Computational Engineering and Technology Innovations

www.ceti.reapress.com

Comput. Eng. Technol. Innov. Vol. 3, No. 3 (2026) 161–175.

Paper Type: Original Article

Mechanism Informed Early Warning of Simulated Dynamical Transitions

Madasu Hanmandlu[1,*], Kharabela Swain[2], Wolfgang H. Müller[3]

[1] Department of Electrical Engineering, Indian Institute of Technology Delhi, New Delhi, India; mhmandlu@ee.iitd.ernet.in.
[2] Department of Applied Sciences, Mangalayatan University, Aligarh, Uttar Pradesh, India; kharabela1983@gmail.com.
[3] Department of Computer Science, Technical University of Munich, Munich, Germany.

Citation:

Received: 06 October 2025 *Revised: 10 December 2025* *Accepted: 21 February 2026*	Hanmandlu, M,. Swain, K., & Müller, W, H. (2026). Mechanism informed early warning of simulated dynamical transitions. *Computational Engineering and Technology Innovations*, *3*(3), 161-175.

Abstract

Background early warning of a dynamical transition is an operational decision problem: a score is useful only when it yields an alarm with actionable lead time and a controlled false-alarm burden. Classical indicators based on critical slowing down are interpretable but can be weak, mechanism dependent and sensitive to observation design. Methods this simulation study evaluates a mechanism-informed feature strategy on a Rosenzweig–MacArthur predator–prey system approaching Hopf onset and a one-state harvesting model approaching a fold. Rolling mean, variance and lag-one autocorrelation are combined with local estimates of equilibrium and vector-field geometry. Logistic Regression (LR), random forest and multilayer-perceptron baselines produce window-level risk scores, while a validation-calibrated cumulative-sum rule converts those scores into trajectory-level alarms. Thirty realizations are reported per model and noise level, with complete trajectories separated into sub-training, validation and test sets. Results At 5% observation noise on the Hopf task, the augmented logistic detector has a reported mean trajectory ROC AUC of 0.901 ± 0.103, compared with 0.690 ± 0.280 for the univariate classical logistic baseline. Seven of 12 test trajectories alarm within the predefined 80-step imminence window, and no far-region alarm is observed in that test sample. However, a matched-channel classical ablation using both prey and predator statistics reaches 0.897 ± 0.115 AUC; the paired comparison between augmented and baseline logistic AUCs is not significant after the stated procedure ($p = 0.064$). On the fold task, augmentation adds little (0.997 versus 0.996 AUC at 5% noise). Conclusion the reported experiments support a restrained conclusion: Local geometry can help when a one-channel warning is weak, but most of the Hopf gain is attributable to observing the second state, and the fold task is nearly saturated. The study clarifies how mechanism-informed features, matched-channel ablations and validation-only sequential calibration can be combined, while also identifying the experiments required for transfer beyond fixed, regularly sampled simulations.

Keywords: Critical slowing down, Early warning, Hopf bifurcation, Fold bifurcation, Mechanism-informed learning, Sequential detection, Trajectory-level validation.

1 | Introduction

Abrupt changes in complex systems are difficult to manage because the observable state can remain apparently regular until the underlying stability margin has become small. Ecological regime shifts, population collapse, oscillatory instabilities and climate transitions have therefore motivated a broad search for Early-

Corresponding Author: mhmandlu@ee.iitd.ernet.in
https://doi.org/10.48314/ceti.v3i3.90
Licensee System Analytics. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0).

Warning Signals (EWS) [1–3]. A central theoretical result is critical slowing down: as the dominant restoring eigenvalue approaches the stability boundary, perturbations decay more slowly, temporal memory increases and stochastic fluctuations may become larger. Rising variance and lag-one autocorrelation are consequently used as generic indicators of diminishing resilience [2], [4]. Their appeal is substantial. They require little mechanistic detail, can be computed in moving windows and have been observed in model systems, laboratory populations and a whole-ecosystem experiment [5–7].

Generality does not imply universality. The observable signature depends on the transition mechanism, the state variable, the noise process, the forcing rate and the sampling design. A fold removes a stable equilibrium and is often associated with reduced damping and basin contraction. A Hopf bifurcation destabilizes an equilibrium through a complex-conjugate eigenvalue pair and creates or destroys oscillatory behavior. Noise-induced and rate-induced tipping can occur without the quasistatic loss of stability assumed by standard critical-slowing-down arguments [8], [9]. Even within bifurcation-induced transitions, detrending, window length, temporal aggregation and spectral content can change the apparent trend of an indicator [10], [11]. A warning method must therefore be judged not only by whether a statistic rises, but also by whether it distinguishes imminent from non-imminent conditions on trajectories that were not used to fit or calibrate the detector.

This operational distinction changes the scientific question. An EWS study is not complete when it presents a monotonic curve before a known event. A prospective detector must define the information available at each decision time, separate training and evaluation at the trajectory level, calibrate an alarm without consulting test outcomes, and report both timely detections and premature alarms. Post-event selection of favorable series can inflate apparent performance, a problem formalized as the prosecutor's fallacy [12]. Finite noisy records can also produce severe sensitivity–specificity trade-offs and systematic false positives [13], [14]. These limitations motivate an approach in which generic statistics are complemented by features tied to a plausible local mechanism, while the final decision is evaluated as a sequential alarm rather than as a retrospective trend.

The present study examines such a hybrid design. It combines rolling statistical summaries with local estimates derived from equilibrium geometry and the fitted vector field. The term mechanism-informed is used deliberately. A prey-to-predator ratio deviation, nullcline slope, local damping estimate or distance to a fold can encode how the current state relates to the governing dynamics, but these quantities are not automatically invariants. A first integral $I(z)$ of an autonomous vector field f would satisfy $\nabla I(z) \cdot f(z) = 0$ along applicable trajectories. The features considered here generally vary with state and forcing, and their scientific value lies in their relation to local stability rather than in conservation.

The study uses two complementary synthetic tasks. The Rosenzweig–MacArthur system provides a two-state oscillatory transition in which a single observed channel may hide information carried by the coupled state. The harvesting system provides a one-state fold in which conventional statistics are already highly discriminative. The contrast is useful: a method should not be judged only on a task where every classifier performs nearly perfectly. Equally important, the design includes a matched-channel ablation that gives the classical model access to both prey and predator statistics. This ablation separates the value of richer observation from the incremental value of the mechanism-informed transformation.

