118 for split-conformal.

The intervals might still adapt in the right direction, widening on turbulent days and narrowing on calm ones, even where the average width is unchanged. They do not. On the highest third of days by realized volatility, the realized absolute residual is 0.964 against 0.513 on the remaining days, a ratio of 1.88. Split-conformal, which cannot adapt at all, has a width ratio of exactly 1.00 by construction. The filtered, calibration-fitted cell reaches only 1.10 (3.39 against 3.09), despite conditioning on the target day's own shock. The forecast-feasible construction reaches 0.97 (2.93 against 3.02), which is to say it is marginally narrower on the volatile days. Regime weighting on this series therefore neither sharpens on average nor redistributes width towards the days where the risk lies.

Figure 3 displays the six cells with and without the correction. They locate the apparent advantage outside the regime-weighting idea itself: what narrows the interval is the information the weights may use, not the act of weighting. Whether that conclusion is an artifact of one arbitrary calibration-test split is the question Section 3.2 takes up.

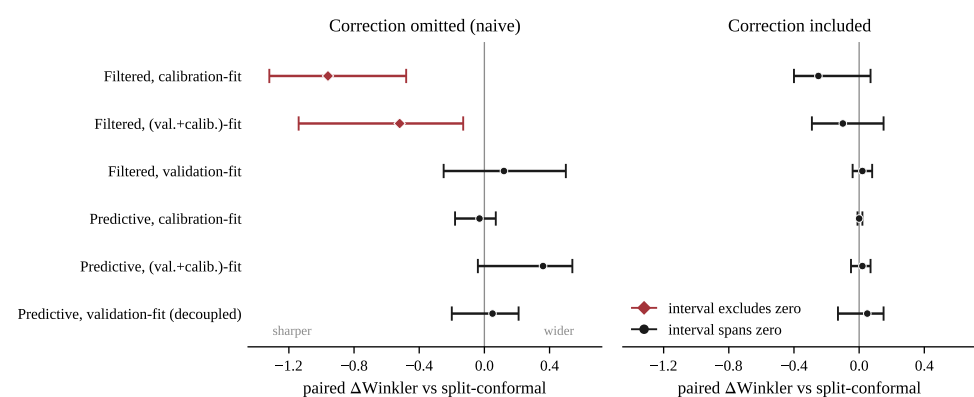


Figure 3. Paired Winkler differences against split-conformal for the six state $\times$ estimation cells, with 95% moving-block-bootstrap intervals; negative values are sharper. The left panel omits the finite-sample $\pm\infty$ correction, as a naive implementation does, and the right panel includes it. Diamonds mark intervals excluding zero. Every apparent gain belongs to a filtered cell with calibration-informed regime estimation, and for the VMD-augmented ridge residuals shown here, none of them survives the correction; Section 3.5 reports a case on the AR(1) residuals where one does. The two panels plot the numbers of Table S2 and, in the right panel, of Table 7.

3.2. Robustness Under Rolling-Origin Re-Estimation

A single test window can flatter or penalize any method by accident of where it falls in the sample. The decomposition therefore repeats at six rolling origins, with every component re-estimated from scratch—expanding-window VMD, ridge with penalty selected on validation, the Markov-switching regime model, and the conformal calibration—comparing the three headline constructions per origin.

Across all six origins, the forecast-feasible construction tracks split-conformal in interval score, at a mean of 4.60 against 4.99 over the five finite origins, while the coupled filtered-and-calibration construction stays ahead at 4.07. Under the finite-sample correction, however, both weighted constructions return unbounded intervals at one of the six origins (Table 8). Over the five origins at which every method stays finite, the ordering matches the

single-window analysis. The generally low coverage at the earliest, most volatile origins (0.50–0.70 for every construction) is itself informative: the series is non-stationary enough that split-conformal itself under-covers materially in places, and regime weighting does not repair this.

Table 8. Full-refit rolling-origin evaluation (six origins).

Construction	Unbounded	Coverage	Width	Winkler
Split-conformal	0/6	0.767	2.51	4.99 ± 1.11
Predictive, validation-fit	1/6	0.812	2.60	4.60 ± 0.96
Filtered, calibration-fit	1/6	0.868	2.74	4.07 ± 0.71

Unbounded is the number of the six origins at which the construction returns an infinite interval. Coverage is over all six origins; width and Winkler (mean ± sd) are over the origins at which the method returns finite intervals. Under the finite-sample $\pm\infty$ correction both weighted constructions are unbounded at one origin, precluding a complete mean-score comparison there. Where finite, the forecast-feasible construction tracks split-conformal and the coupled construction remains ahead.

The rolling evidence therefore reproduces the single-window null and adds a failure mode of its own: once correctly constructed, the regime-weighted interval is not reliably finite. That makes a comparison with methods which are always finite, and which adapt without any regime model at all, the natural next step.

3.3. Comparison with Adaptive Baselines

Failing to beat the simpler adaptive alternatives a practitioner would otherwise reach for is the more demanding test. Those alternatives carry no regime model and adapt only from realized miscoverage, so they set the bar the additional regime structure must clear. The benchmark asks whether the forecast-feasible construction improves on them: Adaptive Conformal Inference (ACI) and its dynamically-tuned aggregate (DtACI), evaluated on a common rolling protocol.

DtACI and ACI attain finite mean interval scores of 3.83 and 4.16, respectively, compared with 4.58 for rolling split-conformal. The finite-sample-corrected regime-weighted interval has no finite mean score in this protocol: under the finite-sample correction, it returns unbounded intervals at 8 of 131 origins (Table 9). A direct comparison of mean interval scores is therefore not meaningful; ACI and DtACI remain finite throughout and offer substantially greater operational reliability. Concentrated weights drive this. Over the single test window, the weight sum stays above the threshold that Section 3.6 states, with a minimum of 19.4 and an effective sample size that falls from 213 to 29 at the most concentrated origins. The shorter rolling windows breach it.

Table 9. Adaptive-conformal baselines on the common rolling protocol (calibration window 220).

Method	Coverage	Width	Winkler
Split-conformal	0.908	3.070	4.578
RWCP (predictive, validation-fit)	0.916	2.825 ‡	∞ ‡
ACI	0.885	2.894	4.164
DtACI	0.908	2.761	3.826

‡ Under the finite-sample $\pm\infty$ correction, the regime-weighted construction returns unbounded intervals at 8 of 131 origins, so its mean interval score is undefined; the reported width is over the finite intervals. ACI and DtACI, which recalibrate online, improve on split-conformal and remain finite throughout.

The additional structure thus fails to clear the bar and forfeits a property both baselines retain, namely an interval that is always finite. What remains to be explained is why the regime signal has so little to contribute to this series; the answer lies one layer down, in the point forecast.

3.4. Point-Forecast Context

The preceding subsections establish that regime weighting does not improve on split-conformal here, but not why. Because the conformal layer calibrates on the base model's residuals, the answer lies in how much predictable structure that model leaves behind. On the test partition, the VMD-augmented ridge model attains a mean absolute error of 0.665 against 0.662 for a no-change benchmark—a mean absolute scaled error of 1.00—with a predicted-to-realized dispersion ratio of 0.37. The ridge regression therefore performs no better than a naive no-change forecast on this window.

That performance is a property of the fitted model rather than of the target. The differenced series is not a martingale. A moving-average specification is the natural alternative for a differenced price. An MA(1) fitted on the same expanding pre-test window attains a test mean absolute error of 0.613, against 0.602 for the AR(1) and 0.662 for the no-change benchmark—a mean absolute scaled error of 0.93 against 0.91. The autoregressive form is therefore retained as the simpler complementary forecaster, though the margin between the two is small enough that neither choice would change what the conformal layer receives. Its first-order autocorrelation is 0.389 over the full sample and 0.325 over the test window (Table 3). A single-parameter AR(1), fitted on an expanding pre-test window, attains a test mean absolute error of 0.602 against 0.662 for the no-change benchmark, a mean absolute scaled error of 0.91. The ridge regression with 64 economic features and six VMD modes does not recover that dependence. Its residuals retain a rank autocorrelation of 0.276 at the first lag ($p < 0.001$, $n = 589$), against 0.388 for Δp_t itself (Table 3, Panel D): under a third of the first-order dependence is removed, and none is left at lags two to five. The residuals passed to the conformal layer therefore stay close to the raw price increments, whereas an AR(1) removes that first-order dependence by construction.

