Temporal Dynamics of World Model Evaluation:
Adaptive Horizons and Failure Type Diagnosis

Régis Rigaud
RQZ Prospective
regis.rigaud@rqz-prospective.fr

Abstract

World models in reinforcement learning must detect when their predictions become unreliable. Two fundamental questions arise: (1) over what temporal horizon should prediction coherence be evaluated? and (2) how can we distinguish transient noise from genuine model failure?

We propose two complementary mechanisms addressing these challenges:

Adaptive Resonance Horizon (ARH): The evaluation horizon N is modulated by system confidence. High confidence extends the horizon (tolerant); low confidence contracts it (reactive). This resolves the noise-latency tradeoff inherent in fixed-horizon approaches. Validation on DeepMind Control Suite shows 65–72% stability under nominal conditions vs. <5% for fixed baselines.

Dual-Horizon Resonance (DH): Two diagnostic signals extend the framework. *DH-σ* (Sign Structure): The sign change rate ρ of prediction errors discriminates failure type—noise causes random sign alternation ($\rho \approx 0.5$), drift causes sign persistence ($\rho \approx 0$). This achieves 76% separation on synthetic benchmarks. *DH-Δ* (Position-Velocity): Combining resonance R with its trend dR/dt detects regime transitions, achieving 76% discrimination on 25 real-world TCPD datasets.

Together, ARH answers "how long to evaluate?" and DH answers "what type of failure?" The Two-Stage architecture separates detection from qualification, enabling both sensitive detection and accurate diagnosis. These mechanisms can integrate with existing model validation frameworks or operate standalone.

Keywords: adaptive horizon, temporal coherence, noise-drift discrimination, sign structure, world model validation.
Code available at: `https://github.com/TheCause/temporal-evaluation-framework`

1 Introduction

Summary. Fixed evaluation horizons create a noise-latency tradeoff. We introduce (1) confidence-modulated horizons that adapt to system state, and (2) sign structure analysis that discriminates noise from drift.

World models in reinforcement learning require mechanisms to detect when predictions become unreliable. A natural approach is to measure *resonance*—the temporal coherence of multi-step predictions. However, resonance-based evaluation raises two fundamental questions.

Question 1: How Long to Evaluate? Resonance is computed over a rollout horizon N. A fixed horizon creates an inherent tradeoff:

- **Small N:** Reactive but noise-sensitive. Transient noise triggers false reconstruction.

- **Large N:** Robust but slow. Genuine failures are detected with unacceptable latency.

Question 2: What Type of Failure? Even with perfect horizon selection, a resonance drop does not reveal *why* the model is failing. Both transient noise and genuine drift collapse resonance, yet require opposite responses:

- **Noise burst:** Wait for conditions to stabilize.
- **Structural drift:** Trigger immediate reconstruction.

Contributions. We propose two mechanisms addressing these questions:

Component	Question	Mechanism	Key Result
ARH	How long to evaluate?	Confidence-modulated N	65–72% DMC stability
DH-σ	What type of failure?	Sign change rate ρ	76% noise/drift separation
DH-Δ	Is something changing?	Position-velocity detection	76% TCPD discrimination

Positioning. Our approach differs from classical change point detection (CUSUM, BOCPD) in two ways: (1) we analyze *prediction errors* from a world model rather than raw observations, and (2) we diagnose *failure type*, not just change occurrence. The mechanisms are model-agnostic and can integrate with any world model architecture.

2 Background

2.1 Resonance as a Validity Signal

Resonance quantifies the temporal coherence of model predictions. Given a world model f_θ and a rollout horizon N, we define:

$$R_t = g(\|f_\theta^{(N)}(z_{t-N}) - z_t\|^2), \tag{1}$$

where $g(\cdot)$ is a monotonically decreasing function (e.g., exponential decay or z-score normalization). High resonance indicates consistent predictions; low resonance signals potential model invalidity.

This paper focuses on *temporal resonance* from multi-step prediction coherence. An alternative formulation uses physics priors Φ to measure model-physics agreement, but we do not assume access to such priors here.

2.2 The Fixed Horizon Problem

Resonance R_t is computed over a rollout horizon N. A common heuristic uses fixed N (e.g., $N = \max(6, W/3)$ for window size W), but this lacks theoretical justification.

Failure Mode 1: Noise Sensitivity. With small N, transient noise causes immediate resonance collapse. A single noisy observation can trigger reconstruction.

Failure Mode 2: Detection Latency. With large N, genuine structural shifts require N steps before detection. This latency can be catastrophic in safety-critical applications.

