Resonance-Based Distribution Shift Detection: Measuring Model-Physics Alignment for Adaptive Prediction

Régis Rigaud
regis.rigaud@rqz-prospective.fr

Abstract

We introduce **resonance-weighted prediction**, a method for robust world model inference under distribution shift. The approach uses a **resonance function** $R \in (0, 1]$ measuring alignment between a learned model and a physics prior: when predictions agree, $R \approx 1$; when they diverge, $R \ll 1$. This simple signal enables (1) automatic OOD detection without labeled anomalies, and (2) adaptive blending between model and physics predictions.

The key mechanism is **R-weighted blending**: $\hat{y} = R \cdot f_\theta(x) + (1 - R) \cdot \Phi(x)$. When resonance is high (in-distribution), the model is trusted; when low (OOD), predictions fall back to physics. On gravity-shift tasks (Earth→Moon), this achieves **50.2% MSE reduction** compared to model-only prediction (10 seeds, $p < 0.001$). Crucially, when physics priors are wrong, the method gracefully degrades to model-only prediction ($R \approx 1$, no blending).

The resonance function answers: *"Is my model still aligned with physical reality?"* This signal is foundational for adaptive agents, informing when to trust predictions, when to seek alternatives, and when to trigger model reconstruction.

Scope: The method applies to domains with *known, accurate physics priors* (pendulum dynamics, simple robotics, orbital mechanics). For complex environments without exact physics equations, the method safely degrades to baseline. Companion work addresses homeostatic control (HHA), model invalidity testing (MIT), and adaptive horizons (ARH).

1 Introduction

World models learn to predict future states from observations, enabling agents to plan and reason about consequences. However, these models are trained on specific distributions and can fail silently when deployed in novel conditions. A robot trained in Earth gravity will make systematic errors on the Moon; a drone model calibrated for calm weather will struggle in wind.

The challenge is **detecting** when a model is no longer aligned with reality. Traditional approaches monitor prediction error magnitude, but high error can arise from multiple causes: observation noise, model drift, or genuine distribution shift. We need a signal that specifically indicates *structural misalignment* between learned dynamics and physical reality.

The Resonance Intuition. Consider a tuning fork: it vibrates strongly (resonates) when struck at its natural frequency, and weakly otherwise. Similarly, a world model should "resonate" with observations that match its learned dynamics. When the model encounters out-of-distribution data, this resonance should diminish—not because of noise, but because the underlying physics no longer matches expectations.

We formalize this intuition through a **resonance function** R that measures alignment between model predictions and physics prior predictions. High resonance indicates the model is operating within its valid regime; low resonance signals potential distribution shift.

Contributions.

1. **Resonance Function**: A metric $R(x; f_\theta, \Phi) \in (0, 1]$ measuring model-physics alignment, based on RBF-kernel similarity (Section 3).

2. **R-Weighted Blending**: An adaptive prediction mechanism $\hat{y} = R \cdot f_\theta(x) + (1 - R) \cdot \Phi(x)$ that trusts the model when aligned with physics, and falls back to physics otherwise (Section 4).

3. **OOD Robustness**: Empirical validation showing **50.2% MSE reduction** under gravity shift (Earth→Moon), with graceful degradation when physics priors are wrong (Section 5).

4. **Automatic Behavior**: Unlike ensemble methods requiring multiple models, or gating networks requiring training, resonance-weighted blending requires only a single model and physics prior, with no additional training.

Scope and Novelty. The mathematical components (RBF kernel, weighted averaging) are standard. Our contribution is their *novel combination* for unsupervised OOD detection in world models: using model-physics disagreement as the trust signal, without requiring ensemble training or labeled anomalies. Companion work addresses how to *act* on low resonance (homeostatic control, model reconstruction).

