MECHANISM LEARNING THROUGH PROTOTYPE-ANCHORED DYNAMICS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study *mechanism learning* for scientific forecasting: constructing an explicit description of local evolution and learning to use it in prediction. Prototype anchoring builds this dynamics descriptor from a finite bank of reusable support patterns. We instantiate this approach using local predictor coefficients in Burgers, observed increment patterns in WeatherBench 2 (WB2), and learned support vectors in Lorenz96. Three sets of experiments assess predictive utility, the contribution of anchoring, and the descriptor's dynamical meaning. First, across 20 shared seeds on a two-variable WB2 task at $+72$ h, the full model reduces temperature RMSE by 0.010 K and Z500 RMSE by $45.6\,\mathrm{m^2\,s^{-2}}$ relative to an approximately parameter-matched direct predictor (paired 95% CIs $[0.007, 0.013]$ and $[42.1, 49.1]$). Both contrasts remain significant after Holm correction. Second, finite support improves temperature forecasts over a learned descriptor generator with 20% and 50% of the training data; its effect varies with target and sample size. Third, Burgers descriptors align with independently computed PDE quantities after history matching and viscosity-pair/time controls (within-stratum partial rank correlation 0.210), and shuffling them increases one-step RMSE by 0.0094 (95% CI $[0.0057, 0.0142]$). Matched swaps provide no clear evidence that descriptor changes reproduce the direction of dynamical change. These findings support an empirical form of mechanism learning: descriptors can organize local dynamics and contribute to prediction, while the value of prototype anchoring depends on the forecasting regime.

1 INTRODUCTION

Scientific forecasting requires a representation of the current state and a description of how it will evolve. We study *mechanism learning*: learning to forecast through an explicit representation of local evolution. An *empirical local mechanism* describes evolution over a specified interval in the observed regime. We represent it by a *dynamics descriptor* $\hat{\theta}_t$ inferred from history, which the forecaster uses alongside observed-state information.

Prototype anchoring constructs this descriptor from a finite bank of reusable support patterns selected by the observed history. We instantiate the support as local predictor coefficients in Burgers, coarse observed forecast increments in WeatherBench 2 (WB2), and learned vectors in Lorenz96. The resulting descriptor is both an input to prediction and an object that can be compared with independently computed dynamics.

Our central claim is that prototype-anchored descriptors can organize local dynamics and contribute to scientific prediction, with the benefit of anchoring depending on the target and data regime. Three comparisons develop this claim: Full versus Direct assesses predictive utility; Full versus NoBank assesses the contribution of finite support; and dynamics probes and descriptor interventions assess dynamical meaning and functional use. WB2 supplies the strongest capacity-controlled forecast evidence, while Burgers connects descriptors to PDE quantities and their role in prediction.

Our contributions are:

1. **An explicit local-evolution representation.** We formulate mechanism learning around a dynamics descriptor used alongside observed history. This gives scientific forecasting an intermediate

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.

representation whose relationship to physical dynamics and contribution to prediction can be examined separately.
2. **Finite-support construction of dynamics descriptors.** We pair history-based retrieval with aggregation over observed or learned prototype support. This introduces a structural constraint on how local evolution is represented, making reusable dynamical patterns available to the forecaster across different histories.
3. **Mechanism-conditioned forecast computation.** We couple retrieved descriptors to local linear prediction in Burgers, feature modulation in WB2, and a learned predictor in Lorenz96. These realizations connect the shared representation to different forecasting architectures, with 20-seed predictive gains over Direct-matched in WB2 and evidence of dynamical alignment and forecast dependence in Burgers.

Impact statement. This work connects forecast learning with the empirical study of local dynamics. By making a dynamics descriptor part of the prediction computation, it provides a practical way to examine what a forecaster represents about change and how that representation affects its output. This can guide the design of scientific models that combine predictive utility with inspectable dynamical structure.

Related work

Forecast operators and latent dynamics. Fourier neural operators learn mappings between function spaces using spectral kernels (Li et al., 2021); DeepONet uses branch and trunk networks to represent nonlinear operators (Lu et al., 2021). Neural ODEs parameterize continuous-time dynamics, while latent ODEs and deep Kalman filters learn temporal structure through latent states (Chen et al., 2018; Rubanova et al., 2019; Krishnan et al., 2015). These approaches offer different representations of evolution. Autoregressive error growth remains a separate issue (McCabe et al., 2023). Our focus is a history-dependent descriptor that explicitly conditions a forecast and can be examined independently of the state features.

Data-driven weather forecasting. FourCastNet uses adaptive Fourier operators, Pangu-Weather uses three-dimensional Earth-specific networks, and GraphCast uses graph neural networks (Pathak et al., 2022; Bi et al., 2023; Lam et al., 2023). NeuralGCM combines learned components with a differentiable atmospheric dynamical core (Kochkov et al., 2024). These systems establish the value of architectural and physical structure for weather prediction. Our two-variable, coarse-grid WB2 experiment examines how a retrieved increment descriptor contributes to a shared forecast architecture.

Structured representations of dynamics. Sparse identification recovers governing equations from candidate libraries (Brunton et al., 2016). Dynamic mode decomposition extracts coherent evolution modes (Schmid, 2010), while Koopman predictors and learned Koopman embeddings seek coordinates in which nonlinear evolution can be modeled linearly (Korda & Mezić, 2018; Lusch et al., 2018). Reservoir computing also supports prediction of high-dimensional chaotic systems from observations (Pathak et al., 2018). These methods already connect representation and dynamics. We study local descriptors built from finite support and assess their content against independently computed dynamics.

Prototype support and feature conditioning. Matching networks predict through attention over labeled support examples, and prototypical networks classify through distances to class prototypes (Vinyals et al., 2016; Snell et al., 2017). FiLM provides feature-wise affine conditioning (Perez et al., 2018). We bring these ideas to local evolution: the retrieved values describe predictor coefficients, observed increments, or learned support vectors, and enter the forecast computation. Prototype construction and the conditioning architecture have distinct roles, motivating the Full–NoBank comparison.

Remaining gap. A remaining challenge is to connect a compact, history-dependent description of evolution with both independently computed dynamics and its functional role in forecasting. Predictive accuracy alone does not establish this connection. Mechanism learning addresses this

challenge by exposing the same descriptor to forecasting, dynamics comparisons, and interventions while retaining access to the observed history.

2 Mechanism learning with prototype anchoring

2.1 From history to a local evolution descriptor

Let $H_t = (x_{t-h+1}, \dots, x_t)$ be a history of length h and $Y_t = x_{t+\Delta}$ its forecast target. A finite bank $\mathcal{B} = \{(k_j, p_j)\}_{j=1}^{K}$ stores keys k_j and support descriptors p_j. For observed-fragment support, each pair is derived from training histories and their local evolution; for learned support, the bank is optimized with the forecaster. Bank construction uses training data only, and retrieval uses the observed history. At inference,

$$
\begin{aligned}
q_t &= Q(H_t), \qquad z_t = E(H_t), \\
\alpha_{t,j} &= \frac{\mathbf{1}\{j \in \mathcal{N}_t\} \exp(\ell_{t,j})}{\sum_{r \in \mathcal{N}_t} \exp(\ell_{t,r})}, \\
\hat{\theta}_t &= \mathcal{A}\big(\alpha_t, \{p_j\}_{j=1}^{K}\big), \qquad \hat{Y}_t = G(H_t, z_t, \hat{\theta}_t).
\end{aligned} \tag{1}
$$

Here Q constructs the retrieval query, E supplies forecast features, $\ell_{t,j}$ scores the similarity between query q_t and key k_j, and $\mathcal{N}_t$ is the selected support neighborhood. The aggregation map $\mathcal{A}$ converts weighted support into a dynamics descriptor. Direct value mixing gives $\hat{\theta}_t = \sum_j \alpha_{t,j} p_j$; Burgers combines direct coefficient mixing with a decoded latent mixture through a learned gate (Appendix H). The query map and forecast encoder can differ: WB2 uses a fixed pooled-history query and a learned spatial encoder.

Controls. We call the prototype-anchored model Full. Direct predicts from history without an explicit dynamics descriptor. In WB2, *Direct-matched* has approximately the same trainable parameter count as Full. The *NoBank* model preserves Full's spatial encoder, decoder, feature-wise linear modulation (FiLM), and output head, but constructs the descriptor with a multilayer perceptron applied to the encoder's global feature. Full retrieves using the fixed pooled-history query. Full versus Direct-matched tests the predictive utility of the complete descriptor-conditioned architecture. Full versus NoBank tests the contribution of anchoring within that architecture: both models use a descriptor, but only Full constructs it from finite support. The NoBank control is specific to WB2 and has a slightly larger parameter budget, as quantified in Section 3.