2.3 The Noise-Law Confusion Problem

Even with optimal N, the system cannot distinguish failure types. Consider:

- **Scenario A:** Observation noise increases from $\sigma = 0.1$ to $\sigma = 0.5$. Underlying dynamics unchanged.
- **Scenario B:** Dynamics change from f to g. Noise unchanged.

Both scenarios produce identical symptoms: resonance drops, stress increases. The correct response differs: wait (A) vs. rebuild (B).

2.4 Relation to Change Point Detection

Classical methods like CUSUM [7] and Bayesian Online Change Point Detection (BOCPD) [6] detect distributional changes in raw observations. Our approach differs in two key ways:

1. We analyze *prediction errors* from a world model, not raw data. This provides a model-centric view of validity.
2. We diagnose *failure type* (noise vs. drift), not just change occurrence. This enables differentiated responses.

BOCPD could be applied to prediction errors as a detection mechanism; DH-σ provides complementary information about failure nature. The two approaches are not mutually exclusive.

3 Method 1: Adaptive Resonance Horizon

3.1 Hypothesis

We investigate the following hypothesis:

> *The validity of a model can be assessed through multi-step temporal coherence, evaluated over an adaptive horizon modulated by internal confidence.*

3.2 Confidence-Modulated Horizon

Let $C_t \in [0, 1]$ denote system confidence at time t. The adaptive horizon is:

$$N_t = N_{\min} + (N_{\max} - N_{\min}) \cdot C_t, \tag{2}$$

where $N_{\min}$ is the minimum horizon (vigilant mode) and $N_{\max}$ is the maximum (tolerant mode).

Behavior.

- High confidence ($C \approx 1$): $N \approx N_{\max}$, tolerant, noise-resistant.
- Low confidence ($C \approx 0$): $N \approx N_{\min}$, reactive, fast detection.

3.3 State-Dependent Confidence Dynamics

The key insight: reactivity should depend on *current state*, not destination. We introduce state-dependent update rates:

$$\beta_t = \beta_{\text{slow}} \cdot C_t + \beta_{\text{fast}} \cdot (1 - C_t). \tag{3}$$

With $\beta_{\text{slow}} = 0.95$ and $\beta_{\text{fast}} = 0.5$:

- Confident mode: slow trust-building (∼200-step half-life).
- Vigilant mode: fast adaptation (∼20-step half-life).

The confidence update:

$$C_{t+1} = \beta_t \cdot C_t + (1 - \beta_t) \cdot R_t. \tag{4}$$

3.4 Phasic Response for Abrupt Failures

Gradual EMA updates are insufficient for abrupt failures. We add a phasic response when resonance drops anomalously:

$$C_t = \begin{cases} \beta_t \cdot C_{t-1} + (1 - \beta_t) \cdot R_t & \text{if } R_t \geq R_{\text{crit}} \\ \alpha \cdot R_t & \text{if } R_t < R_{\text{crit}} \end{cases} \tag{5}$$

where $R_{\text{crit}} = \mu_R - k \cdot s_R$ is an adaptive threshold ($k = 2$, i.e., 2 standard deviations below mean), μ_R and s_R are the running mean and standard deviation of recent resonance values, and $\alpha = 0.3$ controls the magnitude of phasic reset.

3.5 Multi-step Resonance Computation

Given adaptive horizon N_t:

1. Retrieve state z_{t-N_t} from buffer.
2. Rollout model: $\hat{z}_t = f_\theta^{(N_t)}(z_{t-N_t}, a_{t-N_t:t-1})$.
3. Compute error: $e_t = \|\hat{z}_t - z_t\|^2$.
4. Convert to resonance: $R_t = \text{clip}(1 - (e_t - \mu_e)/(\lambda_z \sigma_e), 0, 1)$, where $\lambda_z = 3$ scales the z-score.

The z-score normalization is self-calibrating: no environment-specific tuning required.

3.6 Integration with Model Validation

ARH provides a principled resonance signal that can trigger downstream actions. For instance, a reconstruction decision might combine resonance with other signals:

$$\text{Rebuild}(t) = (\text{Stress} > \Theta) \wedge (\langle R^{\text{ARH}} \rangle_W < \delta), \tag{6}$$

where $\langle R^{\text{ARH}} \rangle_W$ is the windowed average of adaptive resonance. The specific integration depends on the application; ARH is agnostic to downstream usage.

4 Method 2: Dual-Horizon Resonance

ARH resolves the horizon problem but cannot distinguish failure types. We now address noise-drift discrimination.

