SCIDYNBENCH-R: BENCHMARKING ERROR–VALIDITY DISAGREEMENT IN SCIENTIFIC DYNAMICAL FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Scientific dynamical forecasts are usually selected by aggregate point error, but scientific use often depends on regimes, events, oscillations, and constraints. We introduce SciDynBench-R, a benchmark that connects scientifically motivated distribution shifts with task-specific validity criteria across controlled ODEs, Palk inspired biological transport, and leakage safe reduced WeatherBench2 Z500 climate indices. Its reproducible protocols jointly evaluate point error and scientific validity and diagnose disagreement between their model rankings. Across 187 frozen baseline model/split means, SciDynBench-R finds 44 low error scientific failures: ODE forecasts in the wrong regime, biological forecasts with large calcium peak count errors, and climate forecasts selected by NRMSE that behave as event smoothers, missing high variability and amplification events despite low aggregate error. A climate ablation illustrates improvements in event fidelity at higher NRMSE among the evaluated configurations. The benchmark and frozen results provide a reference for examining when favorable point error coexists with failures of task-specific scientific properties.

1 INTRODUCTION

Scientific dynamical forecasts are often evaluated as if their purpose were only to minimize average point error. Yet in many scientific settings, a forecast is useful only if it preserves qualitative structure: an epidemic trajectory should not cross into the wrong outbreak class, a physiological transport model should not erase calcium oscillations, and a climate forecast should not become an event smoother that erases the amplification or high variability events that motivate the prediction. Modern sequence and scientific machine learning models have advanced forecasting and dynamics learning (Chen et al., 2018; Raissi et al., 2019; Li et al., 2021; Zhou et al., 2021; Wu et al., 2021; Nie et al., 2023; Wu et al., 2023), but a trajectory can remain competitive under MSE or NRMSE while failing a regime, event, oscillation, or constraint check that carries the scientific meaning of the forecast.

This mismatch is amplified by evaluation protocols that treat overlapping or nonstationary time series as exchangeable samples. Forecasting studies have warned that validation design and leakage can dominate apparent progress (Hewamalage et al., 2023; Kapoor & Narayanan, 2023). In scientific dynamics, the concern is sharper because leakage or IID style splitting can hide failures that appear only under parameter, boundary condition, temporal, regime, anomaly, or event conditioned shifts. Existing benchmarks such as WeatherBench/WeatherBench 2 (Rasp et al., 2020; 2024), PDEBench (Takamoto et al., 2022), and WILDS (Koh et al., 2021) provide important evaluation infrastructure. SciDynBench-R asks a diagnostic question: under scientifically motivated shifts, when do error-based model rankings disagree with task-specific scientific validity, and which dynamical properties are affected?

We introduce SciDynBench-R, a benchmark connecting scientific forecasting tasks, distribution shifts, and validity criteria through reproducible evaluation protocols (Figure 1). The benchmark is organized around three evidence tiers. Controlled canonical ODEs provide known regimes, a mechanistic biological transport system tests whether calcium and secretion dynamics are preserved, and a leakage safe reduced WeatherBench2 Z500 setting adds train fitted preprocessing and zero overlap

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.

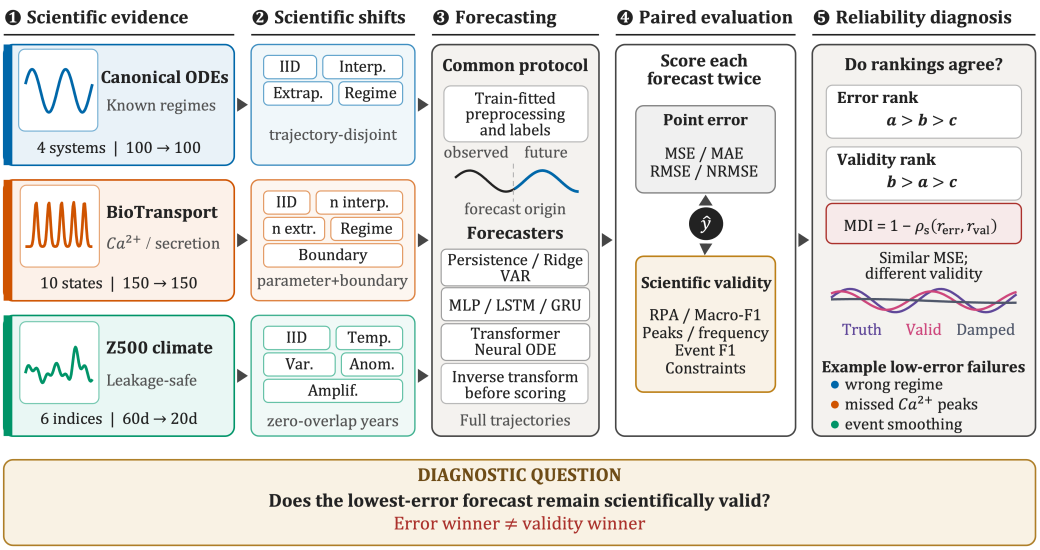


Figure 1: SciDynBench-R benchmark overview. Three evidence tiers are evaluated under IID and scientific shift splits, passed through a common forecasting protocol, and scored twice: once by aggregate point error and once by task specific scientific validity. The benchmark then diagnoses rank disagreement and low error scientific failures, asking whether the forecast with the lowest error remains scientifically valid.

climate split audits. Across these tiers, SciDynBench-R evaluates forecasts by point error, regime preservation, event and oscillation fidelity, constraint violation, and rank disagreement between error based and validity based model rankings.

Our frozen experiments contain 187 baseline model/split means and 44 predefined low error scientific failures spanning all three families. The paper contributes reproducible tasks and protocols that connect scientifically motivated shifts with explicit validity criteria; a cross family characterization of ranking disagreement and scientific-property failures; and a climate ablation illustrating observed error and event-fidelity tradeoffs. Together, these contributions provide a benchmark reference for investigating evaluation disagreement in controlled, mechanistic, and reduced real data settings.

2 Related Work

2.1 Scientific and Time Series Forecasting Benchmarks

Benchmarking has been central to progress in time series forecasting and scientific machine learning. The Monash archive standardizes diverse real world forecasting comparisons (Godahewa et al., 2021). Recent long horizon work compares efficient Transformers, decomposition models, linear baselines, patch models, and temporal two dimensional variation models (Zhou et al., 2021; Wu et al., 2021; Zeng et al., 2023; Nie et al., 2023; Wu et al., 2023). WeatherBench, WeatherBench 2, and PDEBench provide scientific forecasting and simulation tasks (Rasp et al., 2020; 2024; Takamoto et al., 2022). SciDynBench-R complements these resources by connecting scientifically motivated shifts with task-specific validity criteria and jointly analyzing forecast error, scientific-property failures, and disagreement between model rankings.