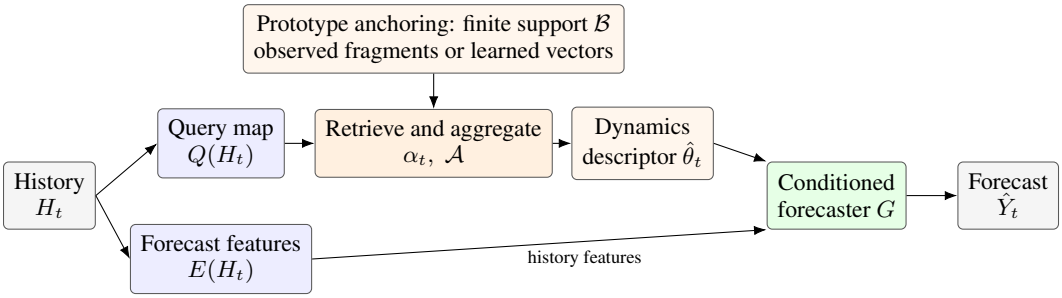


Figure 1: **Mechanism learning with prototype anchoring.** A history supplies both a retrieval query and forecast features. Finite support constrains the dynamics descriptor; the forecaster uses it alongside history features to predict the future. Descriptor interventions replace this input while holding the history features fixed. WB2 uses a fixed query map and an observed-fragment bank; the other systems use the instantiations in Table 1.

Training and inference. We construct observed support from the training split or initialize learned support, then optimize the trainable components to minimize forecast loss. Observed support remains fixed during training. At inference, we compute the query and forecast features from history, aggregate the retrieved support descriptors, and decode the forecast using Eq. (1). Table 1 identifies the components that vary by system.

Table 1: Domain-specific components of the shared computation. Support construction and forecast conditioning vary by system.

Component	Burgers	WeatherBench 2	Lorenz96
History H	20 time points at a spatial location	four time steps of both fields, 6 h apart	four system states
Support descriptor p	local ridge coefficients	standardized 4×4 pooled target increments	learned support vector
Key k	encoded local window	standardized 4×4 pooled history	learned key vectors
Support construction	representative training fragments	deterministic farthest-point selection of paired fragments	trainable finite bank
Conditioning	retrieved local linear predictor	FiLM decoder	descriptor-conditioned MLP
Training objective	local one-step forecast MSE	$+72$ h increment MSE	one-step forecast MSE
Evaluation role	PDE alignment; trajectory splice	capacity and functional controls at $+72$ h	forecasting and dynamics probes

2.2 WeatherBench 2 observed-fragment bank

WB2 keys are standardized, flattened 4×4 pooled histories (128 dimensions); support descriptors are pooled target increments $Y_t - x_t$ (32 dimensions). Deterministic farthest-point selection in standardized descriptor space chooses $K = 256$ paired training fragments. Standardization uses the current training subset, and the bank remains fixed during training. At inference, cosine similarity between the pooled-history query and stored keys, with temperature $\tau = 0.1$, weights the eight nearest entries. Their averaged descriptor modulates three decoder stages through FiLM (Perez et al., 2018). A convolutional encoder supplies spatial features; the decoder predicts an increment added to the latest field. Appendix B gives implementation details.

2.3 Assessing the descriptor's meaning and use

Residual predictability. A ridge probe predicts pooled WB2 residuals $R = Y - \hat{Y}_{\mathrm{DirectMatched}}$ from Full descriptors, NoBank latents, shuffled descriptors, Gaussian features, or pooled-history features. All inputs have 32 dimensions; probes use a fixed penalty $\lambda = 1$, training-only fitting, and disjoint test evaluation (Appendix B). This measures the linear accessibility of forecast residuals.

Descriptor intervention. We hold a trained Full model's history and encoder features fixed and replace only $\hat{\theta}$. WB2 replacements use a within-batch cyclic shift or the farthest observed prototype. Forecast change measures sensitivity; an error increase measures the benefit of the native descriptor. Burgers additionally tests *directional transport*: whether swaps between matched histories shift forecasts in the direction of observed or PDE-derived increment contrasts.

3 Experiments

3.1 Protocols and statistical reporting

WeatherBench 2. We use 64×32, 6-hour WB2 data (Rasp et al., 2024): four history steps of 2-m temperature (T_{2m}, K) and 500-hPa geopotential (Z500, $\mathrm{m^2\,s^{-2}}$), with $+72$ h supervision. A fixed 1,600/400 training/validation split comes from 1979–2016; 2019–2020 supplies 2,907 test samples. Full, NoBank, and Direct-matched share 20 seeds: the original $\{42, 137, 256, 512, 1024\}$ plus $\{0, \ldots, 14\}$, declared before extension results were inspected. They share the optimizer and validation-only checkpoint rule. Direct-matched has 371,072 parameters, within 2.3% of Full's 379,714; NoBank has 400,354 (5.44% more than Full). Data-scarcity experiments retain the original five seeds on nested 20%, 50%, and 100% subsets ($N = 320, 800, 1600$), with fixed validation/test sets. Headline RMSE pools latitude-weighted squared error across test times and grid cells before taking the square root (Appendix B).

Boxplots show seed distributions, and the main table reports medians and quartiles. We estimate mean paired benefits with 95% t-intervals over 20 shared seeds and apply Holm correction to four method-by-target contrasts within each two-sided test family: paired t, Wilcoxon signed-rank, and sign tests. Appendix B reports these tests, bootstrap intervals, cohort sensitivity, and leave-one-seed-out summaries. Intervals quantify training-seed variation on the fixed test set. Data-scarcity and functional diagnostics retain the original five seeds.

Burgers. PDEBench Burgers data (Takamoto et al., 2022) support one-step training and autoregressive evaluation. Descriptors from a frozen $K = 128$ Full checkpoint are compared with independently computed advection $A = -uu_x$, diffusion $D = (\nu/\pi)u_{xx}$, total tendency $T = A + D$, observed tendency, and balance $\rho = |A|/(|A| + |D| + 10^{-12})$. Fixed-penalty ridge probes use 96 training and 96 disjoint test trajectories per in-domain viscosity. Probe intervals cluster trajectory–viscosity realizations; matched-history and profile tests resample initial-condition indices jointly across viscosities. A counterfactual stress test splices constant-viscosity solutions at the same trajectory index and evaluates 128 held-out trajectories. Full and Direct have 207,399 and 268,801 parameters; this test uses one trained pair (Appendices E–H).

Lorenz96. For Lorenz96 (Lorenz, 1995), we evaluate autoregressive forecasts at $F = 8, N = 16$ using five seeds that determine data trajectories and training initialization. Full, Direct, LSTM, neural ODE (NODE), and temporal convolutional network (TCN) models share the data generator and rollout evaluator. For each frozen Full checkpoint, a ridge probe with fixed penalty is fit on its training trajectory and evaluated on an independently generated test trajectory for the nonlinear, damping, total right-hand-side (RHS), and observed-tendency terms; shuffled descriptors, current state, principal-component projections of history and the Full/Direct encoders are controls. Full has more parameters than Direct in this setting; TCN provides an additional architectural comparison.

4 Results

4.1 Predictive utility: Full versus Direct

We first ask whether forecasting through an explicit descriptor improves prediction. The primary comparison is between WB2 Full and Direct-matched, whose trainable parameter counts differ by less than 2.3%.

Table 2: WB2 +72 h capacity and pooled latitude-weighted RMSE (median $[Q_1, Q_3]$, 20 seeds). GMAC estimates billions of multiply–accumulate operations per forward pass; Train h is mean training time per run on one RTX 6000.

Method	Params	GMAC	Train h	T_{2m} RMSE	Z500 RMSE
Direct-matched	371,072	0.057	0.07	3.301 [3.297, 3.304]	987.7 [986.6, 988.1]
NoBank	400,354	0.167	0.02	3.269 [3.260, 3.303]	891.6 [884.1, 936.5]
Full	379,714	0.167	0.05	3.290 [3.286, 3.295]	942.8 [935.6, 946.6]

Mean RMSE is 3.290 for temperature and 941.8 for Z500 with Full, compared with 3.300 and 987.5 for Direct-matched. The paired Direct-matched-minus-Full benefits are 0.010 for temperature (95% t-CI $[0.007, 0.013]$, Full wins $18/20$ seeds) and 45.6 for Z500 (CI $[42.1, 49.1]$, wins $20/20$). Both pass Holm-adjusted paired t-tests ($p_H = 1.72 \times 10^{-5}$ and 4.14×10^{-16}) and the Wilcoxon and sign-test checks. The added 15-seed cohort also favors Full on both targets. Across all 20 leave-one-seed-out summaries, mean benefits remain positive: $[0.009, 0.011]$ for temperature and $[45.1, 46.4]$ for Z500.

Forecasting across systems. In the Burgers trajectory-splice stress test, mean post-splice RMSE is 0.382 for Full and 0.424 for Direct. The paired improvement is 0.042, with trajectory-bootstrap 95% CI $[0.008, 0.080]$; peak-error and area-under-error improvements have the same direction (Appendix F). The interval reflects trajectory variation for this trained model pair. For Lorenz96 with $F = 8, N = 16$, Full improves over the original Direct model at rollout horizons 1, 4, and 16, but the separately trained TCN is substantially stronger at every horizon (Appendix G).