4.1 The Sign Structure Insight

We observe that noise and drift produce fundamentally different *sign patterns* in prediction errors.

Noise Signature. For i.i.d. noise, errors alternate signs randomly:

$$\mathbb{E}[\rho_{\text{noise}}] = 0.5, \tag{7}$$

where ρ is the sign change rate.

Drift Signature. Structural drift causes systematic bias—errors persist with consistent sign:

$$\mathbb{E}[\rho_{\text{drift}}] \approx 0. \tag{8}$$

4.2 DH-σ: Sign Structure Diagnostic

We measure the sign change rate over a sliding window:

$$\rho_t = \frac{1}{W-1} \sum_{i=t-W+1}^{t-1} \mathbf{1}[\text{sign}(e_i) \neq \text{sign}(e_{i+1})]. \tag{9}$$

Interpretation.

- $\rho_t \approx 0.5$: Random sign alternation $\Rightarrow$ **noise**.
- $\rho_t \approx 0$: Consistent sign persistence $\Rightarrow$ **drift**.

Gated Confidence. The sign rate gates the confidence update:

$$\tilde{\rho}_t = \min(1, 2 \cdot \rho_t), \quad C^{\text{DH}}(t) = R_1(t) + \lambda_g \cdot \max(0, \tilde{\rho}_t - R_1(t)), \quad \lambda_g = 0.3. \tag{10}$$

When errors alternate randomly (ρ high), confidence is protected from collapse. When errors persist (ρ low), confidence follows raw resonance.

4.3 DH-Δ: Position-Velocity Detection

Sign structure detects failure *type* but not *timing*. For regime transition detection, we combine position and velocity:

$$P(t) = w_R \cdot (1 - R_t) + |\text{trend}_t|, \tag{11}$$

where trend_t is the slope of recent R values.

Why Position-Velocity? These signals are orthogonal:

- R measures *where* (current state).
- trend measures *where going* (dynamics).

Together they fully characterize a first-order dynamical system.

4.4 Two-Stage Architecture

The key insight: separate detection from qualification.

Stage	Signal	Question	Action
1. Detection	$P = w_R(1 - R) + \lvert\text{trend}\rvert$	Is something happening?	Trigger alert
2. Qualification	ρ (sign rate)	What type of failure?	WAIT/REBUILD

Qualification Thresholds. When P exceeds detection threshold:

- $\rho \geq 0.4$: noise $\rightarrow$ WAIT
- $\rho \leq 0.15$: drift $\rightarrow$ REBUILD
- Otherwise: ambiguous $\rightarrow$ INVESTIGATE

5 Experimental Setup

5.1 Synthetic Benchmarks

Four scenarios isolate temporal evaluation effects:

1. **Noise Burst:** Noise $\sigma : 0.1 \rightarrow 0.5$ for 60 steps. Tests false positive rate.
2. **Law Inversion:** Dynamics flip $a : +0.95 \rightarrow -0.95$. Tests detection latency.
3. **Gradual Drift:** Parameter drifts $a : 0.95 \rightarrow 0.5$. Tests sensitivity.
4. **Compound:** Noise burst then drift. Tests discrimination.

5.2 DeepMind Control Suite

Three environments with trained world models:

- cartpole-swingup, cheetah-run, walker-walk
- 10 seeds, 500 steps per episode
- Scenarios: stable, noise ($\sigma = 0.3$), drift (observation drift at $t = 250$)

5.3 TCPD Real-World Validation

25 datasets from the Turing Change Point Dataset (TCPD) with expert-annotated change points [8].

TCPD Procedure. For each univariate time series, we fit a simple autoregressive model $\hat{x}_{t+1} = f(x_{t-k:t})$ on a sliding window. The prediction error $e_t = x_t - \hat{x}_t$ serves as input to DH-Δ. We compute resonance R_t from normalized error magnitude, trend_t as the linear regression slope of R over the last 10 steps, and detection score P_t per Equation 11. A change point is detected when P_t exceeds the 95th percentile of baseline (warmup) values.

5.4 Physical Simulation

safe-control-gym benchmark:

- CartPole (balancing), Quadrotor 2D (stabilization)
- 10 seeds, 95% bootstrap confidence intervals
- Scenarios: stable, noise injection, drift injection

5.5 Metrics

Stability. We define *stability* as the percentage of timesteps where system confidence exceeds a threshold:

$$\text{Stability} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}[C_t > \tau], \quad \tau = 0.7. \tag{12}$$

This measures how often the system maintains confident operation under nominal conditions.