2.2 Scientific Validity and Structure Aware Modeling

Scientific machine learning has long recognized that predictions should respect structure, not merely fit observations. Neural ODEs learn continuous time dynamics (Chen et al., 2018). Physics informed methods impose equation residuals or priors (Raissi et al., 2019; Karniadakis et al., 2021), and neural operators learn mappings between function spaces (Li et al., 2021). These approaches modify the model or training objective. SciDynBench-R is complementary: it treats scientific validity as

Table 1: SciDynBench-R benchmark inventory. The climate component uses reduced Z500 derived indices and proxy diagnostics, not full field weather prediction.

Family	Systems / variables	Forecast task and splits	Scientific metrics	Evidence role
Canonical ODEs	Van der Pol, Lotka–Volterra, SIR, Lorenz; two or three states	100 observed steps to 100 future steps; IID, parameter interpolation/extrapolation, and regime shift	Regime preservation, Macro F1, oscillation/event errors, constraint violations	Controlled proof of concept: known regimes
BioTransport	Palk inspired ten state transport model with Ca, IP3, ions, and volume	150 observed steps to 150 future steps; IID, n interpolation/extrapolation, regime and boundary shifts	Calcium peak and amplitude fidelity, secretion response, physiological constraints	Mechanistic stress test: calcium/secretion failures matter
Climate indices	WeatherBench2 Z500 derived six index daily sequence	60 observed days to 20 future days; grouped IID, temporal OOD, variability, anomaly, and amplification shifts	Variability/anomaly regimes, event F1, proxy storm track diagnostics, broad plausibility checks	Real data stress test: event fidelity under leakage safe splits

an evaluation target applicable to any forecaster, including simple baselines and generic sequence models.

2.3 Distribution Shift, Leakage, and Forecast Evaluation

Robust evaluation is difficult when observations are temporally dependent or drawn from changing regimes. WILDS shows that IID performance can misrepresent deployment robustness (Koh et al., 2021). Forecasting methodology emphasizes validation design, temporal ordering, horizon choice, and leakage control (Hewamalage et al., 2023), while leakage is a reproducibility threat in machine learning based science (Kapoor & Narayanan, 2023). SciDynBench-R builds these concerns into its splits: parameter extrapolation, boundary condition shift, regime shift, temporal OOD, anomaly shift, and event conditioned climate shifts correspond to scientific hypotheses about when a forecast should remain meaningful.

3 SciDynBench-R

3.1 Design Principles

SciDynBench-R evaluates sequence to sequence forecasts of scientific dynamical systems. Each example contains an observation window, a future target window, and annotations for regimes, events, or constraints. Four design principles guide the benchmark. Point error and scientific validity are both primary outputs, shifted splits have domain meaning, preprocessing and thresholds are fitted only on allowed training or validation data, and cross family synthesis respects evidence strength rather than averaging raw physical errors across unrelated systems.

3.2 Controlled Canonical Systems

The first tier contains Van der Pol, Lotka–Volterra, SIR, and Lorenz dynamics. For each family we sample 5,000 trajectories, save 200 time points over $t = 0$ to 40, and forecast 100 future points from 100 observed points. The systems provide interpretable qualitative structure: oscillation amplitude, population cycles or collapse, outbreak class, and Lorenz variability.

Canonical splits test IID interpolation, parameter interpolation, parameter extrapolation, and held out scientific regimes. Regime labels use trajectory features such as amplitudes, peak counts, population minima, outbreak peaks, and late window variance. Constraint checks flag nonfinite, nonphysical, divergent, or implausible trajectories. This tier gives controlled evidence because failures can be tied to known dynamics rather than ambiguous measurement noise.

3.3 Mechanistic Biological System

The second tier introduces BioTransport-Palk-Full, a Palk inspired salivary acinar cell transport model (Palk et al., 2010). It is not an exact reproduction of every original equation, but it preserves the coupled calcium, ion, secretion, and volume mechanisms needed for a forecasting stress test. We generate 5,000 stiff ten state trajectories over $t = 0$ to 600 with 300 saved points and forecast 150 future points from 150 observed points.

Biological splits mirror the canonical questions and add a boundary condition perturbation. Interpolation and extrapolation hold out regions of the agonist/IP3 production parameter n. Regime shift tests sustained calcium oscillation and high secretion responses, and boundary shift tests low influx or RyR block like trajectories. Validity metrics emphasize calcium peak count, amplitude and frequency errors, secretion response labels, and physiological constraints such as nonnegative concentrations and positive volume.

3.4 Reduced Climate External Validation

The third tier uses WeatherBench2 ERA5 Z500 as real data external validation (Rasp et al., 2024). We reduce daily 500 hPa geopotential from 1960 to 2020 to six indices: two EOF/PCA components and four Z500 derived proxy diagnostics for variability, Eady growth rate, and storm track behavior. These are reduced Z500 only proxies, not exact physical EKE or Eady quantities.

Climate examples use 60 observed days and 20 future days. To avoid overlap leakage from stride one windows, SciDynBench-R assigns complete calendar years before selecting windows, drops boundary crossing windows, and fits climatology, EOF/PCA, standardization, and label thresholds on training years only. Grouped IID uses disjoint year groups. Temporal and event conditioned splits use 1960 to 2002 for training, 2003 to 2011 for validation, and 2012 to 2020 for testing. Event splits test high variability, strong PC1 anomalies, and future variability amplification, and all climate splits record zero source date overlap.

3.5 Metrics and Disagreement Analysis

SciDynBench-R reports point errors (MSE, MAE, RMSE, NRMSE) together with scientific validity metrics. Regime metrics include regime preservation accuracy (RPA), Macro F1, and relabeled prediction confusion matrices. Oscillation and event metrics measure peak count, amplitude, frequency, variability, transition, anomaly, amplification, and storm track proxy fidelity where applicable. Constraint metrics report violation rate, severity, and divergence.