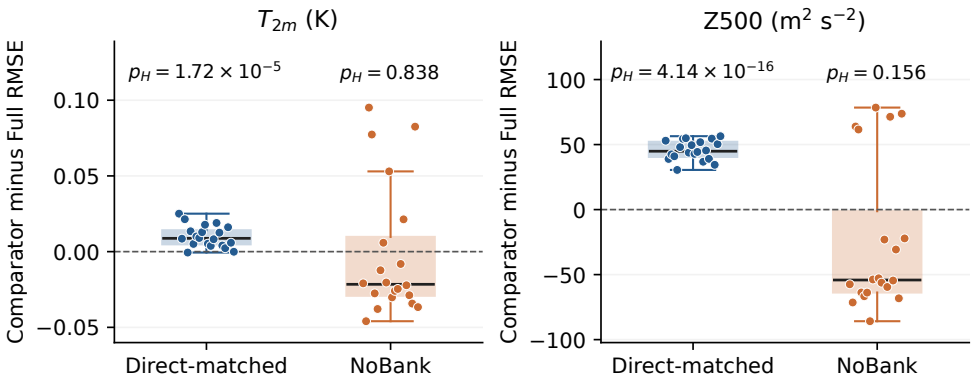


Figure 2: WB2 +72 h paired pooled latitude-weighted RMSE differences across 20 seeds; positive favors Full. Boxes span Q_1–Q_3, center lines mark medians, and whiskers extend to observations within 1.5 IQR. Dots show every seed, including observations beyond the whiskers. Labels give Holm-adjusted paired t-test p_H values. Direct-matched tests the complete architecture; NoBank tests descriptor construction.

4.2 Contribution of anchoring: Full versus NoBank

Full and NoBank share the descriptor-conditioned forecast path; their comparison evaluates the contribution of finite-support construction.

With the full training set, the paired NoBank-minus-Full differences are -0.002 for temperature (95% t-CI $[-0.023, 0.018]$) and -24.0 for Z500 (CI $[-51.1, 3.0]$). Neither mean difference is statistically significant after Holm correction. NoBank has lower Z500 RMSE in 15/20 seeds, but also substantially greater dispersion: its seed SD is 57.6 versus 7.4 for Full, with a paired-bootstrap SD ratio of 7.75 (95% CI $[4.58, 11.61]$). The full-data benefit of anchoring is therefore lower Z500 variability relative to NoBank.

Varying the training-set size reveals a target-dependent pattern (Figure 3). With 20% of the training data, NoBank-minus-Full is $+0.117$ for temperature (CI $[+0.051, +0.183]$, Full wins 5/5) but -37.6 for Z500 (CI $[-67.5, -7.7]$, Full wins 0/5). At 50%, the corresponding differences are $+0.099$ ($[+0.004, +0.194]$, 5/5) and $+19.1$ ($[-19.4, +57.7]$, 3/5). Thus finite support improves temperature forecasts at both smaller training sizes, while NoBank performs better on Z500 at 20%. The benefit of finite support depends jointly on the target and training-set size.

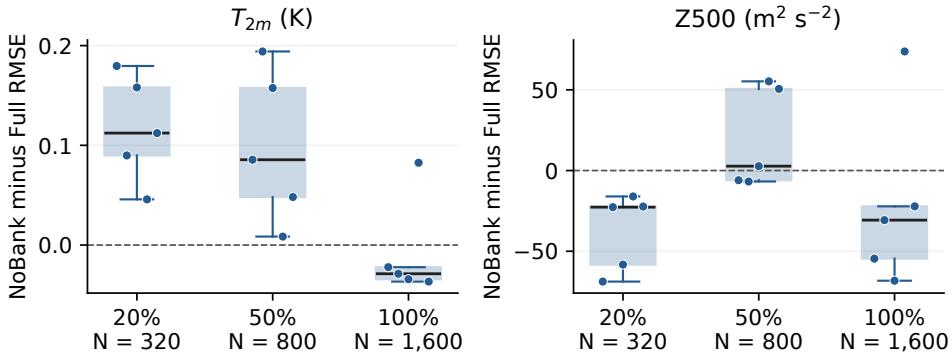


Figure 3: Paired NoBank-minus-Full pooled latitude-weighted RMSE under nested WB2 training subsets. Positive values favor Full. Boxes and whiskers follow Figure 2; dots show the same five original seeds at every fraction, including 100%.

4.3 DYNAMICAL MEANING: ALIGNMENT AND FUNCTIONAL USE

Mechanism learning also requires examining what the intermediate representation captures. We first test alignment with independently computed dynamics, then intervene on the descriptor to assess its role in prediction. Burgers provides the most direct evidence because its PDE quantities are available.

Burgers ground-truth alignment. Probes of the frozen descriptor measure which aspects of the dynamics are linearly accessible (Figure 4). The retrieved descriptor predicts the advection–diffusion balance ρ with test $R^2 = 0.151$ and $\log_{10}\nu$ with $R^2 = 0.189$, versus shuffled-descriptor scores of -0.008 and -0.001. Its probe scores are low for A, D, and $A+D$ ($R^2 \leq 0.007$), for which raw history or encoder controls are stronger. Prototype profiles provide complementary evidence of alignment: after removing viscosity/time means, their explained-variance scores are $\eta^2 = 0.228$ for ρ and 0.469 for $\log|A + D|$. Both give $p = 0.0005$ under an initial-condition trajectory-block null. The raw $\log_{10}\nu$ profile is explained by the preserved viscosity strata.

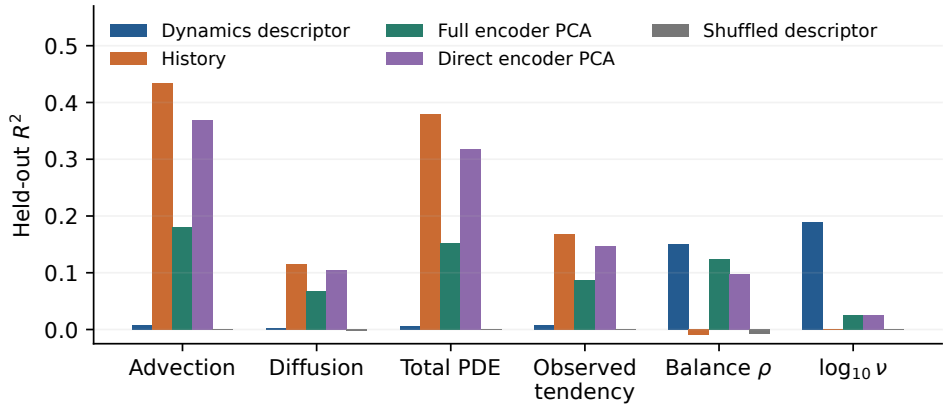
For the matched-history test, we pair samples at the same initial-condition index, time, and spatial location across viscosities. Scalers, history-distance thresholds (calipers, 10–50% training quantiles), and dynamics cut points are set using training trajectories only. At the primary 25% caliper, 7,046 held-out pairs give a Spearman correlation of 0.510 (trajectory CI $[0.394, 0.605]$) between descriptor distance and PDE-based dynamics distance; the partial rank correlation controlling residual history distance is 0.320 ($[0.221, 0.405]$). Using centered normalized ranks within viscosity-pair/time strata, the history-controlled partial correlation remains 0.210 ($[0.142, 0.272]$); all three block-permutation tests give $p = 0.0005$. The association remains positive with intervals excluding zero at each of the five reported calipers.

Forecast dependence and directional transport. On 8,192 sampled test rows, holding history fixed and shuffling descriptors increases RMSE from 0.0191 to 0.0285, a paired change of $+0.0094$ with trajectory–viscosity cluster CI $[+0.0057, +0.0142]$; replacement with a distant descriptor from the test set increases RMSE by 0.0589 ($[0.0444, 0.0751]$), whereas the interval for the nearest alternative includes zero. We then swap descriptors between similar-history pairs with the same initial-condition index and combine the two swap directions symmetrically. For the 2,026 pairs with large dynamics differences at the primary caliper, the swap effect shows no clear association with the observed increment contrast (Spearman -0.029, CI $[-0.133, 0.091]$; sign agreement 0.486, $[0.437, 0.547]$); the PDE-increment and future-state-contrast correlations also have intervals containing zero. This pattern holds at all five calipers. These results connect descriptor alignment with forecast dependence; directional transport remains unresolved.

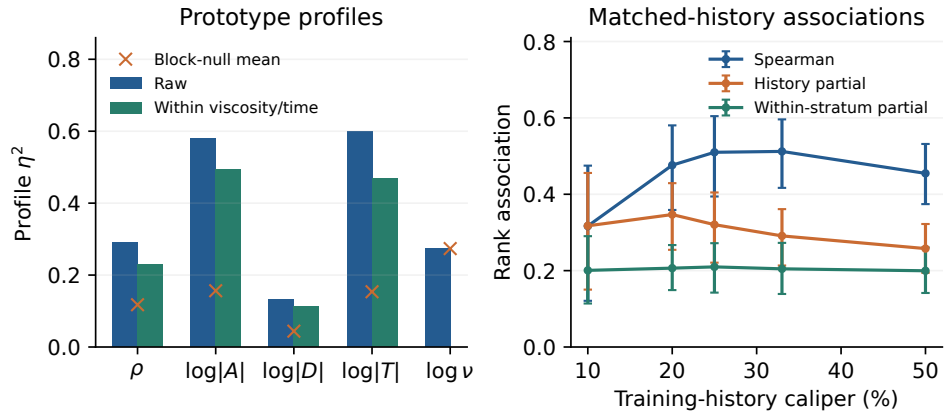
Lorenz96: learned support. Across five frozen Full models, the dynamics descriptor yields mean dynamics-probe $R^2 = 0.106$–0.157 versus shuffled-descriptor scores at or below -0.001. The descriptor carries linearly accessible dynamics information, with lower probe accuracy than the state, history, and encoder controls.