Baseline Selection. Fixed-N baselines span the practical range: $N = 1$ (maximum reactivity), $N = 4$ (commonly used in online learning [9]), $N = 20$ (maximum smoothing within our buffer). We report the *oracle* (best fixed-N per scenario) as an upper bound for non-adaptive methods.

6 Results

6.1 ARH: Adaptive Horizon Validation

Table 1: ARH vs. fixed baselines on synthetic scenarios.

Scenario	Metric	ARH	Fixed-4	Fixed-20
Noise Burst	FPR	0.55	0.40	0.55
Law Inversion	Latency (steps)	**0**	>200	>200
Gradual Drift	Latency (steps)	**2**	100	100
Sensor Dropout	Detection	**10**	>200	>200

Synthetic Benchmarks. Key finding: ARH detects law inversion *immediately* via phasic response, while fixed baselines fail entirely.

Table 2: DeepMind Control Suite stability (10 seeds, 500 steps). Stability = % timesteps with $C > 0.7$.

Environment	Scenario	ARH	Fixed-1	Fixed-4	Fixed-20
cartpole	Stable	**72.3%**	1.2%	2.8%	4.1%
cheetah	Stable	**65.2%**	0.4%	1.1%	1.8%
walker	Stable	**68.1%**	0.5%	1.3%	2.1%

DMC Validation. ARH achieves 65–72% stability vs. <5% for all fixed baselines. The z-score normalization is essential for high-dimensional observations.

6.2 DH-σ: Sign Structure Discrimination

Table 3: Sign structure discrimination.

Scenario	Sign Rate ρ	C (DH)	C (Baseline)
Noise Burst	0.487	**0.54**	≈ 0
Drift	0.014	0.017	0.006
Step	0.015	0.019	0.007

Synthetic Benchmarks. Discrimination: $\tilde{\rho}_{\text{noise}} - \tilde{\rho}_{\text{drift}} = 0.81 - 0.05 = \mathbf{76}\%$.

Table 4: Two-Stage validation (10 seeds, 95% CI).

Environment	ρ_{noise}	ρ_{drift}	NOISE→FREEZE	DRIFT→REBUILD
CartPole	0.51 $[0.45, 0.58]$	0.08 $[0.08, 0.08]$	67.6%	100%
Quadrotor	0.50 $[0.47, 0.53]$	0.02 $[0.02, 0.02]$	67.8%	89.4%

Physical Simulation. 95% CIs do not overlap, confirming statistical significance.

6.3 DH-Δ: Regime Transition Detection

TCPD Validation. Position-velocity combination achieves **76% discrimination** on 25 real-world datasets (19/25 correctly identified).

Bootstrap cross-validation (n=500): 95% CI $[+3.8\%, +30.8\%]$ improvement over sign rate alone, $P = 99\%$.

Table 5: Detection signal comparison on TCPD.

Signal Combination	Discrimination	Improvement
Sign rate alone	35%	–
R + ρ + trend	46%	+11%
R + trend (position-velocity)	**76%**	**+41%**

6.4 Ablation: Component Necessity

Each component addresses a distinct capability:

- ARH: Temporal stability assessment (without it: <5% on DMC).
- DH-σ: Failure type diagnosis (without it: cannot distinguish noise from drift).
- DH-Δ: Regime transition detection (without it: 35% → 76%).

Table 6: Ablation study: each component's contribution.

Configuration	DMC Stability	Noise/Drift Sep.	TCPD Disc.
Fixed-N only	<5%	–	–
ARH (adaptive N)	65–72%	0%	–
ARH + DH-σ	65–72%	76%	35%
ARH + DH-σ + DH-Δ	**65–72%**	**76%**	**76%**

7 Discussion

7.1 Why Adaptive Horizons Work

The signal-to-noise ratio for shift detection improves with horizon length:

$$\text{SNR}(N) = \frac{N \cdot \Delta}{\sqrt{N} \cdot \sigma} = \sqrt{N} \cdot \frac{\Delta}{\sigma}. \tag{13}$$

However, longer horizons increase latency. ARH resolves this tradeoff dynamically: use long horizons when confident (better SNR), short horizons when uncertain (faster reaction).

Detection-Noise Tradeoff. Table 1 shows ARH has higher FPR (0.55) on noise bursts than Fixed-4 (0.40). This is the cost of reactivity: ARH responds to *any* anomaly, including noise. The DH-σ component mitigates this by distinguishing noise from drift *after* detection. In practice, we recommend using ARH for detection sensitivity combined with DH-σ for response qualification.

