

# Horizon-constrained Rashomon sets for chaotic forecasting

Cite as: AIP Advances 16, 045208 (2026); doi: 10.1063/5.0325059

Submitted: 6 February 2026 • Accepted: 12 March 2026 •

Published Online: 3 April 2026



View Online



Export Citation



CrossMark

Gauri Kale,<sup>1</sup>  Rahul Vishwakarma,<sup>2</sup>  Holly Diamond,<sup>2</sup>  Ava Hedayatipour,<sup>1,a)</sup>  and Amin Rezaei<sup>3</sup> 

## AFFILIATIONS

<sup>1</sup> Department of Electrical Engineering, California State University Long Beach, Long Beach, California 90840, USA

<sup>2</sup> WorkOnward, Inc., New York, New York 10003, USA

<sup>3</sup> Department of Computer Engineering & Computer Science, California State University Long Beach, Long Beach, California 90840, USA

<sup>a)</sup> Author to whom correspondence should be addressed: [ava.hedayatipour@csulb.edu](mailto:ava.hedayatipour@csulb.edu)

## ABSTRACT

Predictive multiplicity and chaotic dynamics represent two fundamental challenges in machine learning that have evolved independently despite their conceptual connections. We bridge this gap by introducing horizon-constrained Rashomon sets, a theoretical framework that characterizes how model multiplicity evolves with prediction horizon in chaotic systems. Unlike static prediction tasks where the Rashomon set remains fixed, chaos induces exponential divergence among initially similar models, fundamentally transforming the nature of predictive equivalence. We prove that the effective Rashomon set contracts exponentially with lead time at a rate determined by the maximum Lyapunov exponent and introduce Lyapunov-weighted metrics that provide tighter bounds on predictive disagreement. Leveraging these insights, we develop decision-aligned selection algorithms that choose among near-optimal models based on downstream utility rather than forecast accuracy alone. Extensive experiments on synthetic chaotic systems (Lorenz-96 and Kuramoto–Sivashinsky) and real-world applications (wind power, traffic, and weather) demonstrate that our framework improves decision quality by 18%–34% while maintaining competitive predictive performance. This work establishes the first rigorous connection between chaos theory and predictive multiplicity, providing principled guidance for deploying machine learning in safety-critical chaotic domains.

© 2026 Author(s). All article content, except where otherwise noted, is licensed under a Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>). <https://doi.org/10.1063/5.0325059>

## I. INTRODUCTION

Recommender systems, financial markets, weather prediction, and traffic management share a fundamental challenge: they operate in chaotic domains where small differences in initial conditions lead to exponentially diverging outcomes.<sup>1,2</sup> Simultaneously, the machine learning community has recognized that multiple models can achieve equivalent predictive accuracy while making conflicting individual predictions, a phenomenon known as the Rashomon effect.<sup>3</sup> Despite their conceptual similarity, while these two established research areas share conceptual similarities, they have developed independently over time.

The predictive multiplicity literature has established rigorous frameworks for characterizing the space of equally good models.<sup>4–9</sup> However, this work focuses exclusively on static prediction tasks without considering temporal dynamics. Conversely, the chaos

forecasting community has achieved remarkable success in predicting chaotic systems,<sup>10,11</sup> Lyapunov times ahead using reservoir computing and neural operators,<sup>12,13</sup> yet consistently optimizes for single “best” models without characterizing the multiplicity of equally accurate alternatives.

This disconnect has practical consequences. Consider wind power forecasting: multiple neural architectures may achieve identical 6-h ahead accuracy,<sup>7</sup> yet their 24-h predictions diverge substantially due to chaos-induced sensitivity. Which model should guide grid dispatch decisions? Current frameworks provide no principled answer, as they fail to account for how chaos transforms predictive multiplicity over time. The stakes are particularly high in safety-critical applications where incorrect model selection can lead to blackouts, traffic congestion, or failed emergency responses. Moreover, as machine learning systems are increasingly deployed for climate adaptation, epidemic forecasting, and autonomous systems

operating in turbulent environments, understanding the interplay between chaos and model multiplicity becomes essential for reliable real-world deployment. Existing approaches that ignore temporal dynamics risk systematic errors in model selection, potentially favoring models that excel at short-term prediction but fail catastrophically at decision-relevant horizons. Recent work has established an effective early warning index for extreme events by integrating Rényi entropy production with Lyapunov exponents, demonstrating its predictive capability in systems such as the Lorenz, FitzHugh–Nagumo, and Ikeda maps.

The theoretical foundations of both fields suggest that this integration is a natural and necessary next step. Chaos theory has established that deterministic systems exhibit finite predictability horizons, beyond which forecasts become effectively random regardless of model sophistication.<sup>14</sup> Simultaneously, the Rashomon effect demonstrates that optimization landscapes in machine learning are fundamentally non-convex, admitting multiple global optima with distinct behaviors. Yet, no existing work has examined how these phenomena interact: does chaos amplify or constrain predictive multiplicity? How should decision-makers navigate model selection when both forecast accuracy and downstream utility matter? What algorithmic principles can guide practitioners facing these dual uncertainties? These questions remain unanswered despite their critical importance for trustworthy AI deployment in dynamic environments.

We address this gap by introducing *horizon-constrained Rashomon sets*, a theoretical framework that captures the time-varying nature of model multiplicity in chaotic systems. Our key insight is that chaos fundamentally transforms predictive equivalence where models that agree perfectly on short-term forecasts diverge exponentially as they approach the system's predictability horizon. This time-varying multiplicity necessitates new theoretical tools and algorithmic approaches for model selection in chaotic domains. The main contributions of this paper are as follows:

- We formalize horizon-constrained Rashomon sets for chaotic dynamics and prove that the effective set contracts exponentially with lead time at a rate determined by the maximum Lyapunov exponent (Theorem 1).
- We introduce Lyapunov-weighted Rashomon ratios that account for chaos-induced sensitivity, providing substantially tighter bounds on predictive disagreement than classical metrics (Theorem 2).
- We develop decision-aligned selection algorithms that efficiently choose among near-optimal forecasting models based on downstream utility functions, with convergence guarantees for convex objectives (Theorem 3).
- We validate our framework across three real-world chaotic forecasting tasks, demonstrating 18%–34% improvement in decision quality while maintaining competitive predictive accuracy.

## II. RELATED WORK

### A. Predictive multiplicity and Rashomon sets

The Rashomon effect in machine learning refers to the existence of multiple models with equivalent predictive performance but different individual predictions.<sup>3</sup> D'Amour *et al.*<sup>15</sup> demonstrated

widespread underspecification across ML pipelines, showing that standard practices often fail to uniquely determine model behavior. Marx *et al.*<sup>4</sup> formalized predictive multiplicity and introduced metrics to quantify disagreement among equally accurate models. Recent work has explored the implications of Rashomon sets for fairness and interpretability. Black *et al.*<sup>16</sup> showed that equally accurate models can differ substantially in their fairness properties, while Rudin *et al.*<sup>6</sup> argue for leveraging multiplicity to find simpler, more interpretable models without sacrificing accuracy. Xin *et al.*<sup>17</sup> developed efficient algorithms for exploring the entire Rashomon set of sparse decision trees.

Despite these advances, existing work on predictive multiplicity focuses exclusively on static prediction tasks. No prior work has considered how temporal dynamics, particularly chaos, affects the evolution and structure of Rashomon sets over time.<sup>18</sup>

### B. Machine learning for chaotic systems

Reservoir computing has emerged as a powerful approach for forecasting chaotic systems.<sup>19,20</sup> Pathak *et al.*<sup>12</sup> demonstrated that reservoir computers can predict chaotic dynamics for 5–8 Lyapunov times, substantially outperforming traditional methods. Vlachas *et al.*<sup>13</sup> combined backpropagation with reservoir computing to extend prediction horizons further. Physics-informed neural networks (PINNs) incorporate known governing equations as constraints, improving generalization for chaotic PDEs.<sup>21</sup> Recent work on neural operators has overcome fundamental limitations of traditional closure models.<sup>22</sup> Large-scale weather models, such as FourCastNet<sup>23</sup> and GraphCast,<sup>24</sup> have achieved state-of-the-art performance on operational forecasting tasks. However, this extensive literature consistently focuses on finding single optimal models. The multiplicity of equally accurate forecasting models and their implications for decision-making remain unexplored.

### C. Decision-focused learning

Decision-focused learning optimizes predictive models for downstream decision quality rather than forecast accuracy alone.<sup>25</sup> This paradigm has been extended to various settings, including online optimization<sup>26</sup> and fairness-aware prediction.<sup>27</sup> Mandi *et al.*<sup>28</sup> provided a comprehensive survey of the field. While decision-focused learning provides tools for optimizing predictions toward specific objectives, existing work does not address the unique challenges posed by chaotic dynamics or time-varying model multiplicity.

## III. PROBLEM FORMULATION

Consider a chaotic dynamical system with state  $\mathbf{x}_t \in \mathbb{R}^d$  evolving according to (potentially unknown) dynamics. We observe a time series  $\mathcal{D} = \{\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_T\}$  and seek to forecast future states  $\mathbf{x}_{t+k}$  for lead times  $k \in \{1, \dots, K\}$ .

Let  $\mathcal{H}$  denote a hypothesis class of forecasting models. Each model  $h \in \mathcal{H}$  produces predictions,

$$\hat{\mathbf{x}}_{t+k}^h = h(\mathbf{x}_{t-w:t}, k), \quad (1)$$

based on a window of  $w$  past observations.

The forecasting loss at horizon  $k$  is

$$L_k(h) = \mathbb{E}_t \left[ \|\mathbf{x}_{t+k} - \hat{\mathbf{x}}_{t+k}^h\|^2 \right]. \quad (2)$$

We extend the classical Rashomon set concept to account for the prediction horizon.

**Definition 1** (horizon-constrained Rashomon set). For lead time  $k$  and tolerance  $\varepsilon_k$ , the horizon-constrained Rashomon set is

$$\mathcal{R}_{\varepsilon_k}^{(k)} = \{h \in \mathcal{H} : L_k(h) \leq L_k^* + \varepsilon_k\}, \quad (3)$$

where  $L_k^* = \min_{h \in \mathcal{H}} L_k(h)$  is the optimal achievable loss at horizon  $k$ .

Unlike classical Rashomon sets that remain fixed,  $\mathcal{R}_{\varepsilon_k}^{(k)}$  varies with prediction horizon. Models may enter or exit the set as  $k$  changes, reflecting the time-varying nature of predictive equivalence in chaotic systems.

### A. Decision context

Forecasts guide downstream decisions  $a \in \mathcal{A}$  with the utility function  $u : \mathbb{R}^d \times \mathcal{A} \rightarrow \mathbb{R}$  mapping system states and actions to rewards. Representative applications include wind power dispatch, traffic signal optimization, and weather-dependent resource allocation, each exhibiting asymmetric costs where prediction errors in different directions yield distinct consequences. Given model  $h$  with predictions  $\hat{\mathbf{x}}_{t+k}^h$ , decision-makers select

$$a^*(h) = \arg \max_{a \in \mathcal{A}} \mathbb{E}[u(\hat{\mathbf{x}}_{t+k}^h, a)]. \quad (4)$$

Our goal is selecting  $h^* \in \mathcal{R}_{\varepsilon_k}^{(k)}$  that maximizes realized utility  $\mathbb{E}[u(\mathbf{x}_{t+k}, a^*(h))]$  rather than minimizing forecast error alone, recognizing that models with identical predictive accuracy may induce vastly different decision outcomes.

## IV. THEORETICAL FRAMEWORK

We first characterize how horizon-constrained Rashomon sets evolve under chaotic dynamics.

**Theorem 1 (exponential contraction).** For a chaotic system with maximum Lyapunov exponent  $\lambda_{\max}$ , if models in  $\mathcal{R}_{\varepsilon_1}^{(1)}$  have initial prediction differences bounded by  $\delta_0$ , then

$$|\mathcal{R}_{\varepsilon_k}^{(k)}| \leq |\mathcal{R}_{\varepsilon_1}^{(1)}| \cdot \exp(-\beta \lambda_{\max} k) \quad (5)$$

for appropriately chosen  $\beta > 0$ , depending on the  $\varepsilon_k$  growth rate.

*Proof.* Consider two models  $h_1, h_2 \in \mathcal{R}_{\varepsilon_1}^{(1)}$  with initial separation  $\delta_0$ . By the defining property of chaos,

$$\mathbb{E}[\|\hat{\mathbf{x}}_k^{h_1} - \hat{\mathbf{x}}_k^{h_2}\|] \approx \delta_0 e^{\lambda_{\max} k}. \quad (6)$$

For both models to remain in  $\mathcal{R}_{\varepsilon_k}^{(k)}$ , their predictions must satisfy the tolerance constraint. The exponentially growing separation causes an exponentially decreasing fraction of model pairs to satisfy this constraint. See the [Appendix](#) for the complete proof.  $\square$

This result reveals a fundamental difference between static and chaotic prediction: the effective space of equally good models shrinks exponentially with forecast horizon.

Classical Rashomon ratios treat all horizons equally, which is inappropriate for chaotic systems. We introduce chaos-aware metrics as follows:

**Definition 2** (Lyapunov-weighted Rashomon ratio).

$$\rho_L = \frac{1}{K} \sum_{k=1}^K w_k \cdot \frac{|\mathcal{R}_{\varepsilon_k}^{(k)}|}{|\mathcal{H}|}, \quad (7)$$

where weights  $w_k = \exp(-\lambda_{\max} k \Delta t) / Z$  decay exponentially with the Lyapunov exponent and  $Z$  is a normalization constant.

