Finite-width recurrent kernels: From shared-weight convergence to task-relevant prediction risk and dynamical stability

Xinxin Wang1*

1Department of Precision Instrument, Center for Brain Inspired Computing Research, Tsinghua University, Beijing, 100084, CHINA

*E-mail: awxx_0415@163.com

Abstract

Reservoir computing (RC) uses fixed random recurrent dynamics with a lightweight trained readout, but practical reservoirs operate at finite width and reuse the same random operator across time. We study how finite-width reservoirs with shared recurrent randomness approach the recurrent-kernel limit and when improved kernel fidelity translates into improved prediction performance. Shared weights generate structured off-diagonal two-time correlations that are strongly suppressed by independent temporal resampling, showing that equal-time agreement alone does not capture the full temporal kernel structure. Finite shared-weight reservoirs nevertheless converge systematically toward the analytic kernel as width increases. Bias–variance analysis further shows that orthogonal and SORF constructions reduce finite-width error primarily by lowering realization variance, while SORF offers a favorable accuracy–memory trade-off through its implicit structured transform. Downstream experiments show that global kernel discrepancy only partially explains prediction risk. A task-aligned discrepancy based on regularized readout-relevant directions yields stronger within-method correlations with excess risk. Stateful Lorenz-63 forecasting reveals an additional dynamical constraint: wider reservoirs can extend valid horizons, but the benefit depends strongly on the recurrent operating point. Overall, finite-width reservoir performance depends on temporal covariance, task relevance, and dynamical error amplification in addition to kernel fidelity.

Keywords: reservoir computing; recurrent kernels; finite-width effects; structured random features; SORF; kernel alignment

1. Introduction

RC is a recurrent learning framework in which a high-dimensional nonlinear dynamical system transforms an input sequence into a state trajectory, while training is confined to a comparatively simple readout layer. The recurrent dynamics remain fixed, avoiding backpropagation through time and substantially reducing the optimization burden. This combination of nonlinear temporal representation and lightweight training has made RC useful for nonlinear time-series prediction, physical computing, signal processing, and data-driven modeling of dynamical systems (Jaeger, 2001a; Lukoševičius & Jaeger, 2009; Maass et al., 2002; Tanaka et al., 2019; Verstraeten et al., 2007). Reservoir computing has also become an established tool for data-driven forecasting of nonlinear and chaotic dynamics (Gauthier et al., 2021; Jaeger & Haas, 2004; Pathak et al., 2018; Vlachas et al., 2020). Broad approximation and universality results have been established for echo-state-type systems (Gonon et al., 2024; Gonon & Ortega, 2020, 2021; Grigoryeva & Ortega, 2018).

A useful theoretical perspective views sufficiently wide random reservoirs as implicit kernel machines. Infinite echo-state networks can be formulated as recurrent kernel machines, and later work established a recurrent-kernel limit for random reservoirs, including an $O(N^{-1/2})$ convergence result under restrictive assumptions (Dong et al., 2020; Hermans & Schrauwen, 2012). More recent work extends recurrent-kernel formulations to leaky, sparse, and deep reservoirs and studies their numerical convergence and computational efficiency (D'Inverno & Dong, 2025). This viewpoint also connects RC with wide-network Gaussian-process/kernel limits (Alemohammad et al., 2021; Fermanian et al., 2021; Lee et al., 2018; Matthews et al., 2018; Yang, 2019) and with random-feature and kernel-ridge

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

theory (Rahimi & Recht, 2007; Rudi & Rosasco, 2017; Schölkopf & Smola, 2002). A practical reservoir can therefore be studied through the discrepancy between its empirical recurrent representation and a deterministic infinite-width reference.

The recurrent setting introduces a complication that is absent from standard feedforward random-feature approximations. A conventional reservoir draws one recurrent matrix and reuses that same realization throughout an entire trajectory. By contrast, the convergence analysis of Dong et al. (2020) relies on a time-wise resampling assumption for the random recurrent matrix, which simplifies temporal dependencies and enables the theoretical derivation. The authors note that this assumption is introduced for analytical tractability and does not correspond to the standard RC protocol with a fixed recurrent realization. D'Inverno and Dong (2025) further identify time-wise resampling as a restrictive assumption and complement the theoretical analysis with numerical convergence studies for practical fixed-weight reservoirs. Existing recurrent-kernel results thus establish and extend the infinite-width reference, but they do not by themselves characterize the full finite-width two-time covariance generated by repeatedly reusing a single random recurrent realization. This shared-disorder temporal structure is the first gap addressed here.

Finite-width accuracy also depends on how the random recurrent operator is constructed. In feedforward random-feature models, finite-sample approximation error and variance depend on the sampling scheme, and orthogonalization can reduce redundancy among sampled directions and improve statistical efficiency (Sriperumbudur & Szabó, 2015; Sutherland & Schneider, 2015; Yu et al., 2016). Structured orthogonal random features such as SORF approximate this effect using fast transforms and compact random-sign representations (Yu et al., 2016). Structured reservoir computing based on Hadamard- or Fourier/Hartley-type transforms has already been proposed and shown to reduce the computational and memory cost of large reservoirs (D'Inverno & Dong, 2025; Dong et al., 2020). Accordingly, the novelty pursued here is not the structured transform itself, but a controlled finite-width comparison of how Gaussian, Rademacher, sparse, orthogonal, and SORF constructions distribute error between bias and realization variance, and how that statistical mechanism relates to accuracy-time-memory trade-offs.

A separate issue concerns the relevance of kernel approximation to downstream learning. Global kernel discrepancy measures all representation directions simultaneously, although a supervised readout may depend strongly on only a subset of them. Kernel-target alignment suggests that predictive value depends on the interaction between representation geometry and the target function (Cortes et al., 2012; Cristianini et al., 2001), with recent theory further quantifying how target-kernel alignment affects learning behavior in kernel methods (Wang et al., 2024). A distinct but related literature studies the stability of reservoir dynamics. In particular, Dong et al. (2022) use the recurrent-kernel limit to characterize the asymptotic frontier between stable and chaotic reservoir-state dynamics, while work on chaotic forecasting emphasizes that autonomous rollout quality depends on how local prediction errors are recursively amplified (Pathak et al., 2017, 2018; Platt et al., 2022; Wikner et al., 2024). These results motivate separating representation fidelity, supervised task sensitivity, intrinsic reservoir-state stability, and closed-loop forecast stability rather than treating them as a single notion of model quality.

The closest prior recurrent-kernel studies therefore leave a specific finite-width gap. Dong et al. (2020) establish the recurrent-kernel limit and structured RC, but their proof removes temporal weight sharing by resampling and their numerical convergence diagnostic compares matched-time reservoir/kernel quantities rather than a complete within-trajectory two-time covariance matrix. D'Inverno and Dong (2025) broaden the set of RC topologies and benchmark structured efficiency, but likewise focus on convergence to topology-specific recurrent-kernel limits rather than on shared-realization two-time fluctuations, a bias-variance decomposition, or task-aligned kernel perturbations.

Dong et al. (2022) analyze asymptotic reservoir-state stability, not the survival distribution of finite-width autonomous forecasts after learned outputs are fed back. The present work is positioned as complementary to these studies: it asks what is lost when temporal disorder sharing is removed, what controls finite-width fluctuations when it is retained, which kernel perturbations matter for a trained readout, and how those local errors are subsequently amplified in closed loop.

This study develops a finite-width analysis that follows this progression from recurrent representation to forecasting behavior. We first characterize the two-time covariance induced by reuse of a fixed recurrent realization and compare it with an independently resampled control. We then examine how reservoir width and recurrent construction determine convergence error, bias, realization variance, memory cost, and wall-clock behavior. The downstream analysis measures whether global kernel discrepancy predicts supervised risk and introduces a task-aligned discrepancy derived from regularized readout sensitivity. Finally, strict stateful Lorenz-63 rollouts are used to examine how finite-width prediction errors interact with recurrent operating point and closed-loop dynamical amplification.

The main contributions of this work are summarized as follows:

(1) We formulate and empirically validate a full shared-weight two-time recurrent-kernel diagnostic for a standard fixed reservoir. In contrast with the time-resampled assumption used in prior convergence proofs, the diagnostic preserves one recurrent realization across the entire trajectory and directly isolates the structured off-diagonal covariance that disappears under temporal resampling.

(2) We characterize finite-width convergence across Gaussian, Rademacher, sparse, orthogonal, and SORF reservoirs and decompose kernel MSE into squared bias and realization variance. This extends prior aggregate convergence comparisons by identifying the mechanism of finite-width error: over much of the studied regime, realization variance dominates, and orthogonal/SORF constructions reduce that component most strongly.

(3) Building on existing structured reservoir computing rather than introducing a new structured transform, we place all five recurrent constructions under a common accuracy-time-memory protocol. Dense orthogonal and SORF reservoirs reach comparable low-error regimes, while SORF realizes the near-orthogonal finite-width benefit with substantially lower recurrent-operator storage and competitive large-width wall-clock performance.

(4) We introduce a task-aligned finite-width discrepancy obtained by projecting kernel perturbations through regularized readout-relevant directions. Motivated by kernel-target alignment and first-order kernel-ridge sensitivity, this task-aware diagnostic exhibits stronger within-method correlations with excess prediction risk than a task-agnostic global kernel discrepancy.

(5) We distinguish intrinsic reservoir-state stability from learned closed-loop forecast stability by evaluating strict stateful Lorenz-63 rollouts with valid-horizon survival analysis. The resulting distributions reveal a joint dependence on finite width and recurrent operating point, and show how small local prediction differences that are not captured by global kernel error can be amplified after feedback.

Together, these results connect shared recurrent randomness, two-time covariance, finite-width fluctuations, task-relevant perturbations, and dynamical amplification within a unified framework for analyzing practical reservoir computers.

2. Methods

2.1 Recurrent dynamics and empirical kernels

Let x_t denote the input at time t and $h_t \in \mathbb{R}^N$ the reservoir state. We consider a time-homogeneous recurrent map with a fixed random recurrent matrix W, an input projection U, a bias b, and an elementwise nonlinearity φ, following the standard echo-state/reservoir-computing construction in which recurrent dynamics are fixed and the state is driven by the input history (Jaeger, 2001a; Lukoševičius & Jaeger, 2009; Yildiz et al., 2012):

$$\mathbf{h}_t = \phi(W\mathbf{h}_{t-1} + U\mathbf{x}_t + \mathbf{b}) \tag{1}$$

The same recurrent matrix W is reused for all time steps. Only the readout is trained, as in the standard reservoir-computing paradigm (Lukoševičius & Jaeger, 2009; Maass et al., 2002; Tanaka et al., 2019). Given a reservoir-state trajectory, the finite-width empirical recurrent kernel between two times is defined by the normalized Gram product

$$\widehat{K}_N(t,s) = \frac{1}{N}\mathbf{h}_t^\top \mathbf{h}_s \tag{2}$$

The matrix $\widehat{K}_N(t,s)$ contains both equal-time state correlations on its diagonal and the cross-time temporal covariance induced by recurrent state evolution.