Two consequences follow. The first leaves the diagnostic result untouched. The couplings of Section 3.1 operate on how residuals are weighted by regime, not on how large the residuals are, and the controlled null of Section 3.9 exhibits them with no base model at all. The second bounds the null. The empirical analysis establishes that the residuals of *this* base model carry too little one-step-predictable volatility structure for regime weighting to exploit. That is why the forecast-feasible construction shows no detectable improvement on split-conformal, and why the filtered construction can appear to, by conditioning on the contemporaneous shock. Whether a base forecaster that does recover the conditional-mean dependence would leave exploitable regime structure in its residuals is the question Section 3.5 takes up directly.

3.5. The Decomposition on the AR(1) Residuals

Section 3.1 reported the decomposition on the ridge residuals. Table 10 reports the same decomposition on the AR(1) residuals of Equation (1), which beat the ridge regression on point accuracy (test MASE 0.921 against 1.00). The comparison answers a natural objection—that the null is an artefact of a weak base forecaster—and it also shows where the two base models genuinely differ. The conformal layer, the partitions, the regime specification, the bandwidth grid, and the evaluation protocol are unchanged; only the residuals differ.

The pattern of Table 7 is reproduced in full. In its naive form, the coupled construction attains a paired Winkler difference of -0.66 $[-1.12, -0.34]$, whose bootstrap interval excludes zero—a headline gain of the same order as the one the VMD-augmented base produces. Every gain again belongs to a filtered cell with calibration-informed regime estimation. Coverage on the highest third of days by realized volatility is 0.841 for split-conformal and 0.864 for every regime-weighted cell except the forecast-feasible construction, which matches split-conformal exactly. Replacing the filtered state with the predictive state

eliminates it at each estimation sample, and the forecast-feasible construction shows no statistically detectable improvement.

Table 10. Decomposition of the regime-weighting effect on AR(1) base residuals (test partition, $\alpha = 0.10$, $n = 131$).

Construction	γ	Coverage	Width	Naive	Corrected
Split-conformal	—	0.947	2.962	—	—
Filtered, calibration-fit	4	0.954	2.618	−0.66 [−1.12, −0.34]	−0.54 [−0.86, −0.22]
Filtered, (val.+calib.)-fit	4	0.954	2.622	−0.68 [−1.13, −0.37]	−0.49 [−0.90, −0.18]
Filtered, validation-fit	0.5	0.954	3.083	+0.01 [−0.01, +0.04]	+0.08 [+0.06, +0.12]
Predictive, calibration-fit	2	0.947	3.076	+0.06 [−0.04, +0.17]	+0.10 [+0.02, +0.18]
Predictive, (val.+calib.)-fit	4	0.939	3.066	−0.03 [−0.26, +0.16]	+0.03 [−0.12, +0.26]
Predictive, validation-fit	0.5	0.947	2.962	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]

Base forecaster is an AR(1) fitted on the pre-validation sample (test MASE 0.921); only the base residuals differ from the other tables. Width is that of the corrected construction. The Naive and Corrected columns both report the paired Δ Winkler against split-conformal, without and with the finite-sample $\pm\infty$ correction, respectively; negative is an improvement, and intervals are 95% moving-block-bootstrap. “val.” and “calib.” abbreviate the validation and calibration blocks. No construction returns an unbounded interval on this window.

Two differences from Section 3.1 both cut against the method. First, the finite-sample correction is a weaker defense here than for the ridge residuals. It reduces the coupled cell from −0.66 to −0.54, but the bootstrap interval still excludes zero, because the selected bandwidth leaves the calibration weights less concentrated and the $\pm\infty$ atoms therefore bind at no origin. Applying the correction is thus necessary but not sufficient: on this base model only the removal of the filtered state disposes of the artifact.

Second, the forecast-feasible construction is numerically identical to split-conformal, not merely statistically indistinguishable from it. The regime model is estimated here on the validation block, the calmest of the five (Table 3), and it labels almost the whole of the later sample as stress. Its predictive stress probability over the test window has a mean of 0.789 and a standard deviation of only 0.084, against 0.608 and 0.156 for the calibration-fitted model. The origin regime vector therefore barely moves from one day to the next, the similarity kernel is effectively flat at the selected bandwidth, and the weighted quantile selects the same order statistics as the unweighted one. Figure 4 shows what the coupled construction does instead. That is the transfer failure of Section 4 in its starkest form—an out-of-sample regime model that carries no usable information into the test period collapses the construction back onto the baseline it was meant to improve.

The base-model objection is therefore answered for the direction that matters. A base forecaster that does recover the conditional-mean dependence leaves the diagnostic conclusion intact: the apparent gain still belongs to information timing, and remains associated with calibration-dependent estimation, rather than to regime weighting as such. What a stronger base model changes is the magnitude of the artifact and the adequacy of the finite-sample correction against it, both in the unfavorable direction.

3.6. An Operational Fallback for Unbounded Intervals

The finite-sample correction is not optional, but it carries an operational cost that Sections 3.2 and 3.3 record without resolving: when the calibration weights concentrate, the $\pm\infty$ atoms bind and the interval becomes unbounded. Such an interval is unusable, and a forecasting system that emits one has no answer to give.

The trigger is known in advance and costs nothing to evaluate. An interval is unbounded exactly when the calibration weights concentrate enough that $\sum_i w_i < 2/\alpha - 1$, which is nineteen at $\alpha = 0.10$; the sum is available at the origin, before the interval is formed. The remedy that follows is to abandon regime weighting at those origins and

return the unweighted split-conformal interval. Supplementary Section S3 evaluates that rule against the alternative of weakening the kernel instead, on a rolling protocol swept to drive the weights into concentration. Falling back to split-conformal dominates at every bandwidth, eliminates unbounded intervals entirely, and—this is the finding that matters for the paper’s thesis—confers no detectable advantage in exchange: against rolling split-conformal the paired Winkler difference spans zero at every bandwidth examined. The rule is deterministic, costs one summation, and is auditable after the fact from the fired-origin rate.