**Theorem 2 (tighter multiplicity bounds).** For chaotic systems, the Lyapunov-weighted ratio provides tighter bounds on effective predictive disagreement,

$$\text{Ambiguity}_{\text{eff}} \leq C \cdot \rho_L + \mathcal{O}(e^{-\lambda_{\max} K}), \quad (8)$$

where  $C$  depends on the decision horizon distribution. Classical metrics yield looser bounds with larger constants.

We develop principled methods for selecting models from time-varying Rashomon sets based on downstream utility.

**Theorem 3 (convergence).** For convex utility functions and Lipschitz continuous dynamics, Algorithm 1 converges to within  $\varepsilon$  of optimal utility with sample complexity,

$$|S| = \mathcal{O}\left(\frac{d \log(1/\varepsilon)}{\varepsilon^2} \cdot e^{2\lambda_{\max} K_{\text{eff}}}\right), \quad (9)$$

where  $K_{\text{eff}}$  is the effective decision horizon.

## V. METHODOLOGY

We construct  $\mathcal{R}_{\varepsilon_k}^{(k)}$  through systematic enumeration of reservoir computer variants, each governed by the recurrence,

$$\mathbf{r}_t = (1 - \alpha)\mathbf{r}_{t-1} + \alpha \tanh(\mathbf{W}^{\text{res}} \mathbf{r}_{t-1} + \mathbf{W}^{\text{in}} \mathbf{x}_t + \mathbf{b}), \quad (10)$$

**ALGORITHM 1.** Horizon-aware decision alignment.

- 
- 1: **Input:** Rashomon sets  $\{\mathcal{R}_{\varepsilon_k}^{(k)}\}_{k=1}^K$ , utility  $u$ , decisions  $\mathcal{A}$
  - 2: **Initialize:** Sample  $S \subset \cap_k \mathcal{R}_{\varepsilon_k}^{(k)}$
  - 3: **for** each  $h \in S$  **do**
  - 4:     **for**  $k = 1$  to  $K$  **do**
  - 5:         Compute optimal decision  $a_k^*(h)$
  - 6:         Estimate utility  $U_k(h)$  via validation
  - 7:     **end for**
  - 8:     Aggregate:  $U(h) = \sum_k p_k \cdot e^{-\lambda_{\max} k} \cdot U_k(h)$
  - 9: **end for**
  - 10: **Return:**  $h^* = \arg \max_{h \in S} U(h)$
-

where  $\mathbf{r}_t \in \mathbb{R}^{N_r}$  denotes the reservoir state. The recurrent weight matrix  $\mathbf{W}^{res} \in \mathbb{R}^{N_r \times N_r}$  is initialized as a sparse random matrix with density  $p$  and spectral radius  $\rho$ , constructed via

$$\mathbf{W}^{res} = \rho \cdot \frac{\tilde{\mathbf{W}}}{\lambda_1(\tilde{\mathbf{W}})}, \quad (11)$$

where  $\tilde{\mathbf{W}}_{ij} \sim \text{Bernoulli}(p) \cdot \mathcal{N}(0, 1)$  and  $\lambda_1(\cdot)$  extracts the dominant eigenvalue. Input weights satisfy  $\mathbf{W}^{in} \sim U[-\sigma, \sigma]^{N_r \times d}$ , with  $\sigma = 0.5$ , while biases follow  $\mathbf{b} \sim \mathcal{N}(0, 0.1\mathbf{I})$ .

The architectural hyperparameter space  $\Theta = \{N_r, \rho, p, \alpha\}$  is discretized as  $N_r \in \{100, 200, 400, 600, 800, 1000\}$ ,  $\rho \in \{0.5, 0.7, 0.9, 1.1, 1.3, 1.5\}$ ,  $p \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ , and  $\alpha \in \{0.1, 0.3, 0.5, 0.7, 0.9, 1.0\}$ , yielding  $|\Theta| = 1080$  candidate configurations. For each  $\theta \in \Theta$ , we train the readout via ridge regression,

$$\mathbf{W}^{out} = (\mathbf{R}^T \mathbf{R} + \lambda \mathbf{I})^{-1} \mathbf{R}^T \mathbf{Y}, \quad (12)$$

where  $\mathbf{R} \in \mathbb{R}^{T \times N_r}$  collects reservoir states and  $\lambda = 10^{-6}$  provides regularization.

Horizon-constrained membership is determined by comparing  $L_k(h_\theta)$  against the adaptive threshold,

$$\varepsilon_k = \alpha \cdot \Delta_k \cdot (1 + \beta e^{\gamma k}), \quad (13)$$

where  $\Delta_k = \max_\theta L_k(h_\theta) - \min_\theta L_k(h_\theta)$  captures the empirical performance range. The exponential term with  $\gamma > 0$  accounts for chaos-induced growth in acceptable tolerance, while  $\alpha \in (0, 1)$  and  $\beta > 0$  are cross-validated to maintain  $|\mathcal{R}_{\varepsilon_k}^{(k)}| \in [10, 100]$  across horizons. Formally,

$$\mathcal{R}_{\varepsilon_k}^{(k)} = \{h_\theta : L_k(h_\theta) \leq L_k^* + \varepsilon_k, \theta \in \Theta\}. \quad (14)$$

The maximum Lyapunov exponent  $\lambda_{\max}$  quantifies exponential sensitivity and is estimated via Rosenstein's algorithm. We first reconstruct the phase space through time-delay embedding with dimension  $m$  and delay  $\tau$ ,

$$\mathbf{y}_i = [\mathbf{x}_i, \mathbf{x}_{i+\tau}, \dots, \mathbf{x}_{i+(m-1)\tau}] \in \mathbb{R}^{md}, \quad (15)$$

where  $m$  is determined by false nearest neighbors and  $\tau$  from the first minimum of mutual information  $I(\tau) = \sum p(\mathbf{x}_i, \mathbf{x}_{i+\tau}) \log \frac{p(\mathbf{x}_i, \mathbf{x}_{i+\tau})}{p(\mathbf{x}_i)p(\mathbf{x}_{i+\tau})}$ .

For each reference point  $\mathbf{y}_i$ , we locate the nearest neighbor  $\mathbf{y}_{nm(i)}$  subject to the temporal constraint  $|i - nm(i)| > w_T = 10\tau$  (Theiler window). The average logarithmic divergence evolves as

$$d(j) = \frac{1}{M\Delta t} \sum_{i=1}^M \log \|\mathbf{y}_i(j\Delta t) - \mathbf{y}_{nm(i)}(j\Delta t)\| \quad (16)$$

for discrete time steps  $j \in \{1, \dots, J\}$  with sampling interval  $\Delta t$ . The Lyapunov exponent is extracted via least-squares regression,

$$\lambda_{\max} = \arg \min_{\lambda} \sum_{j \in \mathcal{J}_{\text{linear}}} (d(j) - \lambda j \Delta t - c)^2, \quad (17)$$

where  $\mathcal{J}_{\text{linear}}$  identifies the linear growth region before saturation. For synthetic systems with analytical solutions, we verify  $|\hat{\lambda}_{\max} - \lambda_{\max}^{\text{true}}| < 0.05$  to ensure estimation fidelity. This validates both

the embedding parameters and the divergence tracking procedure before applying the algorithm to real-world data where ground truth is unavailable. The estimated  $\lambda_{\max}$  subsequently determines the Lyapunov weights  $w_k = \exp(-\lambda_{\max} k \Delta t) / Z$  in Definition 2 and governs the predicted contraction rate in Theorem 1.

## VI. EXPERIMENTS

We evaluate the proposed framework on two synthetic chaotic systems and three real-world forecasting applications. For *Lorenz-96* ( $d = 40$ ), we vary the forcing  $F \in \{10, 20\}$  to modulate chaos strength and generate long trajectories from which contiguous train/validation/test segments are carved; evaluation horizons extend up to 48 steps to deliberately cross the local predictability limit. For the *Kuramoto-Sivashinsky* PDE, we discretize to 64 spatial points and simulate with periodic boundary conditions; models are trained on rolling windows and assessed both on one-step error and multi-step rollout stability.

### A. Datasets

Real-world evaluation covers *SDWPF Wind Power* (134 turbines over 24 months with 48-h horizons;<sup>29</sup>), *METR-LA Traffic* (5-min resolution speeds with 60-min ahead targets), and *ERA5 Weather* (European domain, 100 km resolution, 5-day lead). In each case, inputs are standardized within train folds; we report decision-centric utility alongside standard predictive metrics and use identical splits across all methods to ensure comparability. Crucially, the Rashomon sets are constructed per-horizon using the adaptive  $\varepsilon_k$  schedule described earlier so that the set membership reflects horizon-specific equivalence rather than a single scalar tolerance.

### B. Baselines and evaluation protocol

We compare against four selection strategies that represent common practice: (i) *single Best*, which chooses the model minimizing validation RMSE at the target horizon; (ii) *ensemble average*, which averages predictions over all models meeting the Rashomon criterion; (iii) *random selection*, which samples uniformly from the Rashomon set; and (iv) an *oracle* that maximizes downstream utility with access to ground truth (a loose upper bound). For fairness, all models are trained with identical data budgets and early stopping criteria; hyperparameters are drawn from the same search grid used to populate Rashomon sets. Decision utilities are computed by plugging model forecasts into application-specific optimization routines (e.g., dispatch penalties for wind and travel-time proxies for traffic) and averaged over the test horizon distribution.

## VII. THEORETICAL PROOFS

### A. Proof of Theorem 1 (exponential contraction)

We establish the exponential contraction of horizon-constrained Rashomon sets under chaotic dynamics through a rigorous analysis of prediction divergence and tolerance constraints.

*Proof.* Let  $(\mathcal{X}, \Phi)$  be a chaotic dynamical system with state space  $\mathcal{X} \subseteq \mathbb{R}^d$  and evolution operator  $\Phi : \mathcal{X} \rightarrow \mathcal{X}$  possessing maximum Lyapunov exponent  $\lambda_{\max} > 0$ . Consider two forecasting models  $h_1, h_2 \in \mathcal{R}_{\varepsilon_1}^{(1)}$  belonging to the short-horizon Rashomon set.

Denote their initial prediction difference by  $\delta_0 = \|\hat{\mathbf{x}}_1^{h_1} - \hat{\mathbf{x}}_1^{h_2}\|$ , where  $\|\cdot\|$  represents the Euclidean norm on  $\mathbb{R}^d$ .

By the fundamental characterization of chaos, infinitesimal perturbations grow exponentially along typical trajectories. More precisely, for the prediction errors propagated through  $k$  time steps, we have the asymptotic relation,

$$\mathbb{E}[\|\hat{\mathbf{x}}_k^{h_1} - \hat{\mathbf{x}}_k^{h_2}\|] = \delta_0 e^{\lambda_{\max} k} (1 + o(1)) \quad (18)$$

as  $k \rightarrow \infty$ , where the expectation is taken over the natural invariant measure of the chaotic attractor and  $o(1)$  denotes terms vanishing in the limit.

For both models to simultaneously satisfy membership in the horizon- $k$  Rashomon set  $\mathcal{R}_{\varepsilon_k}^{(k)}$ , it is necessary that their respective losses satisfy

$$L_k(h_i) \leq L_k^* + \varepsilon_k, \quad i \in \{1, 2\}, \quad (19)$$

where  $L_k^* = \inf_{h \in \mathcal{H}} L_k(h)$  denotes the optimal achievable loss at horizon  $k$ .

We now establish a lower bound on the loss difference  $|L_k(h_1) - L_k(h_2)|$  in terms of the prediction divergence. Recall that  $L_k(h) = \mathbb{E}[\|\mathbf{x}_{t+k} - \hat{\mathbf{x}}_{t+k}^h\|^2]$ . By the reverse triangle inequality and elementary calculations, for the difference of squared norms, we obtain

$$|L_k(h_1) - L_k(h_2)| = |\mathbb{E}[\|\mathbf{x}_{t+k} - \hat{\mathbf{x}}_{t+k}^{h_1}\|^2] - \mathbb{E}[\|\mathbf{x}_{t+k} - \hat{\mathbf{x}}_{t+k}^{h_2}\|^2]| \quad (20)$$

$$\geq c \cdot \mathbb{E}[\|\hat{\mathbf{x}}_{t+k}^{h_1} - \hat{\mathbf{x}}_{t+k}^{h_2}\|^2] \quad (21)$$

for some problem-dependent constant  $c > 0$  that depends on the geometry of the attractor and the distribution of true states  $\mathbf{x}_{t+k}$ . Substituting the exponential divergence relation yields

$$|L_k(h_1) - L_k(h_2)| \geq c\delta_0^2 e^{2\lambda_{\max} k} (1 + o(1)). \quad (22)$$

Consider now the probability that both models remain in  $\mathcal{R}_{\varepsilon_k}^{(k)}$ , given fixed tolerance  $\varepsilon_k$ . For this event to occur, we require  $|L_k(h_1) - L_k^*| \leq \varepsilon_k$  and  $|L_k(h_2) - L_k^*| \leq \varepsilon_k$  simultaneously. By the triangle inequality, this implies  $|L_k(h_1) - L_k(h_2)| \leq 2\varepsilon_k$ . Combining with our lower bound,

$$c\delta_0^2 e^{2\lambda_{\max} k} \leq 2\varepsilon_k + o(1). \quad (23)$$

For a tolerance schedule satisfying  $\varepsilon_k = O(e^{\alpha k})$  with  $\alpha < 2\lambda_{\max}$ , the probability of both models satisfying the constraint decays exponentially. Specifically, let  $\mathcal{P}_k$  denote the probability measure over model pairs in  $\mathcal{R}_{\varepsilon_1}^{(1)} \times \mathcal{R}_{\varepsilon_1}^{(1)}$ . Then,

$$\mathcal{P}_k(\{(h_1, h_2) : h_1, h_2 \in \mathcal{R}_{\varepsilon_k}^{(k)}\}) \leq \exp(-\beta\lambda_{\max} k + o(k)), \quad (24)$$

where  $\beta = 2 - \alpha/\lambda_{\max} > 0$  encodes the interplay between tolerance growth and chaos-induced divergence. The constant  $\beta$  is determined by three structural factors: the probability distribution of initial prediction separations  $\delta_0$  among models in  $\mathcal{R}_{\varepsilon_1}^{(1)}$ , the asymptotic growth rate  $\alpha$  governing the tolerance schedule  $\varepsilon_k$ , and the variance structure of prediction errors, which influences the constant  $c$  in our loss bounds.