Three contributions are developed: 1) the study places local geometric features within the broader EWS literature and distinguishes them from conserved quantities and unconstrained representation learning, 2) it presents an end-to-end warning pipeline that separates classifier fitting, alarm calibration and final trajectory testing, and 3) it interprets the reported results through matched-channel ablation and operational alarm metrics, thereby narrowing the claim from generic superiority to a model- and observation-specific result. The resulting manuscript does not claim field readiness. It asks a more defensible question: Under controlled simulation, when does local mechanism information add value beyond classical statistics computed from the same observed channels?

2 | Literature Review and Research Positioning

2.1 | Critical Slowing Down and Mechanism Specificity

The modern EWS literature grew from two connected ideas: Ecological systems can possess alternative regimes separated by stability boundaries, and proximity to some boundaries changes the recovery dynamics before the state shifts [1], [15]. Long before EWS became a unified research program, graphical stability analysis of predator–prey systems showed that the local slopes of interacting nullclines determine equilibrium stability [16], while nonlinear harvesting models illustrated how feedback can create folds and abrupt changes [17]. These model-based insights later converged with empirical indicators such as variance, autocorrelation, recovery time, skewness and flickering. Dakos et al. [2] demonstrated increasing autocorrelation before several paleoclimate transitions, and Scheffer et al. [3] synthesized the general critical-slowing-down argument across ecological, climatic and physiological settings.

The mathematical foundation is strongest when slowly varying parameters carry a stable equilibrium toward a local bifurcation. Fast–slow and stochastic dynamical-systems theory explains why variance and autocorrelation can scale with the leading eigenvalue, while also showing that the scaling depends on normal form, noise structure and observation direction [8]. This dependence matters for the two cases studied here. Near a fold, the dominant real eigenvalue approaches zero and the stable branch terminates. Near a Hopf point, the real part of a complex pair approaches zero while an oscillatory frequency remains. Spectral indicators can therefore help distinguish oscillatory and non-oscillatory loss of stability [18]. A purely univariate variance or autocorrelation measure may discard cross-variable and phase information that is especially informative in coupled predator–prey dynamics.

Tipping mechanisms outside quasistatic bifurcation further limit generic interpretation. Ashwin et al. [9] distinguish bifurcation-induced, noise-induced and rate-induced tipping. Noise can drive escape from an attractor even when local stability has not changed, while rapid forcing can cause tracking failure before or without a frozen-system bifurcation. The present simulations use known parameter ramps and model-defined stability crossings; they do not test noise-induced escape or rate-induced loss of tracking. Consequently, their results should be interpreted as warnings of two bifurcation scenarios, not as evidence for a universal tipping detector.

2.2 | Empirical Evidence and Known Failure Modes

Controlled studies provide important proof of concept. Drake and Griffen [5] detected slowing-down signatures before experimental extinction in deteriorating Daphnia populations, while also showing the value of reference populations and stable baselines. Carpenter et al. [6] reported warning signals before a food-web reorganization in a manipulated lake. Veraart et al. directly related recovery rates to distance from a tipping point in a living system [7]. These studies show that dynamic precursors can be measurable, but they do not remove the need for application-specific validation. Field observations usually contain changing drivers, measurement error, missing variables and few replicated transitions.

Methodological work has therefore emphasized uncertainty and negative controls. Dakos et al. [10] catalogued practical choices for detrending, rolling windows and indicator estimation. Boettiger and Hastings [12], [13] showed how conditioning on known transitions can inflate false-positive rates and argued for model comparison and explicit error characterization. Kéfi et al. [19] showed that slowing down can also precede non-catastrophic transitions, so an indicator of reduced local stability is not automatically a predictor of irreversible collapse. Reviews by Dakos et al. [20] and Scheffer et al. [21] stress that resilience indicators are best interpreted as conditional risk evidence rather than event-time forecasts. Jäger and Füllsack [14] further demonstrated that systematic false positives can arise in common EWS analyses.

These findings imply three evaluation requirements adopted here: 1) windows from the same trajectory must not be divided randomly across training and test sets because their overlap and common forcing history would

create leakage. 2) discrimination and alarm behavior must be reported separately: A high ROC AUC does not uniquely determine a useful operating threshold. 3) lead time must be interpreted jointly with the far-alarm burden. A very early alarm can be operationally worse than a later alarm if it occurs in a region declared to be non-imminent.

2.3 | From Generic Indicators to Mechanism-Informed Learning

Between fully specified mechanistic models and generic rolling statistics lies a broad family of hybrid methods. Generalized modeling incorporates partial knowledge of process structure without requiring every functional form or parameter [4]. Nonparametric drift and diffusion estimation can reveal changes in conditional variance while making weaker structural assumptions [22]. Multivariate covariance eigenvalues connect observed fluctuation geometry to the Jacobian of a linearized system and can identify directions that become vulnerable near instability [23]. These approaches share a central idea with the present work: A feature is more interpretable when its behavior can be related to local stability or basin geometry.

Data-driven dynamical-system discovery provides a second connection. Sparse identification seeks parsimonious governing equations from time-series data [24], and physics-informed learning embeds differential constraints or known structure in statistical models [25], [26]. Such methods can improve sample efficiency and interpretability when the assumed mechanism is approximately correct, but they can also fail under model misspecification, poor derivative estimates or inadequate excitation. The local fits used here are far more modest than full equation discovery or physics-informed neural networks. They estimate low-dimensional quantities within rolling windows and use them as classifier inputs. This smaller claim is appropriate for the available evidence.

Machine learning has also entered EWS research directly. Spectral features can distinguish local bifurcation types [18], while deep learning trained across normal forms can identify tipping signatures in systems not seen during training [27]. The advantage of flexible models is their capacity to capture nonlinear and multivariate patterns; the risk is that performance can reflect simulation-specific cues, preprocessing or known event timing rather than transferable dynamics. The present use of Logistic Regression (LR) as the augmented head keeps the final decision surface interpretable, and comparison with random forest and Multilayer Perceptron (MLP) baselines tests whether greater classifier flexibility compensates for a restricted feature set.