To characterize evaluation disagreement, we compare error-based and validity-based model rankings within each system and split. Let r_{err} be the model ranking induced by an error metric, with lower error ranked better, and let r_{val} be the ranking induced by a validity metric or group normalized validity summary, with higher validity ranked better. The Metric Disagreement Index is

$$\mathrm{MDI} = 1 - \rho_s(r_{\mathrm{err}}, r_{\mathrm{val}}),$$

where ρ_s is Spearman rank correlation. MDI near zero indicates similar rankings. Larger values indicate greater disagreement between error-based and validity-based rankings. We use MDI as a diagnostic, not as a new universal leaderboard score.

4 Experimental Protocol

4.1 Baseline Matrix

SciDynBench-R evaluates standard forecasters rather than a new architecture. Canonical and biological families use persistence, ridge regression, MLP, LSTM, GRU, Transformer encoder, and Neural ODE. Climate adds a VAR(1) baseline with ridge regularization. These cover constant, linear, autoregressive, feedforward, recurrent, attention based, and continuous time inductive biases. Hyperparameters are fixed before the final runs and are not tuned separately for IID versus shifted test sets.

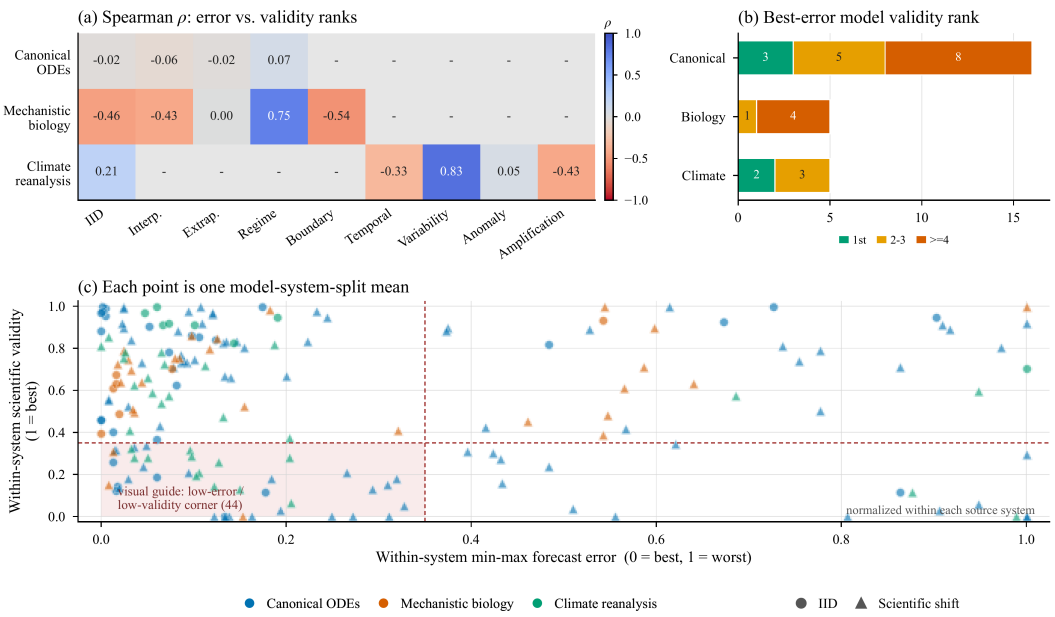


Figure 2: Model selection by error and by scientific validity disagree across all three evidence tiers. (a) Spearman correlation ρ compares model rankings by forecast error and scientific validity. Low or negative values mean that error rank is a poor proxy for validity rank. (b) The best error model often does not rank first by scientific validity. (c) Each point is one model, system, and split mean. Error and validity are min max normalized within each source system. The dashed 0.35 lines and shaded corner are visual guides for the low error and low validity region, not thresholds used to count failures.

4.2 Training and Model Selection

Each run is defined by family, system, split, model, and seed. Normalization is fitted on training windows only, predictions are inverse transformed before scientific scoring, and neural models are trained with Adam on normalized future window MSE with validation MSE early stopping. Deterministic baselines use seed 42. Neural canonical and biological runs use seeds 7, 21, and 42, while neural climate runs use seeds 7, 21, 42, 84, and 123. We report neural mean and standard deviation, while deterministic baselines remain single run estimates.

4.3 Leakage Controls

Canonical and biological partitions are disjoint at the trajectory level. Climate requires stricter handling: complete calendar years are assigned before windowing, boundary crossing windows are dropped, preprocessing and thresholds are fitted on training data, and source date overlap is audited. All frozen climate splits have zero overlap, and all climate outputs from before the fix are excluded from the evidence snapshot. The frozen manifest contains 597 successful raw rows: 272 canonical, 85 biological, 140 leakage safe climate baseline rows, and 100 climate ablation rows.

4.4 Aggregation and Diagnostic Analysis

Scientific metrics are computed on full test sets after concatenating observed prefixes with predicted futures and relabeling trajectories with the ground truth labelers. Cross system summaries use 187 model/split means: 112 canonical, 35 biological, and 40 leakage safe climate rows. Low error scientific failures are identified by fixed rules for each family, and rank disagreement diagnostics are computed within each system and split. The climate fidelity ablation is reported separately as an empirical error and event-fidelity comparison. It is not merged into the baseline leaderboard or used for cross system claims.

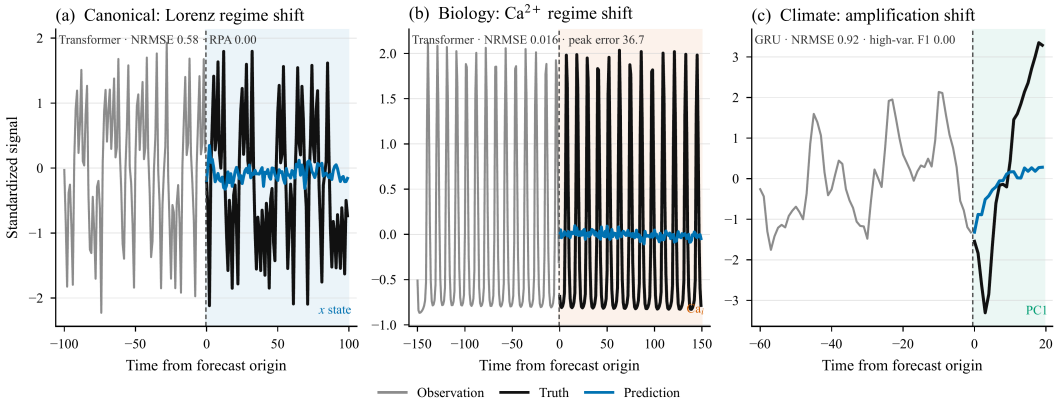


Figure 3: Representative low error scientific failures. Each panel shows an observed prefix, the ground truth future, and a model forecast that remains competitive by aggregate error while failing a scientific property: regime identity in controlled dynamics, calcium oscillation fidelity in BioTransport, or event fidelity in the reduced climate task.