WB2: observed increments and forecast sensitivity. The bank and retrieval rule are fixed across seeds. Descriptors vary across histories and use an effective 149.7 of 256 prototypes (Appendix B). Yet their mean residual-probe test R^2 values are -0.112 for temperature and -0.109 for Z500, both below zero and the shuffled control (Figure 5). A within-batch cyclic shift changes forecasts but leaves aggregate error nearly unchanged: RMSE changes by $+0.000$ and $+0.4$, respectively. Replacing descriptors with the farthest observed prototype changes RMSE by $+3.762$ and $+1355.4$. These diagnostics use pooled unweighted grid RMSE and show that decoder sensitivity alone does not explain the forecasting benefit.

Appendix D visualizes the stored increment patterns, their retrieved test cases, and forecasts selected by fixed quantiles of per-sample forecast benefit. These examples illustrate the spatial patterns stored and retrieved by the bank.



(a) Frozen-descriptor probes against PDE quantities.



(b) Prototype profiles and matched-history robustness.

Figure 4: **Burgers dynamics alignment.** Top: held-out probes show the strongest descriptor scores for balance and viscosity. Bottom: prototype profiles and matched-history associations retain structure under viscosity/time and history controls. The residualized viscosity bar is omitted because its target variance is zero.

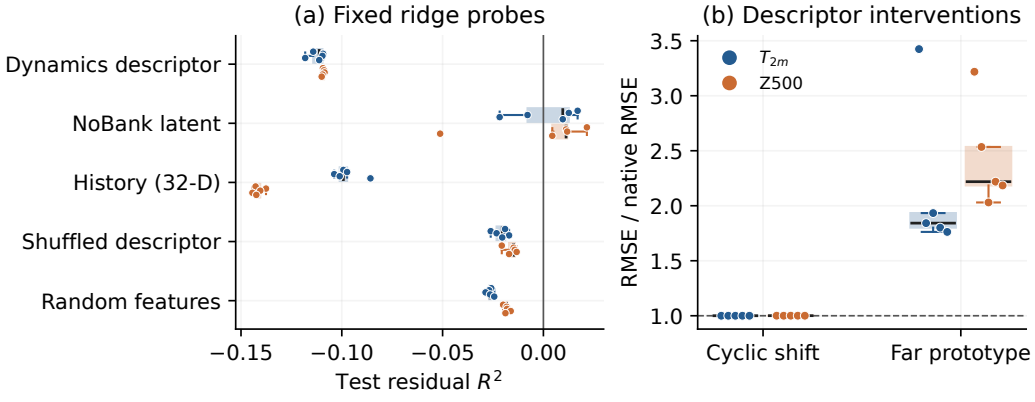


Figure 5: **WB2 descriptor diagnostics.** Left: ridge-probe test R^2. Right: pooled unweighted grid RMSE divided by the native RMSE of the same seed; the dashed line marks a ratio of one. Boxes show quartiles and medians, whiskers use 1.5 IQR, and dots show all five original seeds. Cyclic shifts have little effect on aggregate error, whereas far-prototype replacements increase error.

5 Discussion and limitations

Mechanism learning gives local evolution an explicit representation that can be assessed for predictive utility, anchoring effects, and dynamical meaning. WB2 establishes a Full–Direct-matched forecasting benefit, while Full–NoBank shows that anchoring depends on target and sample size: it reduces full-data Z500 seed variability and improves low-data temperature forecasts. Burgers connects descriptors to PDE quantities and forecast dependence, although matched swaps do not establish directional transport. Together, these results motivate evaluating both what a descriptor represents and how the forecaster uses it.

Limitations and next steps. The strongest capacity-controlled forecast comparison is a two-variable, coarse-grid WB2 task with 20 shared seeds and fixed $+72\,\mathrm{h}$ supervision; low-data and functional diagnostics use five seeds. NoBank shares the forecast path but has 5.44% more parameters, so its comparison also changes generator capacity. Burgers uses one frozen Full/Direct pair and does not include a NoBank comparison; its splice combines constant-viscosity trajectories rather than integrating a physical parameter change. Lorenz96 is not capacity matched, and TCN performs better than Full. Extending the auxiliary studies across seeds, matching generator budgets, and integrating physical regime changes would broaden these tests.

Conclusion. Mechanism learning organizes forecasting around an explicit description of local evolution. Across three systems, prototype-anchored descriptors show predictive utility, regime-dependent anchoring effects, and dynamics alignment with a measurable functional role in Burgers. The resulting framework connects forecast design with direct examination of the representation of local evolution.

AI use statement

Generative AI tools were used to polish the language and improve the readability of this manuscript. The authors reviewed the revised text and take full responsibility for the final content.

Ethics statement

This work uses public scientific benchmark data and simulated dynamical systems; it involves no human subjects or personal data. Applications outside the evaluated regimes could produce unreliable forecasts. We report uncertainty and functional controls to make the empirical scope of the models clear.

Reproducibility statement

Training and evaluation code, all 60 WB2 headline checkpoints, 20 WB2 low-data checkpoints, Burgers Full/Direct weights with the complete prototype memory, and 15 Lorenz96 Full/Direct/LSTM checkpoints are available in the anonymous repository, together with results, provenance, and a file inventory. The appendix specifies protocols and statistics; the repository README documents runtime and public-data dependencies.

REFERENCES

Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, 619:533–538, 2023. doi: 10.1038/s41586-023-06185-3.

Steven L. Brunton, Joshua L. Proctor, and J. Nathan Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 113(15):3932–3937, 2016. doi: 10.1073/pnas.1517384113.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K. Duvenaud. Neural ordinary differential equations. *Advances in Neural Information Processing Systems*, 31:6571–6583, 2018. URL https://papers.nips.cc/paper/2018/hash/69386f6bb1dfed68692a24c8686939b9-Abstract.html.

Dmitrii Kochkov, Janni Yuval, Ian Langmore, Peter Norgaard, Jamie Smith, Griffin Mooers, Milan Klöwer, James Lottes, Stephan Rasp, Peter Düben, Sam Hatfield, Peter Battaglia, Alvaro Sanchez-Gonzalez, Matthew Willson, Michael P. Brenner, and Stephan Hoyer. Neural general circulation models for weather and climate. *Nature*, 632:1060–1066, 2024. doi: 10.1038/s41586-024-07744-y. URL https://www.nature.com/articles/s41586-024-07744-y.

Milan Korda and Igor Mezić. Linear predictors for nonlinear dynamical systems: Koopman operator meets model predictive control. *Automatica*, 93:149–160, 2018. doi: 10.1016/j.automatica.2018.03.046.

Rahul G. Krishnan, Uri Shalit, and David Sontag. Deep kalman filters. *arXiv preprint arXiv:1511.05121*, 2015. URL https://arxiv.org/abs/1511.05121.

Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, Alexander Merose, Stephan Hoyer, George Holland, Oriol Vinyals, Jacklynn Stott, Alexander Pritzel, Shakir Mohamed, and Peter Battaglia. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421, 2023. doi: 10.1126/science.adi2336.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=c8P9NQVtmnO.

Edward N. Lorenz. Predictability: A problem partly solved. In *Seminar on Predictability, 4–8 September 1995*, pp. 1–18, Reading, UK, 1995. ECMWF. URL https://www.ecmwf.int/en/elibrary/75462-predictability-problem-partly-solved.

Lu Lu, Pengzhan Jin, Guofei Pang, Zhongqiang Zhang, and George Em Karniadakis. Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators. *Nature Machine Intelligence*, 3:218–229, 2021. doi: 10.1038/s42256-021-00302-5.

Bethany Lusch, J. Nathan Kutz, and Steven L. Brunton. Deep learning for universal linear embeddings of nonlinear dynamics. *Nature Communications*, 9:4950, 2018. doi: 10.1038/s41467-018-07210-0. URL https://www.nature.com/articles/s41467-018-07210-0.

Michael McCabe, Peter Harrington, Shashank Subramanian, and Jed Brown. Towards stability of autoregressive neural operators. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=RFfUUtKYOG.

Jaideep Pathak, Brian Hunt, Michelle Girvan, Zhixin Lu, and Edward Ott. Model-free prediction of large spatiotemporally chaotic systems from data: A reservoir computing approach. *Physical Review Letters*, 120:024102, 2018. doi: 10.1103/PhysRevLett.120.024102.

Jaideep Pathak, Shashank Subramanian, Peter Harrington, Sanjeev Raja, Ashesh Chattopadhyay, Morteza Mardani, Thorsten Kurth, David Hall, Zongyi Li, Kamyar Azizzadenesheli, Pedram Hassanzadeh, Karthik Kashinath, and Anima Anandkumar. FourCastNet: A global data-driven high-resolution weather model using adaptive fourier neural operators. *arXiv preprint arXiv:2202.11214*, 2022. URL https://arxiv.org/abs/2202.11214.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. doi: 10.1609/aaai.v32i1.11671. URL https://ojs.aaai.org/index.php/AAAI/article/view/11671.

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, Aaron 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. doi: 10.1029/2023MS004019. URL https://doi.org/10.1029/2023MS004019.

Yulia Rubanova, Ricky T. Q. Chen, and David K. Duvenaud. Latent ordinary differential equations for irregularly-sampled time series. *Advances in Neural Information Processing Systems*, 32, 2019. URL https://papers.nips.cc/paper_files/paper/2019/hash/42a6845a557bef704ad8ac9cb4461d43-Abstract.html.