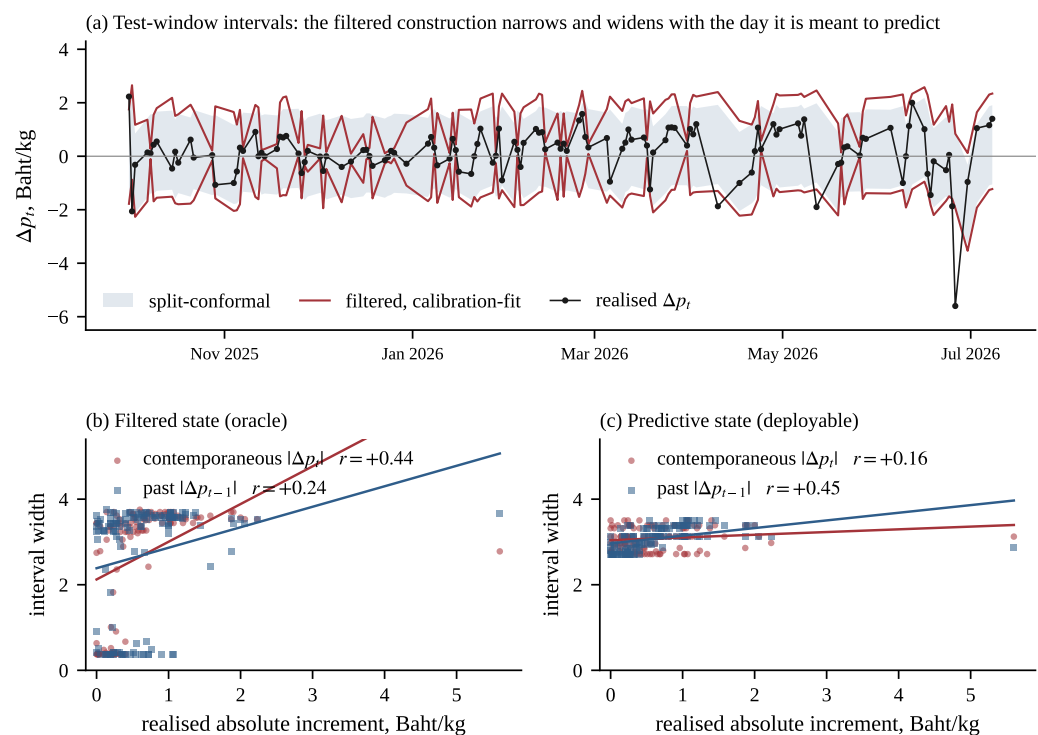


Figure 4. The information-timing artifact made visible, on the AR(1) base residuals. Panel (a) plots the test-window intervals of the coupled construction against the split-conformal band and the realized increments. Panels (b,c) regress interval width on the realized absolute increment of the same day and of the previous day, holding the estimation sample fixed at the calibration block and varying only the state. The filtered construction’s width tracks the contemporaneous increment more closely than the past one ($r = +0.44$ against $+0.24$); the predictive construction reverses the ordering ($r = +0.16$ against $+0.45$), as a one-step-ahead forecast must. The bimodal width in panel (b) also shows the instability of the coupled construction, whose interval ranges from 0.38 to 3.75 Baht/kg across origins.

One caution attaches. Applied to the filtered state rather than the predictive one, the same hybrid appears to beat rolling split-conformal. That apparent gain is the information-timing artifact of Section 3.1 operating inside the fallback rule: a fallback protocol inherits whatever couplings the construction it repairs already contains, and it repairs unboundedness only.

3.7. Robustness: Bandwidth, Signals, Origins, Levels and Regime Cardinality

Three questions remain open. How is the similarity bandwidth chosen, and how much do the results depend on it? How concentrated do the calibration weights become? And does the artifact belong to *latent* state weighting specifically, or to any localizing signal that sees the target period? Table S4 sweeps the full bandwidth grid to answer all three.

The selection rule governs which row of each block is reported elsewhere. On the bandwidth partition, the criterion minimizes

$$\mathcal{C}(\gamma) = \bar{w}(\gamma) + 100 \cdot \max\{0, (1 - \alpha) - \widehat{\text{cov}}(\gamma)\}, \quad (8)$$

where $\bar{w}(\gamma)$ is the mean interval width and $\widehat{\text{cov}}(\gamma)$ the empirical coverage on that block: mean width with a steep penalty on any shortfall below nominal coverage, subject to the constraint that no bandwidth-block interval is unbounded, with ties broken toward the smaller γ . Selection is made separately for each construction and never re-estimated within the test window. It returns $\gamma = 2$ for the filtered, calibration-fitted cell and $\gamma = 4$ for the forecast-feasible construction; the remaining cells are listed in Table S4.

First, the unbounded rate is the operative constraint on the bandwidth. For every signal, the interval score improves as γ rises and the weights concentrate—and the region where it improves most is the region where the construction stops returning finite intervals. The coupled arm reaches its largest paired difference at $\gamma = 8$, where 21.4% of origins are unbounded. Scoring only the finite origins would credit the method for exactly the configurations in which it fails, so no paired difference is quoted where any origin is infinite.

Second, the ranking is stable across the admissible range rather than an artifact of the selected bandwidth: wherever intervals remain finite, the forecast-feasible arm sits at or slightly above split-conformal at every γ , and the coupled arm below it, and no bootstrap interval excludes zero once the correction is in force.

Third, the artifact does not belong to latent-state weighting specifically. A parallel grid in Supplementary Section S3 replaces the latent state with a standardized trailing realized volatility of the residuals, an observed feature of the kind used by Schmitt [19], and varies only its timing. Lagged one day, and so measurable at the forecast origin, it produces no detectable gain at any bandwidth: paired differences run from -0.02 to -0.04 and every interval spans zero. Computed on the same day, it produces a detectable gain wherever intervals remain finite, reaching -0.38 [$-0.57, -0.07$] at $\gamma = 4$. The two blocks differ by one lag and nothing else.

The lag contrast settles a question the two-factor design of Section 3.1 could not. What manufactures the apparent sharpening is the information timing of the localizing signal, not the latent character of the state that carries it. A construction weighting on an observed, pre-specified feature is therefore exempt only if that feature is also past-measurable, a scope Section 4.4 states in full.

3.8. Origins, Nominal Levels, Benchmark Families and Regime Cardinality

The null of Section 3.1 rests on one test window, one nominal level, one benchmark family, and a two-state regime specification, any of which could be carrying the result. Table 11 varies all four.

Panel A extends the rolling evidence from six origins to 314 and adds two further nominal levels, and the pattern is unchanged. The coupled arm improves on split-conformal at every level, with its interval excluding zero wherever a paired difference can be computed; the forecast-feasible arm does not, being worse at $\alpha = 0.05$ and returning unbounded intervals at 10.2% and 1.6% of origins at the other two levels, so its unconditional interval score is undefined. Split-conformal itself under-covers at every level over this longer sequence, reaching 0.873 against a nominal 0.90 at $\alpha = 0.10$ —a property of the series rather than of any construction.

Panel B addresses whether recency alone accounts for the adaptivity that helps on this series. It does not. Rolling-window split-conformal, exponentially time-decayed conformal weighting at two decay rates, and an EnbPI-style bootstrap all attain a higher interval score than the fixed split-conformal baseline, with paired differences from $+0.03$ to $+0.23$ and

every interval spanning zero. Taken with Section 3.3, where methods driven by realized miscoverage do improve on the regime-weighted construction, this narrows the earlier reading: what helps here is error feedback, not recency as such.

Panel C answers whether a two-state specification is too coarse. Adding states reveals no structure a two-state model missed; it enlarges the artifact and leaves the deployable construction untouched. The filtered, calibration-fitted arm improves from -0.22 at $K = 2$ to -1.06 at $K = 3$ and -1.33 at $K = 4$, with intervals excluding zero at both higher cardinalities, while the forecast-feasible arm stays at $+0.01$, $+0.01$ and $+0.06$, with every interval spanning zero. A richer state space gives the filtered probability more room to encode the target day's own shock and the predictive probability nothing further to forecast with, so the audit protocol of Section 5.2 applies unchanged as K grows.

Table 11. Robustness of the null across forecast origins, nominal levels, benchmark families and regime cardinality.

Variation	Construction	Cov.	Width	Winkler	Unbd.	Δ Winkler [95% CI]
Panel A. Rolling origins over the whole out-of-sample sequence (314 origins, window 220)						
$\alpha = 0.05$	Split-conformal	0.917	4.029	5.437	0.0%	—
	Filtered, calibration-fit	0.949	4.282	5.246	9.6%	—
	Forecast-feasible	0.946	4.019	5.381	11.8%	—
$\alpha = 0.10$	Split-conformal	0.873	3.047	4.619	0.0%	—
	Filtered, calibration-fit	0.904	3.311	4.247	0.0%	$-0.37 [-0.50, -0.05]$
	Forecast-feasible	0.898	2.889	4.316	10.2%	—
$\alpha = 0.20$	Split-conformal	0.780	2.180	3.555	0.0%	—
	Filtered, calibration-fit	0.850	2.206	3.200	0.0%	$-0.36 [-0.43, -0.19]$
	Forecast-feasible	0.796	2.237	3.452	1.6%	—
Panel B. Further baselines, test window, $\alpha = 0.10$						
	Rolling-window split-conformal	0.901	3.009	4.604	0.0%	$+0.23 [-0.05, +0.48]$
	Time-decayed conformal, $\lambda = 0.995$	0.901	2.930	4.586	0.0%	$+0.21 [-0.11, +0.49]$
	Time-decayed conformal, $\lambda = 0.98$	0.893	2.911	4.403	0.0%	$+0.03 [-0.36, +0.39]$
	EnbPI-style bootstrap	0.908	2.919	4.613	0.0%	$+0.23 [-0.09, +0.53]$
Panel C. Number of regime states, test window, $\alpha = 0.10$						
$K = 2$	Filtered, calibration-fit	0.939	3.189	—	0.0%	$-0.22 [-0.36, +0.09]$
	Forecast-feasible	0.924	3.050	—	0.0%	$+0.01 [-0.08, +0.05]$
$K = 3$	Filtered, calibration-fit	0.954	2.561	—	0.0%	$-1.06 [-1.16, -0.75]$
	Forecast-feasible	0.924	3.077	—	0.0%	$+0.01 [-0.08, +0.06]$
$K = 4$	Filtered, calibration-fit	0.969	2.411	—	0.0%	$-1.33 [-1.39, -0.94]$
	Forecast-feasible	0.931	3.132	—	0.0%	$+0.06 [-0.05, +0.14]$