2.4 | Sequential Decisions and the Unresolved Gap

Most EWS indicators are computed window by window, whereas interventions are triggered by alarms. Page's cumulative-sum procedure was designed to accumulate small persistent departures while resetting isolated negative evidence [28]. Applied to classifier probabilities, CUSUM can suppress one-window spikes and reward sustained risk. Its threshold, however, must be selected without test information, and its apparent false-alarm control is empirical rather than guaranteed when the score distribution drifts. The validation-only calibration used here is therefore a strength, but it must be assessed on independent trajectories and under varied forcing schedules.

ROC AUC summarizes ranking over thresholds and is useful for comparing classifiers [29], yet rare imminent windows and repeated decisions make it incomplete. Precision–recall analysis is often more revealing for imbalanced classification [30], while operational assessment requires detection fraction, lead time and alarms per unit of non-imminent exposure. The current study reports trajectory-averaged ROC AUC and alarm metrics, but not precision–recall curves or a full threshold frontier. This motivates future evaluation over multiple operating points rather than treating a single validation maximum as definitive.

The unresolved gap is thus specific. Existing work establishes generic, model-based, multivariate, spectral and learned EWS, but fewer studies isolate the incremental value of local mechanism features under matched observation channels and then propagate the resulting scores through an independently calibrated sequential alarm. *Table 1* positions the present design against representative approaches. Its intended advance is not a

claim of universal prediction; it is a transparent comparison of what is gained by local geometry after accounting for the information content of the observed states.

Table 1. Representative early-warning approaches and the gap addressed in this study.

Approach	Primary Information	Main Strength	Main Limitation	Role In This Study
Classical CSD	Variance, autocorrelation, recovery	Low data and model burden	Mechanism and preprocessing dependent	Core baseline
Model-based EWS	Drift, diffusion, local stability	Mechanistic interpretation	Sensitive to specification and data volume	Motivates local fits
Multivariate EWS	Covariance and cross-state structure	Uses coupled observations	Requires aligned channels and stable estimation	Motivates both-state ablation
Spectral EWS	Frequency-specific fluctuation changes	Can distinguish oscillatory transitions	Needs adequate sampling and spectral resolution	Relevant to Hopf onset
Deep learned EWS	Representations learned across simulations	Flexible and transferable in principle	Can learn simulation or preprocessing cues	Context, not executed method
Present pipeline	Classical statistics plus local geometry	Matched-channel attribution and sequential alarm	Evaluated on two fixed synthetic designs	Target contribution

3 | Models and Evaluation Protocol

3.1 | Study Design

The analysis is organized as a nested decision pipeline (see *Fig. 1*). Complete simulated trajectories are allocated to sub-training, validation and test sets. Sliding windows are created within each trajectory. Classifiers are fitted on sub-training windows, but alarm reference values and thresholds are determined only from validation trajectories. Final AUC and alarm metrics are then computed on untouched test trajectories. This ordering prevents the most direct threshold leakage and keeps the test set reserved for performance estimation.

For each model-by-noise condition, the source reports 30 realizations: 12 sub-training, six validation and 12 testing trajectories. Each trajectory contains 600 recorded steps at regular interval $\Delta t = 0.1$. Windows contain 50 successive observations and advance with stride one. Only windows ending no later than the model-defined transition at step 420 are used. A window is labelled imminent when its endpoint lies in the final 80 steps before that transition; earlier windows form the far region. This labeling converts proximity to the known simulation transition into a supervised classification target.

The trajectory, rather than the overlapping window, is the independent evaluation unit. This choice is essential because adjacent windows share 49 of 50 observations and because all windows from a trajectory share the same realized perturbations and forcing history. More generally, validation for temporally structured data should respect dependence rather than use random record-level splitting [31]. The current split satisfies that minimum requirement, although all realizations retain the same ramp schedule and nominal transition time.

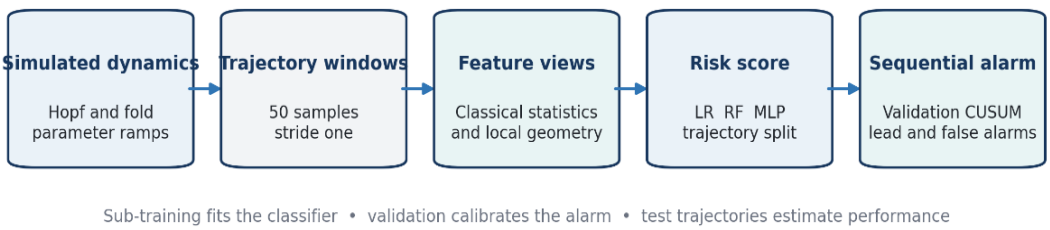


Fig.1. End-to-end workflow.

Classifier fitting, alarm calibration and final testing use disjoint complete trajectories. The figure is conceptual and introduces no results beyond the reported protocol.

3.2 | Rosenzweig–MacArthur Hopf Task

The two-state task uses the Rosenzweig–MacArthur predator–prey model [16]. Prey x grows logistically and is consumed through a saturating functional response; predator y increases through converted consumption and decreases through mortality:

$$\frac{dx}{dt} = rx\left(1 - \frac{x}{K}\right) - \frac{axy}{x + h_0}, \quad (1)$$

$$\frac{dy}{dt} = \frac{baxy}{x + h_0} - my. \quad (2)$$

The reported parameters are $r = a = 1$, $h_0 = 0.6$, $b = 0.5$ and $m = 0.35$. The positive coexistence equilibrium has prey coordinate $x^* = \frac{mh_0}{ba-m} = 1.4$. For this parameterization, the Jacobian trace changes sign at $K_h = 2x^* + h_0 = 3.4$. The carrying capacity is ramped linearly from 2.6 to 3.743 over 600 recorded steps, placing the nominal stability crossing near step 420. The event is therefore onset of oscillatory instability in a known model, not necessarily irreversible ecosystem collapse.

The source describes additive perturbations to the prey dynamics, relative perturbations to the predator dynamics and Gaussian observation noise scaled to either 1% or 5% of reference pre-transition variation. The exact numerical integrator, realized random draws and scaling implementation were not supplied with the document. Those omissions do not change the algebraic location of the deterministic Hopf boundary, but they prevent independent reconstruction of the reported stochastic trajectories.