Peter J. Schmid. Dynamic mode decomposition of numerical and experimental data. *Journal of Fluid Mechanics*, 656:5–28, 2010. doi: 10.1017/S0022112010001217.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://papers.nips.cc/paper_files/paper/2017/hash/cb8da6767461f2812ae4290eac7cbc42-Abstract.html.

Makoto Takamoto, Timothy Praditia, Raphael Leiteritz, Daniel MacKinlay, Francesco Alesiani, Dirk Pflüger, and Mathias Niepert. PDEBench: An extensive benchmark for scientific machine learning. *Advances in Neural Information Processing Systems*, 35: 1596–1611, 2022. URL https://papers.nips.cc/paper_files/paper/2022/hash/0a9747136d411fb83f0cf81820d44afb-Abstract-Datasets_and_Benchmarks.html.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems*, volume 29, 2016. URL https://papers.nips.cc/paper_files/paper/2016/hash/90e1357833654983612fb05e3ec9148c-Abstract.html.

A PROTOCOL CHECKS AND PROVENANCE

The release headline manifest contains all 60 pooled-metric rows (three methods by 20 shared seeds), linked to checkpoint, configuration, result, and source hashes. Historical evaluation columns are named separately because they use mean per-sample RMSE. The original five seeds are retained, and seeds 0–14 extend the same protocol. The extension declaration, per-run logs, cohort summaries, and release hashes accompany the data.

Every headline run uses the verified 2,907-sample $+72$ h test set and validation-only checkpoint selection. Full initializes a fixed bank of paired observed training fragments, saves their source indices, and retrieves with standardized-history cosine top-8 routing at temperature 0.1. Historical horizon-mismatched, uninitialized-bank, and collapsed-retrieval runs are excluded from the headline analysis.

Full and NoBank share the spatial encoder, decoder, FiLM, and output layouts. Full retrieves from fixed pooled-history features, whereas NoBank generates its descriptor from a learned global encoder feature. Their trainable parameter counts are 379,714 and 400,354, respectively (a 5.44% NoBank excess), so this comparison changes both descriptor construction and generator capacity.

B WEATHERBENCH 2 COMPLETE STATISTICS

Training and checkpoint selection. All three models use AdamW with a learning rate of 10^{-4}, batch size 8, a maximum of 100 epochs, and validation-only early stopping. The checkpoint score averages the two validation RMSEs after normalizing each by its epoch-1 value; early stopping uses patience 10.

Headline metric and units. For each target variable, let $e_{n,a,b}$ be the forecast error for test case n, latitude a, and longitude b, and let $w_a = \cos(\varphi_a)$. We report

$$\mathrm{RMSE}_{\mathrm{lat}} = \left(\frac{\sum_{n,a,b} w_a e_{n,a,b}^2}{N_{\mathrm{test}} N_{\mathrm{lon}} \sum_a w_a} \right)^{1/2}.$$

The square root is taken after pooling over test cases and grid cells. Temperature scores are in K; Z500 scores are in $\mathrm{m}^2\,\mathrm{s}^{-2}$ (geopotential, not geopotential height). A paired benefit is comparator RMSE minus Full RMSE. Intervention tables use the analogous pooled metric with $w_a = 1$ and are labeled separately.

Table 3: WB2 pooled latitude-weighted RMSE across 20 seeds. Quartiles describe the seed distribution; confidence intervals describe the mean.

Method	Target	Mean	95% t-CI	Median	$[Q_1, Q_3]$
Direct-matched	T_{2m}	3.300	[3.298, 3.302]	3.301	[3.297, 3.304]
Full	T_{2m}	3.290	[3.288, 3.293]	3.290	[3.286, 3.295]
NoBank	T_{2m}	3.288	[3.268, 3.308]	3.269	[3.260, 3.303]
Direct-matched	Z500	987.5	[986.9, 988.0]	987.7	[986.6, 988.1]
Full	Z500	941.8	[938.4, 945.3]	942.8	[935.6, 946.6]
NoBank	Z500	917.8	[890.8, 944.8]	891.6	[884.1, 936.5]

Table 4: Paired WB2 effects over 20 shared seeds. Positive comparator-minus-Full differences favor Full. Intervals are unadjusted 95% intervals.

Comparison	Target	Mean benefit	t-CI	Bootstrap CI	Wins
Direct-matched - Full	T_{2m}	0.010	[0.007, 0.013]	[0.007, 0.013]	18/20
Direct-matched - Full	Z500	45.6	[42.1, 49.1]	[42.4, 48.8]	20/20
NoBank - Full	T_{2m}	-0.002	[-0.023, 0.018]	[-0.020, 0.018]	6/20
NoBank - Full	Z500	-24.0	[-51.1, 3.0]	[-47.3, 1.4]	5/20

Table 5: Two-sided paired tests for the four WB2 method-by-target contrasts. Holm correction is applied separately within each test family.

Comparison	Target	Raw t p	Holm t p	Holm Wilcoxon p	Holm sign p
Direct-matched - Full	T_{2m}	5.73×10^{-6}	1.72×10^{-5}	2.86×10^{-5}	0.001
Direct-matched - Full	Z500	1.03×10^{-16}	4.14×10^{-16}	7.63×10^{-6}	7.63×10^{-6}
NoBank - Full	T_{2m}	0.838	0.838	0.589	0.115
NoBank - Full	Z500	0.078	0.156	0.589	0.083

Table 6: Cohort sensitivity of the paired WB2 mean benefit (95% t-CI). All 20 seeds enter the main analysis.

Cohort	Comparison	Target	Mean benefit [CI]	Wins
Original 5	Direct-matched - Full	T_{2m}	0.006 [-0.002, 0.013]	4/5
Original 5	Direct-matched - Full	Z500	47.0 [34.9, 59.0]	5/5
Original 5	NoBank - Full	T_{2m}	-0.008 [-0.071, 0.055]	1/5
Original 5	NoBank - Full	Z500	-20.4 [-89.6, 48.8]	1/5
Added 15	Direct-matched - Full	T_{2m}	0.011 [0.008, 0.015]	14/15
Added 15	Direct-matched - Full	Z500	45.2 [41.4, 49.0]	15/15
Added 15	NoBank - Full	T_{2m}	0.000 [-0.024, 0.024]	5/15
Added 15	NoBank - Full	Z500	-25.3 [-58.6, 8.1]	4/15
All 20	Direct-matched - Full	T_{2m}	0.010 [0.007, 0.013]	18/20
All 20	Direct-matched - Full	Z500	45.6 [42.1, 49.1]	20/20
All 20	NoBank - Full	T_{2m}	-0.002 [-0.023, 0.018]	6/20
All 20	NoBank - Full	Z500	-24.0 [-51.1, 3.0]	5/20

Table 7: Leave-one-seed-out mean benefit for Direct-matched minus Full. Each row retains the other 19 paired seeds.

Dropped seed	T_{2m}	Z500
42	0.010	46.0
137	0.010	45.2
256	0.010	46.2
512	0.010	45.4
1024	0.011	45.1
0	0.009	45.2
1	0.010	46.0
2	0.009	45.8
3	0.011	45.9
4	0.010	46.4
5	0.010	45.5
6	0.010	45.2
7	0.010	45.1
8	0.010	45.7
9	0.010	45.4
10	0.010	45.8
11	0.010	45.7
12	0.010	45.3
13	0.010	46.1
14	0.010	45.6

Table 8: Residual-probe test R^2: median $[Q_1, Q_3]$ across the five original seeds.

Representation	T_{2m}	Z500
Dynamics descriptor	-0.111 [-0.114, -0.110]	-0.109 [-0.110, -0.109]
NoBank latent	0.010 [-0.008, 0.013]	0.011 [0.004, 0.012]
History (32-D)	-0.099 [-0.101, -0.097]	-0.142 [-0.143, -0.140]
Shuffled descriptor	-0.020 [-0.023, -0.019]	-0.015 [-0.017, -0.014]
Random features	-0.026 [-0.027, -0.026]	-0.018 [-0.019, -0.018]

Table 9: Descriptor interventions: median $[Q_1, Q_3]$ over five seeds. RMSE pools unweighted grid error; forecast RMS measures change from the native forecast.

Target	Descriptor	RMSE	Forecast RMS change
T_{2m}	Native	3.264 [3.257, 3.267]	0
T_{2m}	Cyclic shift	3.264 [3.257, 3.267]	0.465 [0.436, 0.478]
T_{2m}	Far prototype	6.017 [5.866, 6.290]	5.154 [4.963, 5.493]
Z500	Native	939.9 [935.3, 951.0]	0
Z500	Cyclic shift	940.4 [935.5, 951.5]	330.2 [292.2, 331.9]
Z500	Far prototype	2085.5 [2035.2, 2370.3]	1969.6 [1899.5, 2302.4]

Table 10: Descriptor diagnostics over the five original runs. The bank and retrieval rule are fixed; random-pair sampling varies between runs.

Diagnostic	Median	Range
Mean per-dimension SD	0.404	[0.404, 0.404]
Adjacent/random distance ratio	0.686	[0.685, 0.688]
Global effective prototypes (of 256)	149.7	[149.7, 149.7]
Per-sample effective prototypes	7.7	[7.7, 7.7]
Mean maximum retrieval weight	0.188	[0.188, 0.188]

Distributions, intervals, and significance tests. Boxplots use linearly interpolated quartiles, a median center line, and whiskers ending at the most extreme observations within 1.5 interquartile ranges of the box; every seed is shown as a point. These describe the empirical seed distribution. Mean-effect intervals instead use paired differences and Student's t critical value with 19 degrees of freedom. Bootstrap mean intervals resample 20 paired differences 20,000 times using RNG seed 14119. The NoBank/Full Z500 SD-ratio interval resamples 20 aligned seed triplets 20,000 times using RNG seed 31415. Both types of interval are unadjusted 95% intervals.