2.2 Shared recurrent weights and temporal resampling

The standard RC protocol is shared-weight: one random matrix is drawn and reused throughout the sequence (Jaeger, 2001a; Lukoševičius & Jaeger, 2009). This practical protocol should be distinguished from the time-independence assumption used in the convergence proof of Dong et al. (2020), where the recurrent matrix is resampled at each time step to simplify the temporal dependence of reservoir updates. D'Inverno and Dong (2025) explicitly describe this time-wise resampling assumption as restrictive and study practical convergence numerically instead. We therefore retain the fixed recurrent realization as the primary process and use independent temporal resampling only as a control:

$$\begin{aligned} W_t &= W, \qquad shared - weight \\ W_t &\overset{i.i.d.}{\sim} \mathcal{P}_W, \qquad resampled \end{aligned} \tag{3}$$

These constructions can share an equal-time mean-field recurrence while differing strongly in their joint temporal statistics. In particular, for off-diagonal time pairs we generally have

$$K_{\text{shared}}(t,s) \neq K_{\text{resampled}}(t,s), \qquad t \neq s \tag{4}$$

This distinction is central: independent resampling can reduce variance for selected marginal estimators, but it defines a different stochastic process from a standard fixed reservoir. Prior recurrent-kernel work already identifies temporal resampling as a proof device rather than the practical RC protocol (D'Inverno & Dong, 2025; Dong et al., 2020). Our purpose is to quantify the consequence of that distinction at finite width: diagonal quantities such as equal-time state correlations can agree while the off-diagonal two-time covariance generated by a single shared recurrent realization remains different. Agreement of equal-time marginals is therefore not sufficient evidence of agreement of the recurrent temporal process.

2.3 Analytic shared-weight two-time recurrent kernel

For the Gaussian reference ensemble, recurrent-kernel and wide-network analyses provide deterministic covariance recursions in the infinite-width limit (D'Inverno & Dong, 2025; Dong et al.,

2020; Hermans & Schrauwen, 2012; Lee et al., 2018; Matthews et al., 2018; Yang, 2019). Here we use the corresponding Gaussian covariance recursion as a reference for the full covariance matrix over all time pairs (t,s), rather than only a matched-time or terminal Gram quantity. Let $q_{t,s}$ denote the preactivation covariance. Under width-normalized recurrent and input scaling, the recursion used in the present experiments has the form

$$q_{t,s} = \rho^2 K_{t-1,s-1} + \sigma_{\text{in}}^2 \frac{\mathbf{x}_t^\top \mathbf{x}_s}{d} + \sigma_b^2 \tag{5}$$

where ρ denotes the recurrent weight scaling factor.

The postactivation covariance is obtained by applying the Gaussian activation covariance map

$$K_{t,s} = \Psi_\phi\left(q_{t,t}, q_{s,s}, q_{t,s}\right) \tag{6}$$

where Ψϕ is the covariance map induced by the activation function.

For the erf activation used in the principal theoretical experiments, this map is analytic. The resulting matrix K∞ is evaluated as a full two-time reference, so its off-diagonal entries can be compared directly with the empirical covariance generated by a reservoir that reuses one fixed recurrent realization. This full-matrix use is more comprehensive than convergence diagnostics based only on equal-time or terminal Gram quantities in earlier recurrent-kernel studies (D'Inverno & Dong, 2025; Dong et al., 2020).

We quantify empirical convergence with the relative full-matrix Frobenius error

$$E_{2\text{time}} = \frac{\|\widehat{K}_N - K_\infty\|_F}{\|K_\infty\|_F} \tag{7}$$

and separately report a normalized equal-time discrepancy

$$E_{\text{eq}} = \frac{\sum_t \left(\widehat{K}_N(t,t) - K_\infty(t,t)\right)^2}{\sum_t K_\infty\,(t,t)^2} \tag{8}$$

The equal-time metric is useful for finite-width scaling diagnostics, whereas the two-time metric is the appropriate object for validating the shared recurrent kernel itself.

2.4 Random reservoir constructions

2.4.1 Gaussian and Rademacher reservoirs

The dense Gaussian reference uses independent entries with width normalization

$$W_{ij} \sim \mathcal{N}\left(0, \frac{\sigma_w^2}{N}\right) \tag{9}$$

The Rademacher ensemble replaces Gaussian entries by independent ±1 signs under the same second-moment scaling. Both constructions are unstructured in the sense that pairwise feature directions are not explicitly decorrelated. More generally, random-feature theory relates finite feature dimension and sampling variability to kernel-approximation error (Rahimi & Recht, 2007; Rudi & Rosasco, 2017; Sriperumbudur & Szabó, 2015; Sutherland & Schneider, 2015).

2.4.2 Orthogonal reservoirs

For the orthogonal baseline we draw a random orthogonal matrix and apply a global recurrent scaling γ:

$$W = \gamma Q, \qquad Q^\top Q = I \tag{10}$$

where γ controls the recurrent gain and is matched across reservoir constructions.

Orthogonality explicitly limits directional redundancy among rows of the recurrent map, providing a direct test of whether finite-width variance reduction originates from geometric structure rather than from a particular marginal distribution. In feedforward kernel approximation, orthogonal random features were introduced precisely to reduce approximation error relative to independent Gaussian features (Yu et al., 2016).

2.4.3 Sparse reservoirs

Sparse reservoirs retain only a prescribed fraction of nonzero recurrent connections, a common reservoir-design choice in the ESN literature (D'Inverno & Dong, 2025; Jaeger, 2001a; Lukoševičius & Jaeger, 2009). In the validated implementation, sparse matrices are stored in genuine CSR form rather than as dense matrices with zeros, so the memory comparison reflects the implemented operator rather than only arithmetic sparsity.

2.4.4 Structured orthogonal random features (SORF)

SORF implements a structured randomized orthogonal transform through repeated normalized Hadamard transforms and random-sign diagonal matrices, following the structured orthogonal random-feature construction of Yu et al. (2016) and its use in structured reservoir/kernel methods (D'Inverno & Dong, 2025; Dong et al., 2020):

$$W = \gamma H D_1 H D_2 H D_3 \tag{11}$$

The construction avoids storing a dense N×N recurrent matrix. Only a small number of sign vectors, together with the implicit Hadamard transform, are required. This structured representation is the basis of the reduced storage and fast-transform properties of SORF (Dong et al., 2020; Yu et al., 2016) and is essential in the accuracy–memory analysis.

2.5 Experimental design

2.5.1 Widths and random realizations

The core convergence study uses widths $N \in \{64,128,256,512,1024\}$. Equal-time convergence is evaluated for shared and resampled reservoirs, while the more comprehensive full two-time diagnostic is reported for the standard shared-weight construction. The bias–variance decomposition is evaluated over the same five-width sweep, and the prediction-risk alignment study also uses $N \in \{64,128,256,512,1024\}$. Downstream experiments evaluate RC widths 256 and 512 and aggregate performance over 10 independent data seeds, averaging nested reservoir realizations within each data seed before constructing cross-seed confidence intervals. The autonomous width-survival diagnostic uses $N \in \{128,256,512,1024\}$. This hierarchical aggregation prevents multiple reservoir draws on the same dataset realization from being treated as independent samples.

For convergence-rate analysis, we fit the observed width dependence with an empirical scaling law:

$$E(N) = CN^{-\alpha} \tag{12}$$

Random-feature theory provides the broader finite-sample context for interpreting how approximation error decays with feature dimension (Rudi & Rosasco, 2017; Sriperumbudur & Szabó, 2015; Sutherland & Schneider, 2015):

We distinguish a smaller finite-width prefactor C from a genuinely different asymptotic convergence exponent α; a lower error at a few widths alone is not sufficient evidence for a different asymptotic exponent. This distinction is important in finite random-feature comparisons because structured sampling can change constants and variance without necessarily changing the leading

dimension-dependent rate (Sriperumbudur & Szabó, 2015; Sutherland & Schneider, 2015; Yu et al., 2016).

2.5.2 Kernel approximation metrics

In addition to Eqs. (7)–(8), we evaluate a regularized global spectral discrepancy:

$$D_{\text{global}} = \left\|(K + \lambda I)^{-1/2}\left(\widehat{K} - K\right)(K + \lambda I)^{-1/2}\right\|_2 \tag{13}$$

where K and $\widehat{K}$ denote the reference and empirical kernels, respectively.

The regularization λ matches the prediction setting and stabilizes poorly conditioned directions, but the metric remains task agnostic because it does not use Y.

2.5.3 Stateful autonomous forecasting

2.5.3.1 Strict stateful closed-loop protocol

For Lorenz-63, reservoir state is synchronized once on an observed warm-up segment and then propagated continuously without resetting. The readout uses the persistent hidden state together with a direct input skip connection and predicts the next dynamical-system state. During test rollout, predicted states are fed back into the same reservoir; no ground-truth state is reintroduced and the hidden state is never reset at artificial window boundaries. This strict autonomous protocol follows the closed-loop use of reservoirs for chaotic-system forecasting, where one-step errors are recursively reintroduced and can accumulate over time (Haluszczynski & Räth, 2019; Jaeger & Haas, 2004; Pathak et al., 2017, 2018; Vlachas et al., 2020). The width-survival diagnostic sweeps $N \in \{128,256,512,1024\}$ over recurrent operating points $\chi \in \{0.70,0.90,0.99,1.00,1.03,1.10\}$ using a 512-step valid-horizon observation window. A separate representative rollout uses the paper-profile width N=1024 and follows the normalized Lorenz state for 1,500 autonomous forecast steps.

2.5.3.2 Valid horizon and right censoring

For normalized prediction error e_h and a tolerance ε, the valid forecast horizon is the first threshold crossing

$$H_\varepsilon = \inf\{h : e_h > \varepsilon\} \tag{14}$$

If no crossing occurs before the observation limit, the run is right censored rather than recorded as a failure at the final step. For each method we estimate the Kaplan–Meier survival function (Kaplan & Meier, 1958)

$$S_\varepsilon(h) = \Pr(H_\varepsilon > h) \tag{15}$$

and summarize the survival curve by the restricted mean valid horizon (RMVH), using the restricted-mean-survival-time principle for censored time-to-event data (Royston & Parmar, 2013)

$$\text{RMVH}_\varepsilon(\tau) = \int_0^\tau S_\varepsilon(h)\,\mathrm{d}h \tag{16}$$

The width-dependent survival analysis uses a fixed normalized-error threshold ε=0.25 across $\chi \in \{0.70,0.90,0.99,1.00,1.03,1.10\}$, where χ denotes the recurrent spectral-radius scaling factor. This common threshold isolates how reservoir width and recurrent operating point reshape the distribution of valid horizons. Runs that remain below threshold through the 512-step observation limit are treated as right censored rather than as failures at the final step.