3.3 | Harvesting Fold Task

The second task uses a one-state harvesting model related to classical bistable exploitation systems [17]:

$$\frac{dx}{dt} = rx\left(1 - \frac{x}{K}\right) - \frac{Hx^2}{x^2 + a^2}, \quad (3)$$

with $r = a = 1$ and $K = 8$. For a positive equilibrium, harvesting pressure can be written as

$$H(x) = \frac{\left(1 - \frac{x}{8}\right)(x^2 + 1)}{x}. \quad (4)$$

The upper fold occurs at $x \approx 3.709$ and $H \approx 2.134$. The source increases H linearly from 1.850 to 2.256, again aligning the model-defined transition with approximately step 420. Local damping approaches zero near the fold, while the distance between the current state and an inferred threshold is intended to proxy shrinking safety margin. The model includes process and observation perturbations at the same two observation-noise levels used in the Hopf experiment.

Fig. 2 contrasts the mechanisms. In the Hopf case the equilibrium remains present but loses stability and oscillations emerge; in the fold case a stable branch terminates at a turning point. This distinction motivates different structural features and cautions against assuming that one generic trend will be equally informative in both tasks.

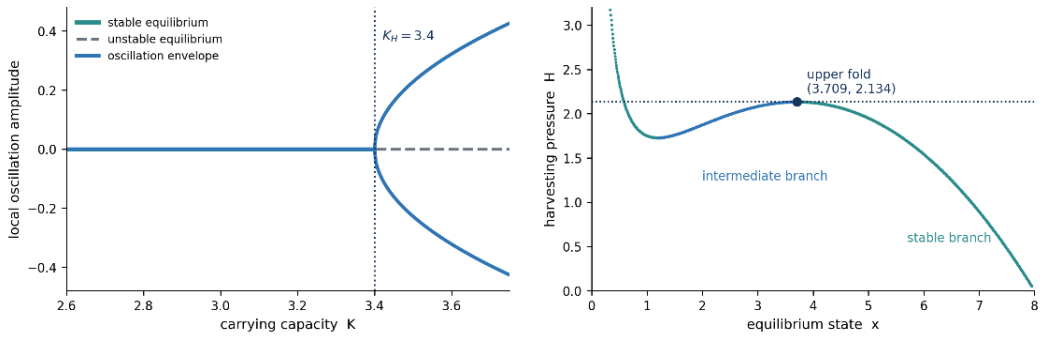


Fig. 2. Schematic of the Hopf bifurcation mechanism and equilibrium relation; a. hopf onset in the predator-prey model, b. FOLD geometry in the harvesting model.

3.4 | Features and Classifiers

The primary classical feature set contains the rolling mean, variance and lag-one autocorrelation of the principal observed state. These measures summarize level, fluctuation magnitude and temporal persistence. Three classifier families are reported on this baseline: LR, a random forest with 200 trees [32], and a MLP with hidden layers of 16 and eight units. Implementations are attributed to scikit-learn [33]. Because the executable configuration is unavailable, unreported choices such as regularization strength, optimizer, feature scaling, stopping rule and random state are not inferred here.

For the Hopf task, the augmented logistic detector adds three quantities estimated from local windowed dynamical fits: Deviation of the prey-to-predator ratio from its local reference, slope of the prey nullcline, and variability of distance to the estimated nullcline. These features are intended to capture changing equilibrium geometry and the coupling between state variables. For the fold task, the augmented set adds a local damping estimate and an inferred distance to the basin threshold. Each is interpretable only to the extent that the local fit is stable and the assumed model geometry is approximately correct.

The central ablation is a bivariate classical model. It adds classical statistics of the predator channel but omits the mechanism-informed transformations. This comparator is scientifically stronger than a one-channel baseline because both it and the augmented model see the same two measured states. A structural-only Hopf variant and incremental fold variants further assess whether individual feature groups are sufficient.

Table 2. Reported simulation, feature and evaluation protocol.

Component	Specification	Interpretive Purpose
Systems	Two-state Hopf and one-state fold models	Contrast oscillatory and non-oscillatory instability
Noise	1% and 5% Gaussian observation noise plus process perturbations	Test sensitivity to observation quality
Realizations	30 per model-noise cell: 12 sub-train, 6 validation, 12 test	Keep fitting, calibration and testing separate
Sampling	600 steps, $\Delta t = 0.1$, fixed crossing near step 420	Regular synthetic monitoring schedule
Windows	Length 50, stride 1, endpoint ≤ transition step	Construct time-local features
Imminence	Final 80 pre-transition steps	Define supervised warning target
Classical features	Mean, variance and lag-one autocorrelation	Represent conventional CSD summaries
Hopf structural features	Ratio deviation, nullcline slope, nullcline-distance variation	Represent local coupled geometry
Fold structural features	Local damping and threshold distance	Represent restoring strength and margin
Alarm	Validation-calibrated one-sided CUSUM	Convert risk scores into persistent alarms

3.5 | Sequential Alarm and Performance Measures

Let p_k be the classifier probability for the kth eligible window. The reported one-sided cumulative-sum statistic is

$$G_k = \max\{0, G_{k-1} + p_k - \nu\}, \qquad G_0 = 0, \tag{5}$$

and an alarm is declared when $G_k > \tau$. The reference value ν is the mean score over validation far-region windows. The threshold τ is the largest G value observed in the validation far region, subject to a minimum value of 2. This calibration deliberately avoids far-region alarms in the six validation trajectories, but it is an empirical choice rather than a bound on false-alarm probability in new data. Its generalization depends on how stable the score distribution remains across trajectories and forcing schedules.

Window discrimination is summarized by ROC AUC computed separately for each complete test trajectory and then averaged across the 12 test trajectories. ROC AUC measures how often an imminent window receives a higher score than a far-region window and is invariant to a particular threshold [29]. The reported dispersion is treated as the standard deviation across test trajectories. Because imminent and far windows are imbalanced, a future release should also include precision–recall curves and average precision [30].

