Stability-Constrained Regime-Aware Forecasting for Heterogeneous Panel Time Series

Dmitry Zaytsev
Lucy Family Institute for Data & Society
Kellogg Institute for International Studies
University of Notre Dame
zaytsevdi2@gmail.com

Valentina V. Kuskova
Lucy Family Institute for Data & Society
University of Notre Dame
vkuskova@nd.edu

Michael Coppedge
Department of Political Science
Kellogg Institute for International Studies
University of Notre Dame
mcoppedg@nd.edu

Abstract

Social science panel time series exhibit multicollinearity, latent unit heterogeneity, and regime-conditional relationships. On V-Dem democratic development data, a linear diagnostic shows that pooled additive predictors cancel regime-conditional signal: pooled coefficients collapse toward zero while regime-specific coefficients straddle zero, a cancellation mechanism that is robust across 24 regime constructions, state summaries, and threshold variants. We study a regime-aware additive forecasting architecture (PANAR-X) under a verifiable contraction certificate. Our main contribution is a rate-based contraction theorem decomposing the global Lipschitz constant of a soft-gated mixture predictor into per-regime, gate-leakage, regime-disagreement, and dwell-time components, each empirically auditable on a trained model. Forecasting outcomes delimit where such structure pays: on V-Dem (151 countries, 1970–2025), no tested model, pooled or stratified, linear or neural, significantly outperforms a random walk under country-clustered inference; on a realized-volatility panel of eight equity indices with 28% headroom over the random walk, a pooled linear model captures the headroom and the regime-aware architecture does not improve on it. The certificate, by contrast, does non-vacuous work on the volatility panel: every per-regime network meets its spectral target, yet $4/5$ seeds fail the contraction certificate because the gating-driven term $L_{\phi,\text{eq}} \cdot M$ dominates, and the dwell-time condition binds non-trivially on the certified seed. The certificate is thus a post hoc, retraining-free audit that detects gating-driven instability that per-edge spectral control misses.

1 Introduction

Forecasting heterogeneous panel time series is a central problem in the social sciences [1]. These data exhibit three persistent challenges: high-dimensional multicollinearity among predictors, latent heterogeneity across units, and regime-dependent relationships that may vary in magnitude or even change sign across states [1; 2]. Classical linear forecasters (ARIMA, vector autoregression [3]) do not address these challenges jointly: their extensions either rely on strong parametric assumptions or scale poorly. Threshold and Markov-switching VAR [4; 5] represent regime structure but inherit linear within-regime dynamics. Recent neural autoregressive approaches [6] offer flexible nonlinear alternatives, but in recursive forecasting settings expressive nonlinear models can amplify small

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

perturbations over time; the standard stability tool [7]- spectral normalization on per-edge layers - turns out to be insufficient for soft-gated mixture architectures, where the gating-driven contribution to the Lipschitz constant is independent of per-edge spectral norms and can dominate.

We present **PANAR-X**, a regime-aware additive forecasting architecture for heterogeneous panel time series with formal stability conditions and an empirical audit protocol. The architecture builds on NAVAR-style per-edge additive networks [6], originally developed for nonlinear Granger-causal structure discovery, and adapts them for panel forecasting in two ways: separating the endogenous lag channel from the predetermined covariate channel, and introducing per-unit intercepts. It further adds a regime-aware soft gating mechanism that captures latent regime structure without hard data partitioning, alongside a contraction analysis to ensure recursive forecast stability. We focus on forecasting a panel of democracy indicators from the V-Dem [8], treating them as country-year time series with heterogeneous, weak-signal dynamics.

Contributions. (i) We introduce a regime-aware additive forecasting architecture for panel data, together with a rate-based contraction theorem that decomposes the global Lipschitz constant into per-regime, regime-disagreement, and dwell-time components, each empirically auditable. (ii) We identify a failure mode of standard stability tools in soft-gated mixture architectures: on a synthetic panel with strong regime contrasts, spectral-norm projection of per-edge layers fails to ensure contraction because the gating term $L_{\phi,\mathrm{eq}} \cdot M$ drives the equilibrium-step Lipschitz constant above unity, even when per-regime $L_Y^{(r)}$ is controlled. A regime-disagreement penalty λ_M on M is therefore required to recover the bound; a 45-run sweep maps the resulting stability–accuracy Pareto frontier. (iii) On a democratic development panel from V-Dem [8], we show that pooled additive models cancel regime-conditional, sign-changing relationships, and that under seed-fair, panel-appropriate evaluation no tested model, pooled or stratified, linear or neural, significantly outperforms a random walk; the architecture's reproducible per-run advantage over its pooled counterpart (unpaired $t > 6.7$ on two seed sets) does not survive ensemble-level comparison, and the contraction certificate holds ($L_{\mathrm{eq,max}} < 1$ in $5/5$ seeds across 9 audited configurations) without explicit projection. On a realized-volatility panel selected under pre-registered criteria, the certificate does non-vacuous work on real data: every per-regime network meets its spectral target, yet $4/5$ seeds fail the contraction certificate through the gating term, with the dwell-time condition binding on the certified seed.

Scope. Our objective is not to maximize accuracy but to show that regime-dependent structure can be modeled in a principled way without sacrificing stability (framing expanded in Appendix I).

2 Related work

Panel forecasting, regime switching, and neural additive autoregression. Sparse VAR [3] and the pooled-versus-per-unit panel literature [1] do not represent regime-conditional sign-changing relationships without explicit regime structure. Markov-switching and threshold VAR [5; 9; 4] do, but inherit linear within-regime dynamics and give stability conditions for individual regime matrices rather than for the switching system as a whole; deep switching state-space models [10] add nonlinear within-regime dynamics without verifiable contraction, and dynamical-systems work on development traps [11] identifies regime structure without a forecasting architecture. NAVAR [6] introduces per-edge additive networks for nonlinear Granger discovery but pools them across the state space, so sign-changing effects cancel; function-valued causal influence [12] shows representationally why scalar summaries collapse such relationships, which motivates the regime-conditional extension here. An extended discussion is in Appendix A.

Positioning relative to general sequence models and hierarchical Bayesian panel models. Our claim is not that PANAR-X outperforms more flexible forecasters, but that it is auditable. Recurrent and attention-based forecasters such as DeepAR [13] and the Temporal Fusion Transformer [14] are the natural expressive alternatives; their one-step Lipschitz constants are neither decomposable by regime nor cheaply estimable, and where recurrent stability has been enforced it is through global constraints on the recurrent weights [15] rather than through regime-wise constants that can be read off a trained model. Hierarchical Bayesian models for time-series cross-sectional data [16; 17] offer partial pooling and uncertainty quantification over unit heterogeneity but no rollout-stability certificate. The per-edge additive structure is what makes the regime-wise constants of Theorem 1 directly computable, so the theorem's premises can be checked rather than assumed. The corrected empirical results also bound what added capacity could buy on the panels studied here: on V-Dem no

tested model separates from the random walk under country-clustered inference (§6.2), and on the volatility panel a pooled linear model captures essentially all measurable headroom (§8), a pattern consistent with recent evidence that linear baselines remain competitive with Transformer forecasters on standard long-horizon benchmarks [18]. None of this implies that such models cannot outperform ours elsewhere; it indicates that on these panels the limiting factor is the available predictive signal rather than model capacity.

Stability of switching systems and neural Lipschitz control. Our stability theorem specializes the switching-system average-dwell-time framework of [19] to soft-gated neural mixture architectures. Where the classical analysis bounds switched-system trajectories in terms of generic per-mode Lipschitz constants, our Theorem 1 expresses the dwell-time threshold α^* in terms of architecture-derived constants ($L_{\text{eq},\max}$, B_{trans}) that are individually computable on the trained model and individually controllable through training. Penalty-based enforcement follows spectral-norm regularization practice [7; 20], with the addition of a regime-disagreement penalty addressing the gating-driven contribution to the bound, which per-edge spectral-norm projection alone cannot control (§7).

3 Architecture

PANAR-X uses NAVAR-style per-edge additive networks [6], a separate small network for each lag-history-to-target edge summed into an additive predictor, with three adaptations for panel forecasting: per-unit intercepts for panel-level heterogeneity, a separate channel for predetermined covariates (X) distinct from the endogenous target-history channel (Y), and the regime-aware soft gating and contraction certificate that are this paper's contribution. We use three variants: **M1** (pooled additive, Y channel only), **M1.5** (pooled additive, Y+X), and **M2** (regime-aware: a soft-gated mixture of R regime-specific M1.5 predictors, which reduces to M1.5 in the single-regime limit). The rest of this section specifies M2; M1 and M1.5 are its pooled specializations.

Panel setup. A panel of U units indexed by u observed at times $t = 0, \ldots, T_u - 1$. The model receives lagged histories $\mathcal{H}_t^{Y,(u)} \in \mathbb{R}^{L \times K_Y}$ and $\mathcal{H}_t^{X,(u)} \in \mathbb{R}^{L \times K_X}$. The state space $\mathcal{Z} = \mathbb{R}^{L \times K_Y} \times \mathbb{R}^{L \times K_X}$ is taken compact, with the Frobenius norm on history matrices and the ℓ_2 operator norm on Jacobians.

Regime function. The regime indicator is a deterministic Lipschitz functional of state. Let $\psi : \mathcal{Z} \to \mathbb{R}$ be a Lipschitz scalar summary, in our applications $\psi(z) = z[L][0]$, the most recent value of the dominant endogenous factor (with $L_\psi = 1$). With fixed regime centers $\{c_r\}_{r=1}^R$ (tertile midpoints of ψ on training data), the regime weights are a tempered softmax

$$\pi_r(\psi) = \frac{\exp(-(\psi - c_r)^2/\tau)}{\sum_{r'} \exp(-(\psi - c_{r'})^2/\tau)}. \tag{1}$$

The composition $\pi(\psi(\cdot))$ has Lipschitz constant L_ϕ (global) on the gating output. We further use the *within-regime* gating Lipschitz $L_{\phi,\text{eq}}^{(r)}$, the maximum gating-output slope inside the regime-r interior, which is empirically 5–10$\times$ smaller than L_ϕ for tempered softmax with separated centers. This distinction matters for Theorem 1.

Per-regime predictors and the soft-gated mixture. For each regime r, a regime-specific predictor takes a NAVAR-style additive form:

$$\mu_\theta^{(r)}(z, e_u)_j = \beta_j^{(r)} + \sum_{i=1}^{K_Y} f_{ij}^{YY,(r)}(\mathcal{H}_t^{Y,(u),i}) + \sum_{p=1}^{K_X} f_{pj}^{XY,(r)}(\mathcal{H}_t^{X,(u),p}) + g_j^{(r)}(e_u), \tag{2}$$

where $f_{ij}^{YY,(r)}$ and $f_{pj}^{XY,(r)}$ are per-edge two-layer ReLU MLPs and e_u is an optional country embedding. The full regime-aware predictor is the soft-gated mixture

$$\mu_\theta(z, e_u) = \sum_{r=1}^{R} \pi_r(\psi(z)) \cdot \mu_\theta^{(r)}(z, e_u). \tag{3}$$

Each per-regime predictor satisfies joint Lipschitz conditions $\|\mu_\theta^{(r)}(\mathcal{H}_1^Y, \cdot) - \mu_\theta^{(r)}(\mathcal{H}_2^Y, \cdot)\|_2 \le L_Y^{(r)} \|\mathcal{H}_1^Y - \mathcal{H}_2^Y\|_F$, with the analogous bound on the X channel ($L_X^{(r)}$).

4 Stability theorem

The architecture's stability properties decompose into three components: within-regime equilibrium dynamics, transition-step expansion at regime punctuations, and the rate at which punctuations occur. We state a rate-based contraction theorem in terms of five constants derived from the trained architecture; a global Lipschitz lemma and a stochastic error-propagation theorem (used to bound recursive predictive variance under covariate-forecast noise) appear in Appendix B, together with all proofs.

Aggregate constants. (i) $L_Y^{(r)}$, the per-regime endogenous Lipschitz of $\mu_\theta^{(r)}$ w.r.t. $\mathcal{H}^Y$, with $L_{\max} := \max_r L_Y^{(r)}$; audited post-hoc by gradient probing on held-out data, max-over-samples. (ii) $L_X^{(r)}$, analogous for the covariate channel, with $L_{X,\max} := \max_r L_X^{(r)}$. (iii) L_ϕ, the global gating Lipschitz, computed analytically from τ. (iv) $L_{\phi,\text{eq}}^{(r)}$, the within-regime gating Lipschitz. (v) $M := \max_{r,r'} \sup_z \|\mu_\theta^{(r)}(z) - \mu_\theta^{(r')}(z)\|_2$, the regime disagreement; the architecture's regime-conditional contribution beyond an additive baseline scales with M. (vi) $\varepsilon^{(r)}$, the *gate leakage* (Definition 1): the maximal probability mass the soft gate assigns outside regime r on states routed to r; $\varepsilon^{(r)} = 0$ under hard routing, and $\varepsilon^{(r)}$ is measured on the trained gate by a single forward pass. These combine into

$$L_{\text{eq}}^{(r)} := \big(1 - \varepsilon^{(r)}\big) L_Y^{(r)} + \varepsilon^{(r)}\, L_{\max} + L_{\phi,\text{eq}}^{(r)} \cdot M, \qquad B_{\text{trans}} := L_{\max} + L_\phi \cdot M, \tag{4}$$

with $L_{\text{eq},\max} := \max_r L_{\text{eq}}^{(r)}$. The leakage term $\varepsilon^{(r)}(L_{\max} - L_Y^{(r)}) \in [0, L_{\max}]$ is the price of soft gating in the equilibrium bound: it vanishes under hard routing, is monotone in $\varepsilon^{(r)}$, and admits the conservative substitution $\varepsilon^{(r)} = 1$, under which $L_{\text{eq}}^{(r)} \le L_{\max} + L_{\phi,\text{eq}}^{(r)} M$ is computable from audit constants alone. Any audited finding of non-contraction ($L_{\text{eq},\max} \ge 1$) is unaffected *a fortiori* by the correction. $L_{\text{eq}}^{(r)}$ is the Lipschitz of one recursion step when the trajectory remains in regime r across consecutive lag-window updates ("equilibrium step"); B_{trans} upper-bounds any single step including those crossing regime boundaries ("transition step" or "punctuation"). A global Lipschitz lemma (Appendix B.1) shows the predictor is globally Lipschitz with constant B_{trans}, but contractivity of recursive trajectories requires the sharper equilibrium-step bound $L_{\text{eq},\max}$.