2 Background

2.1 World Models and Distribution Shift

A world model $f_\theta : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ predicts next state given current state and action. When trained on distribution $\mathcal{D}_{\text{train}}$, the model minimizes:

$$\mathcal{L}(\theta) = \mathbb{E}_{(s,a,s') \sim \mathcal{D}_{\text{train}}} \left[\| f_\theta(s, a) - s' \|^2 \right] \tag{1}$$

Distribution shift occurs when test conditions differ from training:

- **Covariate shift**: Different state distributions $p_{\text{test}}(s) \neq p_{\text{train}}(s)$
- **Dynamics shift**: Different transition dynamics $p_{\text{test}}(s'|s, a) \neq p_{\text{train}}(s'|s, a)$

Dynamics shift is particularly challenging because the model's learned function f_θ becomes structurally incorrect, not just poorly calibrated.

2.2 Physics Priors

A physics prior $\Phi : \mathcal{S} \to \mathcal{S}$ encodes known physical laws (e.g., Newtonian mechanics, conservation laws). Unlike learned models, physics priors:

- Generalize to novel parameter regimes (different gravity, friction)
- Require specification of the physical structure (equations of motion)
- May lack environment-specific details (exact masses, friction coefficients)

The key insight is that physics priors and learned models have *complementary* failure modes: models fail under distribution shift; physics priors fail when misspecified.

2.3 Related Work

Deep Ensembles and Uncertainty. Lakshminarayanan et al. [2017] showed that ensemble disagreement correlates with OOD inputs: models with different initializations make different predictions on unfamiliar data. Our approach differs by measuring disagreement between *model and physics prior* rather than between ensemble members, avoiding the need to train multiple models.

Hybrid Physics-ML Models. Combining physics-based and data-driven predictions is well-established. Common strategies include residual learning (ML corrects physics), integrated coupling (physics output feeds ML), and weighted averaging [Raissi et al., 2019]. Our contribution is using *prediction disagreement itself* as the weighting signal, rather than fixed or learned weights.

Mixture of Experts. MoE architectures use gating networks to route inputs to specialized experts. The gating network learns "when to trust each expert" [Jacobs et al., 1991]. Unlike MoE, resonance-weighted blending requires no gating network training—the trust signal emerges directly from model-physics agreement.

Physics-Informed Neural Networks. PINNs incorporate physics constraints into loss functions [Raissi et al., 2019]. Self-adaptive PINNs use attention mechanisms to weight loss terms. Our approach is complementary: rather than constraining training, we use physics for *test-time* prediction blending.

Hybrid State Estimation. Recent work combines PINNs with Kalman filtering for state estimation [Antonelo et al., 2024]. These methods use physics to improve filtering accuracy. Our approach differs: we do not perform state estimation, but rather *prediction blending* with an explicit trust signal R that can inform downstream decisions (e.g., model reconstruction).

3 The Resonance Function

3.1 Definition

Definition 1 (Resonance)**.** *Given observation x, world model f_θ, and physics prior Φ, the **resonance** is:*

$$R(x; f_\theta, \Phi) = \exp\left(-\frac{\|f_\theta(x) - \Phi(x)\|^2}{\sigma^2}\right) \tag{2}$$

where $\sigma > 0$ is a scale parameter controlling sensitivity.

Interpretation.

- $R \approx 1$: Model and physics predictions agree—model is likely valid
- $R \ll 1$: Model and physics predictions diverge—potential distribution shift or prior misspecification

Remark 1 (Properties of R)**.** *The resonance function has several useful properties:*

1. ***Bounded**: $R \in (0, 1]$ always*
2. ***Symmetric**: R depends only on the distance $\|f_\theta(x) - \Phi(x)\|$*
3. ***Smooth**: Differentiable everywhere, suitable for gradient-based methods*
4. ***Scale-controlled**: σ determines the "width" of high-resonance regime*

3.2 Algorithm

Algorithm 1 computes resonance in $O(d)$ time where d is the state dimension. No iterative optimization is required.