Panel A extends the six-origin exercise of Section 3.2 to every out-of-sample origin with a full trailing calibration window, re-estimating the conformal layer at each; no paired difference is reported where any origin is unbounded. Panel B holds the base forecaster fixed and replaces the conformal layer: the time-decay weight is λ^{age} over a 220-observation window, and the EnbPI-style baseline pools 30 bootstrap resamples of it. Panel C re-estimates the Markov-switching model with K states. Cov. is empirical coverage and Unbd. the fraction of origins at which the interval is infinite. All differences are against split-conformal on the same origins, with 95% moving-block-bootstrap intervals.

Bootstrap inference is also insensitive to the block length. For the filtered, calibration-fitted cell on the test window, the paired difference is -0.220 at every block length, and the 95% interval runs from $[-0.49, +0.01]$ at length one to $[-0.34, +0.16]$ at length twenty-two. The conclusion drawn from it does not turn on the square-root rule.

3.9. Simulation: When Regime Weighting Can and Cannot Help

The controlled simulation of Section 2.6 supplies the counterfactual the application cannot, since the design fixes how predictable the regime is. A falsification test comes first.

Under its exact no-regime null, the regime label is pure noise, so no construction respecting the predictable-weight condition should show a systematic advantage over split-conformal.

Table 12 reports the exact no-regime null results. Neither predictive construction beats split-conformal; both are slightly worse ($\Delta\text{Winkler } +0.04 \pm 0.01$), because weighting on a noise signal only costs efficiency. Both filtered constructions record a lower finite-origin score (-0.07 ± 0.02) while leaving 11–12% of origins unbounded. The direct paired contrasts isolate which coupling is responsible: switching predictive to filtered timing at a fixed estimation sample produces a spurious -0.12 with a Monte-Carlo interval excluding zero, and of near-identical magnitude whether the regime is fit on validation or on calibration, while switching disjoint to calibration estimation at fixed timing produces no detectable change ($-0.004 [-0.016, +0.009]$ predictive; $-0.006 [-0.024, +0.011]$ filtered). The manufactured advantage therefore belongs to filtered state timing, while calibration-score reuse produces none under this null.

Table 12. Exact no-regime null ($F_1 = F_2 = N(0, 1)$; $p = 0.95$; 200 Monte-Carlo replications; symmetric tails $+\pm\infty$ atom, $\alpha = 0.10$), a 2×2 factorial of state timing $\times$ regime-estimation sample.

Arm (State Timing $\times$ Estimation)	$\Delta\text{Winkler}$ [95% MC]	Unbd. Rate
Predictive, validation-fit	$+0.042 [+0.030, +0.053]$	0.052
Predictive, calibration-fit	$+0.039 [+0.028, +0.050]$	0.053
Filtered, validation-fit	$-0.072 [-0.087, -0.057]$	0.106
Filtered, calibration-fit	$-0.077 [-0.093, -0.061]$	0.122

$\Delta\text{Winkler}$ is the paired difference against split-conformal over the common finite-origin set (negative = lower score); the unbounded rate is the mean fraction of test origins returning an infinite interval. The direct paired contrasts are: state timing (filtered – predictive) = $-0.115 [-0.132, -0.099]$ at validation-fit and $-0.117 [-0.134, -0.100]$ at calibration-fit, both excluding zero; regime-estimation (calibration-fit – validation-fit) = $-0.004 [-0.016, +0.009]$ at predictive timing and $-0.006 [-0.024, +0.011]$ at filtered timing, both spanning zero.

Figure 5 traces the paired Winkler difference as persistence varies. The forecast-feasible construction earns no detectable advantage where the regime is weakly predictable: it is slightly worse than split at low persistence ($+0.09 \pm 0.05$ at $p = 0.50$) and shows a lower mean Winkler only once the regime is strongly persistent (-0.28 ± 0.04 at $p = 0.95$, over the common finite-origin set). Because it still returns unbounded intervals at about 5% of origins, that difference is conditional on the shared finite-origin set, not a claim of dominance. The filtered, in-sample artifact construction, by contrast, shows an apparent improvement at every persistence level, from -0.49 at $p = 0.50$ to -0.82 at $p = 0.95$.

Table 13 fixes a genuinely predictable regime ($p = 0.95$, separability 3.5, 300 replications) under the finite-sample-corrected construction, with every difference taken over the common finite-origin set. The oracle filtered state retains the largest apparent gain (-0.80 ± 0.05). Because the regime is genuinely predictable, the forecast-feasible construction fitted out of sample also gains (-0.28 ± 0.04). The predictive calibration-fitted and validation-fitted constructions attain similar conditional finite-origin scores under this stationary process (-0.27 ± 0.04 versus -0.28 ± 0.04). That comparison provides no evidence that calibration-score reuse improves performance here, but it does not establish equivalence either, and without a direct paired factor contrast it cannot separate reuse from estimation-sample size.

Separating the regime-fit data from the scored data leaves no gain (-0.02 ± 0.06) in the calibration-split arm, whose effective calibration size is halved, so that change confounds score separation with the smaller calibration set. Taken together, these results make the apparent calibration-fit advantage in the non-stationary application (Section 3.1) consistent with regime-model recency and transfer under distribution shift—which this stationary design deliberately does not reproduce—rather than with score reuse itself. The simulation

establishes only that score reuse per se produces no detectable advantage here; identifying recency or transfer as the operative cause would require the non-stationary experiments left to future work.

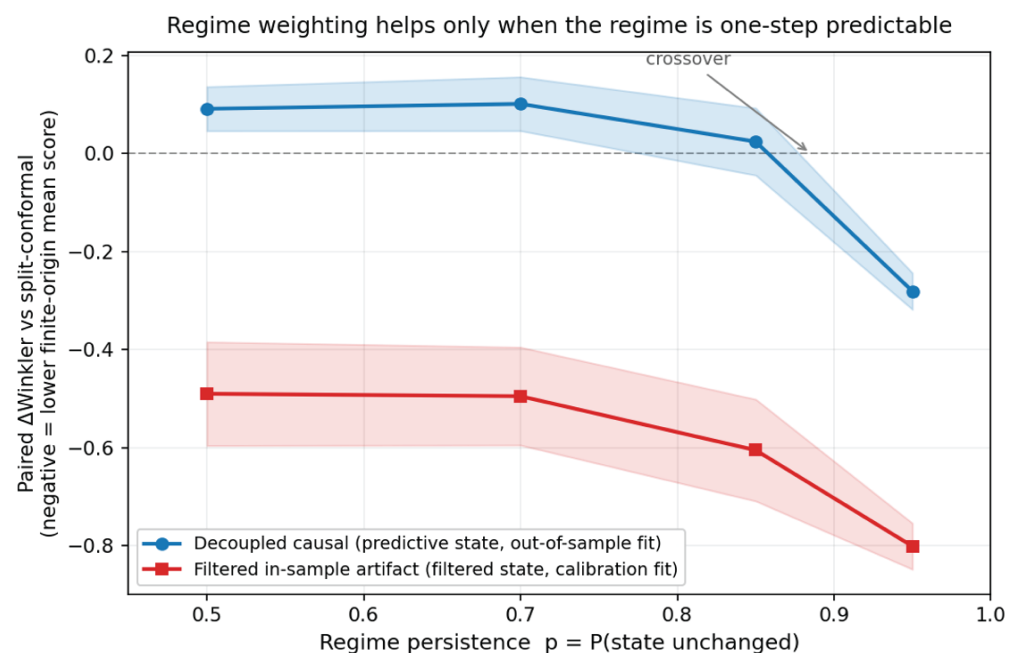


Figure 5. Paired Winkler difference versus split-conformal in the controlled simulation. Regime persistence varies at fixed separability 3.5. The forecast-feasible construction (predictive state, out-of-sample fit) has a lower conditional finite-origin mean Winkler score only as the regime becomes strongly persistent; the filtered, in-sample artifact construction does so at every persistence level. Negative values denote a lower mean Winkler conditional on the finite-origin evaluation set, not a claim of dominance over split-conformal, which is always finite. Replications: 250 at each swept persistence (0.5, 0.7, 0.85) and 300 at the $p = 0.95$ factorial point of Table 13; bands are 95% Monte-Carlo intervals.