2.5.4 Bias–variance decomposition

Let $\widehat{K}_N^{(r)}$ be the empirical kernel generated by random realization r and let $\overline{K}_N = \mathbb{E}_r\left[\widehat{K}_N^{(r)}\right]$ denote the ensemble mean. The expected kernel mean-squared error decomposes into a squared bias term and a realization-variance term:

$$\mathbb{E}_r\left[\left\|\widehat{K}_N^{(r)} - K_\infty\right\|_F^2\right] = \left\|\overline{K}_N - K_\infty\right\|_F^2 + \mathbb{E}_r\left[\left\|\widehat{K}_N^{(r)} - \overline{K}_N\right\|_F^2\right] \tag{17}$$

Bias measures systematic deviation of the finite-width ensemble mean from the infinite-width theory. Realization variance quantifies sensitivity to the particular random reservoir realization. This is the standard bias–variance decomposition of mean-squared error (Geman et al., 1992). The decomposition is especially informative for structured random features because two reservoirs can have comparable total MSE for very different reasons.

2.5.5 Task-aligned kernel discrepancy

2.5.5.1 Task-agnostic global kernel discrepancy

Let K denote a reference kernel and $\hat{K}$ its finite-width approximation. Write

$$\widehat{K} = K + \Delta K \tag{18}$$

where ΔK denotes the finite-width kernel perturbation.

A global norm of ΔK treats all kernel directions as equally relevant. A supervised predictor does not. For kernel ridge regression with training targets Y, the reference dual readout coefficients are defined by the standard regularized kernel solution (Schölkopf & Smola, 2002):

$$A = (K + \lambda I)^{-1} Y \tag{19}$$

The first-order resolvent expansion of the coefficient perturbation gives

$$\hat{A} - A = -(K + \lambda I)^{-1} \Delta K\, A + \mathcal{O}(\| \Delta K \|^2) \tag{20}$$

The leading task-relevant term therefore contains ΔK A. Kernel perturbations whose action on the readout coefficient vector is small can have a large global magnitude while producing little prediction change. This motivation parallels kernel-alignment ideas, which assess kernels through their relation to target-relevant geometry rather than through a target-agnostic matrix discrepancy alone (Cortes et al., 2012; Cristianini et al., 2001).

2.5.5.2 Task-aligned discrepancy and spectral interpretation

We consequently define the normalized task-aligned discrepancy

$$D_{\mathrm{TA}} = \frac{\left\|\left(\widehat{K} - K\right)A\right\|_F}{\| KA \|_F + \varepsilon_0} \tag{21}$$

The readout-sensitive directions are constructed exclusively from training targets. Test labels are never used to select or weight kernel directions. This avoids post-hoc cherry-picking and makes the metric a legitimate diagnostic for generalization.

The same construction also admits a direct spectral interpretation.

If the reference kernel is diagonalized as

$$K = U \Lambda U^{\top} \tag{22}$$

then, for a scalar target, the effective contribution of eigenmode u_i to the regularized readout is weighted approximately by

$$w_i \propto \frac{u_i^\top y}{\lambda_i + \lambda} \tag{23}$$

where w_i denotes the coefficient of the i-th eigenmode in the kernel eigenbasis.

Hence a large kernel approximation error in a mode with $u_i^T y \approx 0$ can dominate a global matrix norm while contributing little to the learned predictor. Task-aligned discrepancy retains all modes but automatically weights them by their actual contribution to the learned predictor, consistent with the principle that task-relevant kernel geometry matters for supervised performance (Cortes et al., 2012; Cristianini et al., 2001).

2.5.6 Runtime and memory accounting

End-to-end wall time is measured after warm-up and summarized by the median across repeated runs for each width $N \in \{64,128,256,512,1024\}$. Operator memory is estimated from the actual implementation: dense Gaussian, Rademacher, and orthogonal operators store N^2 floating-point recurrent weights; SORF stores only a small number of O(N) random-sign vectors together with the implicit Hadamard transform; and sparse operators use CSR data, indices, and indptr storage. The SORF complexity comparison follows its structured sign/Hadamard representation (Yu et al., 2016) and its use in structured reservoir computing (Dong et al., 2020). Runtime therefore reflects both asymptotic operator cost and implementation-dependent library overhead, whereas the memory scaling reveals a more fundamental structural difference between dense, sparse, and implicit structured operators.

2.5.7 Downstream temporal benchmarks

We evaluate five complementary temporal benchmarks under a common stateful protocol. Lorenz-63 and Lorenz-96 probe chaotic dynamics (Chattopadhyay et al., 2020; Lorenz, 1963, 1995), Mackey–Glass probes delayed nonlinear dynamics (Mackey & Glass, 1977), and NARMA10/NARMA30 probe driven nonlinear temporal processing (Jaeger, 2002; Wringe et al., 2025). Nonreservoir baselines are linear ridge regression (Hoerl & Kennard, 1970) and Extra Trees (Geurts et al., 2006). The tasks are intentionally heterogeneous: they test whether a global ordering of reservoir-kernel approximation quality transfers consistently across distinct prediction regimes rather than defining a single universal benchmark score.

2.5.8 Statistical protocol and reproducibility

All hyperparameter selection is performed on training and validation data. The same split, target normalization, and hyperparameter-selection protocol are shared across compared reservoir constructions. For downstream experiments, reservoir seeds are nested within data seeds; reservoir draws are averaged within a data seed before cross-data-seed confidence intervals are computed.

Kernel-convergence figures report uncertainty across independent reservoir realizations. Downstream comparisons use paired statistics tests when data seeds and splits are shared; for heavy-tailed autonomous errors, bootstrap intervals and nonparametric paired tests are preferred (Efron & Tibshirani, 1993; Wilcoxon, 1945).

Task-relevant directions are derived only from the training kernel and Y_train. The test target is used solely for final prediction-risk evaluation. This separation is essential: using test labels to choose aligned modes would constitute information leakage and would invalidate the correlation analysis.

For valid-horizon analysis, event_observed=1 only at a threshold crossing; runs that remain below threshold at the observation limit are coded event_observed=0 and right_censored=1. Kaplan–Meier estimates use this event definition (Kaplan & Meier, 1958), and RMVH integrates survival through a common restriction time (Royston & Parmar, 2013).

3. Results

3.1 Shared recurrent weights determine the full temporal kernel structure

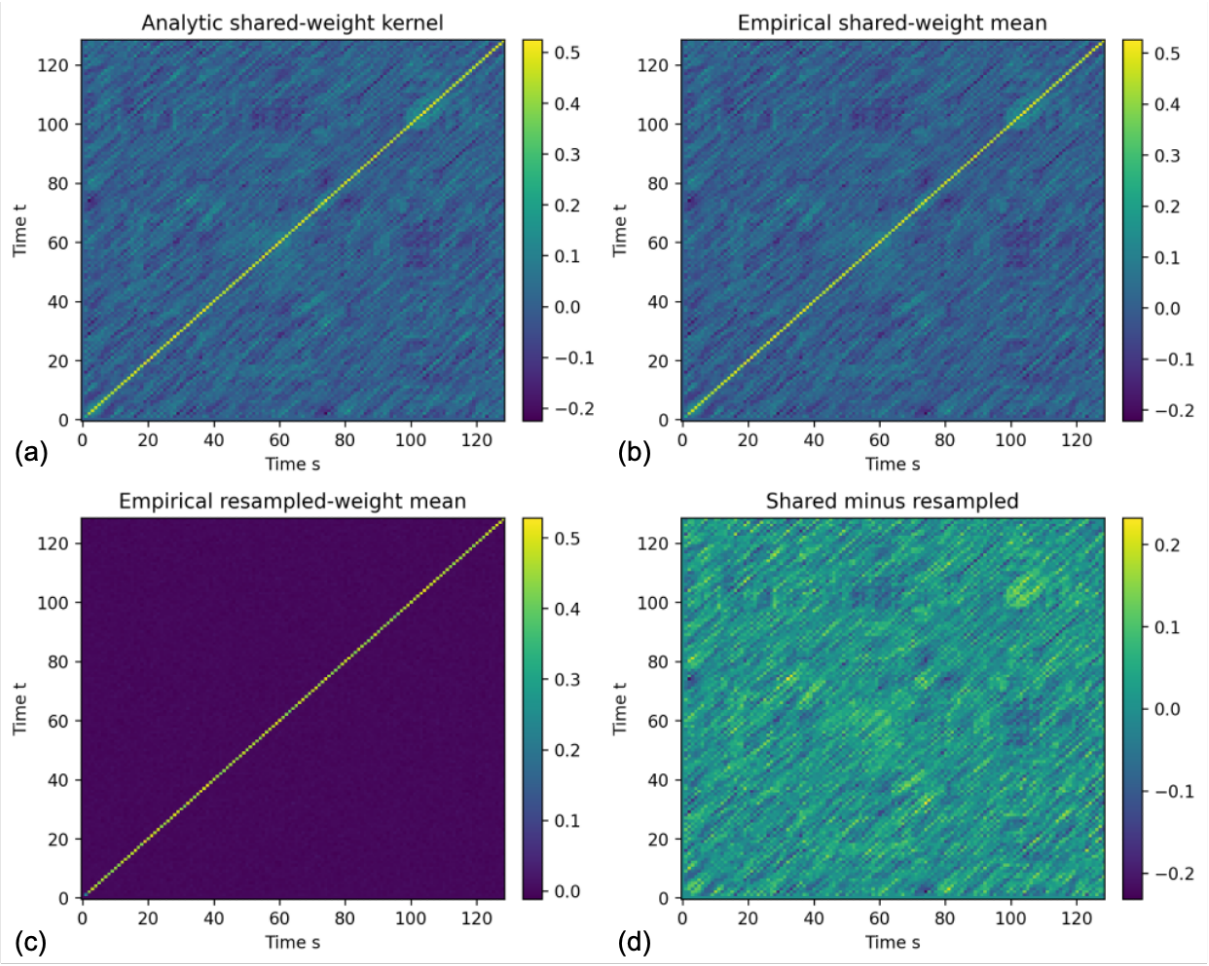

Figure 1. Analytic and empirical temporal kernels under shared and independently resampled recurrent weights. (a) Analytic shared-weight kernel. (b) Empirical shared-weight ensemble mean, reproducing both the dominant main diagonal and the structured positive and negative off-diagonal correlations. (c) Empirical resampled-weight mean, which retains the equal-time diagonal but strongly suppresses cross-time covariance. (d) Shared-minus-resampled difference. The coherent residual demonstrates that weight sharing generates systematic temporal correlations rather than unstructured Monte Carlo noise.

Figure 1 provides the most direct validation of the shared-weight theory. The analytic and empirical shared-weight matrices exhibit close agreement not only along the main diagonal, but also in the structured positive and negative off-diagonal bands spanning widely separated time indices. This agreement demonstrates that the theory captures the full two-time covariance structure, rather than merely equal-time state covariance. Because the same recurrent transformation is repeatedly applied throughout a trajectory, reservoir states at different times remain statistically coupled, giving rise to the characteristic off-diagonal patterns. These structures therefore provide a covariance-level signature of recurrent temporal dependence. Importantly, however, they should not be interpreted as a direct measure of formal memory capacity (Dambre et al., 2012; Jaeger, 2001b).