Theorem 1 (Rate-based regime-conditional contractivity)**.** *Suppose:* ***(A1)*** $L_{eq,\max} < 1$*;* ***(A2)*** $\psi \mapsto \pi(\psi)$ *is Lipschitz with constant* L_ϕ*, giving* $B_{trans} = L_{\max} + L_\phi M$*;* ***(A3)*** *the regime-transition rate* $\alpha(h) := |P(h)|/h$*, where* $P(h)$ *is the set of punctuation steps in horizon* h*, satisfies* $\alpha(h) < \alpha^* := -\log L_{eq,\max} / (\log B_{trans} - \log L_{eq,\max})$ *when* $B_{trans} > 1$*, and is unconstrained when* $B_{trans} \le 1$ *(the rate condition is vacuously satisfied). Then for every horizon* h *and every pair of initial histories,*

$$\|\mathcal{H}_h^Y - \mathcal{H}_h'^Y\|_F \;\le\; C_Y \cdot L_{eq,\max}^{(1-\alpha(h))h}\, B_{trans}^{\alpha(h)h} \cdot \|\mathcal{H}_0^Y - \mathcal{H}_0'^Y\|_F, \tag{5}$$

where C_Y *absorbs the* $\sqrt{L}$ *lag-window-Frobenius conversion. Under (A3), the per-step factor* $\rho := L_{eq,\max}^{1-\alpha} B_{trans}^{\alpha}$ *is strictly less than* 1*, and trajectories contract geometrically.*

The result is a switching-system analog of Hespanha–Morse average-dwell-time stability [19], specialized to neural switching dynamics with architecture-derived constants. The dwell-time intuition, the anchoring of the constants on V-Dem (where every step is contractive and (A3) is vacuous) and on the synthetic panel (where (A1) fails), a trajectory-dependent sharpening (Corollary, Appendix B.3), and a companion stochastic error-propagation theorem (Appendix B.4) are discussed in Appendix C.

Penalty-bridge correspondence. The training loss decomposes as $\mathcal{L} = \mathcal{L}_{\text{forecast}} + \lambda_Y \mathcal{L}_{\text{stab}} + \lambda_\phi \mathcal{L}_{\text{gate}} + \lambda_M \mathcal{L}_{\text{disagree}}$, with the three penalty components targeting $L_Y^{(r)}$, L_ϕ, and M respectively (§5). The correspondence is diagnostic, not constructive: under nonconvex training, no closed-form mapping exists from penalty coefficients to trained Lipschitz constants, and satisfaction of (A1)–(A3) is verified post-hoc by audit. This matches standard practice for spectral-norm regularization [7; 20]. The architecture's contribution is that the constants entering the contraction bound are *computable from the trained model* and *individually controllable* during training, so the failure mode (which constant violates which premise) is identifiable.

5 Training and audit

The trained model is required to satisfy (A1)–(A3) of Theorem 1 after training. Three penalty terms target the relevant constants individually: $\mathcal{L}_{\text{stab}} = \sum_r \mathbb{E}_z[\max(0, \hat{\sigma}_Y^{(r)}(z) - \tilde{L}_Y)^2]$ (soft-max-over-batch ReLU-hinged spectral norm via logsumexp at temperature 5), $\mathcal{L}_{\text{gate}} = (1/\tau)^2$, and $\mathcal{L}_{\text{disagree}} = \sum_{r \neq r'} \mathbb{E}_z \|\mu_\theta^{(r)}(z) - \mu_\theta^{(r')}(z)\|_2^2$.

We compare three enforcement modes for the within-regime $L_Y^{(r)}$: **soft penalty** ($\lambda_Y = 0.20$, target $\tilde{L}_Y = 0.55$); **projection-every-epoch** (light $\lambda_Y = 0.05$ plus hard SVD projection of each per-edge linear layer to spectral norm $\leq \sqrt{0.90}$ each epoch); and **projection-every-batch** (no soft penalty; hard SVD projection to $\leq \sqrt{0.85}$ after every optimizer step [7]). The disagreement penalty λ_M is soft in all three modes; hard projection on M is awkward (it is not a per-layer property). The gating-temperature penalty λ_ϕ is held fixed across modes.

After training, a unified audit computes empirical estimates of all constants from Theorem 1: $\hat{L}_Y^{(r)}$ via gradient probing on held-out data; $\hat{L}_X^{(r)}$ analogously; $\hat{L}_{\phi,\text{eq}}^{(r)}$ by sweeping ψ across each regime interior; $\hat{M}$ as the maximum pairwise regime-output difference on held-out data; $\hat{L}_\phi$ analytically from τ; $\hat{\varepsilon^{(r)}}$ as the maximum of $1 - \pi_r(\psi(z))$ over held-out states routed to regime r (Definition 1). The derived $\hat{L_{\text{eq,max}}}$, $\hat{B_{\text{trans}}}$, $\hat{\alpha}^*$, and $\hat{\rho}$ follow. We treat the bound as satisfied for a trained model when $\hat{L_{\text{eq,max}}} < 1$ at every seed in a multi-seed run. A complementary certified spectral-norm upper bound (Appendix J.2) confirms the gradient-probe estimate is non-vacuous in all our settings; the gap between the two estimators is empirically $\approx 7\%$ on V-Dem democracy indicators.

6 Empirical demonstration: democratic development (V-Dem indicators)

We demonstrate the architecture on predicting year-over-year change in clean elections across 151 countries. The aim is mechanism demonstration first and predictive benchmark second: a diagnostic establishes that regime-conditional signal exists in ΔY and is canceled by pooled additive predictors; the architecture is then evaluated against the random walk and stratified linear baselines under panel-appropriate inference; and the contraction conditions are verified.

Why V-Dem is a hard test. The panel exhibits all three difficulties that motivate this work: multi-collinearity among expert-coded indicators, latent unit heterogeneity across countries, and regime-conditional sign changes that the ΔY target exposes once the trivial pooled persistence ($\rho \approx 0.98$) is removed (details in Appendix F).

Data and setup. We analyze a balanced panel of country-year democracy indicators derived from V-Dem v15 [8]: 151 countries, 1970–2025. Target Y_t = `v2xel_frefair` (clean elections); covariate channel X comprises four substantively-interpretable shock-like indicators (`v2csantimv`, `v2jureform`, `v2psorgs`, `v2mecorrpt`). Lag window $L = 5$; per-country temporal split with the last 10 years held out. Final sample sizes: 6,191 training, 1,510 test observations.

6.1 Diagnostic: regime-dependent signal exists and is canceled by pooling

We test directly whether regime-dependent structure is present in the data, independent of any nonlinear architecture. We define three regimes from training-set Y_{t-1} tertiles (low / mid / high democracy), and compare two linear predictors on ΔY_t: (a) a pooled linear model on the full $(Y_{t-L:t-1}, X_{t-L:t-1})$ lag window; (b) a regime-stratified linear model fit independently on each tertile, with test observations routed to their stratum. The two share identical functional form within each regime; the only difference is whether parameters are estimated jointly (pooled) or separately (stratified).

The diagnostic establishes the substantive social science finding: *regime-conditional predictive signal exists in ΔY and is canceled by pooled additive predictors.* Regime-stratified estimation exposes the structure, but the associated MSE improvement ($+3.02\%$ overall) is not statistically significant under the pre-specified panel test (block bootstrap $p = 0.273$; Table 1). The cancellation mechanism is direct: four of four selected X variables exhibit coefficient sign reversal across regimes at one or more lag, so pooling averages over sign-disagreed regression coefficients and approaches their cancellation

Table 1: Primary V-Dem forecasting comparison on the ΔY test partition (N=1,510; 151 countries). The random walk and the stratified linear model are the primary baselines. Stochastic models report per-run test MSE (mean $\pm$ sd over 20 seeds) and the seed-ensemble MSE (MSE of the mean of the 20 prediction vectors); skill is $1 - \mathrm{MSE}/\mathrm{MSE}_{\mathrm{RW}}$ for the ensemble (or single) MSE. No tested model significantly outperforms the random walk under country-clustered inference (Table 2; Appendix F).

Model	Type	Per-run MSE	Ensemble MSE	Skill vs. RW
Random walk	det.	–	0.00340	0.00%
Stratified linear M1.5	det.	–	0.00337	+0.84%
Linear in Y_{t-1} only	det.	–	0.00346	−1.99%
Pooled linear M1.5	det.	–	0.00347	−2.25%
M1 NN (Y-only)	20 seeds	0.00335 ± 0.00005	0.00333	+1.80%
Pooled NN M1.5	20 seeds	0.00350 ± 0.00004	0.00339	+0.08%
Stratified NN M1.5	20 seeds	0.00344 ± 0.00005	0.00324	+4.60%
M2 (PANAR-X)	20 seeds	0.00342 ± 0.00003	0.00339	+0.26%

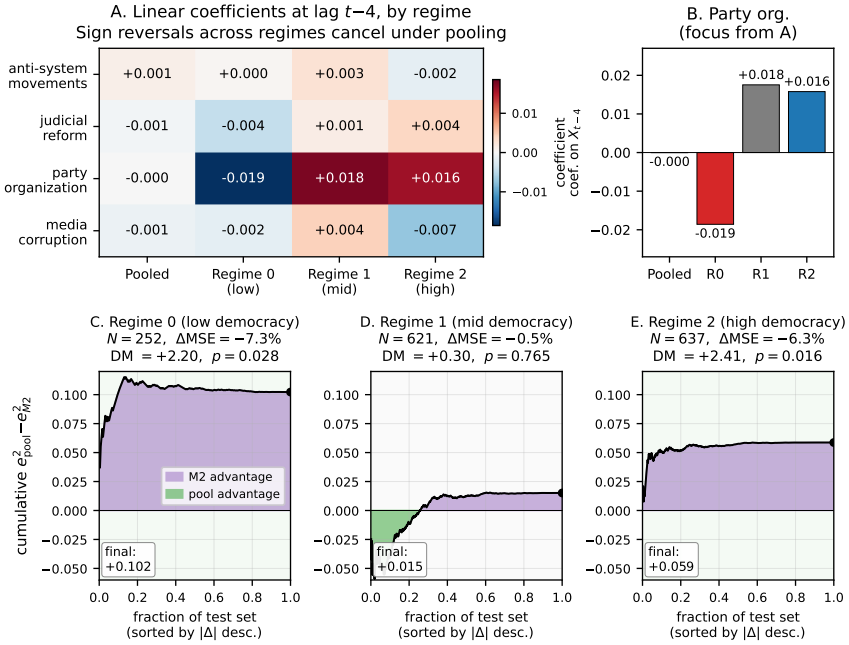


Figure 1: Diagnostic and recovery of regime-conditional signal on V-Dem ΔY. **(A)** Linear coefficients at lag $t-4$ for the pooled model (column 1) and the three regime-stratified models (columns 2–4); sign reversals across regimes drive near-zero pooled coefficients. **(B)** Party organization at lag $t-4$, the cleanest exemplar. **(C–E)** Cumulative paired squared-error advantage of the M2 seed-ensemble over the single-seed pooled NN of the submitted test, sorted by |paired diff| descending; the final value equals $N \cdot \overline{\Delta\mathrm{MSE}}$. This is per-run behavior; the seed-fair ensemble comparison is null (Table 2; reading guide in Appendix F).

limit. Figure 1A–B visualizes this: at lag $t-4$, the pooled-linear coefficient on party organization is essentially zero (-0.000) while the per-regime coefficients are sharply opposed (-0.019 in low democracy, $+0.018$ in mid, $+0.016$ in high). The remaining question is methodological: whether a regime-aware end-to-end architecture recovers the same signal without hard partitioning of the data and while maintaining stability.

6.2 Baselines, architecture sweep, and predictive evaluation

Baselines on the ΔY test partition are in Table 1: the random walk ($\widehat{\Delta Y_t} = 0$) at test MSE 0.00340, pooled linear M1.5 worse than the random walk, stratified linear M1.5 at +0.84% over it (not

significant, $p = 0.830$), and the pooled NN M1.5 at 0.00350 per run and 0.00339 as a 20-seed ensemble (the submitted version described 0.00352 as a 5-seed mean; it was a single seed-0 run, corrected here; further discussion in Appendix F). We sweep PANAR-X (M2) across three audited enforcement modes (soft penalty, projection-every-epoch, projection-every-batch) crossed with $\lambda_M \in \{0, 0.05, 0.10\}$ and five seeds, for 45 audited runs; the primary baselines are the random walk and the stratified linear model, with M1 and M1.5 as the architecture-matched pooled comparisons.

Table 2: Primary inference on V-Dem (overall, N=1,510, 151 country clusters): paired loss-differential tests with country-clustered t (Liang–Zeger, $\mathrm{df} = G-1$) and country-block bootstrap (B=9,999); positive $\bar{d}$ favors the first-named model. Rows 1–3 use the archived 5-seed M2 ensemble (the submitted configuration); row 5 is the seed-fair 20-vs-20 ensemble comparison that replaces the submitted headline test, in which a 5-seed M2 ensemble was compared against a single-seed pooled baseline. Conclusions are identical at both seed counts. No comparison is significant, and per-regime tests are likewise null after Holm correction (closest case: M2 vs. stratified linear in regime 2, uncorrected clustered $p = 0.051$, bootstrap $p = 0.202$, Holm $p = 0.606$).

Comparison	$\bar{d} \times 10^4$	cluster t	cluster p	bootstrap p
M2-ens vs. random walk	-0.056	-0.077	0.938	0.940
M2-ens vs. stratified linear	-0.343	-0.413	0.680	0.690
Stratified linear vs. random walk	$+0.287$	$+0.215$	0.830	0.828
Stratified vs. pooled linear	$+1.050$	–	–	0.273
M2-ens vs. pooled-NN-ens (seed-fair)	$+0.062$	$+0.245$	0.807	0.829

Per-run vs. ensemble performance. The submitted version reported a paired DM test of $+2.41$ ($p = 0.016$) for M2 against pooled NN M1.5. During reanalysis we identified a design flaw in that comparison: it paired a 5-seed M2 ensemble against a single-seed pooled baseline, so part of the gap was the variance reduction of ensembling. At the per-run level the architecture is reproducibly better (unpaired $t = +6.73$ on the primary 20-seed set, $t = +8.67$ on a pre-committed fresh set, both $p < 10^{-5}$; 39 of 40 same-seed comparisons favor M2); at the ensemble level the advantage disappears under the seed-fair 20-vs-20 comparison (Table 2, row 5, clustered $p = 0.807$), because the pooled model's larger seed variance is what ensembling removes. We therefore withdraw the forecasting-improvement claim: the architecture produces better and lower-variance individual fits, and ensembling, the relevant deployment convention, equalizes them. The full account, including per-regime behavior, is in Appendix F.

Per-regime behavior. After Holm correction across regimes, no comparison of M2 against the stratified linear model or the random walk is significant (Table 2, caption); Figure 1C–E therefore shows per-run behavior under the submitted seed-asymmetric comparison, concentrated in regimes 0 and 2, not an ensemble-level gain.