Table 13. Finite-sample-corrected simulation at a predictable regime ($p = 0.95$, separability 3.5; 300 Monte-Carlo replications; symmetric tails $+ \pm \infty$ atom, $\alpha = 0.10$).

Construction (State, Estimation)	Δ Winkler [$\pm 95\%$ MC]	Unbd. Rate
Filtered, calibration-fit (oracle)	-0.80 ± 0.05	0.062
Decoupled (predictive, validation-fit)	-0.28 ± 0.04	0.055
Predictive, calibration-fit (score reuse)	-0.27 ± 0.04	0.053
Predictive, calibration-split (score-separated)	-0.02 ± 0.06	0.072

All four constructions attain finite-origin coverage near the nominal 0.90 (0.91) with finite mean width 5.8–6.3. Because every construction leaves a positive fraction of origins unbounded, its unconditional mean Winkler is infinite, so the Δ Winkler column is the paired difference against split-conformal over the common finite-origin set only (negative = lower score), with 95% Monte-Carlo half-widths. The unbounded rate is the mean fraction of test origins returning an infinite interval.

One alignment check remains. The simulation fits its regime model to a trailing local log realized variance, whereas the empirical study fits the same model to the log squared residual (Section 2.6); the smoothed feature is easier to recover, so the design is a favorable-case existence check rather than a like-for-like replica. Repeating the persistent-regime experiment with the empirical feature over 120 replications at $p = 0.95$ leaves the coupled construction unaffected, at -0.64 ± 0.12 against -0.62 ± 0.08 , while the genuine gain available to the forecast-feasible construction falls from -0.19 ± 0.07 to -0.05 ± 0.03 .

The contrast cuts both ways. The artifact survives the change of feature untouched, whereas the recoverable advantage depends on how legible the regime is in the signal actually used: against the one-point log-chi-square noise of a single squared residual, it is an order of magnitude smaller, and at -0.05 it falls below what a 131-observation test window can resolve (Section 4.4). The empirical null is therefore what the aligned simulation predicts, not evidence against it.

The simulation thus supplies the counterfactual the application cannot. Under the finite-sample-corrected construction, regime weighting can lower the mean Winkler on finite origins—but only under the persistent, one-step-predictable regime examined here, with separability held fixed rather than swept. Even then, the advantage inflates when the state is a filtered oracle, and 5–7% of origins are unbounded, so it does not dominate split-conformal unconditionally. The RSS3 residuals carry only a weak and largely contemporaneous regime signal (Section 3.4), placing them at the low-predictability end of this spectrum—which is what the absence of a detectable gain for the forecast-feasible construction reflects.

4. Discussion

4.1. Two Forms of Information Coupling

The decomposition of Section 3.1 traces the apparent advantage of regime weighting to two couplings. They are not of the same kind, and the distinction matters for what a practitioner should do about each. The first is genuine forecast-origin leakage. The second is score reuse that is operationally feasible but analytically compromising.

The first coupling operates through the state. Weighting calibration residuals by their similarity to a filtered origin conditions the interval on the realized volatility of the very day it is meant to predict (Section 2.3), a quantity unavailable one step ahead. The predictive probability withholds it, which is why substituting it removes the gain. Figure 4 shows the substitution at work: with the estimation sample held fixed, the filtered construction's width tracks the contemporaneous increment more closely than the previous day's, and the predictive construction reverses that ordering.

The second coupling operates through the estimation sample. When the Markov-switching parameters are fitted on the calibration window, the auxiliary model is estimated on the same residuals that supply the conformal scores. The definition of the high-volatility state is then tuned to the calibration-period distribution, which immediately adjoins the test window. Estimating the regime on an earlier validation block breaks that dependence. It also breaks something else: the resulting model does not transfer to the later, more volatile test period, and the gain disappears with it. These two effects cannot be separated by the design used here, which is why calibration fitting is reported throughout as an association rather than a demonstrated causal source. The factorial comparisons identify an implementation-sensitive association; they do not by themselves separate calibration-score reuse from the benefit of fitting a more recent regime model under distribution shift.

Neither coupling is visible in the headline metrics. Both are exposed only by the factorial design, and this is the reason the audit protocol of Section 5.2 is framed as a set of crossed comparisons rather than a single check.

4.2. Why No Detectable Gain Remains

Once both couplings are removed, regime weighting becomes indistinguishable from split-conformal. The reason is that the residuals reaching the conformal layer carry little one-step-predictable regime structure. Section 3.4 shows the linear base forecast sitting at the accuracy of a no-change benchmark, so its residuals stay close to the raw price increments. That is a limitation of the base model rather than a property of the series: a

single-parameter AR(1) attains a mean absolute scaled error of 0.921 on the same window. That limitation nonetheless fixes what the conformal layer receives.

The fitted regime signal is correspondingly weak. Its agreement with a volatility class it cannot influence is barely above chance: the validation-fitted predictive signal, which is what this construction actually uses, attains an area under the curve of 0.542 against next-day volatility, with a bootstrap interval of [0.442, 0.636] that includes 0.50. The filtered signal does no better (0.567), so the timing of the state buys nothing for the quantity a one-step interval must anticipate (Table 6, Panel B). The regime structure that regime weighting is designed to exploit is therefore real enough to be estimated in sample. It is also real enough to sharpen intervals when contemporaneous information is available. What it is not is recoverable one step ahead from past information alone.

The controlled simulation makes the point quantitative. As regime persistence rises, the forecast-feasible construction crosses from no gain to a genuine one; on weakly predictable scores it stays at parity. The RSS3 residuals occupy the second position, and Section 3.5 shows that a base model with better point accuracy does not move them out of it.

4.3. Relation to the Coverage Theory and to Adaptive Baselines

The empirical results are directionally consistent with the information-timing distinction that beyond-exchangeability theory emphasizes. They do not, however, constitute a test of the Barber et al. [11] coverage bound. Both predictive and filtered latent-state weights are data-dependent, and what separates them here is operational causality rather than eligibility for the fixed-weight theorem (Supplementary Section S2). Under that bound, weights fixed in advance enter the guarantee without additional cost, whereas weights depending on the target observation inflate the total-variation coverage gap. The filtered state is precisely such a weight and the predictive state its causal counterpart.

The empirical artifact and the fixed-weight benchmark thus point in the same direction, though no formal coverage penalty for data-dependent weights is established here. Supplementary Section S2.5 sets out the route by which such a penalty might be bounded, through the geometric contraction of the switching filter, and shows why it does not close. The obstacle is instructive rather than technical: the contraction constant that would make the bound tight is small precisely when the regime signal is too fast-mixing to localize anything, and large precisely when the signal carries the predictive content that would justify the method.

The comparison with adaptive baselines sharpens the practical lesson. Adaptive Conformal Inference and its dynamically-tuned variant adjust the interval online from past realized miscoverage, an entirely predictable signal, and both outperform the forecast-feasible construction (Section 3.3). The adaptivity that genuinely helps this non-stationary series is therefore error feedback, not the recovery of a latent volatility regime—and not recency either, a distinction Section 3.7 draws directly. A method that reaches for recency through the current day's state buys only an illusion of error feedback.