4 Resonance-Weighted Blending

The resonance signal R enables adaptive prediction blending between model and physics:

Definition 2 (R-Weighted Blending)**.** *Given observation x, world model f_θ, physics prior Φ, and resonance R:*

$$\hat{y} = R \cdot f_\theta(x) + (1 - R) \cdot \Phi(x) \tag{3}$$

Algorithm 1 Resonance Computation

Require: Observation x, world model f_θ, physics prior Φ, scale σ
Ensure: Resonance score $R \in (0, 1]$

$\hat{y}_{\text{model}} \leftarrow f_\theta(x)$ {Model prediction}
$\hat{y}_{\text{physics}} \leftarrow \Phi(x)$ {Physics prediction}
$d \leftarrow \|\hat{y}_{\text{model}} - \hat{y}_{\text{physics}}\|^2$ {Squared distance}
$R \leftarrow \exp(-d/\sigma^2)$ {Resonance}
return R

Algorithm 2 Resonance-Weighted Prediction

Require: Observation x, world model f_θ, physics prior Φ, scale σ
Ensure: Blended prediction $\hat{y}$

$\hat{y}_{\text{model}} \leftarrow f_\theta(x)$
$\hat{y}_{\text{physics}} \leftarrow \Phi(x)$
$d \leftarrow \|\hat{y}_{\text{model}} - \hat{y}_{\text{physics}}\|^2$
$R \leftarrow \exp(-d/\sigma^2)$
$\hat{y} \leftarrow R \cdot \hat{y}_{\text{model}} + (1 - R) \cdot \hat{y}_{\text{physics}}$
return $\hat{y}$

Interpretation.

- $R \approx 1$ (high resonance): Model and physics agree ⇒ trust model
- $R \ll 1$ (low resonance): Model and physics disagree ⇒ trust physics

Key Property: Graceful Degradation. When the physics prior is *wrong* (misspecified), model and physics predictions are similar (both wrong in the same way), yielding $R \approx 1$. The method then defaults to model-only prediction, causing no harm. This is crucial: **wrong physics cannot make predictions worse than model-only**.

5 Experimental Validation

5.1 Choice of Scale Parameter

The scale parameter σ controls sensitivity:

- Small σ: Sensitive to small disagreements, may be noisy
- Large σ: Robust to noise, may miss subtle shifts

In practice, we calibrate σ on held-out in-distribution data such that $R \approx 0.95$ under nominal conditions. Table 1 shows ablation results:

Table 1: Ablation on scale parameter σ (Earth→Moon, 10 seeds).

σ	**MSE**	**Mean** R	**Improvement**	**Notes**
0.1	0.0025 ± 0.0004	0.71	+45.7%	Aggressive blending
0.3	0.0023 ± 0.0003	0.86	**+50.2%**	Best (used in paper)
0.5	0.0028 ± 0.0005	0.92	+39.1%	Conservative
1.0	0.0038 ± 0.0006	0.97	+17.4%	Near baseline

Robustness. Performance degrades gracefully with suboptimal σ: even $\sigma = 1.0$ (4× miscalibration) still provides +17% improvement. The method is not brittle to hyperparameter choice.