The four headline contrasts are Direct-matched minus Full and NoBank minus Full, each for temperature and Z500. We use two-sided paired t-tests, exact Wilcoxon signed-rank tests, and exact sign tests, applying Holm correction across the four contrasts separately within each test family. Full versus Direct-matched remains significant for both targets in all three families. For NoBank versus Full, no contrast passes this correction; the raw Z500 sign-test $p = 0.0414$ becomes $p_H = 0.0828$. A nonsignificant contrast does not establish equivalence.

The original-five, added-fifteen, and all-twenty summaries distinguish the extension from the initial cohort. Leave-one-seed-out estimates drop each of the 20 seeds in turn, while the main result uses all seeds. These summaries quantify training-seed variation on the fixed test set. Data-scarcity and functional diagnostics retain the five original seeds.

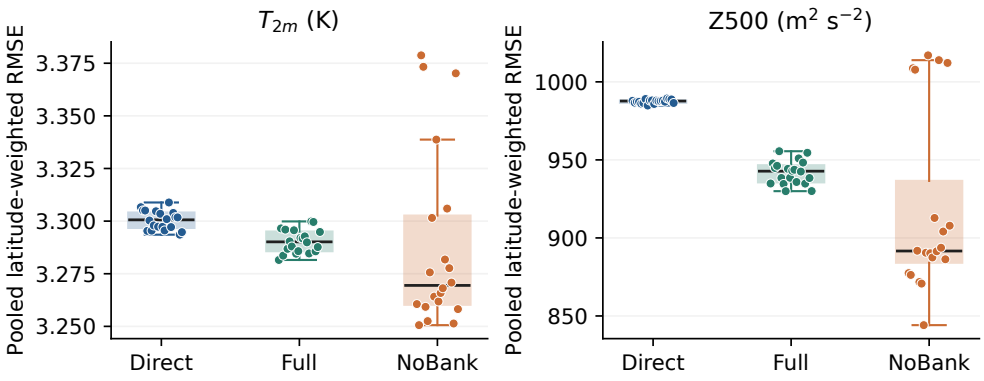


Figure 6: Absolute WB2 pooled latitude-weighted RMSE across 20 seeds. Box and point definitions match Figure 2; Direct denotes Direct-matched. Full is less variable than NoBank, while Direct-matched has the smallest dispersion.

Residual probes. Each probe has 32 input and 32 output dimensions. The Full probe uses the retrieved dynamics descriptor; the NoBank probe uses the learned latent. The history baseline uses 4×4 pooled values from the most recent two-variable field; the shuffled control independently shuffles training and test descriptors; the random control uses independent standard Gaussian features. Features are standardized using training statistics. Multi-output ridge regression with intercept and fixed $\lambda = 1$ is fit on 1,600 training cases and evaluated on all 2,907 test cases. Probe hyperparameters and representations are fixed before test evaluation.

The Full descriptor has negative test R^2 for both variables and scores below the shuffled control. This linear probe does not recover useful residual structure from the descriptor.

Interventions. For each test batch, the encoder features are held fixed while forecasts are recomputed with replacement descriptors. The cyclic-shift control shifts descriptors within the batch. The far-prototype control selects the stored *observed* descriptor farthest in Euclidean distance from the sample's native descriptor. We report pooled unweighted grid RMSE for native and intervened forecasts, plus root-mean-square forecast change. The headline metric additionally applies latitude weights. The far-prototype replacement measures decoder sensitivity to an extreme stored descriptor.

The retrieval descriptor and bank are deterministic functions of the shared training split, so they are identical across training seeds; this accounts for the constant values in the diagnostic table. Random-pair sampling can vary between diagnostic runs. The learned encoder–decoder parameters differ by training seed.

C WeatherBench 2 nested training subsets

The 20%, 50%, and 100% conditions use nested prefixes of the same 1,600-sample training split. They share the architecture, optimizer, schedule, 400 validation samples, 2,907 test samples, +72 h horizon, five paired seeds, and validation-only checkpoint rule. Positive NoBank-minus-Full differences favor Full. The 100% condition here uses the same five seeds for a paired cross-fraction comparison; the separate full-data headline uses all 20 seeds.

Table 11: Paired Full–NoBank comparison across training subsets under pooled latitude-weighted RMSE (mean difference and 95% t-CI over five seeds).

Train fraction (N)	Target	NoBank − Full (95% CI)	Full wins
20% (320)	T_{2m}	$+0.1171$ $[+0.0508, +0.1835]$	5/5
	Z500	-37.6 $[-67.5, -7.7]$	0/5
50% (800)	T_{2m}	$+0.0988$ $[+0.0037, +0.1938]$	5/5
	Z500	$+19.1$ $[-19.4, +57.7]$	3/5
100% (1600)	T_{2m}	-0.0079 $[-0.0710, +0.0553]$	1/5
	Z500	-20.4 $[-89.6, +48.8]$	1/5

For temperature, Full outperforms NoBank in all five seeds at both 20% and 50%; the interval includes zero at 100%. For Z500, NoBank performs better at 20%, while the intervals include zero at 50% and 100%. The difference-in-differences contrast between 20% and 100% is $+0.1250$ for temperature (paired 95% CI $[0.0053, 0.2447]$) and -17.2 for Z500 ($[-69.0, 34.5]$). The effect of reducing the training-set size therefore differs between the two targets.

D PROTOTYPE AND FORECAST EXAMPLES

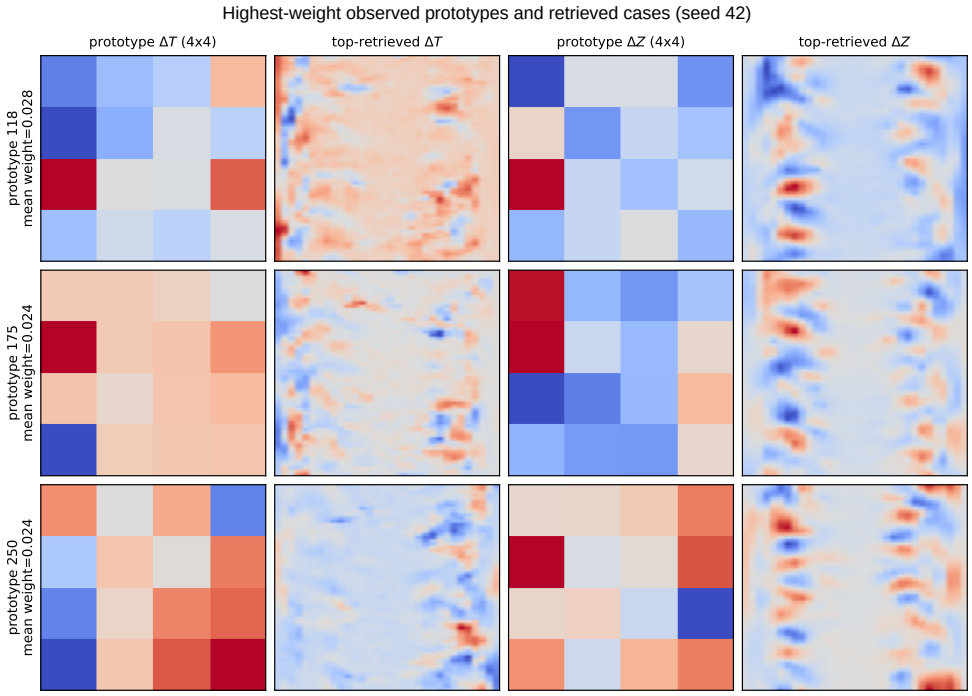


Figure 7: The three bank entries with the highest mean retrieval weight for seed 42. The coarse 4×4 maps are stored observed target-increment descriptors; the neighboring full-resolution maps are the ground-truth increments of the test case with the largest retrieval weight for each prototype. Panels use independent color limits, so their colors compare spatial patterns rather than absolute magnitudes. These examples illustrate spatial correspondence between stored patterns and retrieved cases.

The qualitative forecast examples in Figure 8 are selected by an explicit rule: compute the Direct-matched-minus-Full spatial Z500 RMSE difference for every test case, then select the cases nearest the 25th, 50th, and 75th percentiles of that distribution. We use seed 42, the first seed in the paired list. The panels show lower-, median-, and higher-benefit cases under this rule. All three selected cases have positive benefit.