Alarm performance is described by three complementary quantities. Detection fraction is the proportion of test trajectories with an alarm inside the 80-step imminence window. Lead is the transition index minus the first pre-transition alarm time, with zero assigned when no qualifying alarm occurs. Far alarms are alarm episodes per 1,000 far-region steps. Lead cannot be interpreted alone: a numerically large lead may represent an alarm that occurs too early to be actionable or that belongs to the declared far region. The episode-reset rule and treatment of alarms that persist across the region boundary should be fixed in executable code.

For the 5% noise Hopf condition, the reported inferential analysis uses paired Wilcoxon signed-rank tests on trajectory AUCs, comparing the augmented detector with LR, RF and MLP. Holm's sequential adjustment controls family-wise error across the three contrasts [34]. A Friedman test provides an omnibus comparison. The independent unit is the trajectory ($n = 12$), not the many overlapping windows. Given the small test sample, p values are interpreted alongside effect magnitudes, dispersion and ablation results.

4 | Results

4.1 | Hopf Warning Performance

Table 3 reports the supplied results for the predator–prey task. At 1% observation noise, the augmented detector attains mean trajectory AUC 0.890 ± 0.171 and alarms inside the imminence window on 11 of 12 trajectories. The classical LR, RF and MLP baselines range from 0.470 to 0.605 mean AUC and detect two to five trajectories. At 5% noise, the augmented AUC is 0.901 ± 0.103, compared with 0.690 ± 0.280 for LR, 0.648 ± 0.198 for RF and 0.669 ± 0.218 for MLP.

The alarm results qualify the AUC comparison. At 5% noise, the augmented detector produces an in-window alarm on seven of 12 trajectories and no far-region alarm episode is observed in the test sample. LR detects three of 12 with 0.573 far episodes per 1,000 steps; RF detects five of 12 with 0.859; MLP detects three of 12 with 0.573. The reported median lead of five steps for the augmented detector should be read together with its 0.583 detection fraction: five trajectories receive no qualifying alarm and are assigned zero lead.

Noise does not yield a monotonic performance pattern. The augmented mean AUC is similar at 1% and 5%, yet the in-window detection fraction decreases from 0.917 to 0.583. Because the alarm threshold is recalibrated separately in each condition, this change reflects both the score distribution and the chosen operating point. It does not justify a general claim that the method is noise invariant.

Table 3. Hopf task performance reported over 12 held-out test trajectories per condition.

Noise	Method	AUC Mean ± SD	Median Lead	Far Alarms / 1,000	Detection
1%	LR	0.605 ± 0.263	0	0.286	0.250
1%	RF	0.470 ± 0.210	0	0.286	0.167
1%	MLP	0.558 ± 0.261	60	2.005	0.417
1%	Augmented LR	0.890 ± 0.171	28	0.286	0.917
5%	LR	0.690 ± 0.280	0	0.573	0.250
5%	RF	0.648 ± 0.198	0	0.859	0.417
5%	MLP	0.669 ± 0.218	0	0.573	0.250
5%	Augmented LR	0.901 ± 0.103	5	0.000	0.583

4.2 | Fold Warning Performance

The fold task is nearly saturated for every classifier (see *Table 4*). At 1% observation noise, mean AUC ranges from 0.978 to 0.991 and every method detects all 12 test trajectories inside the imminence window. At 5% noise, LR reaches 0.996 ± 0.013 and the augmented detector reaches 0.997 ± 0.009; both report the same far-alarm rate of 1.145 episodes per 1,000 far-region steps and nearly identical median lead.

These differences are too small to support a practical superiority claim. The result is nevertheless informative: local damping and threshold-distance features do not materially improve a task already solved by classical statistics under the fixed synthetic ramp. A stronger fold evaluation would need harder conditions, such as variable ramp speed, unknown crossing time, non-transitioning controls, coarser observation, stronger process noise and misspecified threshold geometry.

Table 4. Fold task performance reported over 12 held-out test trajectories per condition.

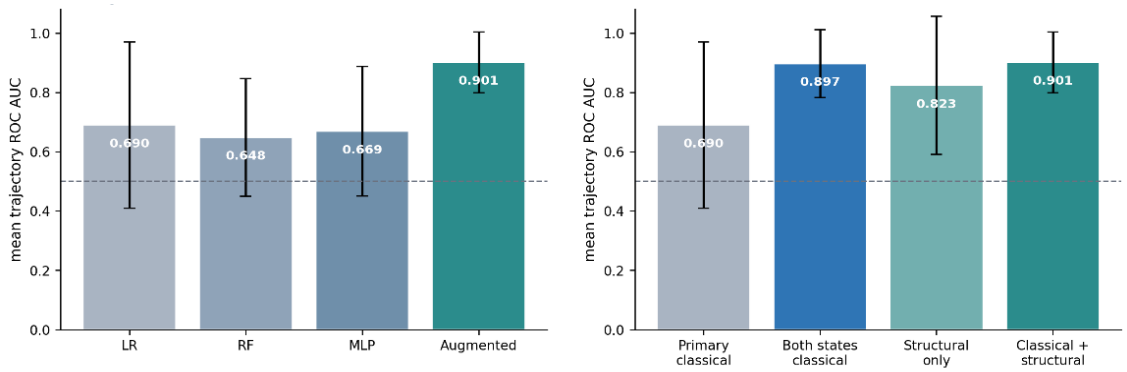
Noise	Method	AUC Mean ± SD	Median Lead	Far Alarms / 1,000	Detection
1%	LR	0.991 ± 0.017	47.5	0.573	1.000
1%	RF	0.978 ± 0.023	51	0.573	1.000
1%	MLP	0.988 ± 0.020	45	0.573	1.000
1%	Augmented LR	0.991 ± 0.018	47.5	0.573	1.000
5%	LR	0.996 ± 0.013	77	1.145	1.000
5%	RF	0.981 ± 0.035	80	2.005	1.000
5%	MLP	0.992 ± 0.016	80	2.005	1.000
5%	Augmented LR	0.997 ± 0.009	76	1.145	1.000

4.3 | Feature Ablation and Paired Comparisons

The matched-channel ablation substantially changes the interpretation of the Hopf result. At 5% noise, adding classical predator-channel statistics raises mean AUC from 0.690 to 0.897 and eliminates observed far-region alarms. Structural features alone reach 0.823. Combining classical and structural features produces 0.901, only 0.004 above the both-state classical model. Thus the main contrast against a one-channel baseline combines two effects: Access to the coupled predator state and transformation of that information into local geometric features.