6.3 Stability verification and interpretation

We verify the contraction conditions of Theorem 1 on the trained models via the audit protocol of Appendix J. All 9 audited cells satisfy $\hat{L}_{\mathrm{eq,max}} < 1$ in $5/5$ seeds; the empirical $\hat{L}_{\mathrm{eq,max}}$ ranges from 0.130 (most enforced) to 0.292 (unconstrained soft mode), all well within the contraction regime; under the conservative gate-leakage substitution $\varepsilon^{(r)} = 1$ in (4), the maximum rises to 0.362 ($\hat{\rho} \leq 0.37$) and every certificate still holds. The certified spectral-norm upper bound $\widetilde{L}_{\mathrm{eq,max}}$ (Appendix J.2), which by construction dominates the gradient-probe estimate and yields a stronger guarantee, exceeds it by only $\approx 7\%$ on V-Dem indicator and remains well below 1. Combined with the empirical $\hat{\alpha} < \hat{\alpha}^*$, the implied contraction rate $\hat{\rho} \leq 0.36$ across all audited cells: recursive forecasts under the trained PANAR-X are contractive, with small input perturbations decaying over multi-step rollout.

Why the audit holds without projection on this panel, and how this contrasts with the synthetic test, is discussed in Appendix F and Section 9.

6.4 Sensitivity to the regime construction

The diagnostic and the corrected predictive conclusions both rest on a hand-specified regime construction, so we test every element of it under a pre-registered sensitivity design (locked before any

run; design documents in the archived package). The diagnostic is refit across a 24-variant grid: three state summaries ψ (the lagged target, a 3-year smoothed level, and the first principal component of state and covariates), crossed with $R \in \{2, 3, 4, 5\}$ regimes and with ± 5 percentage-point shifts of each tertile cut point at R=3. The stratified linear model is refit under eight configurations, and M2 is fully retrained (5 seeds each) under four representative variants. An independent reimplementation of the regime splits reproduces the canonical coefficients. Configuration-level results, including the eight-configuration stratified-linear table, are in Appendix G; here we report what survives.

The corrected null is construction-invariant. M2, fully retrained under four representative regime definitions (median split, quartiles, the 3-year smoothed ψ, and shifted tertile cuts), remains statistically indistinguishable from the random walk in every variant (block bootstrap $p \geq 0.80$) and from the matching stratified linear model after Holm correction ($p \geq 0.16$). The V-Dem equivalence of Section 6.2 is therefore a property of the panel, not of any particular regime construction.

What kind of signal is recovered. The sensitivity results allow the diagnostic to be read in three parts rather than as a single finding. *Stable structure*: two results meet this standard, the cancellation mechanism itself (a pooled coefficient collapsing toward zero while the regime-specific coefficients straddle it, present in $24/24$ variants) and the low-democracy stratum's negative party-organization coefficient at lag $t-4$ (negative in $24/24$ variants, $|\hat{\beta}| > 2\,\mathrm{se}$ in $13/24$, sign-stable in 94.5% of 200 half-country jackknife subsamples). *Sample-specific pattern*: the symmetric low/mid/high reversal of the submitted version belongs here. Its positively signed stratum moves with the cut points, its significance-filtered outer reversal appears in $0/24$ variants, the complete pattern survives in 1.5% of jackknife subsamples, and the low-stratum association attenuates to approximately zero in the most recent half of the training window. The stratification gain belongs here too: it ranges from -0.5% to $+3.8\%$ across configurations and is not significant under the panel test ($p = 0.273$). *Substantive relationship*: here the design does not license a claim. The robust evidence shows that pooling masks regime-dependent heterogeneity on this panel, a statistical result with a clear methodological implication; whether the persistent low-democracy association reflects a political mechanism cannot be established by a forecasting design, and we make no causal claim. We present it as a robust empirical regularity that motivates regime-aware modeling, not as evidence for a particular mechanism.

7 Synthetic stress test: when the contraction certificate binds

The V-Dem audit holds without explicit projection because the trained M is moderate. To show the certificate is non-vacuous, and to identify what training-time intervention restores contraction when it fails, we construct a synthetic three-regime panel with deliberately strong regime contrasts; Section 8 then shows the same failure mode on a real panel selected under pre-registered criteria.

DGP and design. Univariate three-regime panel with regime determined by the lagged target, $Y_{u,t} = \beta_u + \rho_Y^{(r_t)} Y_{u,t-1} + \beta_X^{(r_t)} X_{u,t-1} + \epsilon_{u,t}$, with $\rho_Y = (0.5, 0.8, 0.7)$ and $\beta_X = (0.05, 0.20, 0.40)$ (an $8\times$ range across regimes), $U = 30$ units, $T = 50$, last 10 observations per unit held out, empirical regime-transition rate $\alpha \approx 0.20$. Full DGP, training details, and the baseline ladder are in Appendix D.

Unconstrained: predictive contribution, broken contraction. Trained without explicit Lipschitz regularization (soft penalty mode, $\lambda_M = 0$, $\lambda_Y = 0.20$, target $\tilde{L}_Y = 0.55$), M2 attains a clean predictive advantage over pooled M1.5: 5-seed mean test MSE drops from 0.1497 to 0.1410, gap $+0.0087$ (95% CI $[+0.0065, +0.0109]$). The audit at this optimum gives $L_{\max} = 0.77$, $L_{\phi,\mathrm{eq}} = 0.27$, $M = 5.94$, and $L_{\mathrm{eq,max}} = 2.38$: *the contractivity premise of Theorem 1 fails*, driven entirely by the gating term $L_{\phi,\mathrm{eq}} \cdot M = 1.6$ while $L_{\max} = 0.77 < 1$. The per-edge networks individually satisfy contraction; their disagreement pushes the mixture-system bound above unity.

The Pareto frontier and the lever. Sweeping the three enforcement modes crossed with $\lambda_M \in \{0, 0.05, 0.10\}$ (45 runs; Tables 4 and 5, Appendix D.4) reveals a monotonic frontier: stronger enforcement reduces $L_{\mathrm{eq,max}}$ and the M2-vs-M1.5 gap together (OLS of gap on $L_{\mathrm{eq,max}}$ across the 45 runs: slope $+0.0075 \pm 0.0011$, $R^2 = 0.54$, $p < 10^{-7}$). Unenforced cells attain the strongest predictive contribution but fail the bound in $5/5$ seeds; per-batch projection at $\lambda_M = 0.10$ attains $L_{\mathrm{eq,max}} < 1$ in $5/5$ seeds with the contribution removed. The decomposition $L_{\mathrm{eq}}^{(r)} = L_Y^{(r)} + L_{\phi,\mathrm{eq}}^{(r)} \cdot M$ identifies the lever: under per-batch projection with $\lambda_M = 0$, $L_Y^{(r)}$ is hard-projected to target yet $L_{\mathrm{eq,max}} \approx 2.59$ because $M \approx 6.95$; activating $\lambda_M = 0.10$ on the same projection collapses M

to 0.47 and $L_{\text{eq,max}}$ to 0.91 with $L_Y^{(r)}$ unchanged. *Hard projection on L_Y alone, the conventional spectral-normalization route, does not deliver contractivity for switching systems with non-trivial regime structure; the regime-disagreement penalty is essential.* The same disagreement that breaks the bound drives the predictive contribution, and λ_M is the controllable knob. On V-Dem, by contrast, $M \approx 0.87$ at $\lambda_M = 0$, nearly an order of magnitude smaller, and $L_{\text{eq,max}} < 1$ holds without any disagreement penalty (Section 6.3); the audit constants tell which case a dataset is in (Section 9).

8 The certificate on real data: a pre-registered realized-volatility panel

The V-Dem audit cannot exercise the certificate: its constants sit far below one, so the dwell-time condition never does any work. To test the certificate where instability is plausible, without selecting a panel on favorable outcomes, we fixed the selection criteria in advance: a candidate panel must show (i) genuine forecast headroom over a random walk and (ii) measurable regime heterogeneity. The criteria, a hierarchy of admissible claims for each possible outcome, and the confirmatory protocol were locked in design documents before any confirmatory training; the documents and all run artifacts are archived (DOI 10.5281/zenodo.23717582).

Panel and screen. We use daily realized volatility for eight international equity indices [21] (2,615 trading days), with the one-day-ahead change in log volatility as the target and training-set tertiles of $\psi = y_{t-1}$ as regimes; the test partition is the final 523 days (N=4,184). Construction, lag structure, inference, and hyperparameter selection are in Appendix H. Both gates passed: pooled linear regression beats the random walk by 28.2% (bootstrap p at the 10^{-4} resolution floor), and regime heterogeneity is strong (block-bootstrap Wald $W = 59.9$, $p = 5.4 \times 10^{-7}$), although no slope changes sign across regimes: the heterogeneity is amplification-type rather than the reversal-type found on V-Dem, an outcome the locked design anticipated.

Certificate audit: spectral compliance without contraction. We train M2 on this panel (hyperparameters selected on a training-internal validation split; 5 seeds) and audit every trained model (Table 7, Appendix H). Every per-regime network meets the soft spectral target (audited $\hat{L}_Y^{(r)} = 0.55$–0.58 against the 0.55 target), yet four of five seeds carry no contraction certificate: the gating term $\hat{L}_{\phi,\text{eq}} \cdot \hat{M} = 0.43$–$0.68$ pushes $\hat{L}_{\text{eq,max}}$ to 1.07–1.25, above one despite full per-edge compliance. This is the failure mode of Theorem 1 realized on real data, previously shown only in the constructed panel of Section 7. The remaining seed is certified with the dwell-time condition binding non-trivially for the first time on real data: $\hat{L}_{\text{eq,max}} = 0.984$, $\hat{B}_{\text{trans}} = 1.010 > 1$, and the empirical switching rate $\hat{\alpha} = 0.271$ lies below the critical $\hat{\alpha}^* = 0.619$, giving $\hat{\rho} = 0.991$. Under the conservative gate-leakage substitution $\varepsilon^{(r)} = 1$ of (4), the four uncertified seeds remain uncertified *a fortiori* (1.07–1.25) and the certified seed survives with a thinner margin ($\hat{L}_{\text{eq,max}} = 0.996$, $\hat{\alpha}^* = 0.305 > 0.271$, $\hat{\rho} = 0.9995$).

Predictive results: the most conservative pre-committed claim level. The locked design specified a hierarchy of admissible claims; the outcome fell at the most conservative level, and we report it as such. M2 beats the random walk decisively ($+8.57$, bootstrap $p = 10^{-4}$) but is significantly worse than both linear baselines out of sample: vs. a pre-specified interaction-augmented pooled linear model ($\bar{d}$-based $t = -3.82$, $p = 0.0014$; calm regime, 70% of test observations, $t = -4.85$, Holm $p = 0.0012$, $5/5$ seeds) and vs. pooled linear ($t = -3.57$, $p = 0.0024$). The interaction model itself adds nothing over pooled linear ($t = -0.06$, $p = 0.95$) despite the strong in-sample heterogeneity: the regime structure is statistically detectable but not exploitable out of sample by any method we tested, linear or neural. Together with V-Dem, the two panels characterize a boundary: without headroom over persistence, all methods tie the random walk; with headroom, a pooled linear model captures it, and regime-aware structure, whether as interactions, stratification, or soft gating, does not add to it at this horizon. The certificate, by contrast, is exactly where this panel is informative: the models the audit flags as non-contractive are also the out-of-sample underperformers, a co-occurrence that is suggestive of diagnostic value, although one panel and five seeds do not establish the link.

9 Discussion: when the certificate costs something, and what to do first

The two-panel trade-off. Reviewers asked why enforcing stability removes the regime-aware gain in the synthetic test and what the trade-off looks like on real data. The two real panels give a two-sided answer. On V-Dem the constraint has no practical bite: across all ten enforcement cells, from penalty-free hard routing to per-batch projection at $\lambda_M = 0.10$, the per-run M2 test MSE is 0.00343

to 0.00345 (seed sd 0.00003 to 0.00005) while $\hat{L}_{\mathrm{eq,max}}$ falls from 0.29 to 0.13; the audited constants sit far below the 0.55 penalty target, the penalties never activate, and the certificate reduces to post hoc confirmation. On the volatility panel (Section 8) the trade-off is real: every per-regime network meets its spectral target, the gating term alone leaves $4/5$ seeds uncertified, and the regime-aware model is significantly worse out of sample than the pooled linear model that captures the panel's headroom. When the dynamics are already stable, the certificate costs nothing; when instability is present, the regime-dependent structure the architecture is built to exploit is the structure the certificate flags, and on this panel no method exploited it out of sample (an untested interpretation of V-Dem's position is in Appendix I).

Audit first. Every constant in Theorem 1 is computable post hoc from a trained model by gradient probing and a single forward pass (Appendix J), in seconds, without retraining or a GPU, so the practical rule is to audit before deciding how much stability control to buy. Constants far below one, as on V-Dem (0.13 to 0.29), mean stability is free and the disagreement penalty can be left off; constants near or above one, as on the volatility panel (1.07 to 1.25 in $4/5$ seeds), mean stability has a price to be paid through λ_M or stronger projection before recursive deployment, and (4) identifies which term is responsible.

Appropriate use for democracy forecasting. At a one-year horizon on measurement-smoothed indices, no tested method, including ours, outperforms persistence under country-clustered inference (Table 2). The appropriate uses of this class of model on democracy data are therefore diagnostic: detecting regime-dependent heterogeneity that pooled models mask (Sections 6.1 and 6.4) and auditing rollout stability before deployment (Section 6.3). The misuse the boundary result warns against is treating one-step forecasts of such indices as operational predictions of democratic change for particular countries: a model indistinguishable from a random walk should not inform such decisions, and that the best models tie persistence is itself the strongest argument against overconfident operational use.

10 Limitations and conclusion

Limitations. Four limitations bound the contribution. (i) V-Dem forecasting performance is statistically indistinguishable from persistence and from stratified linear regression under seed-fair, country-clustered inference (Table 2); the V-Dem result is a mechanism demonstration and a boundary characterization, not a predictive benchmark. (ii) The robust empirical result is narrower than the symmetric sign-reversal claim of the submitted version: what survives the pre-registered sensitivity design is the cancellation mechanism and the low-democracy stratum's negative party-organization coefficient (Section 6.4); the symmetric reversal is withdrawn. (iii) The stability theorem is an architecture-specific, auditable specialization of average-dwell-time analysis [19]; its contribution is that the constants entering the bound are computable on the trained model and individually controllable in training (Section 4), not a new stability technique. (iv) The realized-volatility experiment demonstrates that the certificate is non-vacuous on real data, not that the regime-aware architecture improves forecasting (Section 8). Further limitations, discussed in Appendix I: V-Dem is a single target on a single social-science panel; regime structure is exogenous to the model, and everything finer than the cancellation mechanism depends on it; per-regime parameters scale linearly with R; the penalty-bridge correspondence is diagnostic, not constructive; and function-valued interpretability of the per-regime networks is not re-demonstrated here.