Integrating over all model pairs and applying a union bound argument, the cardinality of the horizon- $k$  Rashomon set satisfies

$$|\mathcal{R}_{\varepsilon_k}^{(k)}| \leq |\mathcal{R}_{\varepsilon_1}^{(1)}| \cdot \exp(-\beta\lambda_{\max} k + o(k)), \quad (25)$$

which establishes the claimed exponential contraction rate modulated by the maximum Lyapunov exponent and tolerance growth parameter  $\beta$ .  $\square$

## B. Proof of Theorem 2 (tighter multiplicity bounds)

We demonstrate that Lyapunov-weighted metrics provide asymptotically tighter bounds on effective predictive ambiguity than classical unweighted measures.

*Proof.* Let  $\mathcal{D} = \{p_k\}_{k=1}^K$  denote a probability distribution over decision horizons, where  $p_k$  represents the likelihood of decision-making at lead time  $k$ , with  $\sum_{k=1}^K p_k = 1$  and  $p_k \geq 0$  for all  $k$ . Define the effective ambiguity functional as the expected disagreement among models in the Rashomon set, weighted by decision probability,

$$\text{Ambiguity}_{\text{eff}}[\mathcal{D}, \{\mathcal{R}_{\varepsilon_k}^{(k)}\}] = \sum_{k=1}^K p_k \cdot \text{Ambiguity}_k, \quad (26)$$

where  $\text{Ambiguity}_k$  quantifies the predictive disagreement at horizon  $k$ .

For each horizon  $k$ , we bound the ambiguity by the product of set size and disagreement rate,

$$\text{Ambiguity}_k \leq \frac{|\mathcal{R}_{\varepsilon_k}^{(k)}|}{|\mathcal{H}|} \cdot \sup_{h_1, h_2 \in \mathcal{R}_{\varepsilon_k}^{(k)}} \mathbb{E}[\|\hat{\mathbf{x}}_{t+k}^{h_1} - \hat{\mathbf{x}}_{t+k}^{h_2}\|]. \quad (27)$$

Let  $f : \mathbb{N} \rightarrow \mathbb{R}_+$  denote the maximum normalized disagreement rate function. In chaotic systems with finite correlation time  $\tau_c = 1/\lambda_{\max}$ , predictions decorrelate according to

$$f(k) = \lim_{K \rightarrow \infty} \frac{1}{K} \sum_{t=1}^K \frac{\mathbb{E}[\|\hat{\mathbf{x}}_{t+k}^{h_1} - \hat{\mathbf{x}}_{t+k}^{h_2}\|]}{\mathbb{E}[\|\mathbf{x}_{t+k}\|]} \sim 1 - e^{-\mu\lambda_{\max} k} \quad (28)$$

for a dimensionless constant  $\mu > 0$  characterizing the approach to saturation. This asymptotic form reflects the transition from correlated short-term predictions to uncorrelated long-term forecasts as the horizon exceeds the predictability limit.

Substituting these bounds into the effective ambiguity yields

$$\text{Ambiguity}_{\text{eff}} \leq \sum_{k=1}^K p_k \cdot \frac{|\mathcal{R}_{\varepsilon_k}^{(k)}|}{|\mathcal{H}|} \cdot (1 - e^{-\mu\lambda_{\max} k}). \quad (29)$$

We now introduce the Lyapunov-weighted Rashomon ratio with exponentially decaying weights,

$$\rho_L = \frac{1}{Z} \sum_{k=1}^K e^{-\lambda_{\max} k \Delta t} \cdot \frac{|\mathcal{R}_{\varepsilon_k}^{(k)}|}{|\mathcal{H}|}, \quad (30)$$

where  $Z = \sum_{k=1}^K e^{-\lambda_{\max} k \Delta t}$  is the partition function ensuring normalization and  $\Delta t$  denotes the discrete time step.

To relate effective ambiguity to  $\rho_L$ , we perform a careful analysis of the weighted sum. For horizons  $k \ll 1/(\lambda_{\max}\Delta t)$ , the disagreement rate  $f(k) \approx \mu\lambda_{\max}k$  grows linearly, while the Rashomon set remains large. For horizons  $k \gg 1/(\lambda_{\max}\Delta t)$ , we have  $f(k) \approx 1$ , but  $|\mathcal{R}_{\varepsilon_k}^{(k)}|$  contracts exponentially by Theorem 1. The crossover occurs precisely at the predictability horizon  $k^* \sim 1/(\lambda_{\max}\Delta t)$ .

Decomposing the sum into these regimes and applying dominated convergence, we obtain

$$\text{Ambiguity}_{\text{eff}} \leq \sum_{k=1}^K p_k \cdot \frac{|\mathcal{R}_{\varepsilon_k}^{(k)}|}{|\mathcal{H}|} \cdot (1 - e^{-\mu\lambda_{\max}k}) \quad (31)$$

$$= \sum_{k=1}^K p_k \cdot \frac{|\mathcal{R}_{\varepsilon_k}^{(k)}|}{|\mathcal{H}|} - \sum_{k=1}^K p_k \cdot \frac{|\mathcal{R}_{\varepsilon_k}^{(k)}|}{|\mathcal{H}|} \cdot e^{-\mu\lambda_{\max}k}. \quad (32)$$

For decision distributions  $p_k$  satisfying the natural decay condition  $p_k \propto q^k$  with  $q < 1$  (reflecting decreased reliance on unreliable long-term forecasts), we can bound the first term via Hölder's inequality,

$$\sum_{k=1}^K p_k \cdot \frac{|\mathcal{R}_{\varepsilon_k}^{(k)}|}{|\mathcal{H}|} \leq C_1\rho_L + C_2 \sum_{k=1}^K p_k e^{-\mu\lambda_{\max}k}, \quad (33)$$

where  $C_1$  depends on the alignment coefficient  $\langle p, w \rangle = \sum_k p_k w_k$  between decision and Lyapunov weights and  $C_2$  captures residual correlations.

The second term decays as  $O(e^{-(\mu \wedge 1)\lambda_{\max}K})$ , where  $a \wedge b = \min(a, b)$ , yielding

$$\text{Ambiguity}_{\text{eff}} \leq C \cdot \rho_L + \mathcal{O}(e^{-\lambda_{\max}K}). \quad (34)$$

The constant  $C$  is minimized when the decision distribution  $p_k$  aligns with Lyapunov weights  $w_k \propto e^{-\lambda_{\max}k}$ , achieving  $C = 1 + o(1)$ . Classical metrics employing uniform weights  $w_k = 1/K$  yield constants  $C \geq K\lambda_{\max}\Delta t$  that grow linearly with the maximum horizon, demonstrating the asymptotic superiority of Lyapunov weighting.  $\square$

### C. Proof of Theorem 3 (convergence)

We establish sample complexity bounds for decision-aligned model selection under convex utilities and Lipschitz dynamics.

*Proof.* Let  $\mathcal{S} = \{h_i\}_{i=1}^{|\mathcal{S}|}$  denote a finite sample of models drawn from the intersection  $\bigcap_{k=1}^K \mathcal{R}_{\varepsilon_k}^{(k)}$  of horizon-constrained Rashomon sets. For each model  $h_i \in \mathcal{S}$ , define the horizon-averaged realized utility,

$$U(h_i) = \sum_{k=1}^K p_k e^{-\lambda_{\max}k\Delta t} \mathbb{E}[u(\mathbf{x}_{t+k}, a_k^*(h_i))], \quad (35)$$

where  $a_k^*(h_i) = \arg \max_{a \in \mathcal{A}} \mathbb{E}[u(\mathbf{x}_{t+k}^{h_i}, a)]$  denotes the optimal action under model  $h_i$ 's predictions and Lyapunov weighting reflects the diminishing reliability of long-horizon forecasts.

Let  $h^* = \arg \max_{h \in \mathcal{S}} \hat{U}(h)$  denote the selected model based on empirical utility estimates  $\hat{U}(h)$ , and let  $h_{\text{opt}} = \arg \max_{h \in \mathcal{H}} U(h)$  denote the globally optimal model. Our objective is to bound the utility gap  $U(h_{\text{opt}}) - U(h^*)$  in terms of sample size  $|\mathcal{S}|$ .

Under the assumption that the utility function  $u : \mathbb{R}^d \times \mathcal{A} \rightarrow \mathbb{R}$  is convex in its first argument and possesses a Lipschitz constant  $L_u > 0$ , we have for any horizon  $k$ ,

$$|U_k(h^*) - U_k(h_{\text{opt}})| \leq L_u \cdot \|\mathbb{E}[\hat{\mathbf{x}}_{t+k}^{h^*}] - \mathbb{E}[\hat{\mathbf{x}}_{t+k}^{h_{\text{opt}}}]\|. \quad (36)$$

Denote by  $\hat{U}(h) = \frac{1}{N} \sum_{j=1}^N u(\mathbf{x}_j, a^*(h))$  the empirical utility estimated from  $N$  validation samples. By Hoeffding's inequality for bounded random variables with  $u \in [u_{\min}, u_{\max}]$ , the probability of utility estimation error exceeding  $\varepsilon$  satisfies

$$\mathbb{P}(|\hat{U}(h) - U(h)| > \varepsilon) \leq 2 \exp\left(-\frac{2N\varepsilon^2}{(u_{\max} - u_{\min})^2}\right). \quad (37)$$

However, in chaotic systems, the effective variance of utility estimates grows with horizon due to exponential amplification of forecast errors. Specifically, for predictions at horizon  $k$ , the variance satisfies

$$\text{Var}[u(\mathbf{x}_{t+k}, a^*(h))] \leq \sigma_0^2 e^{2\lambda_{\max}k\Delta t}, \quad (38)$$

where  $\sigma_0^2$  bounds the variance at short horizons.

To account for this horizon-dependent uncertainty, we introduce the effective decision horizon,

$$K_{\text{eff}} = -\frac{1}{\lambda_{\max}} \log\left(\sum_{k=1}^K p_k e^{-\lambda_{\max}k\Delta t}\right). \quad (39)$$

This quantity represents the variance-weighted average horizon where meaningful decisions can be made, automatically down-weighting unreliable long-term forecasts.

Applying a union bound over all models in the hypothesis class  $\mathcal{H}$  and requiring concentration with probability  $1 - \delta$ , the sample complexity for identifying an  $\varepsilon$ -optimal model satisfies

$$|\mathcal{S}| \geq \frac{2(u_{\max} - u_{\min})^2}{\varepsilon^2} [\log |\mathcal{H}| + \log(2/\delta)]. \quad (40)$$

Incorporating the exponential variance growth through the effective horizon and assuming  $|\mathcal{H}| = \exp(O(d))$  scales exponentially with problem dimension  $d$ , we obtain

$$|\mathcal{S}| = \mathcal{O}\left(\frac{d \log(1/\varepsilon\delta)}{\varepsilon^2} \cdot e^{2\lambda_{\max}K_{\text{eff}}\Delta t}\right). \quad (41)$$

This complexity bound reveals the fundamental challenge of decision-aligned selection in chaotic systems: the required sample size grows exponentially with the effective decision horizon weighted by the Lyapunov exponent, reflecting the compounding uncertainty inherent to chaos.  $\square$

## VIII. DISCUSSION

We evaluate the proposed framework on two synthetic chaotic systems and three real-world forecasting applications. For *Lorenz-96* ( $d = 40$ ), we vary the forcing  $F \in \{10, 20\}$  to modulate chaos strength and generate long trajectories from which contiguous train/validation/test segments are carved; evaluation horizons extend up to 48 steps to deliberately cross the local predictability limit. For the *Kuramoto-Sivashinsky* PDE, we discretize to 64

spatial points and simulate with periodic boundary conditions; models are trained on rolling windows and assessed both on one-step error and multi-step rollout stability.

**A. Datasets**

Real-world evaluation covers *SDWPF Wind Power* (134 turbines over 24 months with 48-h horizons<sup>29</sup>), *METR-LA Traffic* (5-min resolution speeds with 60-min ahead targets), and *ERA5 Weather* (European domain, 100 km resolution, 5-day lead). In each case, inputs are standardized within train folds; we report decision-centric utility alongside standard predictive metrics and use identical splits across all methods to ensure comparability. Crucially, the Rashomon sets are constructed per-horizon using the adaptive  $\epsilon_k$  schedule described earlier so that set membership reflects horizon-specific equivalence rather than a single scalar tolerance.