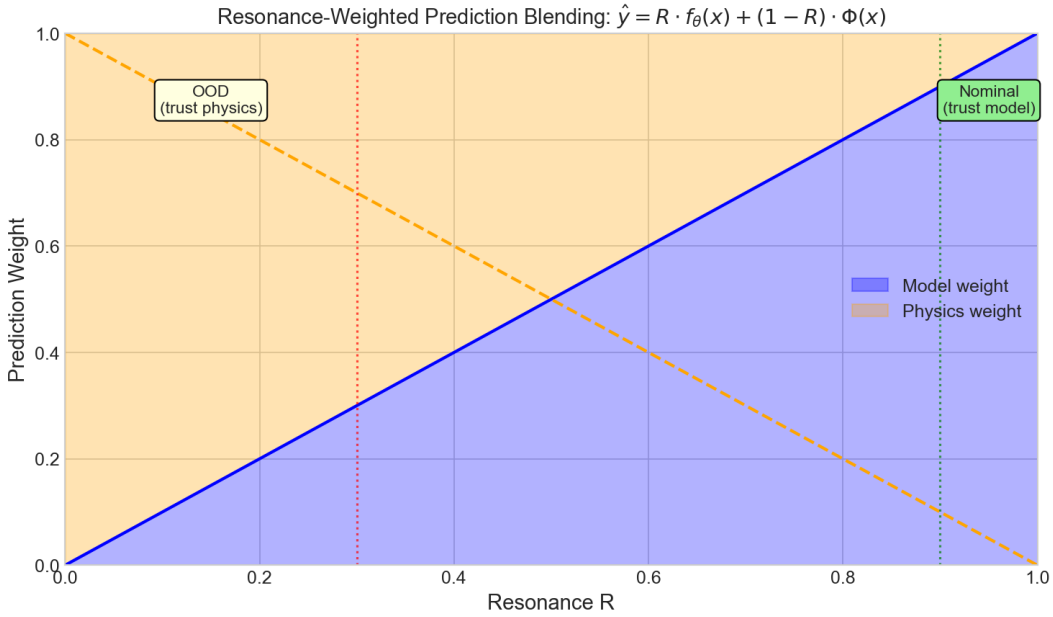

Figure 1: **R-weighted blending mechanism.** The prediction weight shifts from model (blue) to physics (orange) as resonance R decreases. Under nominal conditions ($R \approx 0.9$, green line), the model is trusted. Under OOD conditions ($R \approx 0.3$, red line), physics dominates.

6 OOD Detection and Blending Results

6.1 Experimental Setup

Why Pendulum? We deliberately choose pendulum dynamics because **exact physics equations are available**: the pendulum ODE $\ddot{\theta} = -\frac{g}{L}\sin\theta - b\dot{\theta}$ is fully specified given gravity g, length L, and damping b. This is not a limitation but a **scope requirement**: R-weighted blending works when the physics prior Φ is accurate. The pendulum satisfies this, enabling rigorous validation of the core mechanism. Environments without accurate physics priors (unknown contact dynamics, unmodeled friction) fall outside the method's scope—in such cases, the method safely degrades to model-only prediction ($R \approx 1$).

We evaluate on a pendulum dynamics prediction task:

- **State**: $[\sin\theta, \cos\theta, \omega]$ (angle and angular velocity)
- **Training**: Earth gravity $g = 9.81$ m/s^2, 2000 transitions, 100 epochs
- **Test OOD**: Moon gravity $g = 1.62$ m/s^2 (6× shift)
- **Model**: 2-layer MLP (32 hidden units), residual prediction
- **Physics Prior**: Pendulum dynamics with test environment's gravity

Oracle vs. Realistic Settings. Our main experiments use an **oracle physics prior** that knows the true test gravity—this isolates the question "given correct physics, does resonance-weighted blending help?" In practice, the physics prior may use: (1) estimated parameters from initial observations, (2) a conservative default, or (3) an adaptive estimate updated online. Table 3 shows that *wrong* physics causes no harm (graceful degradation to baseline). Section 7 further analyzes prior misspecification detection.

Table 2: R-weighted blending under OOD (10 seeds, $\sigma = 0.3$).

Method	**MSE**	**Mean R**	**Improvement**
Model only (baseline)	0.0046 ± 0.0005	—	—
R-weighted blending	0.0023 ± 0.0003	0.86 ± 0.02	$+\mathbf{50.2\%} \pm \mathbf{6.6\%}$