5 Results

SciDynBench-R characterizes two patterns of error and validity under the evaluated shifts. In controlled and mechanistic systems, scientific shifts often make forecasts both less accurate and less valid. In the reduced climate setting, several event conditioned shifts are not harder by NRMSE, yet regime and event fidelity collapse. The benchmark shows not only that OOD data are hard, but that the scalar used to decide "harder" can miss the scientific failure.

5.1 Error Rankings Disagree with Scientific Rankings

Across 187 frozen baseline model/split means, Figure 2 shows that rank disagreement is not isolated to one domain. Canonical systems average near zero Spearman correlation between error and validity rankings across IID and shifted settings. Biological IID, interpolation, and boundary shifts are negative, and climate temporal OOD and amplification shifts are also negative. We also identify 44 low error scientific failures: 24 canonical, 10 biological, and 10 climate cases. These are not models that are bad at everything. They remain favorable by error while failing regime preservation, calcium peak count fidelity, or climate event detection.

Failure accounting. Low error failures are assigned by fixed rules before inspecting representative trajectories. A model/split mean must first fall in the favorable error region for its family, so the failure claim is not simply that the model is inaccurate. It is then counted only when a corresponding scientific validity metric fails: regime preservation or Macro F1 in canonical systems, calcium peak count or physiological response fidelity in BioTransport, and event F1 or regime fidelity in climate. This family specific accounting avoids comparing raw calcium errors, climate indices, and ODE state errors on an artificial common scale.

The failure count summarizes cases where favorable NRMSE or MSE coexists with failure under a specified scientific validity criterion. Together with the ranking analysis, it characterizes evaluation disagreement within the tested configurations. SciDynBench-R reports these diagnostics separately so that differences between error rankings and scientific-property preservation remain visible.

5.2 Representative Low Error Scientific Failures

The aggregate disagreement in Figure 2 is accompanied by visible dynamical errors in representative forecasts. Figure 3 shows representative cases from the frozen failure set. In the controlled tier, a forecast can follow the rough amplitude scale while crossing into the wrong qualitative regime. In the biological tier, the future mean level can be close while calcium peak timing and peak count are wrong, changing the mechanistic interpretation of secretion dynamics. In the climate tier, forecasts

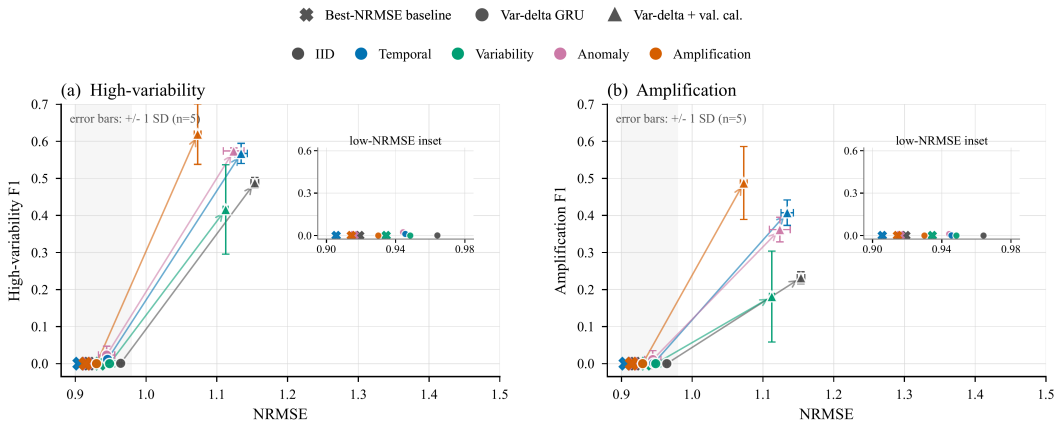


Figure 4: Climate event fidelity tradeoff. Best NRMSE baselines have low NRMSE but near zero high variability and amplification F1. The fidelity aware raw variant is a large GRU trained with a variability delta objective. The calibrated variant applies a validation fitted mean/std affine calibration that is frozen before test evaluation. The inset magnifies the low NRMSE region. Error bars show one standard deviation over five seeds.

with low NRMSE smooth the future index enough to miss high variability or amplification events. These examples explain why SciDynBench-R treats regime, event, oscillation, and constraint scores as scientific outputs rather than post hoc annotations.

5.3 Controlled ODEs: Wrong Regimes Under Low Error

The canonical tier gives the cleanest controlled evidence because qualitative regimes are known by construction. Averaged over Van der Pol, Lotka–Volterra, SIR, and Lorenz, regime shift increases forecast error by about 151% relative to IID and decreases regime preservation and Macro F1 by about 0.20 each. Under regime shift, a Lorenz Transformer forecast has low MSE but zero regime preservation. Lotka–Volterra MLP and Transformer forecasts remain low error while nearly eliminating the target population cycle regime, and SIR extrapolation includes low error GRU and Transformer forecasts with near zero regime preservation and large constraint violations. These are controlled examples of favorable error scores coexisting with poor scientific validity under the evaluated configurations.

5.4 Mechanistic Biology: Calcium Dynamics Are the Stress Test

The biological system turns disagreement into mechanistic consequence. In regime shift, NRMSE increases by about 256% relative to IID, biological RPA decreases by 0.211, and constraint violation rate increases by 0.0023. More importantly, low aggregate error can coexist with severe calcium dynamics errors: a Transformer under regime shift has NRMSE 0.016 but RPA 0.161 and calcium peak count error 36.7, while a Ridge forecast under boundary shift has NRMSE 0.0032, RPA 0.162, nonzero constraint violation, and peak count error 18.0. In this setting, peak count, secretion response, and physiological constraints are the mechanism being forecast, not decorative metrics.

5.5 Climate: Event Fidelity Collapses After Leakage Safe Splitting

The climate tier is deliberately framed as external validation rather than a full weather forecasting benchmark. After the leakage repair, all climate splits use preprocessing fitted on training years and zero audited source date overlap. Under this stricter protocol, reduced Z500 forecasts reveal a validity failure that point error alone would understate. Variability shift increases NRMSE by only 1.2% relative to grouped IID, yet climate RPA drops by 0.583 and Macro F1 drops by 0.238. Amplification shift is even more diagnostic: NRMSE is slightly lower than IID on average, but RPA still drops by 0.582 and Macro F1 by 0.228.