4.4. Where the Diagnostic Applies and Where It Does Not

Because the study reports a null, its scope requires care. One point of language governs everything that follows. A bootstrap interval spanning zero records the absence of a detectable difference at the sample sizes examined; it is not a demonstration of equivalence. The test window carries 131 observations and the high-volatility subset 44. For the paired interval-score difference, the standard deviation is 1.42, so at a 5% level and 80% power the minimum detectable effect is 0.35 Baht/kg if origins are treated as independent and

1.15 Baht/kg once the block dependence used for inference is accounted for. Differences smaller than that would not be resolved by this design.

Claims below are therefore phrased as absence of detectable advantage under the examined design, and an equivalence or non-inferiority framework would be required to say more. A null result invites two opposite misreadings: that regime-aware conformal inference has been refuted in general, and that the finding is a curiosity of one commodity. Neither is right, and the two halves of the contribution have different reach. The mechanism—that a filtered state manufactures apparent sharpness—generalizes to any dependent series, and the controlled null of Section 3.9 demonstrates it with no commodity data at all. Nothing in that demonstration is specific to rubber. The null—that nothing is left once they are removed—is specific to what was studied here. Table 14 separates the two.

Table 14. Scope of the diagnostic: what the evidence settles and what it leaves open.

	Where the Diagnostic Is Robust	Where It Does Not Apply, or Remains Open
Weighting signal	Any weight formed from a signal that sees the target period, latent or observed alike (Section 3.7)	Weights fixed in advance, or formed from a signal that is both pre-specified and past-measurable
Information timing	Universal: a filtered state conditions on the target period on any dependent series	—
Estimation sample	Any auxiliary model fitted on the same residuals that supply the conformal scores, whose contribution is associated with the gain but not separated from regime-model recency	An auxiliary model fitted on a disjoint block that genuinely transfers to the evaluation period
Finite-sample correction	Necessary in every case; its omission inflates sharpness wherever weights concentrate	Not sufficient on its own—Section 3.5 shows an artifact surviving the correction intact, and the unbounded intervals it produces require the fallback of Section 3.6
Target series	The parity result holds for residuals with weak one-step-predictable volatility structure	Series with persistent, well-separated, one-step-predictable regimes, where the simulation shows genuine gains
Base forecaster	Two base models of different point accuracy, giving the same decomposition	Substantially more expressive nonlinear forecasters, whose residuals are not examined here
Regime cardinality	Two, three and four states, over which the artifact grows and the forecast-feasible null is unchanged (Section 3.8)	Continuous or non-Markov volatility states, which are not examined
Coverage guarantee	No finite-sample guarantee is claimed for data-dependent latent-state weights	Exact-validity routes through block exchangeability [12], which this study does not evaluate

The left column states what the evidence supports beyond the present application; the right column states the conditions under which the diagnostic does not bind, or under which the question is left open.

One caution belongs to the corrected construction rather than to the naive one. As Sections 3.2 and 3.3 report, the $\pm\infty$ correction can return unbounded intervals when the calibration weights concentrate on short windows, so the corrected regime-weighted interval is not even reliably finite on this series. That is a further reason to regard it as conferring no usable advantage here, and it is a property of the construction rather than of the data. The failure is repairable: Section 3.6 shows that returning the unweighted split-conformal interval whenever $\sum_i w_i$ falls below $2/\alpha - 1$ removes every unbounded interval, without conferring any advantage in exchange.

4.5. Reconciliation with Prior Point-Forecast Results, and Limitations

The base forecast deserves a word on its relation to earlier work. The present study and the VMD-augmented bidirectional LSTM of Pinitjitsamut [31] target the same quantity, the one-step first difference of the RSS3 price. The deep model attains high point accuracy, with a reported Pearson correlation near 0.82 and preserved dispersion. The two results are

consistent once the roles are separated: the earlier model is an accurate conditional-mean estimator, whereas the present base model is deliberately simple and sits, as Section 3.4 reports, at the accuracy of a no-change benchmark.

The information-coupling result does not depend on that choice. The decomposition concerns how residuals are weighted by regime, not how large they are. A filtered state leaks the contemporaneous shock and an in-sample regime model reuses the calibration residuals it will weight, whatever produced those residuals. What is base-model-dependent is the residual structure a forecast-feasible construction could exploit, and here the evidence goes only so far.

The ridge regression is demonstrably not the best available, since an AR(1) beats it on the test window. Section 3.5 therefore repeats the decomposition on those AR(1) residuals. The couplings survive the change unaltered; indeed, the artifact grows and the finite-sample correction ceases to neutralize it. What remains untested is the far end of the range, namely whether a substantially more expressive forecaster would leave exploitable one-step regime structure in its residuals. The guidance of Section 5.2 is framed accordingly: the couplings must be avoided regardless, and any genuine gain must be demonstrated under the forecast-feasible construction rather than assumed.

Several extensions are left to future work, and none bears on the diagnostic conclusion, which the controlled null and the two-base-model comparison establish as a property of the conformal construction. On the empirical side, the main gaps are a nonlinear base forecaster, extending the comparison of Section 3.5, and formal sensitivity to the high-volatility threshold. On the simulation side, they are a separability sweep and a non-stationary design with parameter drift, which would distinguish calibration-score reuse from regime-model recency and transfer. Finally, the monthly Chinese indicators are aligned by a uniform one-month rule rather than by series-specific release dates, and revised rather than first-print vintages are used (Table 4); a series-level release-date and vintage audit would close both. Section 3.5 reproduces the decomposition on a base model that uses none of these covariates, so neither gap touches the diagnostic.

5. Conclusions and Guidance for Practitioners

5.1. Conclusions

On daily natural-rubber price changes, regime-aware conformal prediction appeared to sharpen intervals by a fifth relative to split-conformal, with a decisively better interval score. That improvement does not survive scrutiny, and this study identifies what produces it.

Three implementation features accompany the gain, and they do not have the same evidential status. A filtered origin state is genuine forecast-origin leakage: it is an oracle quantity, unavailable when the forecast is issued, and a controlled no-regime null reproduces its effect where no regime information exists at all. Omission of the finite-sample correction inflates sharpness wherever the calibration weights concentrate. Calibration-dependent regime estimation is operationally feasible score reuse, and its empirical contribution cannot be separated from regime-model recency and transfer; it is therefore reported as an association rather than a demonstrated cause.

The forecast-feasible construction—which removes the leakage and applies the correction—shows no statistically detectable advantage over split-conformal in the single-window analysis, under full rolling-origin re-estimation, across two to four regime states, or on either the VMD-augmented ridge or the AR(1) residuals. Across 314 rolling origins and three nominal levels, it provides no usable unconditional advantage and returns unbounded intervals at some origins, while adaptive baselines remain finite and attain lower interval scores.

The contribution is diagnostic rather than constructive. Past-measurability is identified as necessary for operational forecasting but not sufficient for a finite-sample guarantee, and the magnitude of the gain that a naive implementation manufactures is quantified. What generalizes is the mechanism, not the null.

5.2. Guidance for Practitioners

The mechanism reaches beyond rubber to any regime-aware conformal construction on dependent data. The protocol below separates genuine gains from artifacts.