Conclusion. PANAR-X is a regime-aware additive forecasting architecture with a verifiable contraction certificate. The theorem decomposes the global Lipschitz constant into per-regime, gate-leakage, regime-disagreement, and dwell-time components, each auditable and individually controllable, and identifies a failure mode of standard spectral normalization on soft-gated mixtures. On V-Dem, pooled additive models cancel regime-conditional signal, the certificate holds in $5/5$ seeds across all 9 audited configurations without explicit projection, and under seed-fair, panel-appropriate evaluation no tested model significantly outperforms a random walk; on a pre-registered realized-volatility panel, $4/5$ spectrally compliant seeds fail certification through the gating term, and the architecture does not beat a pooled linear model there. We claim no predictive gains on these panels. We claim that regime-dependent dynamics can be modeled under formal stability constraints, that the constants needed to tell whether stability binds for a given dataset are computable post hoc without retraining, and that the diagnostic and the certificate together address the gap between pooled models that cancel sign-changing structure and unconstrained nonlinear models whose recursive stability is unverified.

Acknowledgments and Disclosure of Funding

This work was supported in part by funding from the Kellogg Institute for International Studies, the Franco Family Institute for Liberal Arts and the Public Good, and the Lucy Family Institute for Data & Society at the University of Notre Dame.

References

[1] M. Hashem Pesaran, Andreas Pick, and Allan Timmermann. Forecasting with panel data: Estimation uncertainty versus parameter heterogeneity. *arXiv preprint arXiv:2404.11198*, 2024.

[2] Boyuan Zhang. Incorporating prior knowledge of latent group structure in panel data models. *arXiv preprint arXiv:2211.16714*, 2024.

[3] William B. Nicholson, David S. Matteson, and Jacob Bien. BigVAR: Tools for modeling sparse high-dimensional multivariate time series. *arXiv preprint arXiv:1702.07094*, 2017.

[4] Ruey S. Tsay. Testing and modeling multivariate threshold models. *Journal of the American Statistical Association*, 93(443):1188–1202, 1998.

[5] James D. Hamilton. A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 57(2):357–384, 1989.

[6] Bart Bussmann, Jannes Nys, and Steven Latré. Neural additive vector autoregression models for causal discovery in time series. *Discovery Science*, 2021.

[7] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2018.

[8] Michael Coppedge, John Gerring, Carl Henrik Knutsen, Staffan I. Lindberg, Jan Teorell, et al. V-Dem [Country-Year/Country-Date] Dataset v15, 2025.

[9] Hans-Martin Krolzig. *Markov-Switching Vector Autoregressions: Modelling, Statistical Inference, and Application to Business Cycle Analysis*. Springer, 1997.

[10] Xiuqin Xu, Hanqiu Peng, and Ying Chen. Deep switching state space model for nonlinear time series forecasting with regime switching. *International Journal of Forecasting*, 42(1):85–98, 2026. doi: 10.1016/j.ijforecast.2025.05.001.

[11] Shyam Ranganathan, Stamatios C. Nicolis, Viktoria Spaiser, and David J. T. Sumpter. Understanding democracy and development traps using a data-driven approach. *Big Data*, 3(1):22–33, 2015.

[12] Valentina V. Kuskova, Dmitry Zaytsev, and Michael Coppedge. Function-valued causal influence in nonlinear time series. In *Proceedings of the 43rd International Conference on Machine Learning (ICML), PMLR 306:61873-61898*, 2026.

[13] David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 36(3): 1181–1191, 2020. doi: 10.1016/j.ijforecast.2019.07.001.

[14] Bryan Lim, Sercan Ö. Arık, Nicolas Loeff, and Tomas Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4): 1748–1764, 2021. doi: 10.1016/j.ijforecast.2021.03.012.

[15] N. Benjamin Erichson, Omri Azencot, Alejandro Queiruga, Liam Hodgkinson, and Michael W. Mahoney. Lipschitz recurrent neural networks. In *International Conference on Learning Representations (ICLR)*, 2021.

[16] Boris Shor, Joseph Bafumi, Luke Keele, and David Park. A Bayesian multilevel modeling approach to time-series cross-sectional data. *Political Analysis*, 15(2):165–181, 2007. doi: 10.1093/pan/mpm006.

[17] Nathaniel Beck and Jonathan N. Katz. Random coefficient models for time-series–cross-section data: Monte Carlo experiments. *Political Analysis*, 15(2):182–195, 2007. doi: 10.1093/pan/mpl001.

[18] Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 11121–11128, 2023. doi: 10.1609/aaai.v37i9.26317.

[19] João P. Hespanha and A. Stephen Morse. Stability of switched systems with average dwell-time. In *Proceedings of the 38th IEEE Conference on Decision and Control*, pages 2655–2660, 1999.

[20] Henry Gouk, Eibe Frank, Bernhard Pfahringer, and Michael J. Cree. Regularisation of neural networks by enforcing Lipschitz continuity. *Machine Learning*, 110:393–416, 2021.

[21] Bumho Son, Yunyoung Lee, Seongwan Park, and Jaewook Lee. Forecasting global stock market volatility: The impact of volatility spillover index in spatial-temporal graph-based model. *Journal of Forecasting*, 42(7):1539–1559, 2023. doi: 10.1002/for.2975.

[22] Francis X. Diebold and Roberto S. Mariano. Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3):253–263, 1995.

A Extended related work

This appendix expands the condensed survey of Section 2. The regime-disagreement penalty addresses the gating-driven contribution to the bound, a contribution we identify as essential because spectral-norm projection on per-edge layers alone can fail to deliver contraction (§7).

Linear and regime-switching panel forecasting. Classical multivariate panel forecasting relies on vector autoregression with n^2p parameters in an unrestricted $\mathrm{VAR}(p)$ on n series; sparse and structured penalizations [3] address the high-dimensional case but do not impose state-dependent dynamics. The forecasting-with-panel-data literature documents the tension between pooled and per-unit estimators [1], neither of which represents regime-conditional sign-changing relationships without explicit regime structure. Markov-switching VAR (MS-VAR) [5; 9] introduces hidden-Markov regimes with linear within-regime dynamics; threshold VAR (T-VAR) [4] uses an observable threshold variable. Both inherit linear dynamics within regimes and provide stability conditions for individual regime matrices rather than for the switching system as a whole. Recent neural extensions [10] introduce nonlinear within-regime dynamics via deep state-space models but do not address verifiable contraction of the resulting switching system. Dynamical-systems modeling of social science panel data [11] similarly identifies regime structure (development traps) but does not provide a forecasting architecture with stability guarantees.

Neural additive autoregression and our positioning. Neural additive vector autoregression (NAVAR) [6] introduces per-edge nonlinear networks summed in an additive predictor, originally proposed and validated for nonlinear Granger-causal structure discovery on multivariate time series. NAVAR's forecasts were used as a score for Granger links rather than as a forecasting deliverable, and the architecture does not represent regime structure: per-edge functions are pooled across the state space, so sign-changing regime-conditional effects cancel under pooling. Recent work [12] on function-valued causal influence shows representationally that scalar summaries collapse such sign-changing relationships; this motivates our regime-conditional architectural extension. PANAR-X repurposes the NAVAR per-edge structure for forecasting, separates the endogenous lag channel from the predetermined covariate channel, adds per-unit intercepts for panel heterogeneity, and adds regime-aware soft gating with a contraction analysis that NAVAR, MS-VAR, and TVAR each lack.

B Full proofs

We give full proofs of the global Lipschitz lemma, the rate-based contraction theorem (Theorem 1), and a stochastic error-propagation theorem, along with two auxiliary lemmas (equilibrium-step and punctuation-step bounds).

B.1 Global Lipschitz lemma

Lemma 1 (Global Lipschitz of the soft-gated mixture)**.** *The predictor μ_θ defined in Section 3 is globally Lipschitz with constant $L_{global} = L_{\max} + L_\phi \cdot M = B_{trans}$.*

Proof. Let $z_1 = (\mathcal{H}_1^Y, \mathcal{H}^X, e)$ and $z_2 = (\mathcal{H}_2^Y, \mathcal{H}^X, e)$. Using the soft-gating mixture (3), write

$$\mu_\theta(z_1) - \mu_\theta(z_2) = \underbrace{\sum_r \big[\pi_r(\psi(z_1)) - \pi_r(\psi(z_2))\big]\mu_\theta^{(r)}(z_1)}_{\text{(I): gating-driven}} + \underbrace{\sum_r \pi_r(\psi(z_2))\big[\mu_\theta^{(r)}(z_1) - \mu_\theta^{(r)}(z_2)\big]}_{\text{(II): within-regime}} \tag{6}$$

The decomposition is exact: add and subtract $\sum_r \pi_r(\psi(z_2))\mu_\theta^{(r)}(z_1)$.

Bound on (I). Since $\sum_r \pi_r(s) = 1$ for all s, $\sum_r[\pi_r(\psi(z_1)) - \pi_r(\psi(z_2))] = 0$, so we can subtract any reference $\mu_\theta^{(r^*)}(z_1)$ from each summand:

$$\begin{aligned}
\|(\mathrm{I})\|_2 &\le \sum_r |\pi_r(\psi(z_1)) - \pi_r(\psi(z_2))| \cdot \|\mu_\theta^{(r)}(z_1) - \mu_\theta^{(r^*)}(z_1)\|_2 \\
&\le M \cdot \sum_r |\pi_r(\psi(z_1)) - \pi_r(\psi(z_2))| \\
&\le M \cdot L_\phi \cdot |\psi(z_1) - \psi(z_2)| \\
&\le L_\phi M \cdot \|\mathcal{H}_1^Y - \mathcal{H}_2^Y\|_F.
\end{aligned}$$

Bound on (II). A convex combination of per-regime differences:

$$\|(\mathrm{II})\|_2 \le \sum_r \pi_r(\psi(z_2)) \cdot L_Y^{(r)} \|\mathcal{H}_1^Y - \mathcal{H}_2^Y\|_F \le L_{\max} \|\mathcal{H}_1^Y - \mathcal{H}_2^Y\|_F.$$

Combining. By the triangle inequality applied to (6), $\|\mu_\theta(z_1) - \mu_\theta(z_2)\|_2 \le (L_{\max} + L_\phi M)\|\mathcal{H}_1^Y - \mathcal{H}_2^Y\|_F$. The covariate-channel statement follows analogously. □

B.2 Auxiliary lemmas: equilibrium and punctuation step bounds

For the proof of Theorem 1 we need two refined per-step bounds. Consider two recursive trajectories $\{z_j\}_{j=0}^h$ and $\{z'_j\}_{j=0}^h$ updated by $\mathcal{H}_{j+1}^Y = (\mathcal{H}_j^Y[1:], \mu_\theta(z_j))$, $\mathcal{H}_{j+1}^X = (\mathcal{H}_j^X[1:], X_{j+1})$ with the same exogenous covariate sequence. Define the regime indicator $r_j := \arg\max_r \pi_r(\psi(z_j))$, the equilibrium set $E(h) := \{j : r_j = r_{j-1} \text{ on both trajectories}\}$, the punctuation set $P(h) := \{j : r_j \neq r_{j-1} \text{ on at least one trajectory}\}$, and $\delta_j := \mathcal{H}_j^Y - \mathcal{H}_j'^Y$.

Definition 1 (Gate leakage)**.** *For each regime r, define the* gate leakage *as*

$$\varepsilon^{(r)} := \sup_{z \in \mathcal{Z}^{(r)}} \big(1 - \pi_r(\psi(z))\big) \in [0, 1].$$

Under hard routing, $\pi_r(\psi(z)) \in \{0, 1\}$ on $\mathcal{Z}^{(r)}$ and $\varepsilon^{(r)} = 0$ exactly; under tempered-softmax gating, $\varepsilon^{(r)}$ quantifies how far the gate is from concentrating on states routed to regime r, and is computable on the trained model by a single forward pass over states assigned to $\mathcal{Z}^{(r)}$. (Restricting the supremum to a sub-region of $\mathcal{Z}^{(r)}$ bounded away from the regime boundaries yields a smaller $\varepsilon^{(r)}$ and a sharper constant, provided the equilibrium set $E(h)$ is correspondingly restricted; we use the full cell for consistency with the definition of $E(h)$.)

Lemma 2 (Equilibrium-step Lipschitz, corrected)**.** *If step $i \in E(h)$, then $\|\mu_\theta(z_i) - \mu_\theta(z'_i)\|_2 \le L_{eq}^{(r_i)} \|\delta_i\|_F$, where*

$$L_{eq}^{(r_i)} := \underbrace{\big(1 - \varepsilon^{(r_i)}\big) L_Y^{(r_i)} + \varepsilon^{(r_i)} L_{\max}}_{\textit{term (II)}} + \underbrace{L_{\phi,eq}^{(r_i)} M}_{\textit{term (I)}}.$$

Proof. Both z_i and z_i' lie in $\mathcal{Z}^{(r_i)}$, so apply the decomposition (6). Term (I) uses the within-regime gating Lipschitz constant $L_{\phi,\text{eq}}^{(r_i)}$ rather than the global L_ϕ, exactly as before: $\|(\text{I})\|_2 \le L_{\phi,\text{eq}}^{(r_i)} M \|\delta_i\|_F$. Term (II) is a convex combination of per-regime differences under soft gating, so

$$\|(\text{II})\|_2 \;\le\; \sum_r \pi_r(\psi(z_i'))\, L_Y^{(r)} \|\delta_i\|_F \;=\; \Big[\pi_{r_i}(\psi(z_i'))\, L_Y^{(r_i)} + \sum_{r \ne r_i} \pi_r(\psi(z_i'))\, L_Y^{(r)}\Big] \|\delta_i\|_F.$$

Bounding $L_Y^{(r)} \le L_{\max}$ inside the second sum and writing $p := \pi_{r_i}(\psi(z_i'))$, the bracket is at most $p\, L_Y^{(r_i)} + (1-p)\, L_{\max}$, which is non-increasing in p because $L_{\max} \ge L_Y^{(r_i)}$; since $p \ge 1 - \varepsilon^{(r_i)}$ by Definition 1 (as $z_i' \in \mathcal{Z}^{(r_i)}$), the bracket is at most $(1-\varepsilon^{(r_i)})\, L_Y^{(r_i)} + \varepsilon^{(r_i)}\, L_{\max}$. The triangle inequality combines (I) and (II). □

Remark 1 (Relation to the original statement)**.** *The originally stated constant $L_Y^{(r)} + L_{\phi,eq}^{(r)} M$ is recovered exactly when $\varepsilon^{(r)} = 0$, i.e., under hard routing or in the near-hard-gating limit $\tau \to 0$, where the original constant is valid. For soft gating, the original proof step implicitly assumed $\varepsilon^{(r)} = 0$; the corrected constant carries the slack explicitly and is unconditionally valid. When $\varepsilon^{(r)}$ is not estimated, the conservative substitution $\varepsilon^{(r)} = 1$ yields $L_{eq}^{(r)} \le L_{\max} + L_{\phi,eq}^{(r)} M$, computable from the audit constants already reported.*

Remark 2 (Effect on Theorem 1 and its corollary)**.** *Theorem 1 and the trajectory-dependent corollary hold verbatim with $L_{eq}^{(r)}$ redefined as in Lemma 2; no other step of either proof references the gate-concentration property. The audit gains one quantity, $\hat{\varepsilon}^{(r)}$, estimated as the maximum of $1 - \pi_r(\psi(z))$ over held-out states routed to regime r.*

Lemma 3 (Punctuation-step Lipschitz)**.** *If step $i \in P(h)$, then $\|\mu_\theta(z_i) - \mu_\theta(z_i')\|_2 \le B_{trans} \|\delta_i\|_F$.*

Proof. At a punctuation step, the trajectories may straddle a regime boundary and the within-regime restriction does not apply. Lemma 1 gives the global Lipschitz bound directly. □

B.3 Proof of Theorem 1

Step 1: per-step bound. Apply Lemma 2 if $j \in E(h)$ and Lemma 3 if $j \in P(h)$: $\|\mu_\theta(z_j) - \mu_\theta(z_j')\|_2 \le A_j \|\delta_j\|_F$ with $A_j = L_{\text{eq}}^{(r_j)}$ on equilibrium steps and $A_j = B_{\text{trans}}$ on punctuation steps.