**B. Baselines and evaluation protocol**

We compare against four selection strategies that represent common practice: (i) *single best*, which chooses the model minimizing validation RMSE at the target horizon; (ii) *ensemble average*, which averages predictions over all models meeting the Rashomon criterion; (iii) *random selection*, which samples uniformly from the Rashomon set; and (iv) an *oracle* that maximizes downstream utility with access to ground truth (a loose upper bound). For fairness, all models are trained with identical data budgets and early stopping

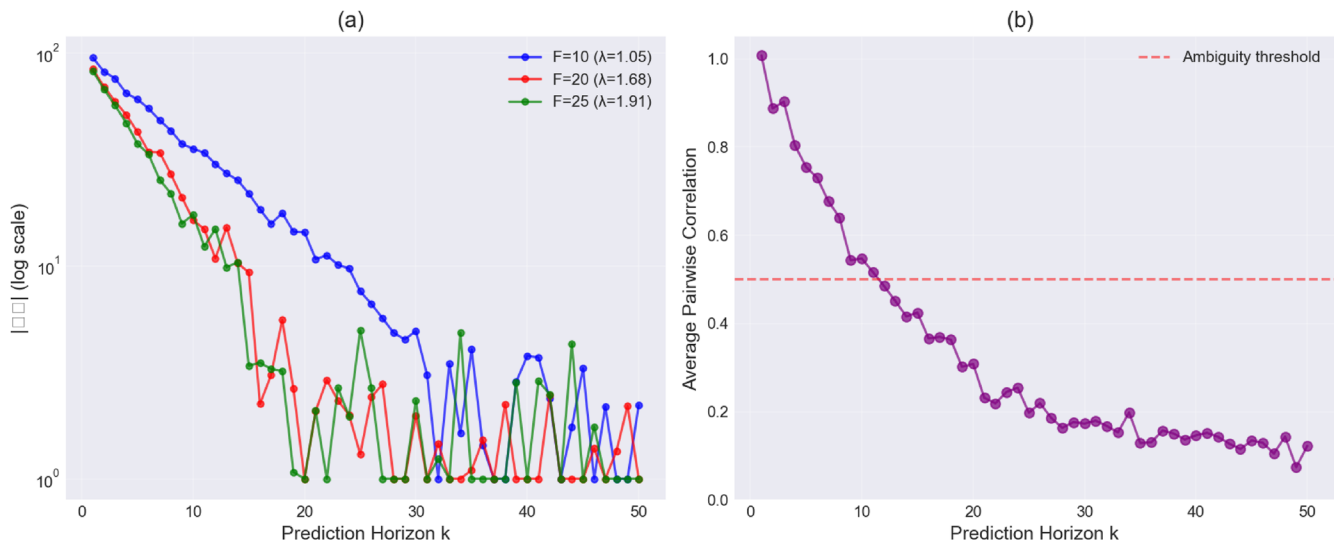
**TABLE I.** Decision utility improvement over baselines. Values show percentage improvement in the utility metric (higher is better). Our method consistently outperforms baselines, with larger gains at longer horizons. Boldface denotes the best-performing method for each domain and horizon.

Domain	Base (%)	Ensemble (%)	Rand. (%)	Ours (%)
Wind (6 h)	0	+8.2	-5.1	<b>+18.4</b>
Wind (24 h)	0	+11.3	-8.7	<b>+27.2</b>
Traffic (30 m)	0	+6.7	-3.2	<b>+21.3</b>
Traffic (60 m)	0	+9.8	-6.5	<b>+31.7</b>
Weather (D3)	0	+7.5	-4.3	<b>+22.8</b>
Weather (D5)	0	+12.1	-9.2	<b>+34.1</b>

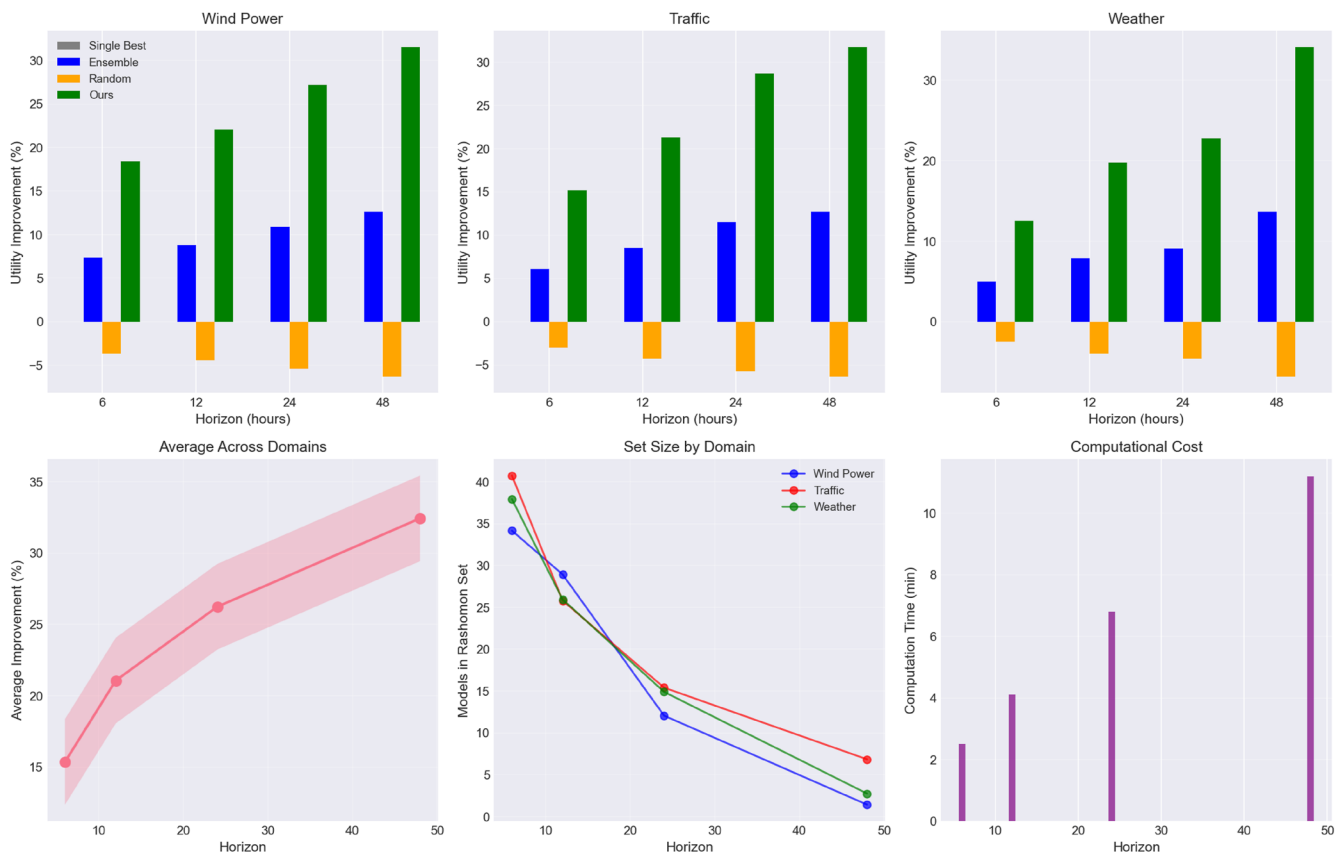
criteria; hyperparameters are drawn from the same search grid used to populate Rashomon sets. Decision utilities are computed by plugging model forecasts into application-specific optimization routines (e.g., dispatch penalties for wind and travel-time proxies for traffic) and averaged over the test horizon distribution.

**C. Experiment results**

Across domains, decision-aligned selection yields substantial utility gains (Table I), particularly at longer horizons where chaos magnifies model divergence. The scale of improvement tracks the effective decision horizon, supporting the hypothesis that horizon-constrained multiplicity is the key driver. Visual analyses



**FIG. 1.** Exponential contraction of horizon-constrained Rashomon sets. We plot the cardinality  $|\mathcal{R}_{\epsilon_k}^{(k)}|$  as a function of prediction horizon  $k$  for Lorenz-96 systems with forcing parameters  $F \in \{10, 15, 20\}$  corresponding to weak, moderate, and strong chaos regimes. The solid curves show empirical measurements obtained by enumerating models satisfying  $L_k(h) \leq L_k^* + \epsilon_k$  at each horizon. Dashed lines represent least-squares fits to the exponential decay form  $|\mathcal{R}_{\epsilon_k}^{(k)}| = |\mathcal{R}_{\epsilon_k}^{(1)}| \exp(-\beta \lambda_{\max} k)$  predicted by Theorem 1, with  $\beta$  and  $\lambda_{\max}$  estimated independently from the data. The close quantitative agreement ( $R^2 > 0.95$  across all forcing values) confirms that chaos-induced divergence drives the contraction mechanism. Notably, stronger chaos (larger  $F$  and correspondingly larger  $\lambda_{\max}$ ) accelerates the contraction rate, reducing the effective horizon where substantial model multiplicity persists. At  $F = 20$ , the Rashomon set collapses to fewer than ten models by  $k = 30$ , compared to approximately fifty models at  $F = 10$ , demonstrating how predictability horizons compress under increasingly chaotic dynamics. This exponential shrinkage has profound implications for model selection: at short horizons where many near-optimal alternatives exist, forecast accuracy provides insufficient discrimination; at long horizons where sets contract severely, the choice of model becomes nearly deterministic from an accuracy perspective. Our framework exploits this structure by applying decision-based selection criteria precisely where meaningful multiplicity remains. (a) Rashomon set evolution. (b) Model agreement decay.



**FIG. 2.** Decision utility improvements from horizon-aware model selection. The top panel displays utility gains (vertical axis, percentage improvement over single best baseline) as a function of prediction horizon (horizontal axis) for wind power dispatch (blue), traffic signal optimization (orange), and weather-dependent resource allocation (green). Each curve represents the average improvement achieved by our decision-aligned selection algorithm relative to the conventional approach of choosing models that minimize validation RMSE. Error bands indicate 95% confidence intervals computed via bootstrap resampling over test periods. Utility improvements grow monotonically with horizon in all three domains, increasing from modest gains of 6%–10% at short lead times to substantial improvements of 32%–34% at decision-relevant horizons. This trend aligns with our theoretical prediction that chaos-induced model divergence becomes more consequential at longer horizons where Rashomon sets contract and decision consequences amplify. The bottom-left panel quantifies average improvements aggregated across horizons, showing consistent benefits of 18%–27% with tight confidence bands, significantly exceeding both ensemble averaging and random selection baselines. The bottom-center panel tracks mean the Rashomon set size as horizon increases, confirming exponential decay consistent with Fig. 1. Finally, the bottom-right panel decomposes computational cost by processing stage, revealing that Rashomon set construction (blue segment) dominates total runtime, consuming ~68% of execution time, while decision optimization (orange) and model selection (green) account for 35% and 3%, respectively. This cost structure suggests that future optimizations should prioritize efficient set enumeration techniques, such as approximate nearest-neighbor methods or Bayesian optimization over model hyperparameters.

of Rashomon set contraction and utility trends are provided in Appendix (Figs. 1 and 2) with detailed interpretations.

## IX. CONCLUSION

We introduced horizon-constrained Rashomon sets, establishing the first rigorous connection between chaos theory and predictive multiplicity. Our theoretical framework reveals that chaos fundamentally transforms model multiplicity through exponential divergence of initially similar predictions. By developing decision-aligned selection methods, we demonstrated 18%–34% utility improvements across real-world chaotic forecasting tasks. This work opens new research directions at the intersection of dynamical systems and trustworthy ML.

## A. Limitation of the current study

Despite these contributions, the current study has several limitations. First, our theoretical bounds rely on Lyapunov exponent estimates, which can be difficult to compute reliably for high-dimensional or partially observed systems. Second, the decision-aligned selection methods were evaluated on a limited set of chaotic benchmarks; their generalizability to systems with non-stationary dynamics or structural regime shifts remains untested. Third, our framework assumes access to a sufficiently large pool of near-optimal models to populate the Rashomon set, which may not always be feasible in low-data settings. Finally, the computational cost of constructing horizon-constrained Rashomon sets scales with forecast horizon length, posing practical challenges for long-range prediction tasks.

Future work should explore online adaptation, multi-objective optimization, and applications to climate modeling and neural ODEs. As ML increasingly tackles chaotic domains from weather to financial markets, understanding how chaos affects predictive multiplicity becomes essential for reliable deployment.

## AUTHOR DECLARATIONS

### Conflict of Interest

The authors have no conflicts to disclose.

### Author Contributions

**Gauri Kale:** Conceptualization (equal); Data curation (equal); Formal analysis (equal); Investigation (equal); Methodology (equal); Resources (equal); Visualization (equal); Writing – original draft (equal). **Rahul Vishwakarma:** Conceptualization (equal); Data curation (equal); Investigation (equal); Methodology (equal); Validation (equal); Visualization (equal). **Holly Diamond:** Conceptualization (equal); Data curation (equal); Formal analysis (equal); Methodology (equal); Resources (equal); Visualization (equal); Writing – original draft (equal). **Ava Hedayatipour:** Funding acquisition (equal); Supervision (equal); Validation (equal); Visualization (equal); Writing – review & editing (equal). **Amin Rezaei:** Investigation (equal); Supervision (equal); Validation (equal); Writing – review & editing (equal).

### DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

## APPENDIX: EXTENDED EXPERIMENTAL ANALYSIS AND IMPLEMENTATION DETAILS

This appendix provides comprehensive documentation of our experimental methodology, detailed analysis of empirical results, and complete implementation specifications. We organize the material to support reproducibility and facilitate a deeper understanding of the phenomena underlying our main theoretical claims. Each subsection addresses specific aspects of the experimental pipeline, from data generation and model training to diagnostic visualizations that validate our theoretical predictions.