The low error failure cases explain why. For the baseline with the best NRMSE in every climate split, high variability F1 and amplification F1 are both zero. The models that look best by NRMSE behave like event smoothers: they forecast reduced indices well enough in aggregate, but miss the high variability and amplification events that define the shifted test set. The climate tier carries weaker absolute regime evidence than controlled systems, but it gives strong external evidence that forecasts optimized for error can be weak on events.

These results describe event fidelity under the specified thresholds and leakage safe protocol. The event labels are fitted from the allowed training or validation period, the year groups are disjoint, and the overlap audit is zero for all frozen climate splits. The failure reflects a choice relevant to deployment: a forecaster can minimize average index error by damping future variability, while a scientific user may care precisely about the tail events that are damped. In this reduced Z500 task, SciDynBench-R does not claim to solve weather forecasting. It documents discrepancies between aggregate error and event fidelity in the evaluated reduced-index forecasts.

The climate ablation illustrates an observed error and event-fidelity tradeoff. Fidelity aware GRU variants calibrated on validation data recover event fidelity, but consistently move rightward in NRMSE. The point is not that one metric should replace another. SciDynBench-R makes the error and event-fidelity changes in these configurations measurable.

We keep this ablation separate from the frozen baseline leaderboard. The enhanced GRU uses a training objective aware of validity and a calibration step fitted on validation data, so it should be read as a constructive probe rather than as another baseline in the error ranking. Its role is to characterize how the evaluated configurations differ in point error and event fidelity. Figure 4 shows higher high variability and amplification F1 at higher NRMSE for the calibrated variants. This is an observed tradeoff among the displayed configurations, rather than evidence that improving event fidelity necessarily requires worse point error.

5.6 Observed Climate Error–Event Tradeoffs

Finally, joint evaluation characterizes the climate ablation configurations. In Figure 4, the best error baseline on temporal OOD has NRMSE 0.906 and zero event F1, while the calibrated variant reaches high variability F1 0.568 and amplification F1 0.408 at NRMSE 1.134. On amplification shift, the baseline has NRMSE 0.914 and zero event F1, whereas the calibrated variant reaches high variability F1 0.620 and amplification F1 0.488 at NRMSE 1.073.

These results illustrate how favorable point error and event fidelity can favor different configurations. The benchmark makes this disagreement explicit without establishing a universal tradeoff or independently identifying the effects of training objectives, calibration, and model selection.

6 Discussion

SciDynBench-R provides reproducible tasks and protocols for studying how forecast error relates to task-specific scientific validity under distribution shift. Its design connects scientific forecasting questions to explicit criteria for preserving regimes, oscillations, events, and constraints. Joint evaluation reveals differences between error-based rankings and scientific-property preservation that aggregate error summaries do not express.

The three evidence tiers play different roles. Canonical ODEs give controlled counterexamples, BioTransport-Palk-Full shows why such failures matter mechanistically, and the reduced climate task adds an external check that forecasts optimized for error can collapse event fidelity after leakage safe splitting. The climate ablation illustrates a use of the benchmark: joint metrics characterize observed differences in event fidelity and point error among the tested configurations.

A useful way to read these results is as a closed audit loop rather than as a model ranking. The canonical tier establishes that the mismatch is possible even when regimes are known and labels are not noisy. The biological tier raises the cost of the mismatch: missing a peak count or secretion response changes the mechanistic conclusion a user would draw. The climate tier then asks whether the same evaluation failure remains visible after leakage safe splitting on real sequences derived from reanalysis. This ordering matters because the benchmark does not rely on climate alone to

prove the phenomenon. Climate supplies external relevance after the controlled and mechanistic tiers have made the failure interpretable.

Practically, SciDynBench-R provides information for assessing whether a forecast meets task-specific scientific requirements alongside its point error. An application could use these diagnostics to specify acceptable validity criteria before comparing errors. The present experiments characterize evaluation disagreement and observed error and event-fidelity tradeoffs; they do not validate a prospective model-selection policy. The benchmark supports such future comparisons by making the properties preserved or violated by each forecast explicit.

This framing also makes negative benchmark results useful. If no baseline reaches acceptable validity at acceptable error, the conclusion is not simply that all models are weak. It points to the missing inductive bias, calibration target, data split, or scientific observable. Conversely, when a simple baseline preserves a validity property better than a neural model with lower error, SciDynBench-R gives a reason to keep that baseline in the comparison rather than treating it as obsolete. The intended output is an auditable evaluation record: which scientific properties each forecast preserves, under which shift, and at what point error.

Limitations. SciDynBench-R is diagnostic rather than exhaustive. The canonical and biological systems are simulated, and their regimes and constraints are chosen for interpretability. The biological system is Palk inspired rather than an exact reproduction of every physiological detail. The climate component uses reduced Z500 derived proxy indices, not full field weather prediction or exact EKE/Eady diagnostics, so it should not be read as a state of the art WeatherBench claim.

The validity metrics are intentionally specific to each task: regime labels, event thresholds, and constraints encode scientific judgments. MDI is a diagnostic summary of ranking disagreement, not a universal model score. The fidelity aware climate ablation is likewise a proof of concept, not a new general architecture. The experiments do not establish whether the observed disagreements persist across broader training objectives or probabilistic forecasters, or independently identify the effect of the selection criterion. Future work should extend SciDynBench-R to additional mechanistic systems, full field geophysical tasks, richer uncertainty estimates, and validity objectives beyond hand designed proxies.

7 Conclusion

SciDynBench-R makes disagreement between forecast error and task-specific scientific validity measurable through reproducible tasks, scientifically motivated shifts, and paired evaluation protocols. Across the evaluated controlled ODEs, biological transport system, and reduced climate indices, favorable point error coexists with failures of regimes, oscillations, events, or constraints. The benchmark documents these cases, differences between model rankings, and observed climate error and event-fidelity tradeoffs. Together, the benchmark and frozen results provide a reference for evaluating scientific forecasting methods beyond aggregate point error.

AI use statement

Generative AI tools assisted with research experiment execution and analysis, literature verification, and drafting and editing manuscript sections. The authors are responsible for verifying the assisted text, implementation, results, and scientific claims.

Reproducibility statement

The supplementary material describes data construction, split policies, metrics, baseline configurations, seed policies, and staged rerun commands. The anonymous code package supplies implementations, configurations, frozen tables, and representative predictions. Its README maps the 187 baseline model/split means, 44 low-error scientific failures, and four manuscript figures to their artifacts and provides commands for verifying results and regenerating figures without retraining.