Step 2: multi-step accumulation. The history shift gives δ_{j+1} as a lag-shift of δ_j with one new entry equal to $\mu_\theta(z_j) - \mu_\theta(z_j')$. After L consecutive steps (one full lag-window refresh), the accumulated $\|\delta_{j+L}\|_F$ is bounded by $\sqrt{L} \cdot \max_{i=j,\ldots,j+L-1} \|\mu_\theta(z_i) - \mu_\theta(z_i')\|_2$. Iterating across L-step blocks: $\|\delta_h\|_F \le C_Y \cdot \prod_{j=0}^{h-1} A_j \cdot \|\delta_0\|_F$, where C_Y absorbs the $\sqrt{L}$ lag-window-Frobenius conversion.

Step 3: bounding the product. Conservatively bound all equilibrium steps by $L_{\text{eq,max}}$ and all punctuation steps by B_{trans}: $\prod_j A_j \le L_{\text{eq,max}}^{|E(h)|} \cdot B_{\text{trans}}^{|P(h)|}$. With $|E(h)| = (1-\alpha(h))h$ and $|P(h)| = \alpha(h)h$, we obtain $\|\delta_h\|_F \le C_Y \cdot L_{\text{eq,max}}^{(1-\alpha(h))h} B_{\text{trans}}^{\alpha(h)h} \|\delta_0\|_F$.

Step 4: contraction under the rate condition. The bound contracts geometrically iff $L_{\text{eq,max}}^{1-\alpha} B_{\text{trans}}^{\alpha} < 1$, i.e., $(1-\alpha)\log L_{\text{eq,max}} + \alpha \log B_{\text{trans}} < 0$, equivalently $\alpha < (-\log L_{\text{eq,max}})/(\log B_{\text{trans}} - \log L_{\text{eq,max}}) = \alpha^*$, which is assumption (A3). □

Corollary 1 (Trajectory-dependent contraction rate)**.** *Let $\bar{L}_Y(\{r_j\}) := \exp\big(\frac{1}{|E(h)|} \sum_{j \in E(h)} \log L_{eq}^{(r_j)}\big)$ be the equilibrium geometric mean realized along a trajectory's regime sequence. Then the bound of Theorem 1 holds with $L_{eq,\max}$ replaced by $\bar{L}_Y(\{r_j\})$, sharpening when the trajectory spends most of its time in low-Lipschitz regimes.*

Proof. Apply the proof of Theorem 1 but bound each equilibrium-step factor by its own $L_{\text{eq}}^{(r_j)}$ rather than $L_{\text{eq,max}}$. The product of equilibrium-step factors becomes $\prod_{j \in E(h)} L_{\text{eq}}^{(r_j)} = \bar{L}_Y(\{r_j\})^{|E(h)|}$ by definition of the geometric mean. □

B.4 Stochastic error propagation

Theorem 2 (Regime-aware stochastic error propagation)**.** *Under (A1)–(A3) of Theorem 1 and additionally **(A4)** covariate-forecast innovations $\{\Delta X_k\}$ are mutually independent with $\mathbb{E}\|\Delta X_k\|_F^2 \le \tau_X^2$, the recursive predictive variance is bounded uniformly in horizon:*

$$\mathbb{E}\|\mathcal{H}_h^Y - \mathcal{H}_h'^Y\|_F^2 \le C_Y^2 \rho^{2h} \|\mathcal{H}_0^Y - \mathcal{H}_0'^Y\|_F^2 + \frac{C_Y^2 L_{X,\max}^2 \tau_X^2}{1-\rho^2}, \tag{7}$$

where $\rho < 1$ is from Theorem 1.

Proof. The recursion under stochastic covariate-forecast innovations is non-homogeneous: $\delta_{j+1}^Y = A_j \delta_j^Y + B_j \Delta X_j$, where A_j is the step-j Y-channel Jacobian (regime-dependent: $L_{\text{eq}}^{(r_j)}$ in equilibrium, bounded by B_{trans} at punctuation) and B_j is the step-j X-channel Jacobian (bounded by $L_X^{(r_j)} \le L_{X,\max}$). The closed-form solution is $\delta_h^Y = \big(\prod_{i=0}^{h-1} A_i\big)\delta_0^Y + \sum_{k=0}^{h-1} \big(\prod_{i=k+1}^{h-1} A_i\big) B_k \Delta X_k$. Take expectations of the squared Frobenius norm. By (A4), the $\{\Delta X_k\}$ are mutually independent, so cross-product terms vanish in the second moment:

$$\mathbb{E}\|\delta_h^Y\|_F^2 \le \big\|\prod_{i=0}^{h-1} A_i\big\|_2^2 \|\delta_0\|_F^2 + \sum_{k=0}^{h-1} \big\|\prod_{i=k+1}^{h-1} A_i\big\|_2^2 \|B_k\|_2^2 \mathbb{E}\|\Delta X_k\|_F^2.$$

The first term is bounded by Theorem 1: $\|\prod_i A_i\|_2 \le C_Y \rho^h$, giving $C_Y^2 \rho^{2h} \|\delta_0\|_F^2$. For the second term, apply Theorem 1 to the sub-trajectory from $k+1$ to $h-1$: $\|\prod_{i=k+1}^{h-1} A_i\|_2 \le C_Y \rho^{h-k-1}$. Combined with $\|B_k\|_2 \le L_{X,\max}$ and $\mathbb{E}\|\Delta X_k\|_F^2 \le \tau_X^2$: $\sum_{k=0}^{h-1} C_Y^2 \rho^{2(h-k-1)} L_{X,\max}^2 \tau_X^2 = C_Y^2 L_{X,\max}^2 \tau_X^2 \sum_{j=0}^{h-1} \rho^{2j} \le C_Y^2 L_{X,\max}^2 \tau_X^2/(1-\rho^2)$ where the geometric sum converges because $\rho < 1$ under (A3). □

The first term is the contraction of initial-condition error; the second is the bounded steady-state variance from covariate-forecast noise accumulation. The bound diverges as $\alpha(h) \to \alpha^*$ ($\rho \to 1$): at the rate-condition limit, variance accumulates without bound. Setting $\alpha = 0$ recovers the additive-baseline NAVAR error-propagation result with $\rho = L_Y$.

C Theorem 1: intuition, anchoring, and companion results

The intuition is dwell-time accounting. Equilibrium steps contract by $L_{\text{eq}}^{(r)}$, punctuation steps expand by up to B_{trans}. Geometric contraction holds whenever the (log-)expansion contributed by punctuations is dominated by the (log-)contraction contributed by equilibrium steps, which gives the threshold α^* on punctuation frequency. The result is a switching-system analog of Hespanha–Morse average-dwell-time stability [19], specialized to neural-network-based switching dynamics with architecture-derived constants. To anchor the constants, consider two settings. On democracy indicators from the V-Dem under an unconstrained training configuration, the audit yields $\hat{L}_{\text{eq,max}} \approx 0.29$ and $\hat{B}_{\text{trans}} \approx 0.30$, both below 1; under the conservative leakage substitution $\varepsilon^{(r)} = 1$, the corrected constant is at most 0.362 across all 45 runs of the 9 soft-gated audited configurations (a shift of at most 0.006, since the per-regime endogenous constants are tightly clustered), so the contraction conclusion on this panel does not depend on the measured leakage. Each step is therefore contractive, and (A3) is vacuously satisfied (the trivial-contraction regime). The audited contraction rate is $\hat{\rho} \le 0.36$ across all 9 audited cells (§6.3). In contrast, on the synthetic stress test (§7), the same audit at $\lambda_M = 0$ gives $\hat{L}_{\text{eq,max}} \approx 2.38$, so the contractivity condition (A1) fails. These two cases illustrate the two regimes of the bound: the empirical panel lies within the trivial-contraction case covered by the edge-case clause, while the synthetic experiment activates the non-trivial rate condition. A trajectory-dependent sharpening (Corollary in Appendix B.3) replaces $L_{\text{eq,max}}$ with the geometric mean realized along a trajectory, tightening the bound when the trajectory spends most of its time in low-Lipschitz regimes. A companion stochastic error-propagation theorem (Appendix B.4) bounds the recursive predictive variance uniformly in horizon under independent covariate-forecast innovations.

D Synthetic data and full Pareto frontier

D.1 DGP parameters

The synthetic panel of §7 is generated as follows. For each unit $u \in \{1, \ldots, 30\}$ and time $t \in \{1, \ldots, 50\}$: per-unit intercept $\beta_u \sim \mathcal{N}(0, 0.5^2)$, sampled once per unit; initial values $Y_{u,0}, X_{u,0} \sim \mathcal{N}(0, 0.3^2)$; covariate evolution $X_{u,t} = 0.9X_{u,t-1} + \nu_{u,t}$, $\nu_{u,t} \sim \mathcal{N}(0, 0.5^2)$; target evolution: regime r_t from $Y_{u,t-1}$ via thresholds $c_1 = -0.6, c_2 = 0.6$; then $Y_{u,t} = \beta_u + \rho_Y^{(r_t)} Y_{u,t-1} + \beta_X^{(r_t)} X_{u,t-1} + \epsilon_{u,t}$ with $(\rho_Y^{(1)}, \rho_Y^{(2)}, \rho_Y^{(3)}) = (0.5, 0.8, 0.7)$, $(\beta_X^{(1)}, \beta_X^{(2)}, \beta_X^{(3)}) = (0.05, 0.20, 0.40)$, and $\epsilon_{u,t} \sim \mathcal{N}(0, 0.3^2)$. The data-generation seed is fixed (`np.random.default_rng(42)`); only model-initialization seeds vary across the 5 replications per cell. On the generated training data, regime occupancy is $28.2\%/37.5\%/34.2\%$ (low/mid/high) and the empirical regime-transition rate is $\alpha_{\text{train}} = 0.217$ ($\alpha = 0.204$ on the full panel; the model audit reports $\alpha_{\text{test}} = 0.156$).

D.2 Training hyperparameters

All 45 sweep runs use Adam with learning rate 3×10^{-3}, weight decay 10^{-4}, batch size 128, and 300 training epochs. Per-edge networks are two-layer ReLU MLPs with hidden width 32. The gating softmax temperature is initialized at $\tau_{\text{init}} = 0.5$. Gating-temperature regularization $\lambda_\phi = 0.005$. Regime centers are fixed at training-data $1/6, 1/2, 5/6$ quantiles of $\psi(z) = z[L][0]$.

D.3 Predictive contribution at the unconstrained optimum

Table 3: Test MSE at the soft-penalty unconstrained ($\lambda_M = 0$) training configuration, 5-seed mean. Decomposition isolates contributions from nonlinear X channel (M1.5 vs. M1) and regime structure (M2 vs. M1.5).

Predictor	Test MSE
RW (random walk)	0.1893
AR(1) + intercept	0.1895
M1 (Y-only PANAR)	0.1895
M1.5 (additive PANAR-X, no regimes)	0.1497
M2 (regime-aware PANAR-X), 5-seed mean	0.1410
Gain decomposition:	
Nonlinear X channel (M1.5 vs. M1)	$+0.0398$ ($+20.98\%$)
Regime structure (M2 vs. M1.5), 5-seed mean	$+0.0087$ ($+5.82\%$)
seed-level 95% CI on gap	$[+0.0065, +0.0109]$

D.4 Full Pareto frontier across 45 training runs

The frontier is monotonic across all 9 cells: stronger enforcement (higher λ_M, transition from soft to per-epoch to per-batch projection) reduces both $L_{\text{eq,max}}$ and the M2-vs-M1.5 gap together. Three regimes are visible: unenforced cells ($\lambda_M = 0$) attain the strongest predictive contribution but fail the contractivity premise in $5/5$ seeds; intermediate enforcement ($\lambda_M = 0.05$, soft penalty) sits on the boundary with a CI-significant predictive gap and partial satisfaction of the bound; full enforcement (per-epoch or per-batch projection at $\lambda_M = 0.10$) attains $L_{\text{eq,max}} < 1$ in $5/5$ seeds, with the predictive contribution removed.

E V-Dem data and preprocessing

E.1 Variable selection

The target Y_t = `v2xel_frefair` (V-Dem "Clean Elections" index) is a high-level democracy indicator constructed by V-Dem from expert-coded sub-indicators via Bayesian item-response models [8]. Range $[0, 1]$; standard deviation ~ 0.30 pooled, ~ 0.06 in first differences.

Table 4: Synthetic stress test: contrast at unconstrained ($\lambda_M = 0$, soft penalty) vs. most-enforced ($\lambda_M = 0.10$, projection-every-batch) training, alongside V-Dem indicators at the same configurations. "Gap" is M2 minus pooled M1.5 in test MSE units (positive favors M2); positive gap with broken bound is the failure mode the certificate identifies. Full 45-run frontier in Table 5.