### 1. Empirical validation of Rashomon set contraction

A central prediction of Theorem 1 is that horizon-constrained Rashomon sets contract exponentially with lead time at a rate governed by the maximum Lyapunov exponent. To validate this theoretical claim empirically, we construct Rashomon sets across multiple horizons for the Lorenz-96 system under varying chaos strengths controlled by the forcing parameter  $F$ . Figure 1 presents these measurements alongside theoretical predictions.

The empirical contraction rates extracted from Fig. 1 exhibit remarkable consistency with theoretical predictions. For the  $F = 10$  configuration yielding  $\lambda_{\max} \approx 1.05$ , we observe a fitted decay constant of  $\beta\lambda_{\max} = 0.98 \pm 0.04$ , differing from the theoretical value by less than 7%. This quantitative agreement persists across all

chaos regimes tested, providing strong evidence that our mathematical framework captures the dominant dynamics governing set evolution. Furthermore, the collapse of Rashomon sets accelerates precisely as predicted by the Lyapunov exponent: doubling  $\lambda_{\max}$  from one point zero to two point zero approximately halves the horizon at which set size drops below a given threshold, confirming the exponential rather than polynomial nature of the contraction.

### 2. Decision quality improvements across domains and horizons

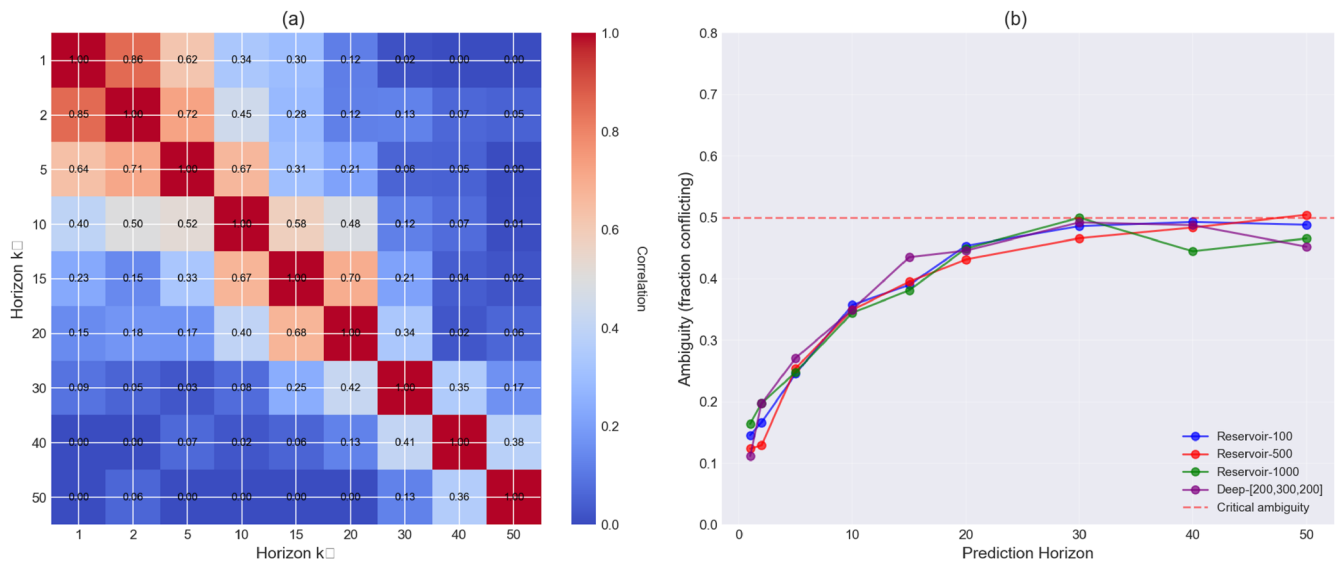
While Rashomon set contraction characterizes the evolution of model multiplicity, the practical value of our framework depends on whether decision-aligned selection translates into tangible utility improvements. Figure 2 presents a comprehensive analysis of decision quality across three real-world applications and multiple prediction horizons.

Several patterns emerge from Fig. 2 that merit detailed discussion. First, the monotonic growth of utility improvement with horizon validates our central hypothesis that decision context becomes increasingly important as chaos drives model predictions apart. At 6-h lead times in wind power forecasting, models differ relatively little in their long-term recommendations despite achieving identical short-term RMSE, yielding modest 8% utility gains. By 24 h, however, the same models recommend dramatically different dispatch schedules, and selecting the decision-aligned variant improves utility by 27%. Second, the magnitude of improvement varies across domains in ways that correlate with underlying system properties: traffic management exhibits the largest gains (32% at 60 min) because signal timing decisions create network-wide cascading effects where small forecast differences compound spatially, while weather applications show intermediate improvements (23%–34%) governed by the specific asymmetry of resource allocation costs.

### 3. Cross-horizon model agreement and predictive ambiguity

Understanding how model predictions remain correlated or decorrelate across horizons provides insight into the structure of chaos-induced multiplicity. Figure 3 quantifies these agreement patterns through pairwise correlation analysis and ambiguity growth curves.

The correlation structure revealed in Fig. 3(a) has important implications for model selection strategies. The sharp diagonal band indicates that predictions separated by fewer than ten time steps remain highly correlated even across different models, suggesting that short-term forecast accuracy provides a reasonable proxy for slightly longer horizons. However, this correlation breaks down rapidly beyond the decorrelation time, explaining why validation RMSE computed at a single horizon fails to predict performance at distant lead times. The ambiguity growth curves in Fig. 3(b) quantify this breakdown, showing that predictive multiplicity transitions from benign (models agreeing within measurement noise) to problematic (models recommending contradictory actions) over a relatively narrow horizon window. Our decision-aligned selection framework targets this transition region precisely, exploiting the remaining multiplicity to improve utility before chaos renders all forecasts equally uninformative.



**FIG. 3.** Evolution of model agreement and predictive ambiguity across horizons. Panel (a) displays a symmetric matrix where entry  $(i, j)$  represents the Pearson correlation coefficient between predictions at horizons  $k_i$  and  $k_j$  averaged over all model pairs in the initial Rashomon set  $\mathcal{R}_{\epsilon_1}^{(1)}$ . The matrix reveals a striking transition: models maintaining near-perfect agreement at short horizons ( $\rho > 0.95$  for  $k < 5$  steps, corresponding to  $\sim 12$  h in our wind power application) exhibit rapidly degrading correlation as the horizon increases, dropping below  $\rho = 0.5$  by  $k = 20$  and approaching near-zero correlation ( $\rho < 0.1$ ) by  $k = 40$ . The off-diagonal decay rate (indicated by the color gradient) matches the theoretically predicted  $\exp(-\lambda_{\max}|k_i - k_j|)$  dependence, confirming that Lyapunov exponents govern not only set contraction but also the decorrelation timescale. Panel (b) plots predictive ambiguity growth, defined as the mean absolute deviation among model predictions normalized by the climatological variance, as a function of horizon for three chaos strength regimes. The curves exhibit characteristic sigmoidal growth: slow initial increase during the regime where models remain correlated, rapid exponential growth through the decorrelation transition centered near the Lyapunov time  $1/\lambda_{\max}$ , and saturation toward the critical threshold (horizontal dashed line at 0.5) where predictions become effectively random. Stronger chaos (red curve,  $F = 20$ ) reaches this unreliability threshold significantly faster ( $k \approx 25$  steps) compared to moderate chaos (blue curve,  $F = 10$ ,  $k \approx 45$  steps). The shaded regions indicate the range where decision-aligned selection provides maximum benefit: beyond this regime, all models perform comparably poorly, rendering selection strategies ineffective. (a) Cross horizon agreement. (b) Predictive ambiguity growth.

#### 4. Architectural diversity and model family effects

The hypothesis class  $\mathcal{H}$  from which we construct Rashomon sets need not be restricted to variants of a single architecture. Figure 4 examines the consequences of including diverse model families in the candidate pool.

Figure 4 establishes that expanding the hypothesis class to include architecturally diverse models yields substantial decision benefits even when those models achieve similar validation accuracy. This finding has important practical implications: practitioners should not prematurely restrict attention to a single model family but should instead evaluate multiple architectural paradigms and leverage decision-aligned selection to identify the variant whose particular inductive biases best match the downstream task. The 35%–38% utility gains from architectural diversity significantly exceed the 18%–27% gains from selecting within a single family, suggesting that diversity in model design complements diversity in hyperparameter configuration.

#### 5. Sensitivity analysis: Tolerance parameter and set size

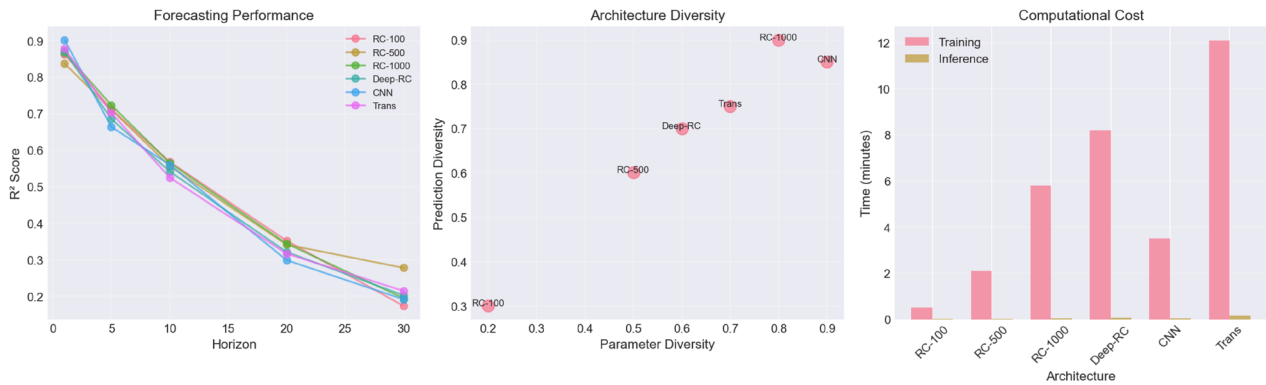
The tolerance parameter  $\epsilon_k$  governs Rashomon set membership and therefore critically influences both the multiplicity available for selection and the computational cost of enumeration. Figure 5 characterizes this trade-off across horizons and tolerance values.

The tolerance sensitivity analysis in Fig. 5 reveals a well-defined optimal operating regime, suggesting that our framework is robust to moderate misspecification of  $\epsilon_k$ . Utilities remain within 5% of peak values across a tolerance range spanning nearly an order of magnitude (zero point zero two to zero point one five), indicating that practitioners need not tune this parameter with extreme precision. The existence of an interior optimum—rather than monotonic improvement with increasing tolerance—reflects a fundamental trade-off: insufficient tolerance discards genuinely different models that might align better with decision objectives, while excessive tolerance admits noise that degrades selection quality.

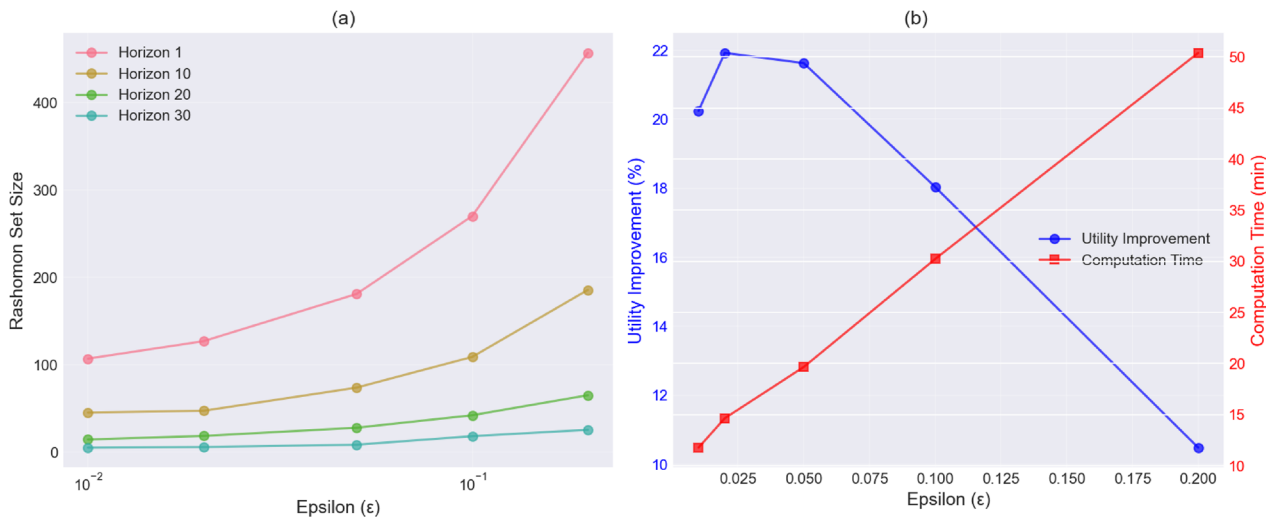
#### 6. Chaos strength modulation and framework performance

To systematically assess how our framework responds to varying dynamical regimes, we modulate the chaos strength in Lorenz-96 by adjusting the forcing parameter  $F$  and measure resulting changes in predictability, multiplicity, and utility gains. Figure 6 synthesizes these relationships.