The independently resampled control exhibits a qualitatively different covariance structure. Its empirical mean preserves the dominant equal-time diagonal, whereas most of the structured off-diagonal covariance is strongly attenuated. Consistently, the shared-minus-resampled difference displays coherent bands and extended matrix patterns whose amplitudes constitute a non-negligible fraction of the overall kernel scale. These residual structures are therefore unlikely to arise from

unstructured finite-sample fluctuations. Instead, they provide a systematic signature of the cross-time correlations induced by repeatedly reusing the same recurrent realization.

3.2 Finite-width convergence to the analytic shared-weight kernel

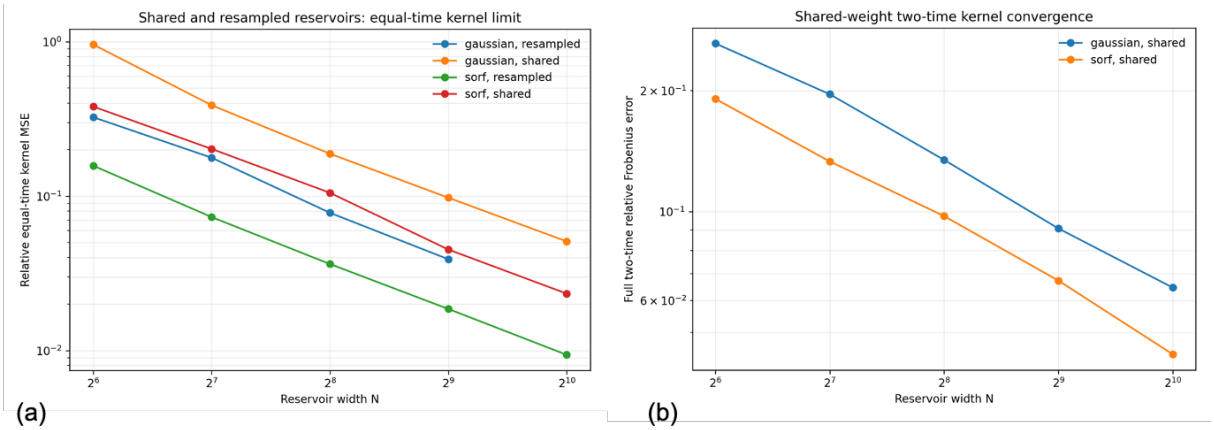

Figure 2. Finite-width convergence of equal-time and full two-time kernels. (a) Relative equal-time kernel MSE at T=64 for Gaussian and SORF reservoirs with shared or independently resampled weights over N=64,128,256,512,1024. Resampling lowers the diagonal-estimation error at matched width, while SORF remains below Gaussian in both protocols. (b) Relative Frobenius error of the complete shared-weight two-time kernel over the same width range. Both curves decrease approximately as N^-1/2; SORF remains consistently lower, indicating a smaller finite-width prefactor rather than a visibly different convergence exponent.

Figure 2(a) establishes a clear finite-width trend in the equal-time error across five reservoir sizes. All curves decrease monotonically as N increases from 64 to 1024, providing evidence of systematic convergence toward the infinite-width limit. Within each random-feature family, the independently resampled control consistently yields lower equal-time error than its shared-weight counterpart, consistent with reduced variance in the Monte Carlo estimator of diagonal kernel entries. This observation is consistent with Figure 1: independent resampling can improve the estimation of equal-time moments while simultaneously suppressing the off-diagonal temporal covariance characteristic of a standard fixed reservoir. SORF further outperforms Gaussian reservoirs under both sampling protocols, consistent with finite-feature variance/error reductions from structured orthogonal features (Sutherland & Schneider, 2015; Yu et al., 2016).

Figure 2(b) provides a more comprehensive convergence test for the full shared-weight temporal kernel. The relative Frobenius error for Gaussian reservoirs decreases from approximately 0.28 at N=64 to 0.065 at N=1024, while the corresponding SORF error decreases from approximately 0.19 to 0.045. Thus, increasing the reservoir width reduces the full two-time kernel error by approximately 75–80% for both constructions, with SORF consistently achieving errors about 23–33% lower than Gaussian across all displayed widths. A log–log fit yields empirical scaling exponents close to 1/2 (approximately 0.53 for Gaussian and 0.51 for SORF), indicating that the two constructions follow essentially the same finite-width convergence rate but with different prefactors. In other words, SORF improves the finite-width approximation primarily by reducing the error magnitude rather than by changing its asymptotic scaling with width. Together with Figure 1, these results support both components of the kernel-limit claim: the analytic theory captures the shared-weight two-time covariance structure, and finite-width reservoirs converge systematically toward this limit as N increases.

3.3 Bias–variance mechanism of structured randomization

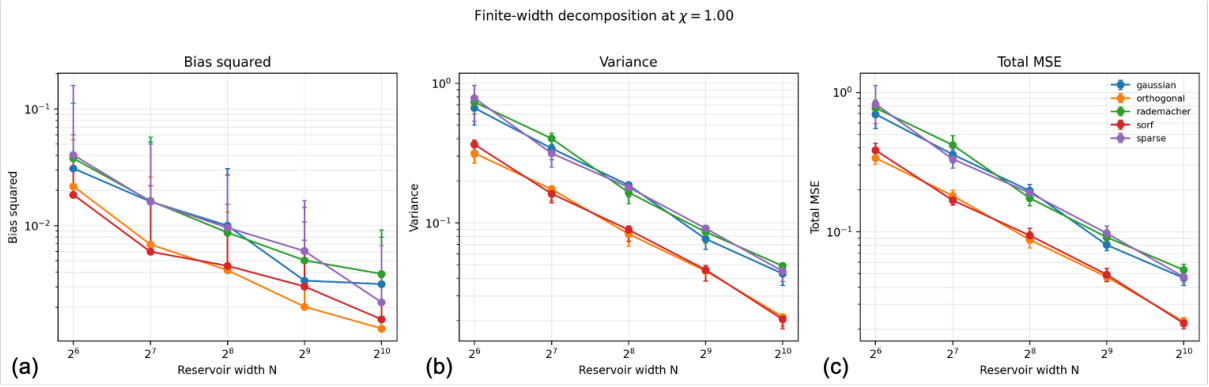

Figure 3. Finite-width bias–variance decomposition at χ=1.00 across N=64,128,256,512,1024. (a) Squared bias of the ensemble-mean kernel. (b) Realization variance across independent reservoirs. (c) Total kernel MSE. Bias decreases with width for all five constructions, whereas the separation in total MSE is dominated by realization variance. Orthogonal and SORF reservoirs remain among the lowest-variance and lowest-MSE constructions throughout the sweep, while Gaussian, Rademacher, and sparse reservoirs retain larger realization-to-realization fluctuations. Error bars summarize independent reservoir realizations.

The expanded bias–variance decomposition clarifies the mechanism underlying finite-width error more effectively than a comparison of total MSE alone, as shown in Figure3. Across all five reservoir constructions, the squared bias decreases systematically with increasing width and becomes small by N=1024. Over most of the width range, however, the realization-variance term remains substantially larger than the corresponding squared-bias term. As a result, the ordering of the total MSE closely mirrors that of the variance component rather than the bias component. These results indicate that, within the finite-width regime studied here, approximation accuracy is governed primarily by realization-to-realization variance rather than systematic bias.

Orthogonal and SORF reservoirs exhibit the strongest reduction in realization variance. Across much of the width range, their variance and total-MSE curves closely track each other and decrease to approximately 2×10^{-2} at N=1024, whereas Gaussian, Rademacher, and sparse reservoirs remain at roughly $4–5\times10^{-2}$. This advantage is already apparent at small widths and persists as N increases. The consistent reduction supports the interpretation that structured orthogonalization reduces redundancy among finite random directions, thereby reducing realization-to-realization fluctuations and stabilizing the empirical recurrent kernel; this interpretation is consistent with the motivation and finite-feature results of orthogonal random features (Yu et al., 2016).

The bias results, however, show that this improvement should not be attributed to a uniformly smaller systematic error. Orthogonal and SORF alternate in achieving the lowest squared bias across widths, while other constructions can attain comparable bias at particular values of N despite having substantially larger total MSE. Together, these results identify realization variance as the dominant factor controlling finite-width kernel accuracy in the studied regime. Conceptually, structured orthogonality ⟶ lower realization variance ⟶ lower finite-width kernel MSE. SORF therefore does not provide a fundamentally different kernel-approximation accuracy advantage over dense orthogonal reservoirs. Its additional practical advantage is computational: it realizes an orthogonal transformation through structured operations, avoiding the need to store and apply a dense N×N recurrent matrix (Dong et al., 2020; Yu et al., 2016).

3.4 Accuracy–efficiency trade-offs

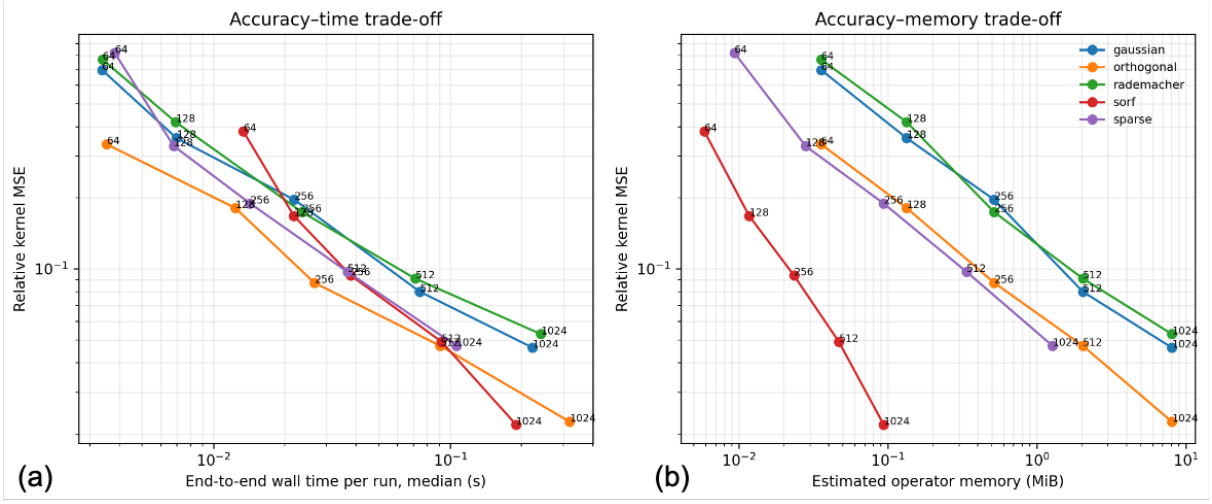

Figure 4. Accuracy–efficiency trade-offs across N=64,128,256,512,1024. (a) Relative kernel MSE versus median end-to-end wall time per run; (b) The same error versus estimated recurrent-operator memory. Point labels indicate reservoir width. Orthogonal and SORF constructions trace the lowest-error frontier, while SORF becomes especially favorable at large width because it achieves kernel MSE comparable to or slightly below dense orthogonal reservoirs without storing an N×N recurrent matrix. Sparse storage reduces memory relative to dense operators but does not match the SORF accuracy–memory frontier.