1. Weight by predictive, not filtered, regime probabilities. The regime signal must condition only on information available at the forecast origin. A filtered state leaks the target-period shock and is an oracle quantity.
2. Separate regime-model estimation from conformal calibration. Fit the regime parameters on a block disjoint from the calibration and test sets; where that is impossible, treat the resulting weights as data-adaptive and validate them through a nested out-of-sample protocol.
3. Apply the finite-sample correction, and do not stop there. The $\pm\infty$ point masses are required for the weighted quantile, and omitting them inflates sharpness—but they are not a substitute for correct information timing, and Section 3.5 exhibits an artifact that survives the correction unchanged.
4. Run the three-factor check before claiming an improvement. Cross-filtered against predictive state, calibration-informed against disjoint regime estimation, and omission against inclusion of the correction. A gain that survives only under oracle state information, calibration reuse, or an uncorrected quantile is not a deployable improvement.
5. Have a fallback ready for unbounded intervals. Compute $\sum_i w_i$ at each origin, and where it falls below $2/\alpha - 1$ return the unweighted split-conformal interval; Section 3.6 shows this to dominate the alternative of weakening the kernel. Report the fired-origin rate: a system that falls back most of the time is not a regime-weighted system.
6. Do not expect a richer state space to rescue the construction. Section 3.7 shows the apparent gain growing monotonically from two to four states while the forecast-feasible construction stays at parity.
7. Benchmark against adaptive baselines. Methods driven by past miscoverage, such as ACI and DtACI, supply predictable and model-free adaptivity. A regime-weighted construction should beat them to justify its additional structure.
8. Assess the residuals' forecastability, and diagnose why it is what it is. Where the base forecaster performs no better than a naive benchmark, little exploitable regime structure reaches the conformal layer once contemporaneous information is withheld. But a scaled error near one has two causes with different remedies—an unforecastable target, or an under-specified base model leaving recoverable structure in its residuals—and a simple benchmark such as an AR(1) separates them.

For the applied problem that motivated the study, the implication is direct. Once the couplings are removed, regime weighting confers no statistically detectable advantage over split-conformal on natural-rubber price changes, under the designs examined here. A plain split-conformal interval, or a lightweight adaptive one, is the appropriate default. The value of the regime-aware apparatus, where it exists, must be sought in markets whose volatility regimes are genuinely predictable one step ahead, and demonstrated under the protocol above.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/forecast8050082/s1>: Supplementary File S1: Expanding-window variational mode decomposition, relation to beyond-exchangeability coverage bounds, and supplementary results, including finite-sample corrections, fallback rules for unbounded intervals, bandwidth sensitivity, and an observed-feature benchmark (Figure S1 and Tables S1–S5).

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Code, a machine-readable data dictionary, and a mapping from every table to the script that generates it are openly available at <https://github.com/talkmcp/Too-Sharp-to-Be-True> (accessed on 17 August 2026), archived at Zenodo (DOI: [10.5281/zenodo.21436513](https://doi.org/10.5281/zenodo.21436513)). A single entry point, `run_all.sh`, regenerates every table and figure in order, and `requirements.txt` pins the package versions used. The repository also distributes the out-of-sample residual series of the base forecaster, from which every conformal result here is reproducible without the underlying prices, since coverage, width and interval score depend on the residual alone. The daily price series were obtained from commercial and institutional providers (CEIC Data; the Rubber Authority of Thailand; and the SGX, SHFE and TOCOM exchanges), are subject to those providers' licensing terms, and are therefore available from the corresponding author on reasonable request rather than redistributed.

Acknowledgments: During the preparation of this manuscript, the author used Claude Anthropic to help with language editing, with the analysis scripts, and with revising the text. All outputs were reviewed and verified by the author, who takes full responsibility for the content of this publication.

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

Abbreviations

The following abbreviations are used in this manuscript:

ACI	Adaptive Conformal Inference
ADF	Augmented Dickey–Fuller (test)
BiLSTM	Bidirectional Long Short-Term Memory
CEIC	CEIC Data (https://www.ceicdata.com)
CQR	Conformalized Quantile Regression
DGP	Data-Generating Process
DtACI	Dynamically-tuned Adaptive Conformal Inference
FOB	Free on Board
IMF	Intrinsic Mode Function
JPX	Japan Exchange Group
MAE	Mean Absolute Error
MASE	Mean Absolute Scaled Error
PMI	Purchasing Managers' Index
RAOT	Rubber Authority of Thailand
RSS3	Ribbed Smoked Sheet No. 3
RWCP	Regime-Weighted Conformal Prediction
SGX	Singapore Exchange
SHFE	Shanghai Futures Exchange
SICOM	Singapore Commodity Exchange
STR20	Standard Thai Rubber 20
TOCOM	Tokyo Commodity Exchange
TSR20	Technically Specified Rubber 20
USS	Unsmoked Sheet (rubber)
VMD	Variational Mode Decomposition

Appendix A. Implementation Details

Table A1. Reproducibility parameters.

Component	Setting
Data and base model	
Study window (daily)	7 May 2018–13 July 2026
Usable Δp observations	1430
Partition (train/val/calib/bw/test)	841/129/287/42/131
Features (economic + VMD modes)	64 + 6
VMD modes M	6, fixed on the training partition
Base forecaster	ridge regression on lagged features
Ridge penalty λ	selected from $\{1, 3, 10, 30, 100, 300, 1000\}$ by minimum validation mean-squared error
Conformal layer	
Regime model	two-state Gaussian Markov-switching model fitted to $\log(\hat{e}_t^2 + 10^{-6})$ by maximum likelihood (Hamilton filter)
Regime signal: state	filtered $\hat{\pi}_t$ versus predictive $\rho_t = \hat{\pi}_{t-1}P$
Regime signal: estimation sample	validation block, calibration block, or validation + calibration
Similarity kernel	$\exp(-\gamma \ \cdot\ _1)$, with $\gamma \in \{0.5, 1, 2, 4, 8, 16, 32\}$ selected per construction on the bandwidth partition
Nominal level	$\alpha = 0.10$ (90% interval)
Adaptive baselines (Section 3.3)	
Trailing calibration window	220 observations, common to all baselines
ACI learning rate η	0.05, fixed
ACI initialization	$\alpha_1 = \alpha = 0.10$
DtACI candidate rates	$\{0.01, 0.02, 0.05, 0.10, 0.20\}$
DtACI expert-weight learning rate	0.02, with equal initial expert weights
Update rule	$\alpha_{t+1} = \alpha_t + \eta(\alpha - \mathbf{1}\{\text{miss at } t\})$, clipped to $[0.001, 0.5]$
Inference	
Moving-block bootstrap	block length $\lfloor \sqrt{n} \rfloor$, 2000 resamples, percentile intervals, seed 0
Coverage intervals	Wilson score intervals

Note. The VMD settings are $M = 6$ modes, bandwidth penalty $\alpha_{\text{VMD}} = 2000$, noise tolerance $\tau = 0$, no DC component, and unit stride; the ridge penalty is selected from $\{1, 3, 10, 30, 100, 300, 1000\}$ by minimum validation mean-squared error. The selected bandwidths for every construction are reported in Table S4. All reported empirical intervals (Sections 3.1–3.3) use the finite-sample $\pm\infty$ correction of Barber et al. [11] and fixed symmetric tails ($\alpha_L = \alpha_U = \alpha/2$); the asymmetric per-regime allocation is a sensitivity variant. The simulation of Section 3.9 uses the same finite-sample-corrected, fixed-symmetric-tail construction as the empirical analysis, with 200–300 independent Monte-Carlo replications per experiment and an unbounded-interval audit.

Appendix B. Coverage Theory and Additional Results

The construction studied here is related to, but not covered by, the non-exchangeable conformal framework of Barber et al. [11]. Their bound holds for weights fixed in advance; the weights used here are generated from a recursively estimated latent-state path and are therefore data-dependent, so the result is used throughout as a benchmark rather than as a guarantee for the implemented procedure. Supplementary Section S2 sets out the construction in the notation of that literature, states the assumptions a conditional guarantee would require, and examines whether the geometric contraction of the switching filter could bound the total-variation penalty. It does not close, and the reason is informative: the geometric filter-stability results it relies on [35,36] give a contraction constant that is small exactly when the regime signal is too fast-mixing to localize anything, and large exactly when the signal carries the predictive content that would justify the method. A finite-sample coverage guarantee for the implemented data-dependent weights is left as an open problem.

Supplementary Section S3 reports three further tables: the six state $\times$ estimation cells with and without the finite-sample correction, the comparison of the two fallback rules for

unbounded intervals, and the full bandwidth grid with weight-concentration diagnostics, including the observed-feature benchmark.