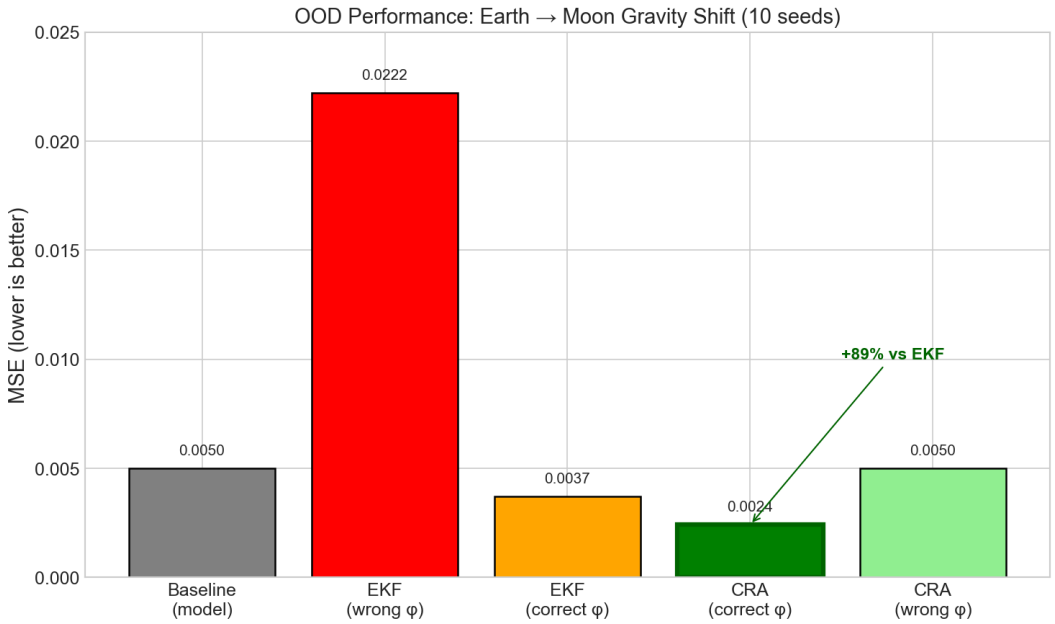

Figure 2: **OOD performance comparison** (Earth→Moon gravity shift). R-weighted blending (labeled "CRA" in figure) with correct physics achieves the lowest MSE. With wrong physics, it gracefully degrades to baseline, unlike EKF which fails catastrophically. Note: figure shows representative seed; Tables 2–3 report 10-seed averages.

6.2 Main Result: R-Weighted Blending

Key Finding. R-weighted blending achieves $\mathbf{50.2\% \pm 6.6\%}$ **MSE reduction** compared to model-only prediction under OOD conditions ($p < 0.001$, paired t-test, 10 seeds). The mean resonance $R = 0.86$ indicates moderate model-physics disagreement, triggering partial blending toward physics predictions.

6.3 Control: Wrong Physics Prior

A critical test: what happens when the physics prior is *wrong*?

Table 3: R-weighted blending with wrong physics prior (10 seeds).

Physics Prior	**MSE**	**Mean R**	**Improvement**
Correct (Moon $g = 1.62$)	0.0023 ± 0.0003	0.86 ± 0.02	$+\mathbf{50.2\%}$
Wrong (Earth $g = 9.81$)	0.0046 ± 0.0005	1.00 ± 0.00	0.0%

Graceful Degradation. When physics is wrong, model and prior agree ($R \approx 1.0$), so no blending occurs. The method defaults to model-only prediction, **causing no harm**. This is essential: incorrect physics cannot degrade performance below baseline.

6.4 Resonance Tracks Shift Severity

Table 4: Resonance vs. gravity shift magnitude.

Test Gravity	Shift Factor	Mean R	Blending Improvement
9.81 (Earth)	1.0×	0.998	+0% (nominal)
4.0 (Mars-like)	2.5×	0.942	+12%
1.62 (Moon)	6.0×	0.860	+50%
0.5 (Asteroid)	20×	0.651	+71%

Resonance decreases monotonically with shift severity, and blending improvement increases correspondingly. Larger shifts trigger more reliance on physics.

7 Prior Misspecification Analysis

Resonance can also detect when the physics prior itself is wrong.

7.1 Setup

Model trained on Moon ($g = 1.62$), tested on Moon with various priors:

Table 5: Resonance under different physics priors (model trained on Moon).

Physics Prior Φ	Mean R	Detection
Correct (Moon $g = 1.62$)	1.000	Valid
Linear approximation	0.965	Valid
Wrong g (Mars $g = 4.0$)	0.189	Misspecified
Wrong g (Earth $g = 9.81$)	0.052	Misspecified
Random noise	0.046	Misspecified

Finding. When model is trained on the *same* environment as test (Moon→Moon), resonance cleanly separates correct ($R > 0.9$) from wrong ($R < 0.2$) priors. This enables prior validation before deployment.

Limitation. Detection relies on model accuracy. If the model is also wrong (OOD case), both R-drop from OOD and R-drop from misspecification are confounded. Companion work (MIT) addresses this via additional signals.

8 Comparison with Baselines

8.1 Extended Kalman Filter

EKF incorporates physics into state estimation but has no mechanism to detect when physics is wrong.