The wall-time analysis (Figure 4) shows that no single reservoir construction uniformly minimizes latency across all widths. At small N, optimized dense linear-algebra routines remain competitive because fixed software overheads associated with execution, memory allocation, and kernel dispatch are comparable to the cost of the recurrent transformation itself. As N increases, however, SORF achieves an increasingly favorable accuracy–time trade-off: at N=1024, SORF reaches one of the lowest kernel errors without incurring the largest runtime among the compared constructions. These results indicate an accuracy–time advantage for SORF in the large-width regime, while emphasizing that the precise wall-time ranking remains dependent on implementation and hardware.

The memory comparison provides a more implementation-independent structural result. Dense Gaussian, Rademacher, and orthogonal reservoirs incur $O(N^2)$ persistent storage for their recurrent operators, causing memory requirements to grow rapidly with width. Sparse CSR reservoirs occupy an intermediate regime, reducing storage relative to dense matrices while retaining a cost proportional to the number of stored nonzero entries. In contrast, SORF represents the recurrent transformation implicitly through $O(N)$ sign vectors and structured fast transforms, allowing operator memory to scale linearly with width; this storage distinction follows directly from the SORF construction (Yu et al., 2016) and structured-reservoir implementations (Dong et al., 2020). At the largest width considered, SORF and dense orthogonal reservoirs achieve essentially the same low-error regime, but SORF does so with substantially lower operator-memory requirements. Thus, SORF achieves the best accuracy–memory balance among the evaluated constructions, whereas the runtime advantage depends on the reservoir width and implementation.

3.5 Downstream temporal prediction

We first examine whether more accurate approximation of the recurrent-kernel limit is associated with better predictive performance. All reservoir constructions are evaluated on the benchmark suite using identical data splits, tuning procedures, and stateful processing settings. We then compare the task-wise predictive rankings with those obtained from global kernel error to assess how consistently kernel fidelity reflects downstream performance.

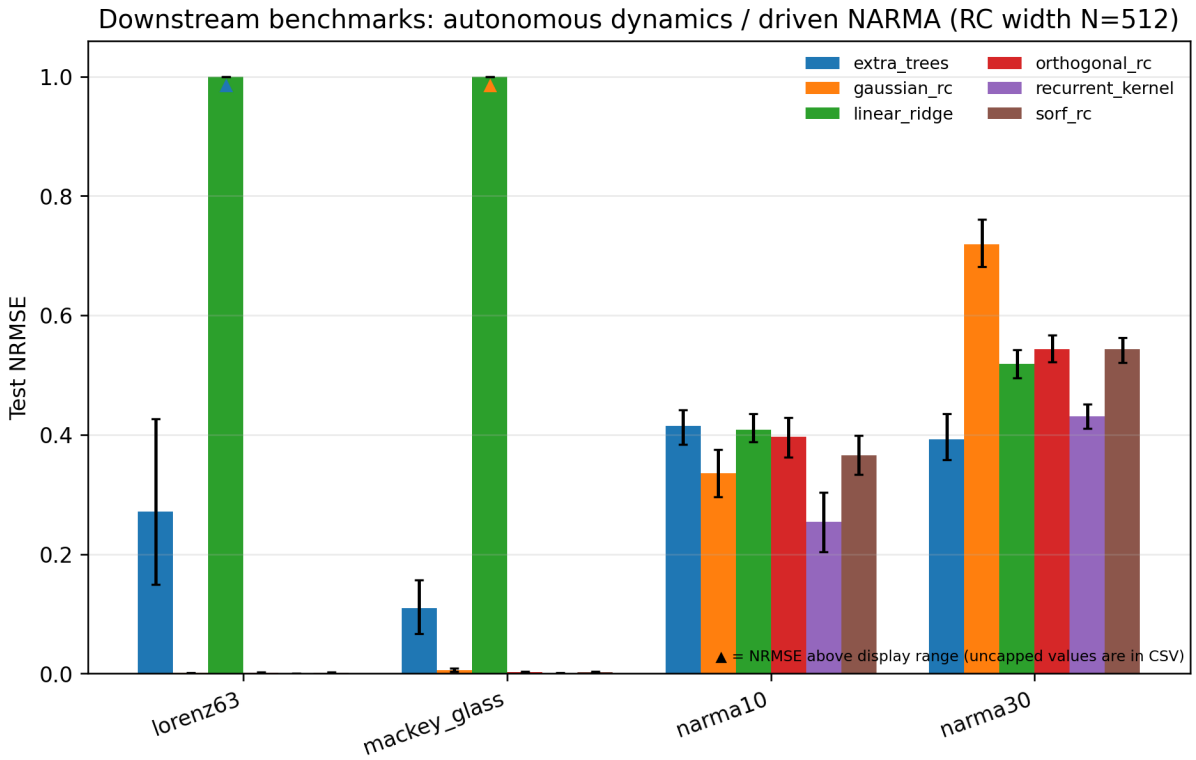

Figure 5. Paper-scale downstream sequence benchmarks at RC width N=512. The expanded suite includes Lorenz-63, Mackey–Glass, NARMA10, and NARMA30. Error bars summarize independent data-seed variation after nested reservoir realizations are averaged within each data seed. The ordering is strongly task dependent: improved finite-width kernel approximation does not translate into a universal SORF advantage.

The downstream benchmark results (Figure 5) show that global kernel MSE does not provide a consistent ranking of predictive performance across tasks. Across Lorenz-63, Mackey–Glass, NARMA10, and NARMA30, the relative ordering of the methods changes substantially from one task to another. Although SORF and orthogonal reservoirs retain lower finite-width kernel errors, these improvements do not consistently correspond to lower test NRMSE than Gaussian RC. Global kernel fidelity therefore provides only a partial indication of downstream predictive quality.

The task dependence is particularly evident on the driven NARMA benchmarks. On NARMA10, the analytic recurrent-kernel reference achieves the lowest NRMSE, while finite recurrent reservoirs show smaller performance differences than suggested by their different global kernel errors. On NARMA30, the ordering changes again: Extra Trees achieves the lowest error among the displayed methods, the recurrent-kernel reference remains competitive, and Gaussian RC performs substantially worse than the orthogonal and SORF variants. These results show that the ranking obtained from kernel approximation error is not preserved as a task-level prediction ranking.

Lorenz-63 and Mackey–Glass provide a complementary view under stateful sequence processing. Several recurrent constructions achieve very low NRMSE under teacher-forced evaluation on these systems, yet their relative performance still does not follow the ordering implied by global kernel discrepancy. Kernel approximation accuracy and predictive accuracy are therefore related indirectly and depend on how the task probes the reservoir representation. Moreover, teacher-forced sequence NRMSE does not characterize long-horizon autonomous stability, a distinction emphasized in work on autonomous reservoir forecasting of chaotic attractors (Gupta et al., 2023; Haluszczynski & Räth, 2019; Pathak et al., 2017, 2018; Platt et al., 2022; Vlachas et al., 2020; Wikner et al., 2024). Section 3.7 therefore evaluates Lorenz-63 separately using strict closed-loop valid-horizon survival.

3.6 Task-aligned discrepancy and excess risk

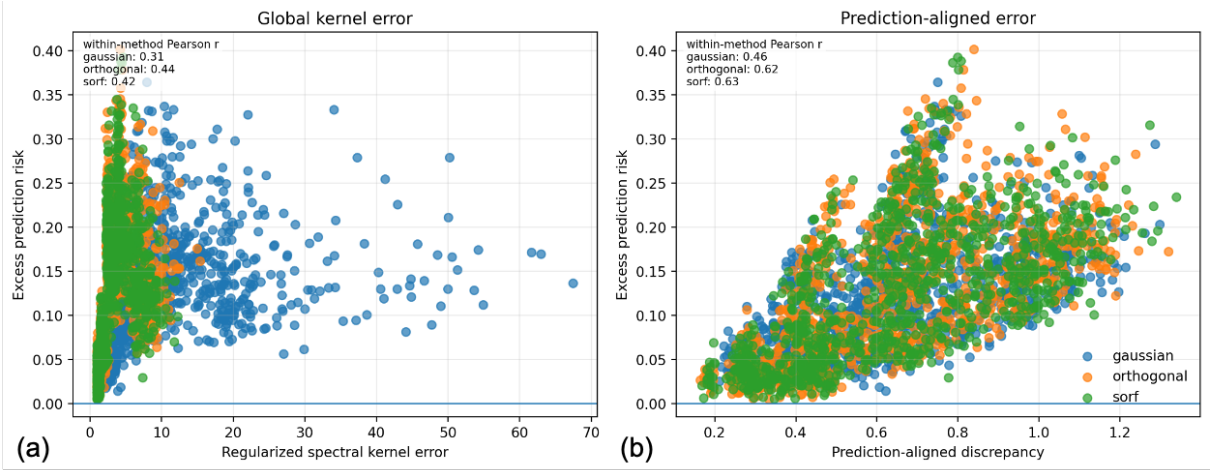

Figure 6. Global versus task-aligned kernel discrepancy. (a) Regularized global spectral kernel error versus excess prediction risk. (b) Task-aligned discrepancy versus excess prediction risk. Within-method Pearson correlations increase from 0.31, 0.44, and 0.42 under the global metric to 0.46, 0.62, and 0.63 under the task-aligned metric for Gaussian, orthogonal, and SORF reservoirs, respectively.

Figure 6 examines whether the task-agnostic global kernel discrepancy can predict downstream excess risk. Using the regularized global spectral error, the Pearson correlations computed within each reservoir family are 0.31 for Gaussian, 0.44 for orthogonal, and 0.42 for SORF reservoirs. These weak-to-moderate positive correlations indicate that global kernel fidelity is related to downstream performance, yet a substantial fraction of the variation in excess risk remains unexplained. Configurations with comparable global kernel error can therefore still exhibit noticeably different prediction risk.

When the same finite-width perturbations are projected onto regularized readout-relevant directions, the correlations increase to 0.46, 0.62, and 0.63 for Gaussian, orthogonal, and SORF reservoirs, respectively. The improvement is consistent across all three reservoir families, corresponding to correlation gains of 0.15, 0.18, and 0.21. The task-aligned discrepancy therefore shows a stronger association with downstream risk than the global spectral norm, although substantial scatter remains. The readout-sensitive directions are constructed exclusively from the training targets and ridge regularization; test labels enter only in the final evaluation of excess prediction risk.

These results strengthen the connection between finite-width kernel approximation and downstream learning. A reservoir can lie closer to the recurrent-kernel limit in a global sense while concentrating part of its residual error in directions that have limited influence on the learned readout. Conversely, a smaller global perturbation can produce larger prediction risk when it is concentrated in target-relevant directions. This observation is consistent with the broader kernel-alignment literature, in which predictive usefulness depends on the relation between kernel geometry and targets (Cortes et al., 2012; Cristianini et al., 2001). The higher correlations observed in Figure 6 indicate that task-aligned discrepancy provides a more informative diagnostic of task-level sensitivity than a purely global kernel norm. The remaining scatter also indicates that this quantity explains only part of the variation in downstream prediction error.