	Synthetic		V-Dem ind.	
	$\lambda_M = 0$	$\lambda_M = 0.10$	$\lambda_M = 0$	$\lambda_M = 0.10$
M (regime disagreement)	5.94	0.47	0.87	0.05
$L_{\text{eq,max}}$	2.38	0.91	0.29	0.13
$L_{\text{eq,max}} < 1$ in 5/5 seeds	0/5	5/5	5/5	5/5
M2 vs. M1.5 gap	$+0.0087$	-0.0112	$+0.00009$	$+0.00007$

Table 5: Full Pareto frontier across enforcement modes and disagreement-penalty strengths ($n = 5$ seeds per cell, 5-seed means). Gap = M1.5 MSE − M2 MSE; positive means M2 better. $L_{\phi,\text{eq}} \cdot M$ is the gating-correction contribution to the contractivity bound from Theorem 1.

Mode	λ_M	$L_{\text{eq,max}}$	M	$L_{\phi,\text{eq}} \cdot M$	gap (95% t-CI)	$L_{\text{eq,max}}<1$
soft penalty	0.00	2.38	5.94	1.62	$+0.0087$ $[+0.0065, +0.0109]$	0/5
	0.05	1.02	0.70	0.19	$+0.0046$ $[+0.0015, +0.0078]$	2/5
	0.10	0.96	0.42	0.11	$+0.0028$ $[+0.0003, +0.0054]$	3/5
proj. per epoch	0.00	2.56	6.79	1.93	$+0.0071$ $[+0.0067, +0.0076]$	0/5
	0.05	1.02	0.71	0.19	-0.0049 $[-0.0076, -0.0022]$	2/5
	0.10	0.92	0.44	0.12	-0.0073 $[-0.0097, -0.0048]$	**5/5**
proj. per batch	0.00	2.59	6.95	1.98	$+0.0066$ $[+0.0058, +0.0075]$	0/5
	0.05	0.99	0.74	0.20	-0.0089 $[-0.0122, -0.0055]$	2/5
	0.10	0.91	0.47	0.12	-0.0112 $[-0.0142, -0.0083]$	**5/5**

The four covariates were selected from the V-Dem indicator set under two criteria: (i) shock-like temporal variation (the fraction of country-years with $|\Delta X_t| > 0.05$ should be of similar magnitude to that of ΔY_t); and (ii) lagged correlation with ΔY above noise floor. `v2csantimv` (anti-system movement strength), `v2jureform` (judicial reforms in the past year), `v2psorgs` (party organization), and `v2mecorrpt` (media corruption / capture). These are V-Dem-internal expert-coded indicators with documented substantive interpretations and are independent of `v2xel_frefair` in the V-Dem index aggregation.

E.2 Sample construction and split

We use the V-Dem v15 country-year panel restricted to countries with complete observation series across all selected variables in the 1970–2025 window. After this restriction, the panel is balanced at 151 countries × 56 years.

The lag window is $L = 5$. For each country, we hold out the last 10 years for testing; earlier years are training. Per-country split ensures every country contributes both training and test observations, which matters because country-level effects are absorbed into per-unit intercepts during training. Final sample sizes after lag-window construction: 6,191 training observations, 1,510 test observations.

E.3 Regime boundaries

Regime centers c_r and regime boundaries b_1, b_2 are computed from training-set Y_{t-1} tertiles. Empirical values: $(c_1, c_2, c_3) = (0.003, 0.41, 0.911)$, $(b_1, b_2) = (0.168, 0.674)$. The regime function is fixed by these training-derived constants and is not retrained on test data.

X normalization in the diagnostic. The linear coefficients plotted in Figure 1A–B are fit on *unstandardized* covariates, so coefficient magnitudes are directly comparable across regimes. The companion diagnostic script that prints sign-disagreement counts standardizes X using training-set mean and standard deviation; the sign-reversal pattern is identical in both, but coefficient magnitudes differ (party-organization at lag $t-4$: unstandardized $-0.019/+0.018/+0.016$ vs. standardized $-0.028/+0.027/+0.024$). Either presentation supports the four-of-four sign-reversal claim.

E.4 Target variable: ΔY vs. Y levels

We forecast $\Delta Y_t = Y_t - Y_{t-1}$ rather than Y_t directly. Pooled persistence in Y is $\rho \approx 0.98$, meaning the random-walk baseline ($\hat{Y}_t = Y_{t-1}$) achieves test MSE of order 10^{-3} on Y-levels, against unconditional variance ~ 0.09. Predicting ΔY removes this trivial persistence and re-anchors the random-walk baseline to "predict zero change," which on V-Dem test data has MSE 0.00340. Under this baseline, beating RW skill becomes a meaningful test of the architecture's contribution beyond persistence.

F V-Dem demonstration: extended material

This appendix holds material moved from Section 6 for space; nothing here changes a conclusion of the main text.

F.1 Why V-Dem is a hard test

The country-year panel of democracy indicators derived from V-Dem [8] (Appendix E) exhibits all three difficulties that motivate this work. *Multicollinearity*: indicators are expert-coded measurements of latent democratic constructs [8], with pairwise correlations exceeding 0.7 for many natural pairs and 80% of variance in the first two principal components. *Latent unit heterogeneity*: the 151 countries exhibit dynamics that grouping models identify as 5 to 9 distinct latent regimes for the income–democracy relationship alone [2]. *Regime-conditional sign changes*: predicting year-over-year change $\Delta Y_t = Y_t - Y_{t-1}$ exposes sign-changing relationships that pooled additive models cancel, while removing the trivial pooled persistence ($\rho \approx 0.98$) that would otherwise drown out any predictive contribution beyond random walk.

F.2 Baselines and the difficulty of forecasting a V-Dem democracy indicator

Baselines on the ΔY test partition: random walk ($\widehat{\Delta Y_t} = 0$) achieves test MSE 0.00340; pooled linear M1.5 achieves 0.00347 (*worse* than RW by -2.25%, because the high-multicollinearity, sign-canceling covariate structure inflates pooled-coefficient variance without compensating bias reduction); stratified linear M1.5 achieves 0.00337 ($+0.84\%$ over RW, not significant under country-clustered inference, $p = 0.830$); and the pooled NN PANAR-X (M1.5) achieves a per-run test MSE of 0.00350 (mean over 20 seeds) and 0.00339 as a 20-seed prediction ensemble. (The submitted version described a value of 0.00352 as a 5-seed mean; it was a single seed-0 run, an error we identified during reanalysis and correct here.) Under panel-appropriate inference, neither nonlinear flexibility nor regime stratification yields a statistically significant gain over the random walk: one-step forecasting of measurement-smoothed democracy indices is a setting in which no tested model, pooled or stratified, linear or neural, separates from persistence (Table 1).

Table 1 in the main text summarizes these baselines. No tested model significantly outperforms the random walk under country-clustered inference (Table 2; the strongest ensemble, stratified NN, has clustered $p = 0.51$ on the primary seed set and $p = 0.54$ on a pre-committed fresh seed set), and the stratified-vs.-pooled linear gain of $+3.02\%$ is not significant under the pre-specified panel test (block bootstrap $p = 0.273$)..

F.3 Per-run versus ensemble evaluation: full account

The submitted version reported a paired DM test of $+2.41$ ($p = 0.016$) for M2 against pooled NN M1.5. During reanalysis we identified a design flaw in that comparison: it paired a 5-seed M2 prediction ensemble against a single-seed pooled baseline, so part of the measured gap was the variance reduction of ensembling rather than the architecture. The two evaluation conventions give different answers, and we report both. At the per-run level the architecture is reproducibly better: on the primary 20-seed set, M2's per-seed test MSE distribution dominates the pooled model's (unpaired $t = +6.73$, $p < 10^{-5}$), the result replicates on a pre-committed fresh seed set ($t = +8.67$, $p < 10^{-5}$), and 39 of 40 same-seed comparisons favor M2 (descriptive, since seeds do not pair across architectures). At the ensemble level the advantage disappears: under the seed-fair 20-vs-20 comparison (Table 2, row 5), clustered $p = 0.807$. The pooled model benefits more from ensembling because of its larger seed variance, which prediction averaging converts into accuracy. We therefore

Table 6: Stratified-linear sensitivity on V-Dem (standardized covariate units; the raw-scale canonical values shown in Figure 1 are $-0.019/+0.018/+0.016$, Appendix E). Regime-specific coefficients straddle zero in every configuration; the gain over pooled linear is configuration-dependent and not significant under the pre-specified panel test (Table 2, row 4).

Config	R	Gain vs. pooled	Party-org $t-4$ per regime (std.)	Straddles 0
Tertiles (canonical)	3	$+3.02\%$	$-.028/+.027/+.024$	yes
Median	2	-0.19%	$-.021/+.040$	yes
Cuts $25/75$	3	$+3.84\%$	$-.027/+.030/-.030$	yes
Cuts $20/80$	3	$+1.30\%$	$-.020/+.009/-.005$	yes
Quartiles	4	$+2.41\%$	$-.027/-.005/+.063/-.030$	yes
Quintiles	5	-0.54%	(5 regimes; straddles 0)	yes
$\psi = Y_{t-2}$, tertiles	3	$+0.33\%$	$-.030/+.035/+.014$	yes
$\psi =$ 3-yr MA, tertiles	3	$+0.07\%$	$-.025/+.021/+.020$	yes

withdraw the forecasting-improvement claim: the architecture produces better and lower-variance individual fits, and ensembling, the relevant deployment convention, equalizes them.

Per-regime behavior. Per-regime fair tests localize no gain: after Holm correction across regimes, no comparison of M2 against the stratified linear model or the random walk is significant (Table 2, caption). Figure 1C–E therefore illustrates per-run behavior under the submitted seed-asymmetric comparison rather than an ensemble-level gain: the cumulative paired squared-error advantage over the single-seed pooled baseline is concentrated in regimes 0 and 2, where individual M2 fits are strongest, but this concentration does not survive the seed-fair ensemble comparison.

F.4 Reading Figure 1

The submitted caption read as follows. Diagnostic and recovery of regime-conditional signal on V-Dem ΔY. **Top (cancellation diagnostic, linear models): (A)** Linear coefficients at lag $t-4$ across pooled (column 1) and the three regime-stratified models (columns 2–4) for each of the four selected covariates. Sign reversals across regimes drive near-zero pooled coefficients. **(B)** Detail on party organization at lag $t-4$, the cleanest sign-reversal exemplar. **Bottom (per-run comparison: M2 seed-ensemble vs. the single-seed pooled NN M1.5 of the submitted test; the seed-fair ensemble comparison is null, Table 2): (C–E)** Cumulative paired squared-error advantage across the test partition, sorted by |paired diff| descending. Final value at $x = 1$ equals $N \cdot \overline{\Delta\text{MSE}}$ (the DM test statistic's numerator). M2 advantage is concentrated in the heavy-hitters in regimes 0 and 2; in regime 1, the heavy-hitters are won by pooled before M2 catches up to a small net positive.

F.5 Why the certificate holds on V-Dem without projection

The decomposition of Theorem 1 clarifies why the audit holds without explicit projection on V-Dem indicators: the regime-disagreement $\hat{M} \approx 0.87$ at $\lambda_M = 0$ and the per-regime $\hat{L}_Y^{(r)}$ are small enough that the gating-correction product $\hat{L}_{\phi,\text{eq}} \cdot \hat{M}$ does not push $\hat{L}_{\text{eq,max}}$ above 1. This is in contrast to the synthetic stress test of the next section, where M reaches ~ 6 and the same architecture violates the contraction condition unless a disagreement penalty is applied. The substantive content of the result is the combination: the diagnostic explains why pooled additive predictors cancel regime-conditional signal, the corrected evaluation shows that on this panel no tested model converts that structure into a significant forecasting gain over persistence, and the contraction analysis ensures that recursive forecasts from the trained architecture remain stable under deployment.

G Sensitivity to the regime construction: configuration-level results

This appendix reports the configuration-level results summarized in Section 6.4. All runs follow the pre-registered sensitivity design in the archived package (DOI 10.5281/zenodo.23717582); the diagnostic grid has 24 variants (three state summaries $\times$ $R \in \{2, 3, 4, 5\}$, plus ± 5 percentage-point cut shifts at R=3), the stratified linear model is refit under the eight configurations of Table 6, and the jackknife draws 200 half-country subsamples.

What is robust. Two findings hold in every variant. First, the cancellation mechanism itself: a pooled coefficient that collapses toward zero while the regime-specific coefficients straddle zero appears in all 24 diagnostic variants, including the learned PC1 summary, and in all eight configurations of Table 6; in the canonical construction, all four covariates straddle zero at one or more lag. Second, the low-democracy stratum's negative party-organization coefficient at lag $t-4$: negative in $24/24$ variants, with $|\hat{\beta}|$ exceeding twice its standard error in $13/24$, and sign-stable in 94.5% of 200 half-country jackknife subsamples.

What is configuration-dependent. The symmetric sign-reversal pattern of the submitted version is not robust, and we withdraw it as a finding. The identity of the positively signed stratum moves with the cut points (the top stratum flips negative at the $25/75$ and $20/80$ cuts); a significance-filtered outer reversal appears in $0/24$ variants; and the complete pattern is retained in only 1.5% of the jackknife subsamples. It is also not temporally stable: the low-stratum association is negative in four of five historical training windows (coefficients -0.024 to -0.015) but attenuates to approximately zero ($+0.001$) in the most recent half of the training window. The stratification gain is likewise configuration-dependent: it ranges from -0.5% to $+3.8\%$, is meaningful only for $R \in \{3, 4\}$ on $\psi = Y_{t-1}$, is approximately zero under the alternative state summaries, and is not significant under the panel test ($p = 0.273$).

H Realized-volatility panel: construction, screen, and audit table

This appendix gives the construction and inference details for the pre-registered realized-volatility experiment of Section 8 and the per-seed certificate audit (Table 7). The screening design, the confirmatory design with its pre-committed claim levels, and all run artifacts are in the archived reproducibility package (DOI 10.5281/zenodo.23717582).

Construction. The panel is daily realized volatility for eight international equity indices [21] (2,615 trading days; 2,611 usable after lagging). With $y_{a,t}$ the log realized volatility of asset a, the target is the one-day-ahead change $\Delta y_{a,t}$; the design per observation comprises four own lags of y, four lags of a leave-one-out cross-asset mean of y (a spillover proxy), and per-asset intercepts. Regimes are training-set tertiles of $\psi = y_{t-1}$ pooled across assets. The split is chronological with an identical boundary for all assets: the final 523 days form the test partition (N=4,184 observations), the preceding days the training partition.

Inference. With eight units, asset-clustered inference is not defensible, so the inference axis follows time: a moving-block bootstrap over days (block length 25 trading days, B=9,999, studentized, two-sided), each resampled block carrying all eight assets' values intact. Per-asset HAC Diebold–Mariano tests (Newey–West, 25 lags) are reported as a cross-check. Holm correction is applied across the three per-regime tests within each comparison family. Screen outcome: pooled linear regression beats the random walk by 28.2% (MSE 0.0746 vs. 0.1039; bootstrap p at the 10^{-4} resolution floor); block-bootstrap Wald test of slope equality across regimes $W = 59.9$, $p = 5.4 \times 10^{-7}$; 0 of 8 own and spillover lag slopes change sign across regimes.