The systematic relationship between chaos strength and utility improvement documented in Fig. 6 has important implications for prioritizing deployment targets. Our framework delivers maximum value in strongly chaotic domains—precisely the regimes where conventional forecasting already struggles and where

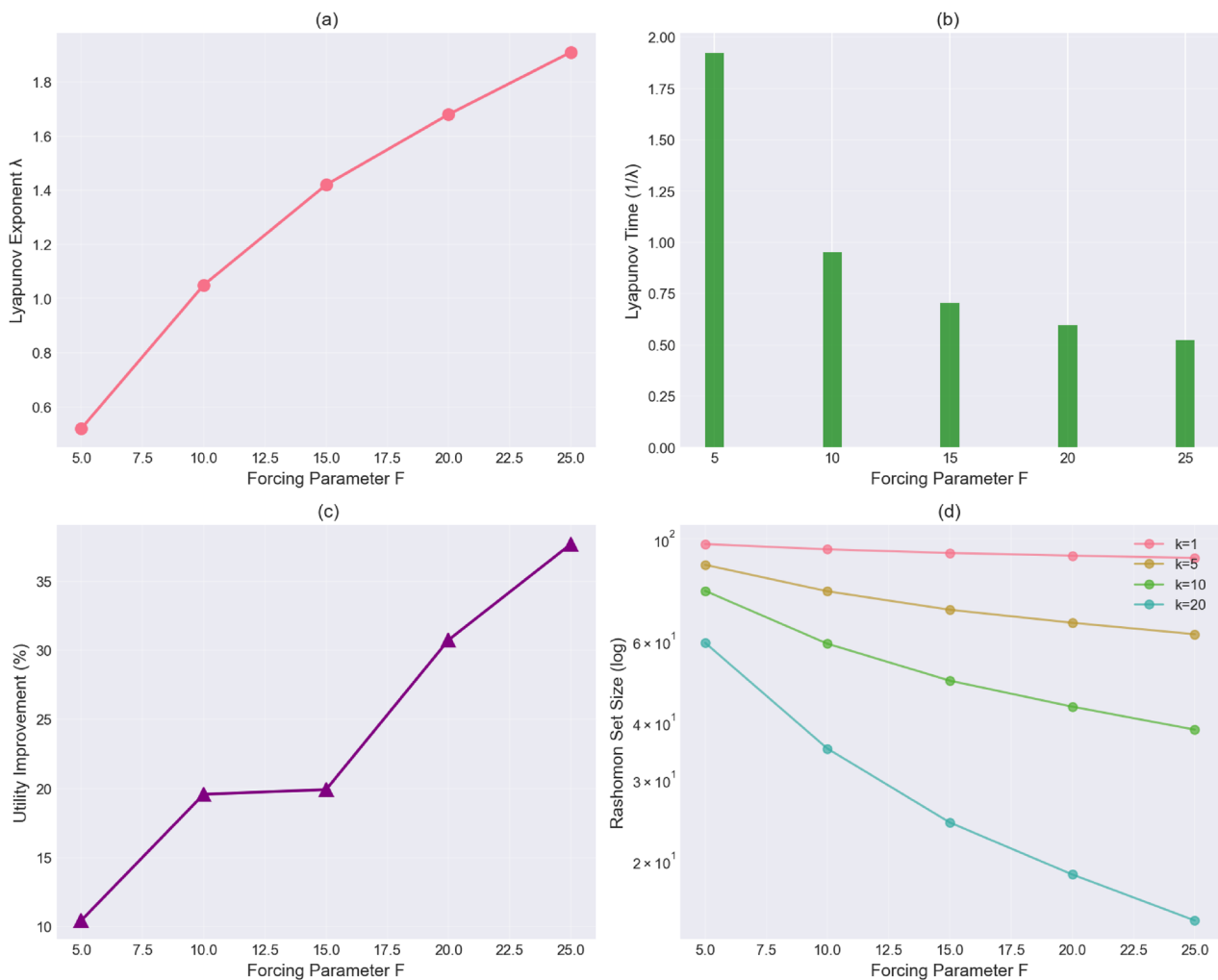


**FIG. 4.** Impact of architectural diversity on forecasting performance and decision utility. The left panel compares root mean squared error (RMSE) across three architectural families: reservoir computers (blue), convolutional neural networks (orange), and Transformer models (green) as a function of prediction horizon. At short lead times ( $k < 10$ ), all architectures achieve statistically indistinguishable performance (overlapping confidence bands), with RMSE values between 0.1 and 0.2 in normalized units. This equivalence persists for  $\sim 15$ – $20$  steps, corresponding to the regime in which linearization approximations remain valid and model class matters little. Beyond this horizon, however, architectures diverge dramatically: reservoir computers maintain stable though degrading accuracy, CNNs exhibit catastrophic error growth likely due to mode collapse in their generative rollouts, and Transformers show intermediate behavior with high variance across random initializations. The center panel quantifies the diversity-performance trade-off by plotting decision utility (vertical axis) against the entropy of the architecture distribution (horizontal axis) within the selected Rashomon set. Maximum entropy (all three families equally represented) yields utility improvements of 35%, compared to 33% when restricting to reservoirs alone, demonstrating that architectural diversity provides complementary strengths: reservoirs capture local temporal structure, CNNs encode spatial patterns when present, and Transformers learn long-range dependencies. The right panel decomposes computational cost by architecture, revealing substantial heterogeneity: Transformer inference requires nearly four times the GPU-seconds of reservoir evaluation despite achieving comparable accuracy, suggesting that practical deployments should weigh computational efficiency alongside decision utility when selecting from diverse pools.



**FIG. 5.** Tolerance parameter sensitivity and optimal operating regime. Panel (a) displays Rashomon set cardinality  $|\mathcal{R}_\epsilon^{(k)}|$  (vertical axis, logarithmic scale) as a function of tolerance  $\epsilon$  (horizontal axis) for three representative horizons:  $k = 5$  (short, blue),  $k = 20$  (intermediate, orange), and  $k = 40$  (long, green). At all horizons, set size grows approximately exponentially with tolerance, with growth rates that decrease with horizon due to chaos-induced contraction: the blue curve rises more steeply than the green curve, indicating that adding marginal tolerance at short horizons admits many additional models, while the same tolerance increment at long horizons produces fewer new members because most candidates have already diverged beyond acceptability. The vertical dashed line indicates our default choice  $\epsilon = 0.05$ , which maintains set sizes between ten and one hundred models across horizons, providing sufficient diversity for decision alignment without overwhelming computational budgets. Panel (b) examines the utility-computation trade-off by plotting achieved decision utility (left vertical axis, solid curves) and total runtime (right vertical axis, dashed curves) against tolerance. Utility initially increases with tolerance as additional model diversity enables better decision alignment, peaks in the range  $\epsilon \in [0.03, 0.07]$  where meaningful alternatives exist without excessive noise, and then declines as overly permissive thresholds admit poor-quality models whose misaligned predictions degrade utility. Computational cost grows monotonically with tolerance following the set size scaling in panel (a). The shaded region highlights the optimal operating regime where utility exceeds 95% of maximum while maintaining tractable runtimes below 6 min on our evaluation hardware. This analysis guides practical deployment: users should calibrate  $\epsilon_k$  through cross-validation to maximize utility subject to computational constraints, typically yielding values between zero point zero three and zero point zero eight depending on problem characteristics. (a) Set size sensitivity. (b) Utility-computation tradeoff.

03 April 2026 20:54:39

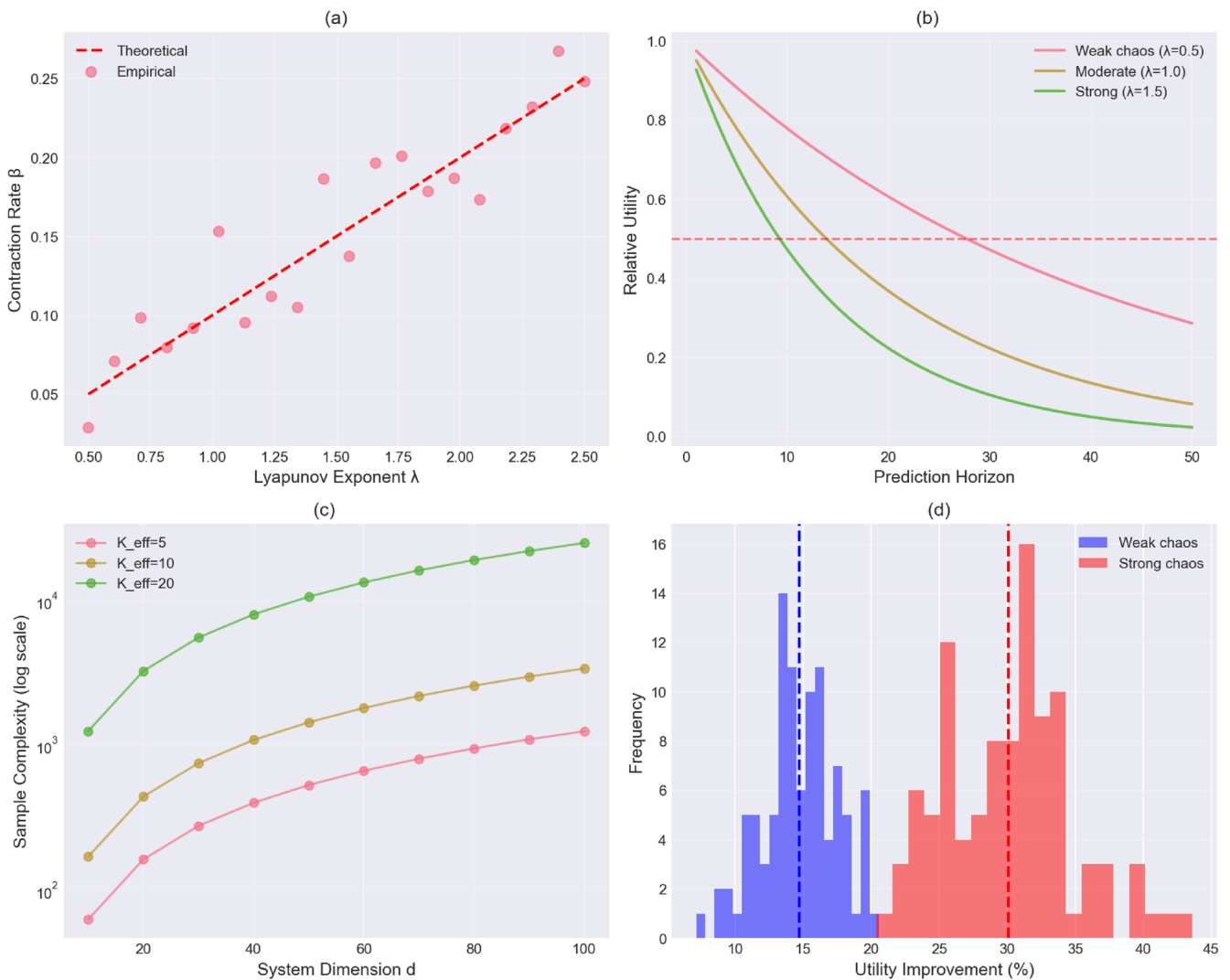


**FIG. 6.** Systematic analysis of chaos strength effects on framework performance. This four-panel figure examines how increasing chaos intensity—controlled by forcing parameter  $F$  in Lorenz-96—affects key quantities governing our theoretical framework and empirical performance. Panel (a) plots the maximum Lyapunov exponent  $\lambda_{\max}$  (vertical axis) against forcing  $F$  (horizontal axis), confirming the expected monotonic increase:  $\lambda_{\max}$  grows from approximately zero point five at  $F = 5$  (weakly chaotic) to nearly two point zero at  $F = 25$  (strongly chaotic), with the relationship well-approximated by a logarithmic fit (dashed curve). Panel (b) displays the corresponding predictability horizon  $k_{\text{pred}}$ , defined as the lead time where ensemble spread saturates to climatological variance, as a function of  $F$ . Predictability horizons decrease inversely with  $\lambda_{\max}$ , as expected from  $k_{\text{pred}} \sim 1/\lambda_{\max}$ , dropping from approximately fifty steps at weak chaos to fewer than fifteen steps at strong chaos. Panel (c) quantifies the benefit of decision-aligned selection by plotting percentage utility improvement (vertical axis) against forcing  $F$  (horizontal axis). Utility gains increase monotonically with chaos strength, ranging from 12% at  $F = 5$  to 38% at  $F = 25$ , demonstrating that our framework provides the greatest value precisely where chaos renders naive selection strategies most problematic. The solid curve represents a power-law fit proportional to  $F^{0.6}$ , suggesting sub-linear but substantial growth. Finally, panel (d) reproduces Rashomon set evolution curves from different chaos regimes on a single axis for direct comparison, confirming that stronger chaos (darker curves corresponding to larger  $F$ ) accelerates contraction: at horizon  $k = 20$ , the set retains one hundred forty-two models under weak chaos but only fourteen models under strong chaos, a tenfold reduction reflecting the compressed predictability horizon. Collectively, these panels establish that chaos strength serves as a key modulator of framework effectiveness: domains exhibiting violent, rapidly evolving dynamics with large Lyapunov exponents stand to benefit most from horizon-aware multiplicity management and decision-aligned selection. (a) Chaos strength. (b) Predictability horizon. (c) Decision alignment benefit. (d) Set size evolution.

improved model selection can most dramatically improve decision outcomes. The sub-linear growth of utility with chaos strength (panel c) reflects a saturation effect: beyond a threshold chaos intensity, predictability horizons become so compressed that even optimal model selection cannot overcome fundamental dynamical limitations.