REFERENCES

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K. Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, 2018.

Rakshitha Godahewa, Christoph Bergmeir, Geoffrey I. Webb, Rob J. Hyndman, and Pablo Montero-Manso. Monash time series forecasting archive. In *Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021.

Hansika Hewamalage, Klaus Ackermann, and Christoph Bergmeir. Forecast evaluation for data scientists: Common pitfalls and best practices. *Data Mining and Knowledge Discovery*, 37:788–832, 2023.

Sayash Kapoor and Arvind Narayanan. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9):100804, 2023.

George Em Karniadakis, Ioannis G. Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3:422–440, 2021.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Bruce Earnshaw, Imran S. Haque, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pp. 5637–5664, 2021.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations*, 2021.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations*, 2023.

Laurence Palk, James Sneyd, Trevor J. Shuttleworth, David I. Yule, and Edmund J. Crampin. A dynamic model of saliva secretion. *Journal of Theoretical Biology*, 266(4):625–640, 2010.

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

Stephan Rasp, Peter D. Dueben, Sebastian Scher, Jonathan A. Weyn, Soukayna Mouatadid, and Nils Thuerey. Weatherbench: A benchmark data set for data-driven weather forecasting. *Journal of Advances in Modeling Earth Systems*, 12(11):e2020MS002203, 2020.

Stephan Rasp, Stephan Hoyer, Alexander Merose, Ian Langmore, Peter Battaglia, Tyler Russell, Alvaro Sanchez-Gonzalez, Vivian Yang, Rob Carver, Shreya Agrawal, Matthew Chantry, Zied Ben Bouallegue, Peter Dueben, Carla Bromberg, Jared Sisk, Luke Barrington, Arka Bell, and Fei Sha. Weatherbench 2: A benchmark for the next generation of data-driven global weather models. *Journal of Advances in Modeling Earth Systems*, 16(6):e2023MS004019, 2024.

Makoto Takamoto, Timothy Praditia, Raphael Leiteritz, Dan MacKinlay, Francesco Alesiani, Dirk Pflueger, and Mathias Niepert. Pdebench: An extensive benchmark for scientific machine learning. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. In *Advances in Neural Information Processing Systems*, 2021.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. Timesnet: Temporal 2d-variation modeling for general time series analysis. In *International Conference on Learning Representations*, 2023.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 11106–11115, 2021.

SUPPLEMENTARY MATERIAL FOR SCIDYNBENCH-R

A REPRODUCIBILITY OVERVIEW

This supplement documents the data construction, split policy, baseline configuration, result aggregation, and artifact layout used by SciDynBench-R. The final manuscript uses a frozen result snapshot dated June 18, 2026 (Asia/Shanghai). The snapshot contains 597 successful raw result rows: 272 canonical rows, 85 biological rows, 140 leakage safe climate baseline rows, and 100 climate enhanced ablation rows. The cross system claims in the paper use 187 baseline model/split means and 44 predefined low error scientific failure cases.

The artifact is organized around two reproducibility levels. A lightweight verification path checks that the tables, figures, checksums, and paper numbers come from the same frozen snapshot. A full rerun path regenerates data, preprocessing, split files, baseline training results, analyses, and figures. The full climate rerun requires access to the public WeatherBench2 ERA5 Z500 store and enough GPU time for neural baselines with multiple seeds.

B BENCHMARK DATA AND SPLITS

B.1 CANONICAL ODE SYSTEMS

The canonical tier contains Van der Pol, Lotka–Volterra, SIR, and Lorenz systems. Each dataset contains 5,000 simulated trajectories over 200 time points. Forecast examples use 100 observation steps and 100 future steps. The split files are disjoint at the trajectory level. IID splits use random trajectory partitions. Parameter interpolation splits hold out internal parameter bins, parameter extrapolation splits hold out high parameter regions, and regime shift splits hold out qualitative labels that are known from the simulated dynamics.

B.2 BIOLOGICAL TRANSPORT SYSTEM

The biological tier uses BioTransport-Palk-Full, a Palk inspired ten state salivary transport model with calcium, IP3, ion, secretion, and volume variables. Calcium and IP3 are represented in nM, major ions in mM, and w is normalized cytosolic volume. The predicted state is

$$x = (C_i, C_e, P, L_i, N_i, K_i, L_\ell, N_\ell, K_\ell, w),$$

corresponding to `Ca_i`, `Ca_er`, `IP3`, `Cl_i`, `Na_i`, `K_i`, `Cl_l`, `Na_l`, `K_l`, and `w`. With $\dot{w} = q_b - q_a$, the ODE used to generate trajectories is

$$\begin{aligned}
\dot{C}_i &= J_{\mathrm{IPR}} + J_{\mathrm{RyR}} + J_{\mathrm{er}} - J_{\mathrm{SERCA}} + J_{\mathrm{in}} - J_{\mathrm{pm}} - C_i\dot{w}/w,\\
\dot{C}_e &= (-J_{\mathrm{IPR}} - J_{\mathrm{RyR}} - J_{\mathrm{er}} + J_{\mathrm{SERCA}})/f_{\mathrm{er}},\\
\dot{P} &= k_n n - \Big(k_{5p} + k_{3K}C_i^2/(C_i^2 + k_{\mathrm{deg}}^2 + \epsilon)\Big)P - P\dot{w}/w,\\
\dot{L}_i &= (2J_{\mathrm{NKCC}} - I_{\mathrm{Cl}})/w - L_i\dot{w}/w,\\
\dot{N}_i &= (J_{\mathrm{NKCC}} - 3J_{\mathrm{NaK}})/w - N_i\dot{w}/w,\\
\dot{K}_i &= (J_{\mathrm{NKCC}} + 2J_{\mathrm{NaK}} - I_{K,b} - I_{K,a})/w - K_i\dot{w}/w,\\
\dot{L}_\ell &= (I_{\mathrm{Cl}} - q_{\mathrm{total}}L_\ell)/w_\ell,\\
\dot{N}_\ell &= (J_{\mathrm{tight,Na}} - q_{\mathrm{total}}N_\ell)/w_\ell,\\
\dot{K}_\ell &= (J_{\mathrm{tight,K}} + I_{K,a} - q_{\mathrm{total}}K_\ell)/w_\ell.
\end{aligned}$$