M2 training. M2 is ported unchanged from the V-Dem code (unconstrained soft gating, R=3, tempered softmax over $\psi = y_{t-1}$, centers at the training tertile midpoints, soft spectral target $\tilde{L}_Y = 0.55$, 300 epochs, batch 128). Hyperparameters were selected on a pre-registered 2×2 grid, learning rate $\{10^{-3}, 3 \times 10^{-4}\}$ $\times$ hidden width $\{16, 32\}$, by mean pooled validation MSE over 5 seeds on the final 20% of training days, with the test partition untouched; the selected cell (learning rate 3×10^{-4}, width 16) was then retrained on the full training partition with 5 seeds. Test MSEs: random walk 0.1039, pooled linear 0.0746, interaction-augmented pooled linear 0.0746, stratified linear 0.0754, M2 ensemble 0.0757.

I Extended discussion, limitations, and conclusion

This appendix holds the full-length versions of Sections 9 and 10, condensed in the main text for space; the conclusions are identical.

I.1 Introductory framing

The Introduction's scope statement and the two-roles framing of the certificate read, in full, as follows.

Table 7: Certificate audit on the realized-volatility panel, per seed. Every per-regime network meets the soft spectral target, yet the gating term leaves $4/5$ seeds without a certificate; the dwell-time condition (A3) binds non-trivially on seed 4. "Corr." is $\hat{L}_{\text{eq,max}}$ under the conservative gate-leakage substitution $\varepsilon^{(r)} = 1$. For reference, the V-Dem unconstrained cell audits at $\hat{L}_{\text{eq,max}} = 0.292$ in the same pipeline.

Seed	$\max_r \hat{L}_Y^{(r)}$	$\hat{L}_{\phi,\text{eq}}\hat{M}$	$\hat{L}_{\text{eq,max}}$	Corr.	$\hat{B}_{\text{trans}}$	$\hat{\alpha}/\hat{\alpha}^*$	Certified
0	0.570	0.499	1.067	1.068	1.085	–	no
1	0.578	0.676	1.252	1.254	1.277	–	no
2	0.577	0.534	1.110	1.110	1.128	–	no
3	0.569	0.560	1.120	1.130	1.148	–	no
4	0.570	0.426	0.984	0.996	1.010	$0.271/0.619$	yes ($\hat{\rho} = 0.991$)

Scope. The forecasting tasks we study are intrinsically difficult: simple baselines such as random walks perform strongly, and absolute predictive gains are modest. Accordingly, our objective is not to maximize accuracy, but to show that regime-dependent structure can be modeled in a principled way without sacrificing stability.

The contraction theorem serves two roles. In synthetic settings with strong regime contrasts, it characterizes when expressive nonlinear forecasters would otherwise diverge under recursive rollout; the experiment shows the constraint is non-vacuous and identifies the training-time controls needed to restore contraction. In real-world panel data, such as V-Dem data [8], where heterogeneity, multicollinearity, and weak signals dominate, the constraint typically holds under standard training configurations; the audit then confirms safe recursive deployment.

I.2 Discussion

The two-panel trade-off. Reviewers asked why enforcing stability removes the regime-aware gain in the synthetic test, and what the stability–accuracy trade-off looks like on real data. The two real panels give a two-sided answer. On V-Dem the constraint has no practical bite, and the sweep makes this precise: across all ten enforcement cells, from penalty-free hard routing through per-batch projection at $\lambda_M = 0.10$, the per-run M2 test MSE is 0.00343 to 0.00345 (seed sd 0.00003 to 0.00005), while $\hat{L}_{\text{eq,max}}$ falls from 0.29 to 0.13. Stability is free because there is no instability to constrain; the audited constants sit far below the penalty target of 0.55, the penalties never activate, and the certificate reduces to post hoc confirmation that an already stable model is contractive. On the volatility panel (Section 8) the trade-off is real: every per-regime network meets its spectral target, the gating term alone leaves $4/5$ seeds uncertified, and the regime-aware model is significantly worse out of sample than a pooled linear model that captures the panel's entire headroom. Taken together: when the dynamics are already stable, the certificate imposes no cost; when instability is present, the regime-dependent structure the architecture is built to exploit is the same structure the certificate flags, and on this panel it was not exploitable out of sample by any method tested. That V-Dem sits on the stable side is consistent with the possibility that expert-coded, measurement-smoothed indices suppress the fast regime-dependent dynamics on which both the gain and the instability depend; the present design does not test that interpretation.

Audit first. The practical rule we draw is to audit before deciding how much stability control to buy. Every constant in Theorem 1 is computable post hoc from a trained model by gradient probing and a single forward pass (Appendix J), in seconds, without retraining or a GPU. The audit tells a practitioner which of the two situations a panel is in. Constants far below one, as on V-Dem ($\hat{L}_{\text{eq,max}} = 0.13$ to 0.29), mean stability is free and the disagreement penalty can be left off. Constants near or above one, as on the volatility panel (1.07 to 1.25 in $4/5$ seeds), mean stability has a price that has to be paid, through λ_M or stronger projection, before recursive deployment, and the decomposition in (4) identifies which term is responsible. The determination does not have to be discovered in production.

Appropriate use for democracy forecasting. The corrected results also sharpen what this class of model should and should not be used for on democracy data. At a one-year horizon on measurement-smoothed indices, no tested method, including ours, outperforms persistence under country-clustered inference (Table 2). The appropriate uses are therefore diagnostic: detecting regime-dependent

heterogeneity that pooled models mask (Sections 6.1 and 6.4), and auditing the rollout stability of a forecasting model before it is deployed (Section 6.3). The misuse the boundary result warns against is treating one-step forecasts of such indices as operational predictions of democratic change for particular countries. A model that cannot be distinguished from a random walk should not inform such decisions, and the finding that the best available models tie persistence is itself the strongest argument against overconfident operational use of democracy forecasts at this horizon.

I.3 Limitations

Limitations. Four limitations bound the contribution. (i) V-Dem forecasting performance is statistically indistinguishable from persistence and from stratified linear regression under seed-fair, country-clustered inference (Table 2); the V-Dem result is a mechanism demonstration and a boundary characterization, not a predictive benchmark. (ii) The robust empirical result is narrower than the symmetric sign-reversal claim of the submitted version: what survives the pre-registered sensitivity design is the cancellation mechanism and the low-democracy stratum's negative party-organization coefficient (Section 6.4); the symmetric reversal is withdrawn. (iii) The stability theorem is an architecture-specific, auditable specialization of average-dwell-time analysis [19]; its contribution is that the constants entering the bound are computable on the trained model and individually controllable in training (Section 4), not a new stability technique. (iv) The realized-volatility experiment demonstrates that the certificate is non-vacuous on real data, not that the regime-aware architecture improves forecasting: M2 is significantly worse than pooled linear regression on that panel (Section 8). Beyond these four, V-Dem remains a single target on a single social-science panel. Regime structure is exogenous to the model: ψ is a fixed Lipschitz functional of state with centers at training-data tertile midpoints, a deliberate choice that keeps the contraction analysis tractable, and the sensitivity study shows that everything finer than the cancellation mechanism depends on it. Per-regime parameters scale linearly with regime count; with $R = 3$ this is not a practical issue on the V-Dem panel, but the architecture is more parameter-hungry than pooled M1.5. The penalty-bridge correspondence is diagnostic, not constructive: practitioners must train, audit, and adjust, with the synthetic Pareto frontier providing the qualitative shape of the trade-off and the audit-first rule of Section 9 the order of operations. Function-valued interpretability of the trained per-regime networks is inherited from the additive structure but not re-demonstrated empirically here; ICE-curve-style inspection per regime is left to future work.

I.4 Conclusion

Conclusion. We presented PANAR-X, a regime-aware additive forecasting architecture for heterogeneous panel time series with a verifiable contraction certificate. The contraction theorem (Theorem 1) decomposes the global Lipschitz constant into per-regime, regime-disagreement, and dwell-time components, each empirically auditable and individually controllable during training, and identifies a failure mode of standard spectral normalization on soft-gated mixture architectures. Empirical demonstration on V-Dem shows that pooled additive models cancel regime-conditional signal, that the contraction certificate holds in $5/5$ seeds across all 9 audited configurations without explicit projection, and that under seed-fair, panel-appropriate evaluation no tested model significantly outperforms a random walk. A synthetic stress test and a pre-registered realized-volatility panel confirm the certificate is non-vacuous when regime contrasts are stronger: on the volatility panel, $4/5$ spectrally compliant seeds fail certification through the gating term, while the regime-aware architecture does not improve on a pooled linear model there either. We claim no predictive gains on these panels. We claim that regime-dependent dynamics can be modeled under formal stability constraints, that the architecture provides the diagnostic constants to detect, post hoc and without retraining, whether stability binds for a given dataset, and that the diagnostic and the certificate together address a methodological gap between pooled additive models that cancel sign-changing regime-conditional structure and unconstrained nonlinear models whose recursive stability is unverified.

J Audit protocol

We describe two complementary estimators of the Theorem 1 constants on trained models: a gradient-probe sample estimator (§J.1) used as the headline audit number throughout the empirical sections, and a certified spectral-norm upper bound (§J.2) that bounds the gradient-probe estimate from above.

Both are deterministic given trained model weights. The certified bound is conservative; the gradient-probe estimate is a sample maximum and may underestimate the true Lipschitz if the empirical batch does not realize the worst-case Jacobian.

J.1 Gradient-probe sample estimator

Per-regime endogenous Lipschitz $\hat{L}_Y^{(r)}$. For each regime r and a held-out batch of 200 samples, compute the per-sample Y-channel Jacobian $J^{Y,(r)}(z) := \partial \mu_\theta^{(r)}(z)/\partial \mathcal{H}^Y$ via automatic differentiation, take the Frobenius norm $\|J^{Y,(r)}(z)\|_F$, and report the maximum over the batch: $\hat{L}_Y^{(r)} = \max_z \|J^{Y,(r)}(z)\|_F$.

Per-regime covariate Lipschitz $\hat{L}_X^{(r)}$. Analogously, with $J^{X,(r)}(z) := \partial \mu_\theta^{(r)}(z)/\partial \mathcal{H}^X$.

Global gating Lipschitz $\hat{L}_\phi$. Sweep ψ across the full empirical range $[\psi_{\min}, \psi_{\max}]$ on a 500-point grid, evaluate $\pi(\psi)$ at each point, and report $\hat{L}_\phi = \max_i \|\pi(\psi_{i+1}) - \pi(\psi_i)\|_1 / |\psi_{i+1} - \psi_i|$.

Within-regime gating Lipschitz $\hat{L}_{\phi,\mathbf{eq}}^{(r)}$. For each regime r, restrict to test observations with regime indicator $= r$, sweep ψ on a 50-point grid covering the regime's empirical 25th–75th percentile range, and apply the same finite-difference estimator as $\hat{L}_\phi$.

Regime disagreement $\hat{M}$. For all pairs $r \neq r'$ and all held-out test observations, compute $|\mu_\theta^{(r)}(z) - \mu_\theta^{(r')}(z)|$, take the maximum over test points, then maximum over regime pairs.

Derived constants. $\hat{L}_{\text{eq,max}} = \max_r[\hat{L}_Y^{(r)} + \hat{L}_{\phi,\text{eq}}^{(r)}\hat{M}]$; $\hat{B}_{\text{trans}} = \max_r \hat{L}_Y^{(r)} + \hat{L}_\phi \hat{M}$; $\hat{\alpha}^* = (-\log \hat{L}_{\text{eq,max}})/(\log \hat{B}_{\text{trans}} - \log \hat{L}_{\text{eq,max}})$ when $\hat{L}_{\text{eq,max}} < 1 < \hat{B}_{\text{trans}}$. When $\hat{B}_{\text{trans}} \leq 1$, the rate condition (A3) is vacuously satisfied and $\hat{\alpha}^*$ is reported as ∞ (no constraint on the empirical punctuation rate).

Empirical punctuation rate $\hat{\alpha}$. For each unit u in the test set, compute the predicted regime sequence $\{\hat{r}_j^{(u)}\}$ via $\arg\max_r \pi_r(\psi(z))$, count transitions, and average the per-unit transition rate.

Empirical contraction rate $\hat{\rho}$. $\hat{\rho} := \hat{L}_{\text{eq,max}}^{\,1-\hat{\alpha}} \hat{B}_{\text{trans}}^{\,\hat{\alpha}}$. We treat the trained model as certifiably contractive when $\hat{L}_{\text{eq,max}} < 1$ and either $\hat{B}_{\text{trans}} \leq 1$ (in which case $\hat{\rho} \leq \hat{L}_{\text{eq,max}} < 1$ for all $\hat{\alpha}$) or $\hat{\alpha} < \hat{\alpha}^*$ (in which case $\hat{\rho} < 1$ by Theorem 1).

J.2 Certified spectral-norm upper bound

The per-edge functions in PANAR-X are two-layer ReLU MLPs with weight matrices $W_1^{(r,e)} \in \mathbb{R}^{h \times d}$ (input-to-hidden) and $W_2^{(r,e)} \in \mathbb{R}^{1 \times h}$ (hidden-to-output) for each regime r and edge e. Since ReLU is 1-Lipschitz, the Lipschitz constant of each per-edge function $f^{(r,e)} : \mathbb{R}^d \to \mathbb{R}$ admits the analytical upper bound $\text{Lip}(f^{(r,e)}) \leq \sigma(W_1^{(r,e)}) \cdot \sigma(W_2^{(r,e)})$, where $\sigma(\cdot)$ denotes the spectral (ℓ_2 operator) norm. This is a pointwise bound: the inequality holds for all inputs, not only for the held-out batch. Spectral norms are computed by power iteration on $W_1^{(r,e)}(W_1^{(r,e)})^\top$ and analytically for the output layer.

The per-regime endogenous certified bound is the sum over Y-channel edges: $\widetilde{L}_Y^{(r)} := \sum_{e \in E_Y} \sigma(W_1^{(r,e)}) \cdot \sigma(W_2^{(r,e)})$ for the additive predictor. The certified equilibrium-step bound is $\widetilde{L}_{\text{eq,max}} = \max_r[\widetilde{L}_Y^{(r)} + \hat{L}_{\phi,\text{eq}}^{(r)}\hat{M}]$, replacing the gradient-probe $\hat{L}_Y^{(r)}$ with the certified $\widetilde{L}_Y^{(r)}$.

By construction $\widetilde{L}_Y^{(r)} \geq \hat{L}_Y^{(r)}$ for all regimes, so $\widetilde{L}_{\text{eq,max}} < 1$ implies $\hat{L}_{\text{eq,max}} < 1$ and provides a stronger guarantee.