References

1. Vovk, V.; Gammerman, A.; Shafer, G. *Algorithmic Learning in a Random World*; Springer: New York, NY, USA, 2005. [\[CrossRef\]](#)
2. Shafer, G.; Vovk, V. A tutorial on conformal prediction. *J. Mach. Learn. Res.* **2008**, *9*, 371–421.
3. Angelopoulos, A.N.; Bates, S. Conformal prediction: A gentle introduction. *Found. Trends Mach. Learn.* **2023**, *16*, 494–591. [\[CrossRef\]](#)
4. Stocker, M.; Małgorzewicz, W.; Fontana, M.; Ben Taieb, S. A Gentle Introduction to Conformal Time Series Forecasting. *arXiv* **2025**, arXiv:2511.13608.
5. Papadopoulos, H.; Proedrou, K.; Vovk, V.; Gammerman, A. Inductive confidence machines for regression. In *Proceedings of the Machine Learning: ECML 2002*; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2002; Volume 2430, pp. 345–356.
6. Lei, J.; G'Sell, M.; Rinaldo, A.; Tibshirani, R.J.; Wasserman, L. Distribution-free predictive inference for regression. *J. Am. Stat. Assoc.* **2018**, *113*, 1094–1111. [\[CrossRef\]](#)
7. Romano, Y.; Patterson, E.; Candès, E.J. Conformalized quantile regression. In *Proceedings of the Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*; Curran Associates, Inc.: Red Hook, NY, USA, 2019.
8. Lima, J.F.; Pereira, F.C.; Gonçalves, A.M.; Costa, M. Bootstrapping State-Space Models: Distribution-Free Estimation in View of Prediction and Forecasting. *Forecasting* **2024**, *6*, 36–54. [\[CrossRef\]](#)
9. Winkler, R.L. A Decision-Theoretic Approach to Interval Estimation. *J. Am. Stat. Assoc.* **1972**, *67*, 187–191. [\[CrossRef\]](#)
10. Tibshirani, R.J.; Barber, R.F.; Candès, E.J.; Ramdas, A. Conformal prediction under covariate shift. In *Proceedings of the Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*; Curran Associates, Inc.: Red Hook, NY, USA, 2019.
11. Barber, R.F.; Candès, E.J.; Ramdas, A.; Tibshirani, R.J. Conformal prediction beyond exchangeability. *Ann. Stat.* **2023**, *51*, 816–845. [\[CrossRef\]](#)
12. Nettasinghe, B.; Chatterjee, S.; Tipireddy, R.; Halappanavar, M. Extending conformal prediction to hidden Markov models with exact validity via de Finetti's theorem for Markov chains. In *Proceedings of the 40th International Conference on Machine Learning (ICML 2023)*; PMLR: New York, NY, USA, 2023; Volume 202, pp. 25890–25903.
13. Diaconis, P.; Freedman, D. de Finetti's theorem for Markov chains. *Ann. Probab.* **1980**, *8*, 115–130. [\[CrossRef\]](#)
14. Gibbs, I.; Candès, E.J. Adaptive conformal inference under distribution shift. In *Proceedings of the Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*; Curran Associates, Inc.: Red Hook, NY, USA, 2021.
15. Gibbs, I.; Candès, E.J. Conformal inference for online prediction with arbitrary distribution shifts. *J. Mach. Learn. Res.* **2024**, *25*, 1–36.
16. Xu, C.; Xie, Y. Conformal prediction interval for dynamic time-series. In *Proceedings of the 38th International Conference on Machine Learning (ICML 2021)*, Virtual, 18–24 July 2021; Volume 139, pp. 11559–11569.
17. Lu, E.D.; Findling, C.; Clausel, M.; Leite, A.; Gong, W.; Kersaudy, P. Adaptive Regime-Switching Forecasts with Distribution-Free Uncertainty: Deep Switching State-Space Models Meet Conformal Prediction. In *Proceedings of the NeurIPS 2025 Workshop on Recent Advances in Time Series Foundation Models*, San Diego, CA, USA, 7 December 2025. Available online: <http://arxiv.org/abs/2512.03298> (accessed on 16 August 2026).
18. Zarnani, A.; Karimi, S.; Musilek, P. Quantile Regression and Clustering Models of Prediction Intervals for Weather Forecasts: A Comparative Study. *Forecasting* **2019**, *1*, 169–188. [\[CrossRef\]](#)
19. Schmitt, M. Taming Tail Risk in Financial Markets: Conformal Risk Control for Nonstationary Portfolio VaR. *arXiv* **2026**, arXiv:2602.03903.
20. Heurich, M.; Granz, M.; Landgraf, T. RareCP: Regime-Aware Retrieval for Efficient Conformal Prediction. *arXiv* **2026**, arXiv:2605.08857.
21. Cuonzo, S.; Deliu, N. Conformal Prediction Intervals with Tail-Specific Guarantees. *arXiv* **2026**, arXiv:2606.18199.
22. El Bakkali, Y.; Krami, N.; Rochdi, Y.; Boukaibat, A. DG-TFT-CQR: A Dynamic Graph–Temporal Fusion Transformer with Conformalized Quantile Regression for Wind Power Forecasting. *Forecasting* **2026**, *8*, 55. [\[CrossRef\]](#)
23. Hamilton, J.D. A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica* **1989**, *57*, 357–384. [\[CrossRef\]](#)
24. Cifarelli, G.; Paladino, G. Oil price dynamics and speculation: A multivariate financial approach. *Energy Econ.* **2010**, *32*, 363–372. [\[CrossRef\]](#)
25. Burger, K.; Smit, H.P. Long-Term and Short-Term Analysis of the Natural Rubber Market. *Rev. World Econ. (Weltwirtsch. Arch.)* **1989**, *125*, 718–747. [\[CrossRef\]](#)

26. Kummaraka, U.; Srisuradetchai, P. Time-Series Interval Forecasting with Dual-Output Monte Carlo Dropout: A Case Study on Durian Exports. *Forecasting* **2024**, *6*, 616–636. [[CrossRef](#)]
27. Khin, A.A.; Thambiah, S. Forecasting analysis of price behavior: A case of Malaysian natural rubber market. *Am.-Eurasian J. Agric. Environ. Sci.* **2014**, *14*, 1187–1195.
28. Zahari, F.Z.; Khalid, K.; Roslan, R.; Sufahani, S.; Mohamad, M.; Rusiman, M.S.; Ali, M. Forecasting natural rubber price in Malaysia using ARIMA. *J. Phys. Conf. Ser.* **2018**, *995*, 012013. [[CrossRef](#)]
29. Dragomiretskiy, K.; Zosso, D. Variational mode decomposition. *IEEE Trans. Signal Process.* **2014**, *62*, 531–544. [[CrossRef](#)]
30. Phoksawat, K.; Phoksawat, E.; Chanakot, B. Forecasting smoked rubber sheets price based on a deep learning model with long short-term memory. *Int. J. Electr. Comput. Eng.* **2023**, *13*, 688–696. [[CrossRef](#)]
31. Pinitjitsamut, M. Multi-Scale Forecasting of Natural Rubber Prices Using VMD-Augmented BiLSTM: A Hybrid Architecture Ablation Study. *Forecasting* **2026**, *8*, 43. [[CrossRef](#)]
32. Rubber Authority of Thailand (RAOT). *Thailand Rubber Statistics: Registered Growers and Production*; Technical Report; Rubber Authority of Thailand: Bangkok, Thailand, 2024.
33. Künsch, H.R. The Jackknife and the Bootstrap for General Stationary Observations. *Ann. Stat.* **1989**, *17*, 1217–1241. [[CrossRef](#)]
34. Zhou, Y.; Qin, D.; Wang, Y. How Shall We Evaluate Load Forecasts? *Power Energy Future* **2026**, *1*, 9650004. [[CrossRef](#)]
35. Del Moral, P.; Guionnet, A. On the stability of interacting processes with applications to filtering and genetic algorithms. *Ann. Inst. Henri Poincaré Probab. Stat.* **2001**, *37*, 155–194. [[CrossRef](#)]
36. Le Gland, F.; Oudjane, N. Stability and uniform approximation of nonlinear filters using the Hilbert metric and application to particle filters. *Ann. Appl. Probab.* **2004**, *14*, 144–187. [[CrossRef](#)]

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.