3.7 Autonomous forecasting and dynamical stability

3.7.1 Width-dependent valid-horizon survival

We next move from stateful supervised prediction to strictly closed-loop forecasting, where model outputs are recursively reintroduced as inputs. Lorenz-63 provides a stringent test because small local errors can be amplified by chaotic dynamics (Lorenz, 1963; Pathak et al., 2017; Platt et al., 2022;

Wikner et al., 2024). We therefore characterize autonomous forecast stability using Kaplan–Meier valid-horizon curves, which retain information from trajectories that remain accurate up to the finite observation limit (Kaplan & Meier, 1958).

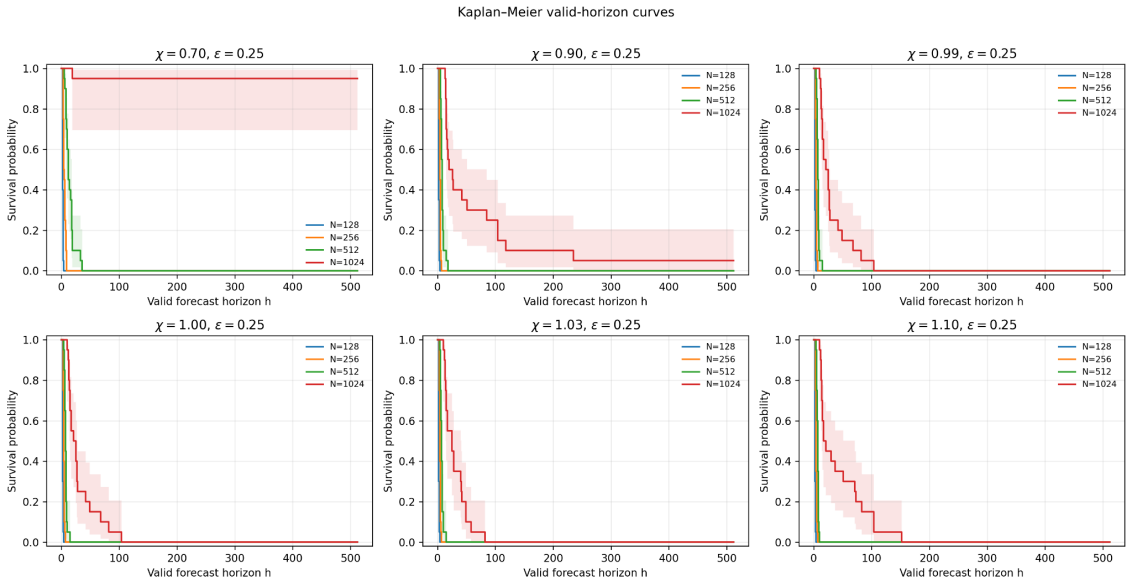

Figure 7. Width-dependent Kaplan–Meier valid-horizon survival for Lorenz-63 at ε=0.25. Reservoir widths N∈{128,256,512,1024} are compared across χ∈{0.70,0.90,0.99,1.00,1.03,1.10}. Increasing width shifts the survival distribution strongly to longer horizons. The effect is most pronounced at χ=0.70, where most N=1024 runs remain valid through the 512-step observation limit and are therefore right censored. Near χ≈1, N=1024 still produces a substantial long-horizon tail, whereas smaller widths generally fail much earlier.

Figure 7 shows a clear width dependence of autonomous forecast stability. At N=128 and N=256, the survival probability drops rapidly within the first few tens of forecast steps for nearly all displayed recurrent gain values χ. Increasing the reservoir width to N=512 produces a modest extension of the valid horizon, whereas the most pronounced improvement appears at N=1024, where the survival distributions are consistently shifted toward longer forecast times. The effect is particularly strong at χ=0.70: following a small initial decline, the N=1024 survival probability remains close to 0.95 throughout the observation window.

The impact of reservoir width also depends strongly on the recurrent operating point. Near χ=0.90, the N=1024 survival distribution develops a broad right tail that reaches the end of the observation window. At χ=0.99, 1.00, 1.03, and 1.10, the larger reservoir still extends the typical valid horizon, although the survival probability eventually falls to zero within approximately 100–150 forecast steps. Autonomous forecast stability is therefore jointly controlled by reservoir width and recurrent gain. Increasing N extends the range of operating conditions with longer valid horizons, while substantial sensitivity to χ remains.

Figure 7 adds a dynamical perspective to the preceding kernel-convergence results. Larger reservoirs more closely approximate the recurrent-kernel limit at the representation level, whereas closed-loop forecast stability additionally depends on how local prediction errors are propagated and amplified by autonomous dynamics (Pathak et al., 2017, 2018; Wainrib & Galtier, 2016; Wikner et al., 2024). This distinction helps explain why smaller kernel errors do not necessarily produce uniformly longer valid horizons. Right censoring is especially relevant for the χ=0.70, N=1024 condition, where many trajectories remain valid at the end of the observation window. Treating these censored trajectories as failures at the final time point would underestimate forecast stability, which is precisely the type of incomplete-observation issue addressed by survival analysis (Kaplan & Meier, 1958; Royston & Parmar, 2013).

3.7.2 From task alignment to dynamical amplification

Task-aligned discrepancy quantifies the effect of kernel perturbations on the trained readout, but autonomous deployment introduces repeated dynamical amplification. Linearizing the learned transition around a reference trajectory gives

$$\delta \mathbf{x}_{t+h} \approx J_{t+h-1} J_{t+h-2} \cdots J_t \, \delta \mathbf{x}_t \tag{24}$$

Therefore, long-horizon forecast deviations are governed by the propagation of local errors through products of subsequent Jacobians. In chaotic systems, even a small perturbation aligned with an expanding direction can eventually dominate a much larger perturbation aligned with a contracting direction; Lyapunov-based analyses formalize this local amplification geometry (Pathak et al., 2017; Wainrib & Galtier, 2016; Wikner et al., 2024). This motivates a hierarchy of discrepancy measures: global kernel error, task-aligned error, and ultimately dynamics-aligned error. The present study shows that global kernel error alone is insufficient to explain downstream behavior, while task-aligned discrepancy provides a more informative account of task-level risk. Extending this framework to an explicitly dynamics-aligned metric offers a natural direction for future work.

3.7.3 Representative autonomous rollout and forecast-error growth

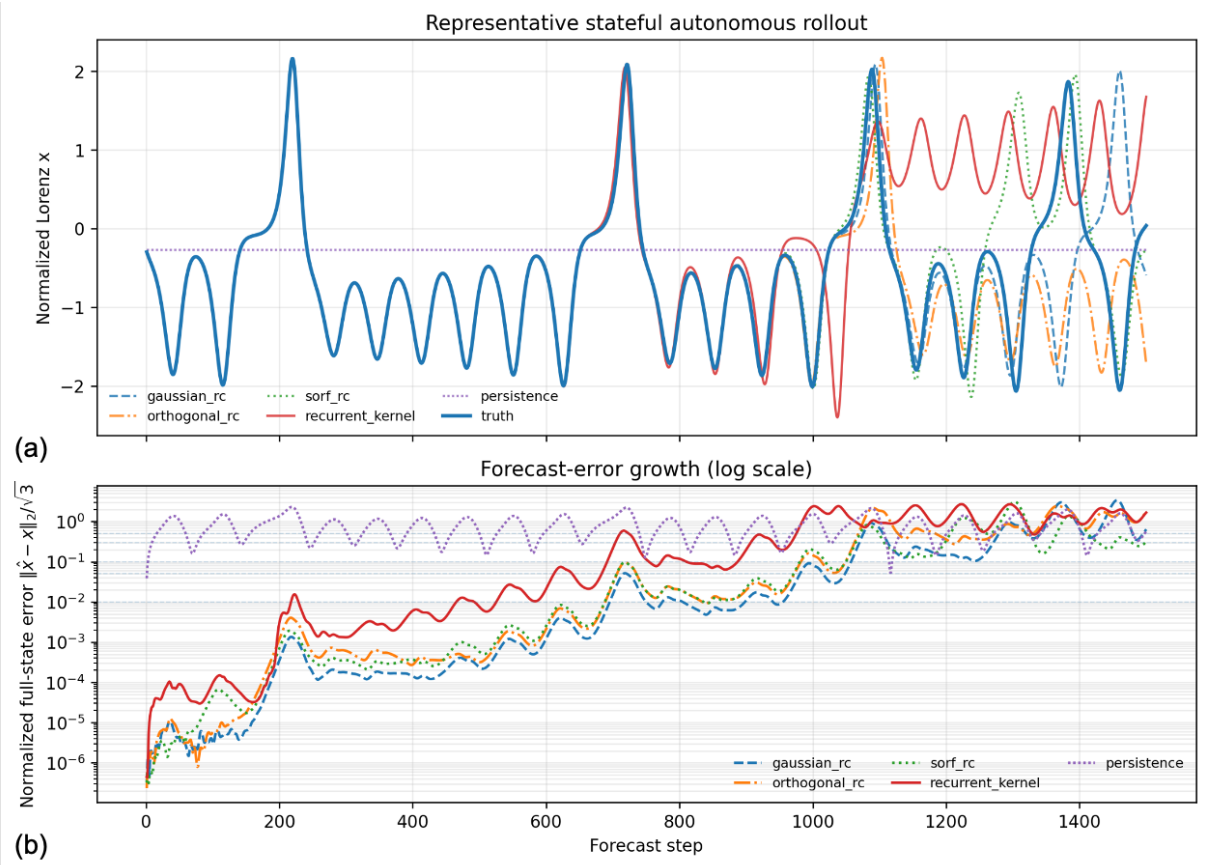

Figure 8. Representative stateful autonomous Lorenz-63 rollout and full-state error growth. (a) Normalized Lorenz x trajectory for Gaussian RC, orthogonal RC, SORF RC, the recurrent-kernel predictor, a persistence baseline, and the ground truth. (b) Normalized full-state error on a logarithmic scale over the same 1,500-step rollout. The finite RC variants remain close to the reference trajectory for an extended interval before chaotic separation, whereas persistence is inaccurate almost immediately and the recurrent-kernel predictor accumulates larger error earlier in the rollout.

The representative trajectory illustrates the behavior summarized by the survival statistic. Over the first several hundred forecast steps, Gaussian, orthogonal, and SORF reservoirs exhibit similarly low prediction errors, with normalized full-state errors remaining several orders of magnitude below the persistence baseline. Their differences become pronounced only after repeated closed-loop feedback

amplifies initially small local discrepancies. The recurrent-kernel reference model also follows the reference trajectory closely at early times, but its error begins to grow earlier and more rapidly, preceding the eventual macroscopic separation of the trajectories. Such divergence is expected in chaotic systems and does not by itself imply poor one-step prediction accuracy (Haluszczynski & Räth, 2019; Lorenz, 1963; Pathak et al., 2017; Wikner et al., 2024). Together, Figures 7 and 8 show that long-horizon autonomous performance depends on the interaction among reservoir width, recurrent operating point, task-relevant local error, and dynamical error amplification.