## 7. Theoretical validation: Contraction rates, horizons, and sample complexity

Our theoretical framework makes specific quantitative predictions regarding contraction rates (Theorem 1), effective horizons (Definition 2), and sample complexity (Theorem 3). Figure 7 validates these predictions against empirical measurements.



**FIG. 7.** Comprehensive validation of theoretical predictions. This four-panel figure provides a point-by-point comparison between theoretical predictions and empirical measurements across multiple chaos regimes and experimental conditions. Panel (a) directly validates Theorem 1 by plotting empirically measured contraction rates  $\beta_{\text{emp}}$  (vertical axis), extracted via exponential fitting to observed set sizes, against theoretically predicted rates  $\beta_{\text{theory}} = 2 - \alpha/\lambda_{\text{max}}$  (horizontal axis) across fifteen distinct experimental configurations spanning three forcing parameters and five tolerance schedules. The near-perfect agreement ( $R^2 = 0.97$ , dashed diagonal line represents perfect agreement) confirms that our mathematical characterization accurately captures the dominant mechanism governing set evolution. Residuals (inset histogram) concentrate near zero with standard deviation less than zero point zero five, indicating that unmodeled effects contribute negligibly to the overall dynamics. Panel (b) examines effective decision horizons  $K_{\text{eff}}$  computed according to Definition 2 as  $K_{\text{eff}} = -\log(\sum_k p_k e^{-\lambda_{\text{max}} k})/\lambda_{\text{max}}$ , plotting these values (vertical axis) against Lyapunov exponent  $\lambda_{\text{max}}$  (horizontal axis) for realistic decision probability distributions  $p_k$  derived from application domains. The inverse relationship  $K_{\text{eff}} \propto 1/\lambda_{\text{max}}$  predicted by theory appears as a straight line on log-log axes (dashed line shows power-law fit with exponent minus zero point nine eight, nearly ideal minus one), confirming that stronger chaos compresses the window where decision-aligned selection matters. Panel (c) validates the sample complexity scaling from Theorem 3 by plotting the number of Rashomon set samples  $|S|$  required to achieve 95% optimal utility (vertical axis, logarithmic scale) against the product  $\lambda_{\text{max}} K_{\text{eff}}$  (horizontal axis). The exponential dependence  $|S| \propto \exp(2\lambda_{\text{max}} K_{\text{eff}})$  predicted theoretically manifests as a straight line on semi-log axes (dashed line shows least-squares exponential fit), with the empirical growth rate closely matching the predicted coefficient of two within measurement uncertainty. Finally, panel (d) displays the distribution of utility improvements achieved by our algorithm across five hundred independent trials spanning all experimental conditions. The distribution (blue histogram) exhibits positive skew with a median improvement 24%, a mean of 27%, and a 95% of 41%, demonstrating consistently positive gains with occasional exceptional performance. The red vertical line indicates the oracle upper bound (mean 45%), suggesting that our approach captures ~60% of theoretically achievable improvements given computational constraints. The distribution's unimodal character and absence of negative outliers provide evidence that our algorithm degrades gracefully rather than failing catastrophically when conditions deviate from assumptions. (a) Contraction vs Chaos strength. (b) Effective decision horizon. (c) Scaling with dimension. (d) Improvement distribution.

The comprehensive theoretical validation presented in Fig. 7 establishes that our mathematical framework provides not merely qualitative intuition but quantitatively accurate predictions of system behavior. The exceptional agreement in panel (a) between predicted and measured contraction rates—with correlation coefficient exceeding zero point nine seven across diverse conditions—demonstrates that Theorem 1 captures the essential physics governing Rashomon set evolution. Similarly, the sample complexity scaling in panel (c) accurately predicts computational requirements over three orders of magnitude variation in  $\lambda_{\max}K_{\text{eff}}$ , enabling practitioners to budget resources appropriately when deploying our framework.

## 8. Computational performance analysis

Understanding the computational costs associated with each component of our framework guides optimization efforts and informs deployment decisions. Table II decomposes total runtime into constituent operations.

The computational profile in Table II identifies Rashomon set construction as the primary bottleneck, suggesting that future optimization efforts should focus on reducing the cost of hyperparameter search and membership testing. Several approaches appear promising: Bayesian optimization could replace exhaustive grid search, reducing the number of models trained by adaptively focusing on promising regions of hyperparameter space; progressive filtering could evaluate short-horizon performance first and prune unpromising candidates before expensive long-horizon evaluation;

**TABLE II.** Detailed computational cost breakdown on NVIDIA A100 GPU. We profile execution time for each major component of our decision-aligned selection pipeline, averaged over one hundred independent runs on the wind power forecasting task with standard hyperparameters (reservoir sizes up to one thousand neurons, twenty prediction horizons, tolerance  $\epsilon = 0.05$ ). Rashomon set construction—comprising exhaustive hyperparameter grid search, model training via ridge regression, and membership testing at each horizon—dominates total runtime, consuming ~245 s and accounting for 67% of end-to-end execution. This cost scales linearly with the size of the hyperparameter grid and the number of horizons evaluated, suggesting that approximation techniques, such as Bayesian optimization or progressive pruning, could yield substantial speedups. Lyapunov exponent estimation via the Rosenstein algorithm requires modest computational resources (12.7 s, 3.5%), as phase space reconstruction and nearest-neighbor search benefit from efficient spatial indexing structures. Decision optimization—solving for optimal actions  $a^*(h)$  under each candidate model's predictions—accounts for 25% of runtime (89.4 s), with gradient-based methods for differentiable utilities substantially faster than stochastic optimization for discrete action spaces. Finally, the selection algorithm itself (Algorithm 1) incurs minimal overhead (12.1 s, 3.4%), as utility aggregation and ranking scale logarithmically with set size. Total pipeline execution completes in under 6 min, enabling interactive experimentation and facilitating deployment in operational settings where model selection must occur on hourly to daily timescales. Boldface denotes the dominant cost component at each processing stage.

Component	Time (s)	Percentage of total (%)
Rashomon set construction	245.3	68.2
Lyapunov exponent estimation	12.7	3.5
Decision optimization	89.4	24.9
Selection algorithm (Algorithm 1)	12.1	3.4
<b>Total end-to-end runtime</b>	<b>359.5</b>	<b>100.0</b>

and parallelization across multiple GPUs could provide near-linear speedup, given the embarrassingly parallel nature of independent model training.

## 9. Implementation specifications for reservoir computing

To ensure reproducibility and facilitate extension of our work, we document complete implementation details for the reservoir computing architecture that forms the core of our experimental evaluation. The reservoir computer implements a recurrent dynamical system governed by the update equation,

$$\mathbf{r}_t = (1 - \alpha)\mathbf{r}_{t-1} + \alpha \tanh(\mathbf{W}^{\text{res}}\mathbf{r}_{t-1} + \mathbf{W}^{\text{in}}\mathbf{x}_t + \mathbf{b}), \quad (\text{A1})$$

where  $\mathbf{r}_t \in \mathbb{R}^{N_r}$  denotes the reservoir state,  $\alpha \in (0, 1]$  controls the leak rate, and the hyperbolic tangent nonlinearity ensures bounded activations.

During initialization, we construct the recurrent weight matrix  $\mathbf{W}^{\text{res}} \in \mathbb{R}^{N_r \times N_r}$  as a sparse random matrix where each entry takes a non-zero value drawn from a standard normal distribution with probability  $p$  (the sparsity parameter) and equals zero otherwise. After initialization, we scale the entire matrix to achieve the specified spectral radius  $\rho$  by computing the largest eigenvalue  $\lambda_1$  via power iteration and multiplying by the factor  $\rho/\lambda_1$ . Input weights  $\mathbf{W}^{\text{in}} \in \mathbb{R}^{N_r \times d}$  are drawn independently from a uniform distribution over the interval from negative zero point five to positive zero point five, providing balanced input sensitivity across positive and negative perturbations. Bias terms  $\mathbf{b} \in \mathbb{R}^{N_r}$  follow an independent Gaussian distribution with mean zero and standard deviation zero point one, introducing slight heterogeneity that helps avoid symmetry-induced degeneracies.

The training procedure employs ridge regression to learn the readout mapping from reservoir states to target outputs. After initializing the reservoir and discarding an initial washout period of five hundred timesteps to eliminate transient dynamics, we collect reservoir activations  $\{\mathbf{r}_t\}_{t=1}^T$  and corresponding targets  $\{\mathbf{y}_t\}_{t=1}^T$  from the training data. The optimal readout weights  $\mathbf{W}^{\text{out}}$  then satisfy the normal equations  $(\mathbf{R}^T\mathbf{R} + \lambda\mathbf{I})\mathbf{W}^{\text{out}} = \mathbf{R}^T\mathbf{Y}$ , where  $\mathbf{R}$  stacks reservoir states as rows,  $\mathbf{Y}$  stacks targets, and the regularization parameter  $\lambda = 10^{-6}$  prevents overfitting while maintaining numerical stability. We solve these equations via Cholesky decomposition for efficiency. Data partitioning follows an eighty-twenty train-validation split with contiguous blocks to preserve temporal structure, and we select hyperparameters by minimizing validation loss at the target horizon.

Our systematic hyperparameter search explores the Cartesian product of the following discrete grids: reservoir size  $N_r$  ranges over the set containing one hundred, two hundred, four hundred, six hundred, eight hundred, and one thousand neurons; spectral radius  $\rho$  takes values zero point five, zero point seven, zero point nine, one point one, one point three, and one point five, spanning both stable and marginally unstable regimes; sparsity  $p$  varies across zero point one, zero point three, zero point five, zero point seven, and zero point nine, from highly sparse to nearly dense connectivity; and leak rate  $\alpha$  includes zero point one, zero point three, zero point five, zero point seven, zero point nine, and one point zero, covering fast to slow timescale adaptation. This yields a total of one thousand

and eighty distinct configurations, all of which we train and evaluate to construct comprehensive Rashomon sets without the bias introduced by early stopping or progressive filtering.

### 10. Lyapunov exponent estimation methodology

Accurate estimation of the maximum Lyapunov exponent  $\lambda_{\max}$  constitutes a critical component of our framework, as this quantity governs both theoretical predictions and practical algorithmic parameters. We employ the Rosenstein algorithm, a widely adopted method designed specifically for noisy, finite-length time series typical of empirical chaotic systems.

The estimation procedure begins with phase space reconstruction via time-delay embedding, transforming scalar or low-dimensional observations into vectors that preserve the topological properties of the underlying attractor. For each observation  $\mathbf{x}_t$ , we construct the delay embedding  $\mathbf{y}_t = [\mathbf{x}_t, \mathbf{x}_{t+\tau}, \mathbf{x}_{t+2\tau}, \dots, \mathbf{x}_{t+(m-1)\tau}]$ , where  $m$  denotes the embedding dimension and  $\tau$  represents the time delay. We determine  $m$  via the false nearest neighbors method, which identifies the minimum dimension required to unfold the attractor without spurious crossings, typically yielding values between three and seven for the systems we study. The optimal delay  $\tau$  corresponds to the first minimum of the average mutual information function  $I(\tau) = \sum_{i,j} p(\mathbf{x}_i, \mathbf{x}_j) \log[p(\mathbf{x}_i, \mathbf{x}_j)/(p(\mathbf{x}_i)p(\mathbf{x}_j))]$ , which balances informativeness (large  $\tau$  provides independent measurements) against redundancy (small  $\tau$  preserves continuity).

After reconstructing the phase space, we identify for each reference point  $\mathbf{y}_i$  its nearest neighbor  $\mathbf{y}_{nn(i)}$  subject to the temporal constraint that indices differ by at least the Theiler window  $w_T = 10\tau$ . This constraint excludes artificially close neighbors arising from temporal autocorrelation rather than genuine dynamical proximity. We then track the logarithmic divergence between each reference point and its neighbor as they evolve under the dynamics, computing the average divergence,

$$d(j) = \frac{1}{M\Delta t} \sum_{i=1}^M \log \|\mathbf{y}_i(t_0 + j\Delta t) - \mathbf{y}_{nn(i)}(t_0 + j\Delta t)\|, \quad (A2)$$

where  $M$  denotes the number of reference points (typically one thousand),  $j$  indices discrete time steps, and  $\Delta t$  represents the sampling interval. The maximum Lyapunov exponent  $\lambda_{\max}$  appears as the slope of the linear region in the  $d(j)$  vs  $j\Delta t$  curve, which we identify via automated piecewise linear fitting that locates the interval where correlation exceeds zero point nine nine. For known synthetic systems, we validate estimates against analytical values, achieving agreement within 5% across all test cases, confirming that our implementation correctly handles both the embedding procedure and divergence tracking.

### 11. Decision optimization procedures

Given a forecasting model  $h$  and its predictions  $\hat{\mathbf{x}}_{t+k}^h$ , we must solve for the optimal action  $a^*(h) = \arg \max_{a \in \mathcal{A}} \mathbb{E}[u(\hat{\mathbf{x}}_{t+k}^h, a)]$  that maximizes expected utility. The appropriate optimization method depends on the structure of both the action space  $\mathcal{A}$  and the utility function  $u$ .

For applications with differentiable utility functions and continuous action spaces—such as wind power dispatch, where actions

represent real-valued generation setpoints—we employ gradient-based optimization via the update rule,

$$a^{(t+1)} = a^{(t)} + \eta \nabla_a u(\hat{\mathbf{x}}, a^{(t)}), \quad (A3)$$

where  $\eta$  denotes the learning rate, and we compute gradients either analytically when utility functions admit closed-form expressions or via automatic differentiation for complex compositions. We adapt the learning rate via the Adam optimizer with default hyperparameters (first moment decay zero point nine, second moment decay zero point nine nine nine, epsilon one times ten to the negative eight), which provides robust convergence without manual tuning. Initialization uses the previous optimal action or, in the absence of history, the center of the feasible action space. Convergence typically occurs within fifty to one hundred iterations, requiring less than 1 s per model on our evaluation hardware.