The main fluxes are reduced Hill and electrochemical transport laws. Let $H(z; K, h) = z^h/(K^h + z^h + \epsilon)$ and $(u)_+ = \max(u, 0)$. Then

$$\begin{aligned}
P_{\mathrm{IPR}} &= H(P; K_{\mathrm{IP3}}, h_{\mathrm{IP3}}) H(C_i; K_{\mathrm{act}}, h_c) \\
&\quad \times \frac{K_{\mathrm{inh}}^{h_c}}{K_{\mathrm{inh}}^{h_c} + C_i^{h_c} + \epsilon}, \\
P_{\mathrm{RyR}} &= H(C_i; K_{\mathrm{RyR}}, h_r), \quad P_{\mathrm{Cl}} = H(C_i; K_{\mathrm{Cl}}, h_{\mathrm{ch}}), \quad P_K = H(C_i; K_K, h_{\mathrm{ch}}), \\
J_{\mathrm{IPR}} &= k_{\mathrm{IPR}} s_{\mathrm{IPR}} P_{\mathrm{IPR}} (C_e - C_i)_+, \\
J_{\mathrm{RyR}} &= k_{\mathrm{RyR}} s_{\mathrm{RyR}} P_{\mathrm{RyR}} (C_e - C_i)_+, \\
J_{\mathrm{SERCA}} &= V_{\mathrm{SERCA}} s_{\mathrm{SERCA}} C_i^2/(K_{\mathrm{SERCA}}^2 + C_i^2 + \epsilon), \\
J_{\mathrm{in}} &= s_{\mathrm{in}}(J_{\mathrm{base}} + J_{\mathrm{agonist}} n), \\
J_{\mathrm{pm}} &= V_{\mathrm{pm}} C_i^3/(K_{\mathrm{pm}}^3 + C_i^3 + \epsilon), \\
J_{\mathrm{NKCC}} &= k_{\mathrm{NKCC}} \log R_{\mathrm{NKCC}}, \\
R_{\mathrm{NKCC}} &= N_e K_e L_e^2/(N_i K_i L_i^2 + \epsilon), \\
q_{\mathrm{total}} &= q_a + q_{\mathrm{tight}}.
\end{aligned}$$

The apical and basal K/Cl currents, tight junction current, osmotic flows, and bounded algebraic membrane potential proxies V_a, V_b are recomputed from the same state using the code in `systems/biotransport_palk_full.py`. This is also the authoritative executable specification used by all labels and metrics.

We generate 5,000 trajectories on $t \in [0, 600]$ with 300 time points using the BDF solver and seed 42. The default initial condition, before the small random perturbation, is

$$(50, 50000, 25, 61, 11, 145, 124, 119, 5.6, 1.0).$$

Each trajectory perturbs every initial component independently by a uniform factor in $[0.98, 1.02]$. Unless sampled, parameters use the defaults in the system file. The sampled parameters are uniform over:

$n : [3500, 7500]$	$s_{\mathrm{RyR}} : [0, 2]$	$s_{\mathrm{IPR}} : [0.5, 1.5]$
$s_{\mathrm{SERCA}} : [0.5, 1.5]$	$s_{\mathrm{in}} : [0, 1.5]$	$a_K : [0, 1]$
$s_{\mathrm{LPa}} : [0.5, 2]$	$s_{\mathrm{LPb}} : [0.5, 2]$	$R_{\mathrm{tight}} : [0.5, 2]$

The generated trajectories are labeled by calcium and secretion response, then checked for physiological constraints before splitting. The primary parameter splits hold out regions of the agonist/IP3 production parameter n. The regime shift split holds out sustained calcium oscillation and high secretion response trajectories. The boundary condition split holds out `Jin_scale`≤ 0.25 or `k_RyR_scale`≤ 0.25 examples, while training/validation examples require both quantities to be at least 0.50.

All biological forecasts use 150 observed steps and 150 future steps. Standard point errors are computed only on the ten primary state variables after inverse normalization. Derived quantities are not included in NRMSE. Specifically,

$$\mathrm{NRMSE} = \frac{\sqrt{\mathrm{mean}((Y - \hat{Y})^2)}}{\max(\mathrm{std}(Y), 10^{-8})},$$

where $Y, \hat{Y} \in \mathbf{R}^{N \times 150 \times 10}$ are the future true and predicted primary state tensors for the test split. The neural models use statewise normalization fitted on the training trajectories only. Metrics are computed after converting predictions back to physical units.

Biological validity is computed on the full trajectory formed from observations and predictions. For each prediction, we concatenate the observed prefix and predicted future, then recompute V_a, V_b, transport fluxes, channel probabilities, and q_{total} from the predicted primary states and sampled parameters. Regime validity uses the relabeled full trajectory: `Ca_i` and w define calcium/volume features, and q_{total} defines secretion response features. We report biological and abstract RPA/Macro

F1 against the true labels. Oscillation validity uses future `Ca_i` to compute amplitude, peak count, and dominant frequency errors. Additional mechanistic diagnostics use `IP3`, w, the six ion states, and the recomputed q_{total} for signaling, volume, ion, and secretion flow errors. Constraint validity checks all primary states plus derived $q_{\text{total}}, V_a, V_b$ for nonfinite values, negative concentrations, nonpositive volume, numerical explosion, broad voltage range, and absurdly large total flow. It reports violation rate, violation severity, and divergence rate.

For biological model rankings, the Scientific Validity Score (SVS) is computed within each split from min max utilities: biological RPA (0.30), biological Macro F1 (0.20), abstract RPA (0.15), low constraint violation rate (0.15), low divergence rate (0.10), and calcium oscillation utility (0.10). The calcium oscillation utility averages min max utilities of amplitude, peak count, and dominant frequency errors, where lower error is better. In the cross system Figure 2 summary, the biological validity metric uses biological RPA (0.30), biological Macro F1 (0.20), calcium event fidelity (0.25), constraint fidelity $1 - \text{CVR}$ (0.15), and divergence fidelity $1 - \text{divergence}$ (0.10), renormalizing weights when a component is missing.

When the paper refers to the error picked or best error model, the selection is made within each source system and split after aggregating raw runs to model/split means. The error metric is NRMSE when available and MSE otherwise. The best error model is the row with the smallest error metric, with ties broken by better validity rank and then model name. Its validity rank is obtained by sorting the same candidate rows by the corresponding validity metric in descending order. Low error biological failure cases are selected separately: within each system/split, a forecast is "low error" if it falls in the best 30% by MSE or NRMSE, and it is flagged when it also has weak regime agreement, high constraint violation, high calcium peak count error, or high calcium frequency error according to the rules specific to each split in `analysis/find_biological_low_mse_wrong_regime.py`.