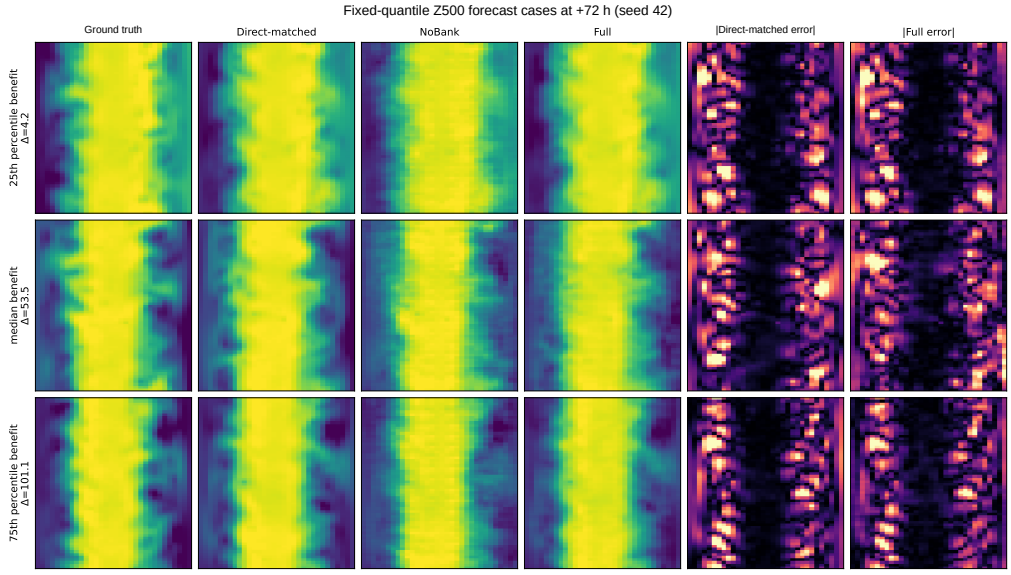


Figure 8: Fixed-quantile +72 h Z500 cases from seed 42. Rows are chosen by fixed quantiles of paired per-sample benefit; columns show the ground truth, three forecasts, and two error maps. Within each row, the ground-truth and forecast fields share color limits set by the ground truth's 2nd and 98th percentiles and the two errors share a zero-based scale; limits differ across rows. Axes index grid cells.

E BURGERS DYNAMICS ALIGNMENT AND INTERVENTIONS

All diagnostics use the same frozen $K = 128$ Full checkpoint. The original split is reproduced with seed 20260228. Ridge probes with fixed penalty ($\lambda = 1$) are fit on 27,648 rows from 288 training trajectory–viscosity realizations and evaluated on 27,648 rows from 288 disjoint in-domain test realizations. Each in-domain viscosity contributes 96 train and 96 test trajectories; the additional $\nu = 0.002$ set is held out as an OOD diagnostic. Following the PDEBench file convention, the evaluated PDE quantities are

$$A = -uu_x, \qquad D = (\nu/\pi)u_{xx}, \qquad T = A + D, \qquad \rho = \frac{|A|}{|A| + |D| + 10^{-12}}.$$

Probe intervals use trajectory–viscosity clusters (2,000 bootstrap draws). Forecasts recomputed from the retrieved coefficients and histories match the native model predictions with a maximum absolute error of 3.55×10^{-7}.

Table 12: Burgers in-domain test R^2 for frozen representations. Principal component analysis (PCA) controls retain 20 components.

Representation	A	D	$A + D$	observed tendency	ρ	$\log_{10} \nu$
Descriptor $\hat{\theta}$	.007	.002	.006	.007	**.151**	**.189**
Shuffled $\hat{\theta}$	−.001	−.002	−.001	−.001	−.008	−.001
Raw history	.433	.115	.381	.168	−.009	.000
Full encoder PCA	.180	.068	.152	.087	.124	.026
Direct encoder PCA	.370	.105	.318	.147	.097	.026

The descriptor probe is most informative about advection–diffusion balance and viscosity. Assigning each sample to its highest-weight prototype uses 95 of 128 bank entries. Raw between-prototype explained variance η^2 is .292 for ρ, .579 for $\log|A|$, .133 for $\log|D|$, .600 for $\log|A + D|$, and .274

for $\log_{10}\nu$. The block-permutation null moves complete initial-condition trajectory blocks while preserving viscosity, time, and spatial positions (2,000 permutations). For the first four targets, raw and viscosity/time-residualized profiles remain significant ($p = .0005$); residual η^2 values are .228, .493, .113, and .469. The raw $\log_{10}\nu$ profile gives $p = 1.0$ under this null. Its residualized profile is undefined because viscosity strata account for all target variance. The other profiles retain within-stratum variation in PDE quantities, beyond differences in viscosity and time.

The training and test trajectory sets are disjoint. History standardization, calipers, and dynamics cut points are fit on training trajectories only; held-out trajectories are used to evaluate the resulting fixed diagnostics. The five reported calipers are set at the 10th, 20th, 25th, 33rd, and 50th training-history-distance percentiles. For the matched-history and prototype-profile tests, bootstrap and permutation move complete initial-condition trajectories jointly across viscosity pairs, times, and spatial samples. The observed tendency reconstructed from the final history value and next target agrees with the independently stored value with a maximum error of 2.84×10^{-9}.

Distances and trajectory units. Matched pairs share the initial-condition index, time, and spatial location across different in-domain viscosities. History distance is Euclidean distance after fitting per-coordinate standardization on training trajectories. Dynamics distance is Euclidean distance in the standardized four-vector $(A, D, A + D, \rho)$; descriptor distance is Euclidean distance between retrieved coefficient vectors. The dynamics vector is a joint diagnostic with correlated components. The 288 trajectory–viscosity realizations correspond to 96 initial-condition indices across three viscosities. Resampling for the matched-history and prototype-profile tests keeps these initial-condition blocks together across viscosities. The overall partial Spearman statistic residualizes ranks with respect to history-distance rank. The within-stratum variant ranks each distance separately inside each viscosity-pair/time stratum, divides by stratum size, centers these normalized ranks, and residualizes descriptor and dynamics ranks against history rank before computing their pooled Pearson correlation. Profile logarithms use $\log_{10}(|\cdot| + 10^{-8})$. Reported permutation $p = 0.0005$ is the rounded minimum attainable with 2,000 permutations and the add-one correction.

At the primary 25% caliper, 7,046 held-out pairs give a Spearman correlation of .510 (trajectory-bootstrap CI $[.394, .605]$), partial Spearman .320 after controlling residual history distance ($[.221, .405]$), and within-stratum partial rank correlation .210 ($[.142, .272]$); all corresponding block-permutation values are $p = .0005$. Table 13 shows stability across all calipers. The difference in mean descriptor distance between the highest and lowest dynamics-distance strata is .206 ($[.162, .253]$), but its stricter block-permutation value is $p = .702$ because the null preserves viscosity-pair/time structure; we therefore base the alignment conclusion on the rank associations, which remain significant under this null.

Table 13: Held-out matched-history robustness. Calipers and cut points are training-derived; intervals cluster-bootstrap complete initial-condition trajectories.

History caliper	Test pairs	Spearman (95% CI)	Partial (95% CI)	Within-stratum partial (95% CI)
10%	2,498	.317 [.121,.475]	.317 [.150,.456]	.201 [.114,.290]
20%	5,338	.476 [.359,.580]	.347 [.255,.429]	.207 [.149,.267]
25%	7,046	.510 [.394,.605]	.320 [.221,.405]	.210 [.142,.272]
33%	9,380	.512 [.417,.596]	.291 [.213,.361]	.205 [.139,.273]
50%	14,116	.455 [.374,.532]	.258 [.194,.322]	.200 [.142,.261]

Table 14: Burgers descriptor-only sensitivity with frozen history features. CIs resample trajectory–viscosity clusters (288 groups) on 8,192 sampled rows.

Descriptor	RMSE	Forecast RMS change	ΔRMSE (95% CI)
Native	.019107	.000000	.000000
Shuffled	.028514	.022170	$+.009407$ $[+.005672, +.014212]$
Mean	.020699	.011237	$+.001592$ $[-.000075, +.003640]$
Zero	.554780	.555291	$+.535673$ $[+.497135, +.585451]$
Nearest alternative	.019107	.001574	$+.000000$ $[-.000137, +.000124]$
Distant test descriptor	.078052	.075984	$+.058945$ $[+.044371, +.075070]$

The matched-swap test exchanges retrieved coefficient vectors between histories with the same initial-condition index that satisfy a training-derived distance caliper. It holds each receiver's history fixed and combines the two swap directions symmetrically. Among the 2,026 matched pairs above the training-derived upper-quartile cutoff for dynamics distance at the primary caliper, the Spearman correlation with the observed one-step increment contrast is $-.029$ (cluster CI $[-.133, .091]$), sign agreement is .486 ($[.437, .547]$), and the Spearman correlation with the independently computed PDE increment is $-.038$ ($[-.144, .083]$). The future-state-contrast correlation is .026 ($[-.071, .137]$). The correlation intervals include zero at every reported caliper. Figure 9 reports the association and transport results across all five calipers. The probes and interventions support dynamics alignment and forecast dependence, while matched swaps provide no clear evidence of directional transport. The Burgers experiments do not include a NoBank comparison.

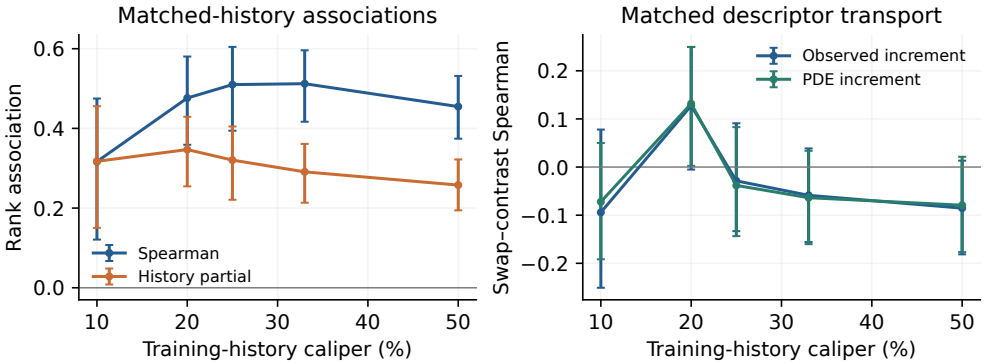


Figure 9: Matched-history robustness and matched descriptor transport across five training-derived calipers. Bars are trajectory-bootstrap 95% intervals. Association is stable; directional transport remains unresolved.