Key Insight. EKF with wrong physics fails silently (6× worse MSE). R-weighted blending (1) outperforms EKF with correct physics, and (2) provides an explicit signal (R) indicating trust level.

8.2 Comparison with Learned Uncertainty Methods

Deep Ensembles. Ensemble disagreement [Lakshminarayanan et al., 2017] detects OOD by measuring variance across N independently trained models. Key trade-offs vs. resonance:

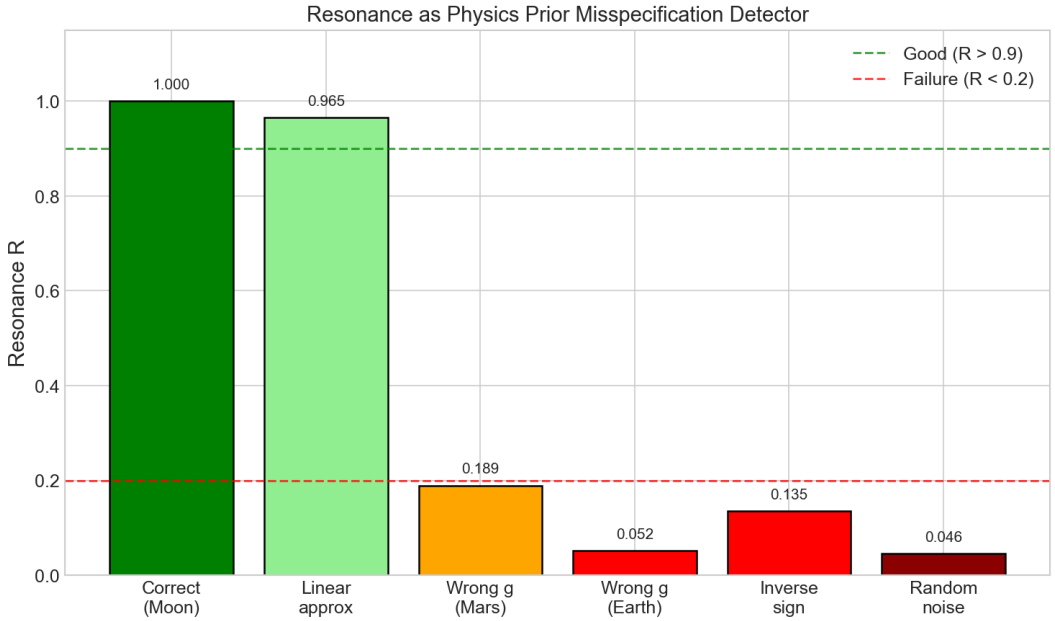

Figure 3: **Resonance as physics prior misspecification detector.** When the model is accurate (trained on Moon, tested on Moon), resonance cleanly separates correct priors ($R > 0.9$, green threshold) from misspecified ones ($R < 0.2$, red threshold). This enables automated prior validation.

Table 6: EKF vs. R-weighted blending under OOD (Earth→Moon).

Method	**Physics**	**MSE**	**OOD Signal?**
EKF	Correct	0.0037	No
EKF	Wrong	0.0222	No (just fails)
R-weighted blending	Correct	0.0023	Yes ($R = 0.86$)
R-weighted blending	Wrong	0.0046	Yes ($R = 1.0$)

When to prefer resonance. Resonance is preferable when: (1) physics structure is known (robotics, control, simulation); (2) computational budget is limited; (3) interpretable trust signal is needed for downstream decisions. Ensembles are preferable when physics is unknown or hard to specify.

MC Dropout. Dropout-based uncertainty [Gal & Ghahramani, 2016] is cheaper than ensembles (1 model, K stochastic passes) but provides epistemic uncertainty only. Resonance provides a *different* signal: model-physics alignment, which specifically detects dynamics shift rather than general unfamiliarity.

9 Discussion

9.1 What's Novel, What's Not

Not novel:

- RBF kernel similarity: $R = \exp(-d^2/\sigma^2)$ is standard
- Weighted averaging: $\hat{y} = \alpha \cdot y_1 + (1 - \alpha) \cdot y_2$ is textbook
- Hybrid physics-ML: combining learned and physics models is established

Table 7: Resonance vs. Deep Ensembles for OOD detection.