4. Discussion

4.1 Meaning of shared-weight recurrent-kernel convergence

The shared-versus-resampled comparison clarifies an important distinction from prior recurrent-kernel convergence studies. Dong et al. (2020) establish convergence results under a time-wise resampling assumption and numerically demonstrate that practical fixed reservoirs can also approach the recurrent-kernel limit; D'Inverno and Dong (2025) explicitly identify time-independence as a restrictive theoretical assumption and extend numerical convergence studies to additional RC topologies. Our results are consistent with these convergence findings. Figure 1 shows that the stochastic construction of the finite-width reservoir affects the observed convergence behavior: reusing one realization induces structured off-diagonal two-time covariance that is largely removed by temporal resampling, even when equal-time quantities remain similar. Figure 2 further shows that the fixed shared-weight process converges systematically toward the analytic reference as width increases. Thus, the finite-width behavior of practical reservoirs is also shaped by temporal weight sharing.

This distinction also clarifies the role of the infinite-width limit as a design criterion. A smaller global discrepancy from the recurrent kernel indicates closer agreement at the level of representation statistics, but it does not indicate whether the residual error lies primarily in directions relevant to a particular supervised task. This limitation is consistent with kernel-alignment theory, which emphasizes target-relevant geometry rather than task-agnostic kernel similarity (Cortes et al., 2012; Cristianini et al., 2001). Therefore, agreement with the recurrent-kernel limit reflects representation-level similarity, but does not by itself determine downstream prediction performance.

4.2 Structured randomization and finite-width design

Structured reservoir computing and its favorable asymptotic complexity were introduced by Dong et al. (2020), and D'Inverno and Dong (2025) further benchmarked structured RC against vanilla and sparse implementations. In our experiments, the advantage of structured transforms is mainly observed in finite-width accuracy and memory scaling. The expanded bias–variance analysis provides evidence that orthogonal and SORF constructions improve finite-width accuracy by reducing realization-to-realization fluctuations: both constructions exhibit substantial variance reduction and achieve comparable kernel MSE at matched width. Figure 4 further compares this statistical behavior with implementation cost. Dense orthogonal reservoirs retain $O(N^2)$ recurrent storage, whereas SORF achieves a similar low-error regime through an implicit structured transform requiring only $O(N)$ sign-vector storage (D'Inverno & Dong, 2025; Dong et al., 2020; Yu et al., 2016). The wall-time comparison remains dependent on implementation details, whereas the memory advantage follows directly from the structured representation.

These results also suggest that reservoir design involves three separate objectives: representation accuracy, task alignment, and dynamical stability. In the studied regime, structured randomization is most beneficial when finite-width variance or recurrent-operator memory is the primary bottleneck (Dong et al., 2020; Yu et al., 2016). For a fixed supervised task, task-aligned diagnostics provide a more informative basis for comparing representations than global kernel error alone, consistent with

kernel-alignment principles (Cortes et al., 2012; Cristianini et al., 2001; Wang et al., 2024). Autonomous deployment introduces an additional requirement, since even small local prediction errors can be amplified through closed-loop dynamics (Pathak et al., 2017, 2018; Wainrib & Galtier, 2016; Wikner et al., 2024). Long-horizon performance therefore depends on an explicit stability criterion in addition to representation fidelity and task-level alignment.

4.3 Global fidelity, task alignment, and dynamical amplification

Task-aligned discrepancy focuses on the predictive relevance of residual kernel errors in finite-width reservoirs. The construction is related to kernel-target and centered-alignment frameworks (Cortes et al., 2012; Cristianini et al., 2001; Wang et al., 2024), but the diagnostic here is formulated for finite-width RC by weighting kernel perturbations through regularized readout-relevant directions. In Figure 6, the higher within-method correlations obtained after this projection show a stronger association between representation error and supervised prediction risk. The remaining scatter, however, indicates that task relevance alone does not fully determine downstream behavior.

Figure 7 reveals the dependence of autonomous forecasting on reservoir width and recurrent operating point. At N=1024, the survival distribution develops a pronounced long-horizon tail, particularly at χ=0.70, whereas the same reservoir width produces substantially shorter valid horizons near $\chi\approx1$. Figure 8 provides a dynamical view of this behavior: Gaussian, orthogonal, and SORF reservoirs remain close over an extended early forecasting interval, and their trajectories separate only after repeated closed-loop feedback amplifies initially small local discrepancies.

Long-horizon autonomous performance depends on the magnitude and direction of local prediction errors, together with reservoir width, operating point, and the amplification geometry of the learned closed-loop dynamics. Although related, valid-horizon stability and asymptotic reservoir-state stability characterize different aspects of reservoir dynamics. Dong et al. (2022) use recurrent kernels to characterize whether two reservoir states driven by the same input converge or diverge, whereas the present valid-horizon analysis measures when a trained predictor loses forecast accuracy after its own outputs are recursively fed back. Together with Lyapunov- and attractor-based analyses of reservoir forecasting (Haluszczynski & Räth, 2019; Pathak et al., 2017; Wainrib & Galtier, 2016; Wikner et al., 2024), these results suggest that autonomous stability requires diagnostics beyond global kernel discrepancy and task-aligned error.

4.4 Limitations and outlook

The analytic recurrent-kernel recursion considered here is derived for a controlled class of random recurrent networks. A remaining theoretical challenge is obtaining rigorous finite-width concentration results for a single recurrent matrix reused indefinitely, which is more difficult than the time-resampled setting used in existing convergence proofs (D'Inverno & Dong, 2025; Dong et al., 2020). The shared-weight analysis presented here combines analytic covariance construction with empirical ensemble validation. A general nonasymptotic concentration theorem for arbitrary shared recurrent realizations remains an open theoretical problem. The present formulation also does not directly extend to arbitrary trained recurrent dynamics or to the broader range of nonlinear physical reservoirs, whose implementations and dynamical substrates can differ substantially from standard random ESNs (Gonon et al., 2024; Tanaka et al., 2019). Extending the shared-weight analysis beyond these assumptions will help determine which aspects of the observed two-time structure persist under learned, heterogeneous, or hardware-specific dynamics.

Task-aligned discrepancy is motivated by a first-order perturbation analysis of regularized kernel prediction. Although it exhibits a stronger empirical association with excess prediction risk than global kernel error, the remaining scatter indicates that it does not fully characterize task-level sensitivity.

Kernel-alignment theory likewise shows that target-aware kernel criteria provide useful information without fully characterizing prediction performance (Cortes et al., 2012; Cristianini et al., 2001; Schölkopf & Smola, 2002). The gap between representation perturbation and prediction risk may become larger when finite-width perturbations are large or when higher-order effects become important. Developing higher-order or nonlinear task-aligned diagnostics may provide a closer connection between representation error and downstream prediction behavior.

The autonomous Lorenz analysis further reveals a pronounced dependence of forecast stability on reservoir width and recurrent operating point. A systematic dynamics-aligned characterization remains to be developed. Such a framework could incorporate finite-time Lyapunov geometry, products of rollout Jacobians, and more systematic statistical comparisons across reservoir constructions and operating points, building on Lyapunov-based analyses of driven reservoirs and data-driven chaotic attractors (Pathak et al., 2017; Wainrib & Galtier, 2016; Wikner et al., 2024). These extensions could help connect local prediction errors more directly to long-horizon closed-loop stability.

Wall-clock performance remains dependent on software implementation and hardware architecture. Over the N=64–1024 range considered here, SORF becomes competitive in the large-width accuracy–time regime, while optimized dense linear-algebra routines can remain faster at smaller widths. The precise crossover therefore depends on the computational environment. By contrast, the operator-memory advantage of the implicit SORF construction follows directly from its structured representation (Dong et al., 2020; Yu et al., 2016) and remains a consistent property across implementations.

Finally, the benchmark suite focuses primarily on controlled synthetic dynamical systems. This setting enables direct analysis of kernel convergence, finite-width fluctuations, and closed-loop instability, but does not capture the full complexity of real physical data. Reservoir computing has been applied more broadly to physical, experimental, and high-dimensional dynamical systems (Gauthier et al., 2021; Gupta et al., 2023; Platt et al., 2022; Tanaka et al., 2019; Vlachas et al., 2020). Extending the framework to high-dimensional experimental and real-world time series will be important for assessing whether the distinction among global kernel discrepancy, task-aligned discrepancy, and dynamics-aligned stability remains valid in broader settings.

5. Conclusion

This work studies reservoir computing from a finite-width approximation perspective and examines how recurrent-kernel approximation relates to prediction performance. The experiments show that reusing the same recurrent realization across time generates structured two-time correlations. These correlations are strongly attenuated under independent temporal resampling, even when equal-time statistics remain similar. Matching equal-time statistics alone therefore does not recover the temporal statistics of a standard fixed reservoir. Finite shared-weight reservoirs nonetheless approach the analytic recurrent-kernel limit systematically as width increases.

Structured randomization also changes the finite-width error profile. Orthogonal and SORF reservoirs substantially reduce realization-to-realization variance and reach comparable low kernel-error levels at large width. Their computational implementations differ more clearly. SORF uses an implicit structured transform and achieves a clear accuracy–memory advantage, while its wall-clock performance remains competitive at larger widths. The downstream benchmarks show that global kernel fidelity provides only partial information about predictive performance. Task-aligned discrepancy, constructed from regularized readout-relevant directions, exhibits stronger within-method correlations with excess prediction risk for Gaussian, orthogonal, and SORF reservoirs. This result

indicates that task-aware diagnostics can provide additional information when evaluating finite-width representations.

Stateful Lorenz-63 forecasting provides a complementary view of autonomous stability. Increasing the reservoir width from N=128–512 to N=1024 can substantially extend the valid-horizon distribution, with the magnitude of the improvement depending strongly on the recurrent operating point χ. Representative closed-loop trajectories remain close to the reference dynamics for extended intervals before accumulated errors are amplified by chaotic feedback. Differences among the finite RC predictors can remain small at early times and become pronounced only later in the rollout. Approximation to the infinite-width kernel therefore characterizes convergence at the representation level. Long-horizon prediction additionally depends on how representation error couples to the supervised task and how the resulting local errors propagate through the closed-loop dynamics.

Together, these results connect recurrent weight sharing, temporal covariance, finite-width fluctuations, task-relevant kernel perturbations, and closed-loop error amplification within a common analysis framework. They also distinguish three aspects of finite reservoir design that require separate evaluation: representation fidelity, task relevance, and autonomous dynamical stability.

Data and code availability

The numerical data underlying all main figures, including per-seed outputs and the source data corresponding to each plotted panel, are available from the corresponding author upon reasonable request.

The custom simulation and analysis code used to generate the reported results is available from the corresponding author upon reasonable request. The exact scripts and parameter settings used for each figure will be provided with the code.