B.3 Leakage Safe Climate Construction

The climate tier was rebuilt after detecting overlap leakage caused by stride one windows followed by random splitting at the window level. The frozen protocol assigns complete calendar years before selecting windows, drops any window whose source dates cross the assigned year boundary, and audits source date overlap between train, validation, and test partitions. The grouped IID climate split uses disjoint year groups. Temporal and event conditioned climate splits use chronological training, validation, and test year groups, then apply support filters for high variability, PC1 anomaly, or future variability amplification.

Climate preprocessing is specific to each split. For grouped IID, climatology, EOF/PCA, standardization, and label thresholds are fitted on grouped IID training years. For temporal and event conditioned splits, the same quantities are fitted on chronological training years only. The split files record the dataset path for the matching preprocessing variant. All frozen climate splits have zero audited source date overlap.

C Evaluation Metrics

SciDynBench-R reports point error metrics and scientific validity metrics together. Point errors include MSE, MAE, RMSE, and NRMSE. Regime metrics include regime preservation accuracy (RPA), Macro F1, and relabeled confusion matrices. Event and oscillation metrics are specific to each family. Canonical systems use amplitude, peak count, frequency, and regime diagnostics. The biological system emphasizes calcium peak count, amplitude, dominant frequency, secretion response, and physiological constraints. The climate tier reports high variability F1, amplification F1, transition F1, anomaly sign accuracy, variance preservation, and storm track proxy errors.

Constraint metrics report violation rate, violation severity, and divergence. For cross system synthesis, validity signals specific to each family are normalized only within comparable system groups. The Metric Disagreement Index (MDI) is computed within each system and split as

$$\text{MDI} = 1 - \rho_s(r_{\text{err}}, r_{\text{val}}),$$

where r_{err} ranks models by error, r_{val} ranks models by a validity metric or validity summary, and ρ_s is Spearman rank correlation. MDI is a diagnostic of ranking disagreement, not a universal leaderboard score.

D Baselines and Seeds

SciDynBench-R uses standard forecasting baselines rather than a new architecture. Canonical and biological experiments use persistence, ridge regression, MLP, LSTM, GRU, Transformer encoder, and Neural ODE baselines. The climate baseline matrix adds a VAR(1) model with ridge regularization. Neural models are trained with Adam on normalized future window MSE and selected by validation MSE early stopping. Predictions are inverse transformed before scientific scoring. Canonical and BioTransport neural baselines use seeds 7, 21, and 42. Reduced climate baselines and climate enhanced ablations use seeds 7, 21, 42, 84, and 123. Deterministic baselines are run once with seed 42. The standard neural policy is batch size 128, max 100 epochs, patience 10, and learning rate 10^{-3}. The climate enhanced ablation uses batch size 768, max 160 epochs, patience 20, and the same learning rate.

E Full Rerun Commands

The following commands are intended to be run from the project root. They describe the reproducible workflow rather than a single monolithic script. The repository's `experiments/run_all.py` entry point is intentionally not used because it is only a placeholder.

Environment.

```
conda env create -f environment.yml
conda activate scidynbench
pip install -e .
pytest
```

Canonical data and splits.

```
python -m preprocessing.generate_canonical_data
  --config configs/canonical.yaml
python -m preprocessing.label_canonical_regimes
python -m preprocessing.annotate_canonical_constraints
python -m preprocessing.create_canonical_splits
  --seed 42
```

Biological data and splits.

```
python -m preprocessing.generate_biological_data
  --config configs/biological.yaml
python -m preprocessing.label_biological_regimes
python -m preprocessing.annotate_biological_constraints
python -m preprocessing.create_biological_splits
  --seed 42
```

Leakage safe climate data and splits.

```
CLIM_SPLITS=data/splits/climate/weatherbench2_z500_daily
python -m preprocessing.download_weatherbench2_z500
  --config configs/climate_weatherbench2.yaml
python -m preprocessing.preprocess_climate
  --config configs/climate_weatherbench2.yaml \
  --data-dir data/processed/climate_benchmark \
  --split-dir $CLIM_SPLITS \
  --t-obs 60 --t-future 20 --stride 1 --seed 42
```

Baseline training. For the frozen tables with multiple seeds, repeat the following commands for the seeds specified above, writing one output per seed and then aggregating them into all seed, mean, and standard deviation tables.

```
python -m experiments.run_canonical_baselines
  --seed SEED
  --output results/tables/canonical_seedSEED.csv
python -m experiments.run_biological_baselines
  --seed SEED --device cuda \
  --output results/tables/biological_seedSEED.csv
python -m experiments.run_climate_baselines
  --seed SEED --device cuda \
  --output results/tables/climate_seedSEED.csv
```

Analysis and figures.

```
python -m analysis.cross_system_synthesis
  --table-dir results/tables \
  --output-dir results/tables \
  --figure-dir results/figures/cross_system
python analysis/generate_aaai_final_figures.py
  --snapshot results/final_aaai27 \
  --output-dir results/figures/aaai27_final
python results/final_aaai27/build_manifest.py
```

F Frozen Artifact Map

The frozen quantitative evidence is kept in a dated result snapshot. The manifest files record file sizes, row and column counts, seeds, status counts, image dimensions, and SHA-256 hashes. The core tables contain raw, mean, and standard deviation results for canonical, biological, leakage safe climate baseline, and climate enhanced ablation runs. The paper ready tables contain cross system difficulty, MDI, model rank, and low error failure summaries. Enhanced climate summaries are kept separately so the tradeoff ablation is not mixed into the baseline leaderboard. Exact mapping from results to files is provided in the reproducibility package.

G Known Artifact Boundaries

Three boundaries are important for interpreting the artifact. First, the climate component is a reduced Z500 index task with proxy diagnostics. It should not be read as a state of the art WeatherBench or full field numerical weather prediction claim. Second, the enhanced climate experiment is an ablation about the tradeoff between error and fidelity and is not merged into the baseline leaderboard. Third, the final climate enhanced table records the architecture, objective, calibration, seeds, device string, and dataset paths specific to each split in metadata. The corresponding enhanced ablation training entry point and RTX 4090 remote launcher are included alongside the baseline and analysis scripts. The project environment is defined by `environment.yml` and `pyproject.toml`. Full neural sweeps were run with GPU support on a rented NVIDIA RTX 4090 GPU. Lightweight table checks, manifest generation, and figure regeneration can be performed on CPU after the frozen tables are available.