7.2 Why Sign Structure Works

The theoretical basis is straightforward:

- i.i.d. noise: $P(\text{sign change}) = 0.5$ per timestep.
- Systematic bias: $P(\text{sign change}) \approx 0$ once bias exceeds noise floor.

This holds regardless of noise magnitude—a 5× noise increase has the same sign pattern as baseline.

7.3 Limitations

ARH Limitations.

- Requires buffer of past states (memory overhead).
- Phasic threshold R_{crit} needs warm-up period.
- Does not distinguish failure types (addressed by DH).

DH-σ Limitations.

- Assumes approximately i.i.d. observation noise. Strongly autocorrelated noise (e.g., pink noise) would reduce discrimination accuracy.
- Cannot detect regime changes (level shifts with alternating errors).
- Requires sufficient window length for stable ρ estimates.

Empirical Validation of i.i.d. Assumption. In DMC environments, physics simulation introduces mild temporal correlation. We observe $\rho \approx 0.45$–0.51 under stable conditions (Table 4), close to the theoretical $\rho = 0.5$ for i.i.d. noise. This suggests the approximation holds sufficiently for discrimination, though users should verify noise structure in their specific domains.

DH-Δ Limitations.

- 24% failure rate on TCPD (6/25 datasets).
- Fails when model performs *better* during transitions.
- Requires trend estimation (smoothing delay).

7.4 Practical Guidelines

Recommended Defaults.

Parameter	Default	Derivation
$N_{\min}$	$\max(4, W/10)$	Noise averaging bound
$N_{\max}$	$\max(N_{\min} + 6, W/3)$	MIT heuristic at $C = 1$
β_{slow}	0.95	Slow trust-building
β_{fast}	0.5	Fast vigilance response
W_ρ	20	Sign rate window
w_R	0.9	Position-velocity weight

When to Use Which.

- **Always use ARH:** Required for meaningful stability assessment.
- **Add DH-σ:** When noise-drift distinction affects policy.
- **Add DH-Δ:** When early regime change detection is critical.

8 Conclusion

We proposed two complementary mechanisms for temporal evaluation of world models:

Adaptive Resonance Horizon (ARH) modulates the evaluation horizon based on system confidence. This resolves the noise-latency tradeoff: 65–72% stability on DMC vs. <5% for fixed baselines.

Dual-Horizon Resonance (DH) provides two diagnostic signals:

- *DH-σ*: Sign structure discriminates noise from drift (76% separation).
- *DH-Δ*: Position-velocity detects regime transitions (76% TCPD accuracy).

The **Two-Stage Architecture** separates detection (position-velocity) from qualification (sign structure), enabling both sensitive detection and accurate diagnosis.

Together, these address the two fundamental questions of temporal model evaluation: *how long to evaluate* (ARH) and *what type of failure* (DH). The mechanisms are model-agnostic and complement existing approaches to distribution shift detection.

Acknowledgments

This work was conducted independently. The author acknowledges the use of AI-assisted tools (Claude Opus 4.5) for brainstorming, code development, and manuscript editing. All experimental design, theoretical contributions, and scientific claims are solely the responsibility of the author.

References

[1] R. Rigaud. *Structural Uncertainty as a Control Problem: When to Adapt, Freeze, or Reconstruct World Models*. 2025.

[2] R. Rigaud. *Homeostatic Hamiltonian Agent: Inter-temporal Energy Surprise as a Minimal Principle for Adaptive Inference Under Non-Stationarity*. 2025.

[3] R. Rigaud. *Resonance-Based Distribution Shift Detection: Measuring Model-Physics Alignment for Adaptive Prediction*. 2025.

[4] Y. Tassa, Y. Doron, A. Muldal, T. Erez, Y. Li, D. de Las Casas, D. Budden, A. Abdolmaleki, J. Merel, A. Lefrancq, T. Lillicrap, and M. Riedmiller. DeepMind Control Suite. *arXiv preprint arXiv:1801.00690*, 2018.

[5] A. Yuan, M. Golowich, A. Majumdar. safe-control-gym: A Unified Benchmark for Safe Learning-based Control and Reinforcement Learning. *IEEE Robotics and Automation Letters*, 2022.

[6] R. P. Adams and D. J. C. MacKay. Bayesian Online Changepoint Detection. *arXiv preprint arXiv:0710.3742*, 2007.

[7] E. S. Page. Continuous Inspection Schemes. *Biometrika*, 41(1/2):100–115, 1954.

[8] G. J. J. van den Burg and C. K. I. Williams. An Evaluation of Change Point Detection Algorithms. *arXiv preprint arXiv:2003.06222*, 2020.

[9] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2nd edition, 2018.