F BURGERS TRAJECTORY-SPLICE STRESS TEST

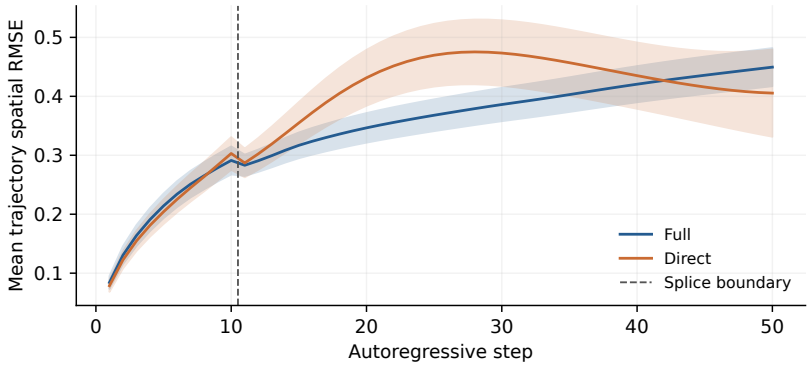


Figure 10: Counterfactual Burgers trajectory-splice stress test. Lines are means over 128 held-out trajectory indices and bands are 95% normal intervals for the trajectory mean. The vertical line marks the splice boundary.

Table 15: Burgers trajectory-splice metrics. Paired benefits are Direct minus Full; intervals use 10,000 trajectory-bootstrap resamples.

Metric	Full	Direct	Paired benefit (95% CI)
Mean post-splice RMSE	0.382	0.424	0.042 [0.008, 0.080]
Peak post-splice RMSE	0.458	0.538	0.080 [0.023, 0.155]
Post-splice AUC	14.922	16.631	1.709 [0.401, 3.183]

Full has 207,399 trainable parameters and Direct has 268,801 (29.6% more). The archived code label `Direct-matched` is retained in the raw records for traceability; the manuscript calls this model Direct.

At each rollout step, spatial RMSE is computed for each trajectory. The mean post-splice score averages these RMSEs over steps 11–50 and then over trajectories. AUC uses the trapezoidal rule with unit step spacing.

The raw PDEBench files contain the same initial condition at the same sample index across viscosity values. We choose 128 indices without replacement with seed 14119, use 20 low-viscosity history steps and 10 low-viscosity rollout targets, then splice the same-index high-viscosity trajectory at the corresponding absolute time for the remaining 40 targets. The two constant-viscosity trajectories have evolved differently by the splice time, so the constructed target can be discontinuous. A physically continuous parameter-change test would require integrating the PDE from the boundary state after changing viscosity.

G LORENZ96 FORECASTS AND DYNAMICS PROBES

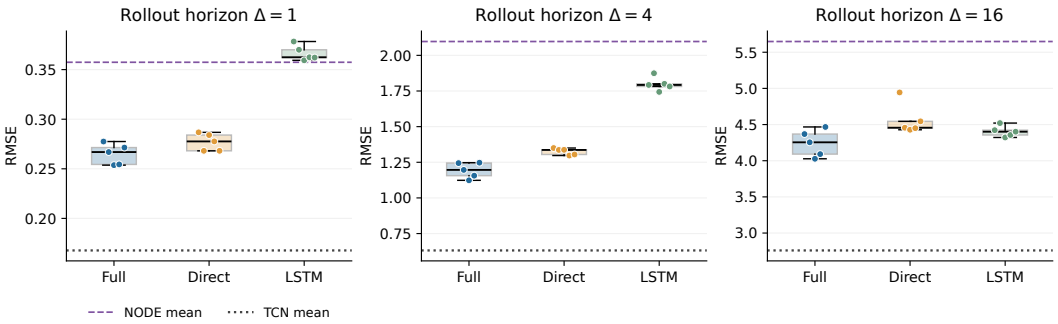


Figure 11: Lorenz96 ($F = 8, N = 16$) autoregressive RMSE. Full, Direct, and LSTM boxes show quartiles and medians across five seeds controlling generated trajectories and training initialization; whiskers use 1.5 IQR, and dots show all seeds. Dashed and dotted lines mark archived NODE and TCN means, respectively. Lower is better.

Full has lower mean RMSE than Direct at all three reported rollout horizons, while TCN has the lowest reported means (Figure 11). Full has more parameters than Direct; this comparison evaluates the two architectures with their respective parameter budgets. NODE means at horizons 1, 4, and 16 are 0.3574, 2.0971, and 5.6482; TCN means are 0.1677, 0.6316, and 2.7596. Both use the same data generator and rollout evaluator.

Frozen dynamics probes. For every Full seed, ridge probes ($\lambda = 1$) are fit on the generated training trajectory and evaluated on its independent test trajectory. Five-seed mean R^2 for the dynamics descriptor is .111 for the nonlinear term, .157 for damping, .106 for total RHS, and .108 for observed tendency; shuffled-descriptor means are $-.001$, $-.005$, $-.001$, and $-.001$. The descriptor has higher mean probe R^2 than the shuffled control for all four targets. For comparison, current-state, history-PCA, Full-encoder, and Direct-encoder scores are respectively (.305, .290, .352, .368) for the nonlinear term and (1.000, .905, .979, .970) for damping. The Lorenz96 experiments do not

include a NoBank comparison. Highest-weight assignments use 5–7 of the 256 prototypes across seeds, indicating concentrated use of the learned bank.

H BURGERS SPARSE SUPPORT AND SYSTEM DETAILS

The selected Burgers model combines coefficient-space and latent-space mixing. With prototype coefficient vectors p_j, trainable maps g and h, and a learned entropy gate γ_t, it computes

$$\theta_S = \sum_j \alpha_{t,j} p_j, \qquad \theta_C = h\left(\sum_j \alpha_{t,j} g(p_j)\right), \qquad \hat{\theta}_t = \theta_S + \gamma_t(\theta_C - \theta_S).$$

Here $\gamma_t = \sigma(a(\mathcal{H}(\alpha_t) - b))$, with learned a, b and attention entropy $\mathcal{H}$. The first 20 descriptor components multiply the observed history; the last is a bias. The original memory arrays, configuration, and selected weights are distributed together. This instantiates the aggregation map in Eq. (1).

In Burgers, a local ridge predictor supplies a $d_m = h + 1$ coefficient descriptor. The archived prototype scan reports values labeled near-split RMSE@20 of $(0.347, 0.348, 0.331, 0.395)$ at $K = (32, 64, 128, 2048)$, respectively. The main findings use the frozen $K = 128$ model; the archived scan provides context for this configuration.

Lorenz96 uses history length four and a trainable finite bank. WB2 uses observed-fragment support as detailed in Section 2.2. The shared formulation is the retrieval-conditioned forecast in Eq. (1); the semantics and construction of a descriptor remain system-dependent.

I REPRODUCIBILITY MANIFEST

The anonymous research repository contains:

- the corrected WB2 model and explicit observed-fragment initialization;
- launchers and checkpoints for all 20 paired Direct-matched, Full, and NoBank headline seeds, plus the five-seed low-data runs;
- raw seed-level CSV files, WB2 checkpoints, logs, and runtime files for corrected runs;
- analysis entry points that regenerate aggregate, paired, bootstrap, LOO, data-scarcity, probe, intervention, complexity, and figure artifacts;
- the Burgers Full/Direct checkpoints, complete $K = 128$ memory bank, dynamics and trajectory-splice evaluators, five-caliper matched-history robustness, block-permutation prototype-profile tests, row-level matched-swap results, selected trajectory indices, and cluster-bootstrap/permutation summaries;
- the Lorenz96 generator and training entry, all 15 Full/Direct/LSTM checkpoints at $F = 8, N = 16$, and five-seed dynamics probes;
- machine-readable headline-provenance manifests, including checkpoint, configuration, result, and source hashes. The original provenance hashes identify the experimental versions; a separate release manifest hashes the files actually distributed.

The corrected WB2 reruns used two NVIDIA Quadro RTX 6000 GPUs (24 GB each), at most one training process per GPU, and four data-loading workers per run. Runtime, epochs, checkpoints, and full training logs are recorded for every fresh Direct-matched, Full, and NoBank run. The analysis records parameter and forward-GMAC estimates separately from wall time.

J DATASETS AND SOFTWARE DEPENDENCIES

WeatherBench 2 and PDEBench are public research datasets (Rasp et al., 2024; Takamoto et al., 2022); their raw files are external dependencies. Lorenz96 trajectories follow the published dynamical system (Lorenz, 1995).

The repository inventory lists the released code, checkpoints, figures, and result tables. Evaluation entry points specify model checkpoints, configurations, and data inputs; WB2 evaluation uses prepared caches and split-index files. Runtime requirements and setup instructions are documented in the repository README, and the software license accompanies the repository.