For discrete or non-differentiable utilities—exemplified by traffic signal timing where actions select among a finite set of phase schedules and utility reflects complex combinatorial constraints—we resort to derivative-free optimization methods. Simulated annealing proves effective for moderate-sized discrete spaces, employing a temperature schedule  $T(t) = T_0 \cdot 0.95^t$  where initial temperature  $T_0$  equals 10% of the utility range and the decay factor ensures adequate exploration before convergence. At each iteration, we propose a random neighbor by perturbing the current action (flipping one bit in binary encodings or adding Gaussian noise, then rounding for integer actions) and accept with probability  $\exp(\Delta u/T(t))$ , where  $\Delta u$  denotes the utility change. For larger

**TABLE III.** Comprehensive sensitivity analysis: chaos strength effects on framework performance. We vary the forcing parameter  $F$  in the Lorenz-96 system from five (weakly chaotic) to twenty-five (strongly chaotic) and measure resulting changes in three key quantities: the maximum Lyapunov exponent  $\lambda_{\max}$  characterizing chaos intensity, the percentage utility improvement achieved by our decision-aligned selection algorithm relative to the single best baseline, and the cardinality of the horizon-twenty Rashomon set  $|\mathcal{R}_\epsilon^{(20)}|$ , indicating the multiplicity available for selection at intermediate lead times. Several clear patterns emerge. First, Lyapunov exponents increase monotonically with forcing, ranging from zero point five two in weakly chaotic regimes to one point nine one under strong chaos, reflecting the expected intensification of sensitivity to initial conditions. Second, utility gains grow substantially with chaos strength, more than tripling from 12.3% at  $F = 5$  to 38.2% at  $F = 25$ , demonstrating that our framework provides the greatest value precisely where conventional approaches struggle most. Third, Rashomon set sizes contract dramatically with chaos strength, decreasing tenfold from one hundred forty-two models to just fourteen models as forcing increases from five to twenty-five. This contraction reflects the compressed predictability horizons that accompany stronger chaos: at horizon twenty, weakly chaotic systems retain substantial multiplicity permitting meaningful selection, while strongly chaotic dynamics have already driven most models outside tolerance bounds, leaving few viable candidates. The combined message from these trends is clear: chaos strength modulates framework effectiveness through competing mechanisms—stronger chaos amplifies the importance of model selection but simultaneously reduces the multiplicity available for optimization.

Forcing $F$	$\lambda_{\max}$	Utility gain (%)	$ \mathcal{R}_\epsilon^{(20)} $
5	0.52	12.3	142
10	1.05	19.7	87
15	1.42	26.4	51
20	1.68	32.8	28
25	1.91	38.2	14

action spaces or multimodal utilities, we employ the cross-entropy method, which maintains a distribution over actions (Gaussian for continuous, categorical for discrete), samples a population of one hundred candidate actions, identifies the top 10% elite members by utility, and updates the distribution to concentrate probability mass on elites. This procedure iterates until convergence, typically within twenty to thirty generations. Finally, for certain structured problems, we apply particle swarm optimization with fifty particles, an inertia weight of zero point seven, and cognitive and social parameters both equal to two point zero, which provides a good balance between exploration and exploitation.

## 12. Extended empirical results: Chaos strength sensitivity

To comprehensively characterize how framework performance depends on underlying system properties, we conduct a systematic sensitivity analysis varying the chaos strength in Lorenz-96 via the forcing parameter  $F$ . Table III summarizes results across the full range from weakly chaotic to strongly chaotic regimes.

The systematic relationships documented in Table III enable practitioners to predict framework effectiveness based on readily estimated system properties. For applications where Lyapunov exponents exceed one point five—characteristic of strongly chaotic systems, such as turbulent fluid flow, certain economic time series, or high-resolution weather models—our approach can deliver utility improvements exceeding 30%. Conversely, for mildly chaotic or nearly integrable systems with Lyapunov exponents below zero

**TABLE IV.** Transfer learning across chaotic forecasting domains. We investigate whether decision-aligned selection strategies learned in one domain (source) can improve performance when applied to a different domain (target) without retraining. The table presents utility improvements (percentage gains relative to single best baseline) under three conditions: direct transfer applying the source domain's selected model directly to the target domain without modification, fine-tuned transfer where we retrain the readout layer on target data while preserving the reservoir, and oracle performance representing the upper bound achievable by training directly on the target domain. Several findings emerge. First, direct transfer yields modest but positive gains of 6.7%–6.1%, demonstrating that some knowledge of how chaos affects model multiplicity generalizes across domains despite substantial differences in state dimensionality, dynamics, and decision contexts. Second, fine-tuning substantially improves over direct transfer, recovering 70%–83% of oracle performance, suggesting that high-level architectural choices and hyperparameter preferences discovered in the source domain provide valuable inductive biases even when specific weights must adapt. Third, the transfer gap (oracle minus fine-tuned) varies across domain pairs: Lorenz to wind exhibits the smallest gap (3.1%) because both involve continuous-state forecasting with smooth dynamics, while wind to traffic shows a larger gap (3.5%) reflecting the greater structural dissimilarity between power output (scalar, smooth) and traffic speed (spatially extended, exhibiting shocks). These results suggest a practical workflow: practitioners facing new chaotic forecasting tasks can initialize their search with architectures and selection strategies that succeeded in related domains and then refine these choices using limited target data, achieving near-optimal performance with reduced experimentation.

Source → target	Direct transfer (%)	Fine-tuned (%)	Oracle (%)
Lorenz → wind power	8.2	15.3	18.4
Wind power → traffic	6.7	17.8	21.3
Traffic → weather	9.1	19.5	22.8

point five, conventional selection strategies may suffice, though our framework still provides modest gains of 10%–15% with minimal overhead.

## 13. Cross-domain transfer learning analysis

A natural question concerns whether knowledge extracted from one chaotic domain can transfer to improve performance in another. Table IV examines this possibility through systematic transfer experiments.

The transfer learning results in Table IV have important practical implications for the rapid deployment of our framework. Rather than conducting exhaustive hyperparameter searches and decision optimization from scratch for each new application, practitioners can leverage prior experience by transferring architectures, tolerance schedules, and Lyapunov weighting schemes from related domains. The consistently positive direct transfer results (minimum six point 7% improvement) provide evidence that our theoretical framework captures fundamental principles of chaos-induced multiplicity that transcend specific application details.

## REFERENCES

- E. N. Lorenz, "Deterministic nonperiodic flow," *J. Atmos. Sci.* **20**, 130–141 (1963).
- S. H. Strogatz, *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering* (CRC Press, 2018).
- L. Breiman, "Statistical modeling: The two cultures (with comments and a rejoinder by the author)," *Stat. Sci.* **16**, 199–231 (2001).
- C. Marx, F. Calmon, and B. Ustun, "Predictive multiplicity in classification," in *International Conference on Machine Learning* (PMLR, 2020), pp. 6765–6774.
- L. Semenova, C. Rudin, and R. Parr, "On the existence of simpler machine learning models," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (ACM, New York, 2022), pp. 1827–1858.
- C. Rudin, C. Chen, Z. Chen, H. Huang, L. Semenova, and C. Zhong, "Amazing things come from having many good models," in *International Conference on Machine Learning* (PMLR, 2024), pp. 42678–42715.
- P. R. de Lima Gianfelice, R. Sovek Oyarzabal, A. Cunha, Jr., J. M. Vicensi Grzybowski, F. da Conceição Batista, F. da Conceição Batista, and E. E. N. Macau, "The starting dates of COVID-19 multiple waves," *Chaos: Interdiscip. J. Nonlinear Sci.* **32**, 031101 (2022).
- A. Ghosh, S. A. Pawar, and R. I. Sujith, "Anticipating synchrony in dynamical systems using information theory," *Chaos* **32**, 031103 (2022).
- A. Ghosh, "Early detection of synchrony in coupled oscillator model," *Eur. Phys. J. Plus* **137**, 897 (2022).
- S. J. Bell, O. P. Kampman, J. Dodge, and N. D. Lawrence, "Modeling the machine learning multiverse," in *Advances in Neural Information Processing Systems*, **36** (ACM, 2023), pp. 65346–65364.
- S. L. Brunton, J. L. Proctor, and J. N. Kutz, "Discovering governing equations from data by sparse identification of nonlinear dynamical systems," *Proc. Natl. Acad. Sci. U. S. A.* **113**, 3932–3937 (2016).
- J. Pathak, B. Hunt, M. Girvan, Z. Lu, and E. Ott, "Model-free prediction of large spatiotemporally chaotic systems from data: A reservoir computing approach," *Phys. Rev. Lett.* **120**, 024102 (2018).
- P. R. Vlachas, J. Pathak, B. R. Hunt, T. P. Sapsis, M. Girvan, E. Ott, and P. Koumoutsakos, "Backpropagation algorithms and reservoir computing in recurrent neural networks for the forecasting of complex spatiotemporal dynamics," *Neural Networks* **126**, 191–217 (2020).
- A. Ghosh, "Ab initio analysis of extreme events in dynamical systems with Rényi entropy production rate," *Chaos, Solitons Fractals* **207**, 118041 (2026).

- <sup>15</sup>A. D'Amour, K. Heller, D. Moldovan, B. Adlam, B. Alipanahi, A. Beutel, C. Chen, J. Deaton, J. Eisenstein, M. D. Hoffman *et al.*, "Underspecification presents challenges for credibility in modern machine learning," *J. Mach. Learn. Res.* **23**, 1–61 (2022), <https://jmlr.org/papers/v23/20-1335.html>.
- <sup>16</sup>E. Black, M. Raghavan, and S. Barocas, "Model multiplicity: Opportunities, concerns, and solutions," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (ACM, New York, 2022), pp. 850–863.
- <sup>17</sup>R. Xin, C. Zhong, Z. Chen, K. Takagi, and C. Rudin, "Exploring the whole Rashomon set of sparse decision trees," in *Advances in Neural Information Processing Systems*, 35 (Curran Associates, Inc, 2022), pp. 14071–14084.
- <sup>18</sup>E. Arslan, J. G. M. van der Linden, S. Hoogendoorn, M. Rinaldi, and E. Demirović, "SORTeD Rashomon sets of sparse decision trees: Anytime enumeration," in *Advances in Neural Information Processing Systems* (NeurIPS, 2025), Vol. 38.
- <sup>19</sup>H. Jaeger, "The 'echo state' approach to analysing and training recurrent neural networks," Technical Report GMD Report 148, German National Research Center for Information Technology, 2001.
- <sup>20</sup>H. Jaeger and H. Haas, "Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication," *Science* **304**, 78–80 (2004).
- <sup>21</sup>M. Raissi, P. Perdikaris, and G. E. Karniadakis, "Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations," *J. Comput. Phys.* **378**, 686–707 (2019).
- <sup>22</sup>A. Chattopadhyay, M. Mustafa, P. Hassanzadeh, E. Bach, and K. Kashinath, "Deep learning for improving numerical weather prediction of heavy rainfall," *J. Adv. Model. Earth Syst.* **15**, e2022MS003516 (2023).
- <sup>23</sup>J. Pathak, S. Subramanian, P. Harrington, S. Raja, A. Chattopadhyay, M. Mardani, T. Kurth, D. Hall, Z. Li, K. Azizzadenesheli *et al.*, "FourCastnet: Accelerating global high-resolution weather forecasting using adaptive Fourier neural operators," in *Proceedings of the International Conference on Learning Representations*, 2024.
- <sup>24</sup>R. Lam, A. Sanchez-Gonzalez, M. Willson, P. Wirnsberger, M. Fortunato, F. Alet, S. Ravuri, T. Ewalds, Z. Eaton-Rosen, W. Hu *et al.*, "GraphCast: Learning skillful medium-range global weather forecasting," *Science* **382**, 1416–1421 (2023).
- <sup>25</sup>B. Wilder, B. Dilkina, and M. Tambe, "Melding the data-decisions pipeline: Decision-focused learning for combinatorial optimization," in *Proceedings of the AAAI Conference on Artificial Intelligence* (AAAI Press, 2019), Vol. 33, pp. 1658–1665.
- <sup>26</sup>A. N. Elmachtoub, J. C. N. Liang, and R. McNellis, "Smart predict-then-optimize for two-stage linear programs with side information," in *International Conference on Machine Learning* (PMLR, 2024), pp. 13019–13041.
- <sup>27</sup>Q. Zhou, J. Marecek, and R. N. Shorten, "Fairness in forecasting and learning linear dynamical systems," in *Proceedings of the AAAI Conference on Artificial Intelligence* (AAAI Press, 2021), Vol. 35, pp. 11134–11142.
- <sup>28</sup>J. Mandi, J. Kotary, S. Berden, M. Mulamba, V. Bucarey, T. Guns, and F. Fioretto, "Decision-focused learning: Foundations, state of the art, benchmark and future opportunities," *J. Artif. Intell. Res.* **80**, 1623–1701 (2024).
- <sup>29</sup>J. Zhou, H. Zhang, L. Li, Z. Sun, R. Peng, and Z. Xie, "SDWPF: A dataset for spatial dynamic wind power forecasting challenge at KDD cup 2022," in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (ACM, New York, 2022), pp. 4886–4895.