Criterion	Deep Ensembles	Resonance
Training cost	$N\times$ baseline	$1\times$ baseline
Inference cost	N forward passes	2 forward passes
OOD signal	Disagreement (learned)	Disagreement (physics)
Domain knowledge	None required	Physics prior Φ required
Failure mode	Confident on OOD	Defaults to physics

Novel combination: Using model-physics *prediction disagreement* as the adaptive weighting signal for test-time blending, without requiring:

- Multiple models (unlike ensembles)
- Gating network training (unlike MoE)
- Labeled OOD examples (unsupervised)

9.2 Role in Adaptive Systems

Resonance provides a **signal** for downstream decisions:

- **MIT (Model Invalidity Test)**: Uses $\langle R \rangle_W < \delta$ as the 3rd pillar for triggering model reconstruction
- **ARH (Adaptive Resonance Horizon)**: Uses R statistics for z-score anomaly detection
- **HHA (Homeostatic Hamiltonian Agent)**: Combines R with stress signal for adaptation control

9.3 Limitations and Scope

1. **Requires accurate physics prior**: The method's benefit is proportional to prior accuracy. When Φ exactly matches test dynamics, blending helps significantly ($+50\%$). When Φ is approximate or wrong, the method safely degrades to baseline. This fundamentally limits applicability to **domains with known, accurate physics equations** (pendulum, simple robotics, orbital mechanics).
2. **Oracle assumption**: Main experiments assume correct physics parameters are known at test time. In practice, parameters may need estimation from initial observations, introducing additional error.
3. **Not applicable to complex environments**: Environments with unknown contact dynamics, unmodeled friction, or high-dimensional interactions (DMC Suite, MuJoCo) generally lack accurate physics priors. For such domains, alternative approaches (ensemble uncertainty, learned adaptation) are more appropriate.
4. **Scale sensitivity**: σ requires calibration (Table 1 shows $\pm 3\times$ range is tolerable). Automatic σ selection remains future work.
5. **Simple environments by design**: We validate on pendulum (3-dim state) because it satisfies the method's requirements. This is not a scalability limitation but a scope constraint: the method applies where physics is known.
6. **No formal guarantees**: We provide empirical evidence, not theoretical bounds on when resonance reliably detects shift.

10 Conclusion

We introduced **resonance-weighted blending**, a simple method for robust world model prediction under distribution shift:

$$R = \exp\left(-\frac{\|f_\theta(x) - \Phi(x)\|^2}{\sigma^2}\right) \tag{4}$$

$$\hat{y} = R \cdot f_\theta(x) + (1 - R) \cdot \Phi(x) \tag{5}$$

Key Results:

- **50.2% MSE reduction** under OOD (Earth→Moon gravity shift)
- **Graceful degradation**: Wrong physics $\Rightarrow R \approx 1 \Rightarrow$ no blending $\Rightarrow$ no harm
- **Automatic OOD signal**: R indicates trust level without labeled anomalies

Scope and Positioning: The method applies to domains with **known, accurate physics priors**—pendulum dynamics, simple robotics, orbital mechanics. For complex environments without exact physics equations (DMC Suite, MuJoCo), the method safely degrades to baseline but provides no benefit. This scope constraint is explicit: resonance-weighted blending is a niche method for physics-rich domains, not a general-purpose solution.

The mathematical components are standard (RBF kernel, weighted average). The contribution is their combination for *unsupervised OOD detection and adaptive blending* using model-physics disagreement as the trust signal, specifically in domains where such priors exist. Companion work (HHA, MIT, ARH) addresses how to act on the resonance signal.

Acknowledgements

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

References

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *NeurIPS*, 2017.

Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991.

Maziar Raissi, Paris Perdikaris, and George E. Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019.

Danijar Hafner et al. Mastering diverse domains through world models. *arXiv:2301.04104*, 2023.

Karl Friston. The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11(2):127–138, 2010.

Eric A. Antonelo, Eduardo Camponogara, and Leo O. Tong. Hybrid state estimation: Integrating physics-informed neural networks with adaptive UKF for dynamic systems. *Electronics*, 13(11):2208, 2024.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, 2016.