The alarm metrics show a small timing difference between the two strongest Hopf variants. The bivariate classical model has reported median lead 0.5 steps and the combined model five steps, with zero observed far alarms for both. Per-trajectory alarm times are required to determine whether this difference is stable or driven by a few trajectories. On the fold task, all feature variants remain within 0.001 AUC of one another and have identical reported far-alarm rates.

Fig. 3 summarizes the central result without hiding the ablation. *Fig. 3.a* reproduces the improvement over univariate baselines; *Fig. 3.b* shows that both-state classical statistics nearly recover the same discrimination. Error bars display the reported between-trajectory standard deviation rather than uncertainty in the mean.



a. **b.**

Fig. 3. Comparison of classifier performance and matched-channel ablation under 5% observation noise; a. hopf task at 5% observation noise, b. matched-channel feature ablation.

Table 5. Feature ablation at 5% observation noise. All variants use a logistic head and the reported CUSUM calibration.

Task	Feature set	AUC mean ± SD	Median lead	Far alarms / 1,000
Hopf	Primary-state classical	0.690 ± 0.280	0	0.573
Hopf	Both-state classical	0.897 ± 0.115	0.5	0.000
Hopf	Structural only	0.823 ± 0.233	0	0.286
Hopf	Classical + structural	0.901 ± 0.103	5	0.000
Fold	Classical	0.996 ± 0.013	77	1.145
Fold	Classical + damping	0.997 ± 0.009	76	1.145
Fold	Classical + threshold	0.996 ± 0.013	76	1.145
Fold	Classical + both	0.997 ± 0.009	76	1.145

The reported paired tests compare the augmented Hopf detector with the three classifier baselines at 5% noise (*Table 6*). The contrast with LR has $W = 15$ and raw $p = 0.0640$, which remains 0.0640 after Holm adjustment. Contrasts with RF and MLP pass the nominal 0.05 level after adjustment. The Friedman omnibus test is reported as $\chi^2 = 12.2$, $p = 0.0067$. These results establish heterogeneity among the four methods in the supplied test set, but they do not test the most important incremental contrast between the combined feature set and the both-state classical ablation.

The inferential scope is also limited by $n = 12$ trajectories and by reuse of one simulation schedule. Statistical significance against weaker univariate baselines should not be presented as evidence that the structural transformations are universally superior. The ablation effect size, paired uncertainty and external validation are more consequential for the study's substantive claim.

Table 6. Reported paired AUC tests for the Hopf task at 5% noise.

Contrast	Wilcoxon W	Raw p	Holm p	Adjusted Decision
Augmented vs LR	15	0.0640	0.0640	Not significant
Augmented vs RF	2	0.0015	0.0029	Significant
Augmented vs MLP	0	0.0005	0.0015	Significant
Friedman omnibus	$\chi^2 = 12.2$	0.0067	—	Methods differ

5 | Discussion

5.1 | What the Hopf Result Does and Does Not Show

The most credible positive finding is that the full augmented pipeline performs better than the one-state classical baselines on the reported Hopf simulations. This is dynamically plausible. Oscillatory onset is governed by a coupled two-dimensional stability structure, and a single prey channel may not expose the phase relation, predator response or nullcline geometry that changes as the equilibrium loses stability. Mechanism-informed features can concentrate this information into low-dimensional quantities that a linear classifier can use.

The matched-channel result prevents overstatement. When the classical model receives statistics from both prey and predator, its AUC rises to 0.897, almost identical to the augmented value of 0.901. The appropriate interpretation is not that local geometry is useless; it is that most of the measured gain comes from observing the second state. The remaining potential value may lie in timing, robustness across conditions or interpretability, but those claims require paired alarm outputs and broader simulations. This distinction is essential for sensor design. If a second channel is available, simple multivariate summaries may be competitive. If a channel is expensive or missing, the structural fit cannot recover information that was never observed.

The result also illustrates why transition type matters. Conventional CSD statistics can be less stable around oscillatory transitions, and spectral or multivariate features may be better aligned with the complex eigenstructure [18], [23]. Future Hopf experiments should therefore compare the current geometry features against frequency-domain EWS, vector autoregression, covariance eigenvalues and models trained across variable oscillation frequencies—not only against univariate rolling statistics.

5.2 | Why the Fold Task Saturates

The fold task yields a useful negative result. Classical features already rank imminent windows almost perfectly, leaving little headroom for local damping or threshold-distance estimates. In a one-state system with a fixed slow ramp and known crossing schedule, mean level, variance and autocorrelation can jointly encode proximity to the fold. The augmented features may be redundant transformations of the same trajectory information.

Near-perfect AUC does not necessarily imply deployable warning. A classifier can exploit the deterministic progression of the state under a fixed ramp even if it has not learned a transferable stability signal. Varying the initial pressure, ramp slope, crossing time and process-noise realization would reveal whether predictions are tied to an absolute trajectory phase. Including non-transitioning ramps that approach but turn away from the fold would be especially valuable: such controls test whether the detector confuses high pressure with inevitable transition.

A more difficult fold benchmark should also separate interpolation from extrapolation. Training on some forcing rates and testing on unseen rates, on irregular sampling, or on a moderately misspecified harvesting function would test structural robustness. Under those conditions, a reliable damping estimate might become more useful—or its sensitivity to derivative noise might become apparent.

5.3 | Operational Meaning of the Sequential Alarm

The use of CUSUM is conceptually stronger than thresholding each window independently because a transition precursor should persist. Page's recursion accumulates evidence above a reference value and resets when evidence weakens [28]. The validation-only choice of ν and τ also respects the separation between tuning and final testing. Nevertheless, using the maximum validation far-region statistic sets a threshold from only six trajectories and cannot guarantee a stable false-alarm rate. Zero observed far alarms among 12 test trajectories is compatible with a nonzero underlying risk.

Operational studies should estimate a lead–false-alarm frontier. For each candidate threshold, investigators can report the fraction detected, distribution of first-alarm lead, number and duration of far alarms, and time under alarm. Confidence intervals should be obtained by resampling trajectories rather than windows. Threshold selection can then be linked to the relative cost of missed transitions, premature interventions and delayed action. Such analysis would make the decision problem explicit rather than treating one validation-derived operating point as universally optimal.