Acknowledgements

The authors acknowledge funding from National Key R&D Program of China (2021ZD0200300, 2018YFE0200200), National Natural Science Foundation (grant nos. 61974082, 61704096, 61836004), Youth Elite Scientist Sponsorship (YESS) Program of China Association for Science and Technology (CAST) (no. 2019QNRC001).

References

Alemohammad, S., Wang, Z., Balestriero, R., & Baraniuk, R. G. (2021). The recurrent neural tangent kernel. International Conference on Learning Representations.

Chattopadhyay, A., Hassanzadeh, P., & Subramanian, D. (2020). Data-driven predictions of a multiscale Lorenz 96 chaotic system using machine-learning methods: Reservoir computing, artificial neural network, and long short-term memory network. Nonlinear Processes in Geophysics, 27, 373–389. https://doi.org/10.5194/npg-27-373-2020

Cortes, C., Mohri, M., & Rostamizadeh, A. (2012). Algorithms for learning kernels based on centered alignment. Journal of Machine Learning Research, 13, 795–828.

Cristianini, N., Shawe-Taylor, J., Elisseeff, A., & Kandola, J. (2001). On kernel-target alignment. Advances in Neural Information Processing Systems, 14, 367–373.

D'Inverno, G. A., & Dong, J. (2025). Comparison of reservoir computing topologies using the recurrent kernel approach. Neurocomputing, 611, Article 128679. https://doi.org/10.1016/j.neucom.2024.128679

Dambre, J., Verstraeten, D., Schrauwen, B., & Massar, S. (2012). Information processing capacity of dynamical systems. Scientific Reports, 2, Article 514. https://doi.org/10.1038/srep00514

Dong, J., Börve, E., Rafayelyan, M., & Unser, M. (2022). Asymptotic stability in reservoir computing. 2022 International Joint Conference on Neural Networks (IJCNN), 1–8. https://doi.org/10.1109/IJCNN55064.2022.9892302

Dong, J., Ohana, R., Rafayelyan, M., & Krzakala, F. (2020). Reservoir computing meets recurrent kernels and structured transforms. Advances in Neural Information Processing Systems, 33, 16785–16796.

Efron, B., & Tibshirani, R. J. (1993). An introduction to the bootstrap. Chapman & Hall/CRC.

Fermanian, A., Marion, P., Vert, J.-P., & Biau, G. (2021). Framing RNN as a kernel method: A neural ODE approach. Advances in Neural Information Processing Systems, 34, 3121–3134.

Gauthier, D. J., Bollt, E., Griffith, A., & Barbosa, W. A. S. (2021). Next generation reservoir computing. Nature Communications, 12, Article 5564. https://doi.org/10.1038/s41467-021-25801-2

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. Neural Computation, 4(1), 1–58. https://doi.org/10.1162/neco.1992.4.1.1

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. Machine Learning, 63(1), 3–42. https://doi.org/10.1007/s10994-006-6226-1

Gonon, L., & Ortega, J.-P. (2020). Reservoir computing universality with stochastic inputs. IEEE Transactions on Neural Networks and Learning Systems, 31(1), 100–112. https://doi.org/10.1109/TNNLS.2019.2899649

Gonon, L., & Ortega, J.-P. (2021). Fading memory echo state networks are universal. Neural Networks, 138, 10–13. https://doi.org/10.1016/j.neunet.2021.01.025

Gonon, L., Grigoryeva, L., & Ortega, J.-P. (2024). Infinite-dimensional reservoir computing. Neural Networks, 179, Article 106486. https://doi.org/10.1016/j.neunet.2024.106486

Grigoryeva, L., & Ortega, J.-P. (2018). Echo state networks are universal. Neural Networks, 108, 495–508. https://doi.org/10.1016/j.neunet.2018.08.025

Gupta, V., Li, L. K. B., Chen, S., & Wan, M. (2023). Model-free forecasting of partially observable spatiotemporally chaotic systems. Neural Networks, 160, 297–305. https://doi.org/10.1016/j.neunet.2023.01.013

Haluszczynski, A., & Räth, C. (2019). Good and bad predictions: Assessing and improving the replication of chaotic attractors by means of reservoir computing. Chaos, 29(10), 103143. https://doi.org/10.1063/1.5118725

Hermans, M., & Schrauwen, B. (2012). Recurrent kernel machines: Computing with infinite echo state networks. Neural Computation, 24(1), 104–133. https://doi.org/10.1162/NECO_a_00200

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12(1), 55–67. https://doi.org/10.1080/00401706.1970.10488634

Jaeger, H. (2001a). The echo state approach to analysing and training recurrent neural networks (GMD Report 148). German National Research Center for Information Technology.

Jaeger, H. (2001b). Short term memory in echo state networks (GMD Report 152). GMD Forschungszentrum Informationstechnik. https://doi.org/10.24406/publica-fhg-291107

Jaeger, H. (2002). Adaptive nonlinear system identification with echo state networks. Advances in Neural Information Processing Systems, 15, 593–600.

Jaeger, H., & Haas, H. (2004). Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication. Science, 304(5667), 78–80. https://doi.org/10.1126/science.1091277

Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. Journal of the American Statistical Association, 53(282), 457–481. https://doi.org/10.1080/01621459.1958.10501452

Lee, J., Bahri, Y., Novak, R., Schoenholz, S. S., Pennington, J., & Sohl-Dickstein, J. (2018). Deep neural networks as Gaussian processes. International Conference on Learning Representations.

Lorenz, E. N. (1963). Deterministic nonperiodic flow. Journal of the Atmospheric Sciences, 20(2), 130–141.

Lorenz, E. N. (1995). Predictability: A problem partly solved. In Seminar on Predictability (pp. 1–18). European Centre for Medium-Range Weather Forecasts.

Lukoševičius, M., & Jaeger, H. (2009). Reservoir computing approaches to recurrent neural network training. Computer Science Review, 3(3), 127–149. https://doi.org/10.1016/j.cosrev.2009.03.005

Maass, W., Natschläger, T., & Markram, H. (2002). Real-time computing without stable states: A new framework for neural computation based on perturbations. Neural Computation, 14(11), 2531–2560. https://doi.org/10.1162/089976602760407955

Mackey, M. C., & Glass, L. (1977). Oscillation and chaos in physiological control systems. Science, 197(4300), 287–289. https://doi.org/10.1126/science.267326

Matthews, A. G. de G., Rowland, M., Hron, J., Turner, R. E., & Ghahramani, Z. (2018). Gaussian process behaviour in wide deep neural networks. International Conference on Learning Representations.

Pathak, J., Hunt, B., Girvan, M., Lu, Z., & Ott, E. (2018). Model-free prediction of large spatiotemporally chaotic systems from data: A reservoir computing approach. Physical Review Letters, 120(2), 024102. https://doi.org/10.1103/PhysRevLett.120.024102

Pathak, J., Lu, Z., Hunt, B. R., Girvan, M., & Ott, E. (2017). Using machine learning to replicate chaotic attractors and calculate Lyapunov exponents from data. Chaos, 27(12), 121102. https://doi.org/10.1063/1.5010300

Platt, J. A., Penny, S. G., Smith, T. A., Chen, T.-C., & Abarbanel, H. D. I. (2022). A systematic exploration of reservoir computing for forecasting complex spatiotemporal dynamics. Neural Networks, 153, 530–552. https://doi.org/10.1016/j.neunet.2022.06.025

Rahimi, A., & Recht, B. (2007). Random features for large-scale kernel machines. Advances in Neural Information Processing Systems, 20, 1177–1184.

Royston, P., & Parmar, M. K. B. (2013). Restricted mean survival time: An alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. BMC Medical Research Methodology, 13, 152. https://doi.org/10.1186/1471-2288-13-152

Rudi, A., & Rosasco, L. (2017). Generalization properties of learning with random features. Advances in Neural Information Processing Systems, 30, 3215–3225.

Schölkopf, B., & Smola, A. J. (2002). Learning with kernels: Support vector machines, regularization, optimization, and beyond. MIT Press.

Sriperumbudur, B. K., & Szabó, Z. (2015). Optimal rates for random Fourier features. Advances in Neural Information Processing Systems, 28, 1144–1152.

Sutherland, D. J., & Schneider, J. (2015). On the error of random Fourier features. In Proceedings of the 31st Conference on Uncertainty in Artificial Intelligence (pp. 862–871). AUAI Press.

Tanaka, G., Yamane, T., Héroux, J. B., Nakane, R., Kanazawa, N., Takeda, S., Numata, H., Nakano, D., & Hirose, A. (2019). Recent advances in physical reservoir computing: A review. Neural Networks, 115, 100–123. https://doi.org/10.1016/j.neunet.2019.03.005

Verstraeten, D., Schrauwen, B., D'Haene, M., & Stroobandt, D. (2007). An experimental unification of reservoir computing methods. Neural Networks, 20(3), 391–403. https://doi.org/10.1016/j.neunet.2007.04.003

Vlachas, P. R., Pathak, J., Hunt, B. R., Sapsis, T. P., Girvan, M., Ott, E., & Koumoutsakos, P. (2020). Backpropagation algorithms and reservoir computing in recurrent neural networks for the forecasting of complex spatiotemporal dynamics. Neural Networks, 126, 191–217. https://doi.org/10.1016/j.neunet.2020.02.016

Wainrib, G., & Galtier, M. N. (2016). A local echo state property through the largest Lyapunov exponent. Neural Networks, 76, 39–45. https://doi.org/10.1016/j.neunet.2015.12.013

Wang, C., He, X., Wang, Y., & Wang, J. (2024). On the target-kernel alignment: A unified analysis with kernel complexity. Advances in Neural Information Processing Systems, 37, 40434–40485. https://doi.org/10.52202/079017-1280

Wikner, A., Harvey, J., Girvan, M., Hunt, B. R., Pomerance, A., Antonsen, T., & Ott, E. (2024). Stabilizing machine learning prediction of dynamics: Novel noise-inspired regularization tested with reservoir computing. Neural Networks, 170, 94–110. https://doi.org/10.1016/j.neunet.2023.10.054

Wilcoxon, F. (1945). Individual comparisons by ranking methods. Biometrics Bulletin, 1(6), 80–83.

Wringe, C., Trefzer, M., & Stepney, S. (2025). Reservoir computing benchmarks: A tutorial review and critique. International Journal of Parallel, Emergent and Distributed Systems, 40(4), 313–351. https://doi.org/10.1080/17445760.2025.2472211

Yang, G. (2019). Tensor programs I: Wide feedforward or recurrent neural networks of any architecture are Gaussian processes. Advances in Neural Information Processing Systems, 32.

Yildiz, I. B., Jaeger, H., & Kiebel, S. J. (2012). Re-visiting the echo state property. Neural Networks, 35, 1–9. https://doi.org/10.1016/j.neunet.2012.07.005

Yu, F. X., Suresh, A. T., Choromanski, K. M., Holtmann-Rice, D. N., & Kumar, S. (2016). Orthogonal random features. Advances in Neural Information Processing Systems, 29, 1975–1983.