Empirical comparison on V-Dem. 3 seeds, vanilla MSE training, 100 epochs. Per-regime L_Y values from the same trained model:

	Probe ($\hat{L}_Y^{(r)}$)	Certified ($\widetilde{L}_Y^{(r)}$)	Gap (cert−probe)
Regime 0	0.129 ± 0.011	0.136 ± 0.014	$+0.007$
Regime 1	0.152 ± 0.022	0.157 ± 0.023	$+0.004$
Regime 2	0.024 ± 0.015	0.030 ± 0.012	$+0.006$
Max over regimes	0.155 ± 0.019	0.164 ± 0.014	$+0.009$ ($+7\%$)

The certified upper bound exceeds the gradient-probe sample maximum in all 9 regime-seed cells, confirming the bound is non-vacuous, and the gap is small ($\approx 7\%$ in absolute terms at the maximum-regime level). Both $\hat{L_{\text{eq,max}}} \ll 1$ and $\widetilde{L}_{\text{eq,max}} \ll 1$ hold on V-Dem, so the contraction conclusion of Theorem 1 holds whether we use the conservative certified bound or the gradient-probe estimate.

K Hyperparameter sensitivity and reproducibility

Gating temperature $\tau_{\text{init}} = 0.5$. Earlier configurations used $\tau_{\text{init}} = 3.0$, producing soft gating with broad π_r overlap that empirically averaged across regimes. A diagnostic showed that hard tertile stratification recovered regime-conditional signal that soft gating with broad τ did not. Setting $\tau_{\text{init}} = 0.5$ (sharper gating) and $\lambda_\phi = 0.005$ (lower temperature regularization, keeping gating sharp) recovered this signal under the soft-gating architecture. We did not do a formal sensitivity sweep on τ_{init}.

Stratified-baseline training epochs. The stratified-NN baseline (a separate M1.5 PANAR-X trained per regime) is trained for 900 epochs per regime, $3\times$ the 300 epochs of the pooled NN baseline, to roughly compensate for the smaller per-regime training sample (about $1/3$ of the pooled sample). Validation-based early stopping was deferred. The stratified NN baseline serves as a diagnostic of NN-based regime-stratification overfitting on small per-regime samples.

Spectral-norm targets. The soft penalty target $\tilde{L}_Y = 0.55$, projection-every-epoch target 0.90 (with light soft penalty target 0.85), and projection-every-batch target 0.85 are conservative choices ensuring the trained $L_Y^{(r)}$ stays below the unit-Lipschitz threshold by a margin. Tighter targets (e.g., 0.4) have been tried in synthetic preliminaries and produce qualitatively similar Pareto behavior with shifted absolute MSE values.

Reproducibility. Random seeds are set via `torch.manual_seed` and `np.random.seed` immediately before each model instantiation; data generation uses `np.random.default_rng(42)` (synthetic) or fixed CSV input (V-Dem). Audit functions are deterministic given trained weights. Test MSE values reproduce on independent T4 GPU hardware to within $\sim 10^{-5}$; audit constants based on max-over-samples Jacobian probes drift by up to $\sim 10^{-3}$ across hardware due to non-deterministic CUDA matmul kernels, but every substantive claim in the paper (paired-DM significance, the contraction certificate holding in $5/5$ seeds across all 9 audited cells, the sign-reversal diagnostic, and the synthetic stress-test ordering) reproduces independent of hardware.

The reproducibility package supports two reproduction paths. The audit-only path executes 141 numerical checks against archived artifact pickles in under 30 seconds on any machine, with no GPU required, and reproduces every numerical claim in the paper. Full retraining on a single T4 GPU takes approximately 3.5 hours: ~ 1 minute for synthetic baselines, ~ 17 minutes for the synthetic Pareto sweep, ~ 175 minutes for the V-Dem sweep (the dominant cost), ~ 23 minutes for postanalysis, ~ 2 minutes for the certified-bound audit, and ~ 15 seconds for figure generation. The V-Dem sweep dominates because its per-batch Jacobian penalty, required for the contraction certificate, scales with training-set size ($N_{\text{train}} = 6{,}191$ vs. $1{,}050$ for the synthetic panel). CPU-only retraining is approximately 12 CPU-hours and is supported but not recommended for reviewer use. Total runtime exceeds the free-tier Colab idle timeout (90 minutes); reviewers using Colab should use Colab Pro or run the audit-only path.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and §1 state the contributions that the paper supports: a rate-based contraction theorem with an explicit, auditable gate-leakage term (§4); a failure mode of per-edge spectral normalization on soft-gated mixtures, shown on a synthetic panel (§7) and on a real realized-volatility panel where $4/5$ spectrally compliant seeds fail certification (§8); and a V-Dem diagnostic showing that pooled additive models cancel regime-conditional signal, together with the corrected result that under seed-fair, country-clustered inference no tested model significantly outperforms a random walk (§6). The forecasting-improvement claim of the submitted version is withdrawn in the abstract, §6.2, and §10 ("We claim no predictive gains on these panels").

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: §10 contains a dedicated Limitations paragraph that states, with cross-references: (i) V-Dem forecasting performance is statistically indistinguishable from persistence and from stratified linear regression; (ii) the robust empirical result is narrower than the symmetric sign-reversal claim of the submitted version, which is withdrawn; (iii) the stability theorem is an architecture-specific, auditable specialization of average-dwell-time analysis; (iv) the realized-volatility experiment demonstrates certificate relevance, not forecasting improvement. It further discusses single-target dependence on V-Dem, exogenous regime structure, parameter scaling in R, the diagnostic-not-constructive penalty bridge, and deferred function-valued interpretability. §9 discusses when the certificate is practically useful and the appropriate and inappropriate uses of democracy forecasts at this horizon.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.

- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Theorem 1 is stated in §4 with explicit assumptions (A1)–(A3), including the edge case $B_{\text{trans}} \leq 1$. Complete proofs appear in Appendix B: Lemma 1 (global Lipschitz) with proof; auxiliary Lemmas 2 and 3 (equilibrium-step and punctuation-step bounds) with proofs; full proof of Theorem 1 in §B.3 broken into four steps; Corollary 1 (trajectory-dependent sharpening) with proof; Theorem 2 (stochastic error propagation) with proof. The relationship to Hespanha–Morse average-dwell-time stability [19] is stated explicitly.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper?

   Answer: [Yes]

   Justification: The architecture is fully specified in §3 (Equations 1–3). Training procedure, penalty terms, three enforcement modes, and the audit quantities including gate leakage are specified in §5. V-Dem preprocessing is in Appendix E; the synthetic DGP and all 45-run sweep hyperparameters are in Appendix D; the two-estimator audit protocol is in Appendix J. The panel-appropriate inference (country-clustered t and country-block bootstrap, B=9,999, Holm correction) is specified in §6.2; the realized-volatility panel construction, split, regime definition, moving-block bootstrap, hyperparameter-selection grid, and audit are specified in §8; the 24-variant sensitivity grid is specified in §6.4. The four pre-registered design documents that fixed these protocols before any confirmatory run, and all resulting artifacts, are included in the supplementary material.

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.

- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Code, data, and all artifacts are released in the archived reproducibility package, DOI 10.5281/zenodo.23717582 (Zenodo archive of the public repository `https://github.com/vkuskova/NeurIPS-18877`). The repository contains scripts to reproduce all 45 synthetic-sweep cells, all 9 V-Dem audited cells, the diagnostic linear stratification analysis, the full audit protocol, the panel-appropriate inference of Table 2, the 24-variant sensitivity grid and the four M2 retraining variants of §6.4, and the realized-volatility screen, training, tests, and audit of §8, together with the four locked design documents and the per-run artifacts every reported number traces to. The V-Dem v15 dataset [8] is publicly available under CC BY 4.0 at `v-dem.net`; the preprocessed CSV is included. The realized-volatility panel follows [21] and the processed file used is included with its SHA-256 hash recorded in the run metadata. The synthetic DGP is fully specified in Appendix D.1 with its generation seed, and the repository regenerates it bit-exact.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.

- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Optimizer (Adam), learning rate (3×10^{-3}), weight decay (10^{-4}), batch size (128), training epochs (300 for the sweep, 900 for the stratified-NN baseline with rationale in Appendix K), hidden width (32), and lag window ($L = 5$) are stated in §5 and Appendix D.2. The V-Dem split is per-country temporal (last 10 years held out, 6,191 training /1,510 test observations), Appendix E; the 20-seed per-run and ensemble conventions are defined in Table 1. Spectral-norm enforcement targets ($\tilde{L}_Y = 0.55$ soft, $\sqrt{0.90}$ per-epoch projection, $\sqrt{0.85}$ per-batch projection) are stated and justified in Appendix K. For the realized-volatility panel, the split (final 523 days per asset), lag structure, regime cut points, and the pre-registered 2×2 hyperparameter grid selected on a training-internal validation split are stated in Appendix H and in the locked design document in the supplementary material.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: V-Dem predictive comparisons use paired loss-differential tests with country-clustered t statistics (Liang–Zeger, $\mathrm{df} = G-1$, G=151) and a country-block bootstrap (B=9,999), with Holm–Bonferroni correction across the $m = 3$ per-regime tests (Table 2); stochastic models report per-run mean $\pm$ sd over 20 seeds and the seed-ensemble MSE (Table 1). The per-run vs. ensemble distinction, including the seed-asymmetric comparison of the submitted version and its seed-fair replacement, is stated explicitly in §6.2. The realized-volatility panel uses a moving-block bootstrap over days (block 25, B=9,999) with per-asset HAC Diebold–Mariano tests [22] as a cross-check (§8). Synthetic gaps report seed-level 95% t-confidence intervals across 5 seeds (§7, Appendix D.4). Sensitivity results report jackknife sign-stability fractions over 200 half-country subsamples (§6.4). Bound satisfaction is reported per seed in every audited cell.

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix K ("Reproducibility" paragraph) reports compute. The audit-only reproduction path (`verify_paper_numbers.py` against the archived artifacts) executes the numerical checks in under 30 seconds on any machine and requires no GPU. Full retraining is approximately 30 minutes on a single GPU (free Colab T4 sufficient) for each of the synthetic and V-Dem main sweeps, plus approximately 25 additional minutes on GPU for postanalysis, certified-bound audit, and figure generation. The camera-ready additions cost: the panel-appropriate inference and the 24-variant linear sensitivity grid, minutes on CPU; the four M2 retraining variants of §6.4, 20 runs of the same cost as a V-Dem sweep run; the realized-volatility experiment, 25 runs (20 validation, 5 final) of about 2 minutes each on a T4. CPU-only retraining is supported but not recommended; the per-batch Jacobian penalty is the dominant cost. Random seeds are fixed per run; results are deterministic on CPU and reproduce on GPU up to non-deterministic CUDA matmul kernels (drift bounded at $\sim 10^{-6}$ in test MSE). paragraph) reports compute. The audit-only reproduction path (`verify_paper_numbers.py` against the archived artifacts in `artifacts/`) executes 141 numerical checks in under 30 seconds on any machine and requires no GPU. Full retraining is approximately 30 minutes on a single GPU (free Colab T4 sufficient) for each of the synthetic and V-Dem main sweeps, plus approximately 25 additional minutes on GPU for postanalysis, certified-bound audit, and figure generation. CPU-only retraining is approximately 8 hours total and is supported but not recommended for reviewer use; the per-batch Jacobian penalty required for the contraction certificate is the dominant cost and benefits substantially from GPU. Random seeds are fixed via `torch.manual_seed` and `np.random.seed` per run; results are deterministic on CPU and reproduce on GPU up to non-deterministic CUDA matmul kernels (drift bounded at $\sim 10^{-6}$ in test MSE).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics?

   Answer: [Yes]

   Justification: The work uses the publicly released V-Dem v15 country-year panel [8] (CC BY 4.0 license), a public realized-volatility panel of eight equity indices following [21], and synthetic data the authors generated. No human subjects research is conducted; expert coders are part of V-Dem's existing publication, not part of this submission. No personally identifiable information is used. The work complies with the NeurIPS Code of Ethics.

   Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: §9 discusses the appropriate uses of this class of model on democracy data (diagnosing regime-dependent heterogeneity that pooled models mask, and auditing rollout stability before deployment) and the misuse the paper's boundary result warns against: treating one-step forecasts of measurement-smoothed democracy indices as operational predictions of democratic change for particular countries, when no tested model, including ours, is distinguishable from a random walk at this horizon. The architecture itself is generic for heterogeneous panel time series; beyond the use and misuse discussion above, its societal impact is that of statistical forecasting methods in general.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse?

    Answer: [N/A]

    Justification: The paper does not release a pretrained model, generative system, or scraped dataset. The V-Dem dataset used is already publicly released by the V-Dem Institute under CC BY 4.0. No safeguards are required.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.

- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The V-Dem v15 dataset [8] is cited with the standard V-Dem citation and is available under the Creative Commons Attribution 4.0 International (CC BY 4.0) license at `v-dem.net`; the variables used (`v2xel_frefair`, `v2csantimv`, `v2jureform`, `v2psorgs`, `v2mecorrpt`) are listed in Appendix E. The realized-volatility panel of eight equity indices is credited to [21]; it is not redistributed with our package, and readers are referred to that source for the data and its terms of use. The package records the SHA-256 of the file we used, so a panel reconstructed from the source can be checked against it before the realized-volatility cells are run. NAVAR [6] is cited as the architectural starting point; PANAR-X is a re-implementation rather than a derivative of released code.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The PANAR-X reference implementation is released in the archived reproducibility package, DOI 10.5281/zenodo.23717582 (`https://github.com/vkuskova/NeurIPS-18877`). The repository includes a `README` with installation instructions, a documented architecture module, training scripts for the three enforcement modes, the audit protocol (both gradient-probe and certified spectral-norm estimators), the synthetic DGP generator (seed-fixed), the V-Dem preprocessing pipeline, the camera-ready inference and sensitivity scripts, the realized-volatility pipeline, the four locked design documents, a paper-to-code traceability file, and a verification script that checks every number reported in the paper against the archived artifacts. Hyperparameters used in the paper match those in Appendix D.2 and Appendix H and are exposed as configurable arguments.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.

- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects. The expert coding underlying V-Dem was conducted by the V-Dem Institute as part of their existing data product, not as part of this research.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve research with human subjects. No IRB approval is required for the secondary analysis of publicly released aggregate panel data.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research?

    Answer: [N/A]

    Justification: LLMs are not a component of the core methods. The architecture (PANAR-X) is a regime-aware additive forecasting model with per-edge MLPs of hidden width 32; it does not use, train, or fine-tune any large language model. LLM usage was limited to standard editing and code-assistance tasks during paper preparation, not to the scientific contribution itself.

    Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.