Calibration under distribution shift is equally important. The far-region score distribution can change with ramp rate, baseline abundance, sensor noise or model mismatch. Prospective monitoring may therefore require rolling recalibration, conformal or hierarchical thresholds, or explicit false-alarm constraints. Those extensions are outside the current experiment but follow directly from the sequential framing.

5.4 | Limitations and Reproducibility Requirements

The evidence is bounded by five design choices. First, only two low-dimensional simulated systems are evaluated. Second, all trajectories use regular sampling and a fixed nominal transition time. Third, window labels are derived from known simulator distance to that time, so the task does not represent unknown event timing. Fourth, only 12 test trajectories are available per condition. Fifth, the described code, raw trajectories and per-trajectory outputs were not present in the supplied materials, so the numerical results have not been independently reproduced.

Several implementation details must be preserved in a reproducibility package: numerical integration and step size, exact process-noise equations, observation-noise scaling, all realized seeds, trajectory allocation, feature normalization, derivative estimation, local-fit regularization and conditioning, classifier hyperparameters, validation-score construction, CUSUM reset logic, far-region denominator, per-trajectory AUCs and alarm episodes. An environment manifest and archived source should regenerate every table and figure from raw simulation output. Without those artifacts, aggregate values can be edited and interpreted but not computationally verified.

The structural features also assume that relevant state channels and approximate mechanism are known. Missing predators, delayed observations, asynchronous sampling and unmodeled feedback may destabilize ratio and nullcline estimates. Mechanism-informed learning can reduce variance when its assumptions are correct, but misspecification can introduce systematic bias. This trade-off should be tested explicitly by perturbing functional forms and withholding channels.

Finally, the present study assesses bifurcation-induced warning. It does not justify claims about noise-induced escape, rate-induced tipping or conservation-law learning. Those phenomena require distinct simulation designs and target definitions [9]. Likewise, a high AUC before a model-defined transition should not be described as proof of prevention, resilience recovery or field utility.

5.5 | Research Agenda

A decisive next study would use a factorial simulation benchmark. Transition mechanism, ramp rate, crossing time, sampling interval, missingness, process noise, observation noise and model mismatch should vary independently. Training should occur on one subset of mechanisms and parameter ranges, while testing includes unseen schedules and non-transitioning controls. Comparators should include univariate and

multivariate classical statistics, spectral EWS, local Jacobian or damping estimates, covariance eigenvalues and learned representations. All methods should share the same observed channels.

Evaluation should report trajectory-level ROC AUC and average precision, calibration, lead–false-alarm curves, detection fraction, missed-event rate and confidence intervals. Paired tests should target prespecified contrasts, especially combined features versus matched-channel classical features. External empirical examples can then be introduced only after the synthetic benchmark identifies where the structural assumptions help and where they fail.

This staged program would convert the current proof of concept into a transferable contribution. Its central hypothesis is testable: mechanism-informed features should provide the greatest benefit when a generic indicator is weak, the relevant coupled state is observed, the local model is approximately correct, and the evaluation shifts enough to penalize memorization of trajectory phase.

6 | Conclusion

This study integrates classical early-warning statistics, local dynamical geometry and validation-calibrated sequential alarms in two controlled simulation tasks. The reported Hopf experiment shows strong improvement over one-channel baselines, but the matched-channel ablation reveals that most of the gain follows from observing both interacting states. The incremental discrimination attributable to structural transformations is small in the current test set. The fold experiment is nearly saturated and provides no meaningful evidence of added value from augmentation.

The defensible contribution is therefore methodological and conditional. Local mechanism features can be embedded in a transparent warning pipeline, evaluated on complete held-out trajectories and interpreted through alarm burden as well as AUC. Their advantage must be demonstrated against classical features with identical information access and under forcing schedules that differ from training. With executable code, raw trajectories and the broader validation program specified above, the approach could become a useful bridge between generic CSD indicators and fully mechanistic forecasting.

Data and Code Availability

The source manuscript names an experiment script (exp_01.py) and an aggregate results file (exp_01_results.json), but those files and the raw trajectories were not included in the materials available for this revision. Publication should be accompanied by an archived package containing source code, realized seeds, trajectory splits, raw and processed data, per-trajectory predictions, alarm events and environment metadata.

Ethics Approval

The reported study uses simulated dynamical systems and no human participants, animals or identifiable personal data.

Funding

No funding information was included in the source manuscript.

Competing Interests

No competing-interest declaration was included in the source manuscript.

Author Contributions

Individual contributor roles were not specified in the source manuscript.

References

[1] Scheffer, M., Carpenter, S., Foley, J. A., Folke, C., & Walker, B. (2001). Catastrophic shifts in ecosystems. *Nature*, *413*(6856), 591–596. https://doi.org/10.1038/35098000

[2] Dakos, V., Scheffer, M., Van Nes, E. H., Brovkin, V., Petoukhov, V., & Held, H. (2008). Slowing down as an early warning signal for abrupt climate change. *Proceedings of the National Academy of Sciences*, *105*(38), 14308–14312. https://doi.org/10.1073/pnas.0802430105

[3] George, S. V, Kachhara, S., & Ambika, G. (2023). Early warning signals for critical transitions in complex systems. *Physica Scripta*, *98*(7), 72002. https://doi.org/10.1038/nature08227

[4] Lade, S. J., & Gross, T. (2012). Early warning signals for critical transitions: A generalized modeling approach. *PLoS Computational Biology*, *8*(2), e1002360. https://doi.org/10.1371/journal.pcbi.1002360

[5] Drake, J. M., & Griffen, B. D. (2010). Early warning signals of extinction in deteriorating environments. *Nature*, (467), 456–459. https://doi.org/10.1038/nature09389

[6] Cline, T. J., Seekell, D. A., Carpenter, S. R., Pace, M. L., Hodgson, J. R., Kitchell, J. F., & Weidel, B. C. (2014). Early warnings of regime shifts: Evaluation of spatial indicators from a whole-ecosystem experiment. *Ecosphere*, *5*(8), 1–13. https://doi.org/10.1890/ES13-00398.1

[7] Veraart, A. J., Faassen, E. J., Dakos, V., Van Nes, E. H., Lürling, M., & Scheffer, M. (2012). Recovery rates reflect distance to a tipping point in a living system. *Nature*, *481*(7381), 357–359. https://doi.org/10.1038/nature10723

[8] Kuehn, C. (2011). A mathematical framework for critical transitions: Bifurcations, fast-slow systems and stochastic dynamics. *Physica D: Nonlinear Phenomena*, *240*(12), 1020–1035. https://doi.org/10.1016/j.physd.2011.02.012

[9] Ashwin, P., Wieczorek, S., Vitolo, R., & Cox, P. (2012). Tipping points in open systems: Bifurcation, noise-induced and rate-dependent examples in the climate system. *Philosophical Transactions: Mathematical, Physical and Engineering Sciences*, 1166–1184. https://www.jstor.org/stable/41348437

[10] Dakos, V., Carpenter, S. R., Brock, W. A., Ellison, A. M., Guttal, V., Ives, A. R., ..., & Scheffer, M. (2012). Methods for detecting early warnings of critical transitions in time series illustrated using simulated ecological data. *PloS one*, *7*(7), e41010. https://doi.org/10.1371/journal.pone.0041010

[11] Lenton, T. M., Livina, V. N., Dakos, V., Van Nes, E. H., & Scheffer, M. (2012). Early warning of climate tipping points from critical slowing down: Comparing methods to improve robustness. *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences*, *370*(1962), 1185. https://doi.org/10.1098/rsta.2011.0304

[12] Boettiger, C., & Hastings, A. (2012). Early warning signals and the prosecutor's fallacy. *Proceedings of the Royal Society B: Biological Sciences*, *279*(1748), 4734. https://doi.org/10.1098/rspb.2012.2085

[13] Boettiger, C., & Hastings, A. (2012). Quantifying limits to detection of early warning for critical transitions. *Journal of the Royal Society Interface*, *9*(75), 2527. https://doi.org/10.1098/rsif.2012.0125

[14] Jäger, G., & Füllsack, M. (2019). Systematically false positives in early warning signal analysis. *PLoS one*, *14*(2), e0211072. https://doi.org/10.1371/journal.pone.0211072

[15] Scheffer, M., Bascompte, J., Brock, W. A., Brovkin, V., Carpenter, S. R., Dakos, V., ..., & Sugihara, G. (2009). Early-warning signals for critical transitions. *Nature*, *461*(7260), 53-59. https://doi.org/10.1038/nature08227

[16] Rosenzweig, M. L., & MacArthur, R. H. (1963). Graphical representation and stability conditions of predator-prey interactions. *The American Naturalist*, *97*(895), 209–223. https://doi.org/10.1086/282272

[17] Ludwig, D., Jones, D. D., Holling, C. S., & others. (1978). Qualitative analysis of insect outbreak systems: The spruce budworm and forest. *Journal of Animal Ecology*, *47*(1), 315–332. https://doi.org/10.7208/chicago/9780226125534-035

[18] Bury, T. M., Bauch, C. T., & Anand, M. (2020). Detecting and distinguishing tipping points using spectral early warning signals. *Journal of the Royal Society Interface*, *17*(170), 20200482. https://doi.org/10.1098/rsif.2020.0482

[19] Kéfi, S., Dakos, V., Scheffer, M., Van Nes, E. H., & Rietkerk, M. (2013). Early warning signals also precede non-catastrophic transitions. *Oikos*, *122*(5), 641–648. https://doi.org/10.1111/j.1600-0706.2012.20838.x

[20] Dakos, V., Carpenter, S. R., van Nes, E. H., & Scheffer, M. (2015). Resilience indicators: Prospects and limitations for early warnings of regime shifts. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *370*(1659), 20130263. https://doi.org/10.1098/rstb.2013.0263

[21] Scheffer, M., Carpenter, S. R., Dakos, V., & van Nes, E. H. (2015). Generic indicators of ecological resilience: Inferring the chance of a critical transition. *Annual Review of Ecology, Evolution, and Systematics*, *46*(1), 145–167. https://doi.org/10.1146/annurev-ecolsys-112414-054242

[22] Brock, W. A., & Carpenter, S. R. (2012). Early warnings of regime shift when the ecosystem structure is unknown. *PLOS ONE*. https://doi.org/10.1371/journal.pone.0045586

[23] Chen, S., O'Dea, E. B., Drake, J. M., & Epureanu, B. I. (2019). Eigenvalues of the covariance matrix as early warning signals for critical transitions in ecological systems. *Scientific Reports*, *9*(1), 2572. https://doi.org/10.1038/s41598-019-38961-5

[24] Brunton, S. L., Proctor, J. L., & Kutz, J. N. (2016). Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, *113*(15), 3932–3937. https://doi.org/10.1073/pnas.1517384113

[25] Raissi, M., Perdikaris, P., & Karniadakis, G. E. (2019). Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, *378*, 686–707. https://doi.org/10.1016/j.jcp.2018.10.045

[26] Karniadakis, G. E., Kevrekidis, I. G., Lu, L., Perdikaris, P., Wang, S., & Yang, L. (2021). Physics-informed machine learning. *Nature Reviews Physics*, *3*(6), 422–440. https://doi.org/10.1038/s42254-021-00314-5

[27] Bury, T. M., Sujith, R. I., Pavithran, I., Scheffer, M., Lenton, T. M., Anand, M., & Bauch, C. T. (2021). Deep learning for early warning signals of tipping points. *Proceedings of the National Academy of Sciences*, *118*(39), e2106140118. https://doi.org/10.1073/pnas.2106140118

[28] Page, E. S. (1954). Continuous inspection schemes. *Biometrika*, *41*(1/2), 100–115. https://doi.org/10.2307/2333009

[29] Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

[30] Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PloS One*, *10*(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

[31] Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., ..., & Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, *40*(8), 913-929. https://doi.org/10.1111/ecog.02881

[32] Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

[33] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ..., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *the Journal of Machine Learning Research*, *12*, 2825-2830. https://www.jmlr.org/papers/v12/pedregosa11a.html

[34] Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 65–70. https://www.jstor.org/stable/4615733