Hénon–Heiles Escape Prediction from Noisy Position Histories

George Weale
University of California, Santa Barbara (2024)
gweale@ucsb.edu

***Abstract*—Can four noisy positions forecast when a Hénon–Heiles trajectory will escape and which of three exits it will take? We sample 3,900 exact-energy states and evaluate every landmark survivor in a fold withholding both its energy and initial phase region. Pooled noisy-history escape AUC is 0.816 with integrated Brier score 0.160; adding oracle maximal FTLE does not improve cold transfer. A second-cycle atlas shows where the signal lives. At position-noise standard deviation 0.015, the equal-cell Brier gain of history over energy alone is 0.0566 at horizon 10, with a two-way energy-by-region interval of** $[0.0311, 0.0834]$**, but falls to 0.0066** $[-0.0078, 0.0249]$ **at horizon 60. The conditional FTLE increment includes zero at all four horizons. Raw cell-median FTLE correlation with cell escape prevalence is 0.895** $[0.450, 0.949]$ **at horizon 10 and 0.863** $[0.398, 0.939]$ **at 60, but FTLE and horizon-10 risk are themselves strongly ordered by energy. Partial Spearman controlling energy is 0.291** $[-0.569, 0.812]$ **at 10 and** -0.089 $[-0.776, 0.677]$ **at 60. The raw ecological structure is therefore largely energy-driven and provides no resolved within-energy mechanism. A target-free support gate also fails. The evidence is fixed-seed simulation for one Hamiltonian, sensor, sampling measure, and model family.**

***Index Terms*—Hamiltonian chaos, finite-time Lyapunov exponent, competing risks, calibration, domain shift, negative result**

I. The Question Before the Machinery

The Hénon–Heiles system is a two-coordinate Hamiltonian introduced through numerical experiments on an additional isolating integral [1]. Its dimensionless energy is

$$H = \frac{p_x^2 + p_y^2}{2} + \frac{x^2 + y^2}{2} + x^2 y - \frac{y^3}{3}. \tag{1}$$

At $E_{\mathrm{esc}} = 1/6$, three equal-height saddles open. Above that energy a trajectory may leave upward, lower-right, or lower-left. Their basin boundaries can be Wada: one boundary borders all three outcomes [2].

That geometry makes a useful conceptual trap. Nearby trajectories may separate rapidly, yet "separates rapidly" does not specify "leaves soon" or "takes this exit." Nor does good interpolation imply transfer to a new energy or part of phase space. The study asks one bounded question:

> From four noisy position observations and the known energy setpoint, how well can a model forecast later escape time and exit channel in an unseen energy-region cell, and does a correctly computed maximal FTLE add information?

The FTLE comparison is deliberately generous. The sensor sees only (x, y), but the FTLE uses the latent simulator and its tangent map. It is therefore an oracle mechanistic feature, not something claimed to be directly measured.

II. From Trajectories to a Fair Forecast

A. Accessible states and an honest landmark

Velocity Verlet integrates Eq. (1) at $h = 0.02$. For each $E \in \{0.17, 0.18, 0.19, 0.20, 0.21, 0.22\}$, configuration points are drawn uniformly inside the saddle triangle *and rejected unless* $V(x, y) \le E$. Momentum direction is uniform, while its magnitude is $\sqrt{2(E - V)}$. Thus every accepted state lies on its declared energy shell; the maximum executed discrepancy is 8.33×10^{-17}.

The particle is observed through $t_L = 6$. Positions are recorded at $0, 2, 4, 6$ and independently perturbed by Gaussian noise with standard deviation 0.015. Momentum remains hidden. Of 3,900 starts, 360 (9.2%) cross radius 2.5 before the landmark. They are reported as early competing events, not silently relabeled or used for the later conditional task. The remaining 3,540 trajectories enter the forecast cohort; 2,161 escape and 1,379 remain inside through $t_L + 60$.

B. The maximal finite-time exponent

One arbitrary perturbation direction need not find the fastest-growing direction over a finite window. The discrete Verlet step has an exact tangent Jacobian J_k. We propagate the full 4×4 map

$$M_T = J_{N-1} J_{N-2} \cdots J_0, \qquad \lambda_{\max}(T) = \frac{1}{T} \log \sigma_1(M_T), \tag{2}$$

where σ_1 is the largest singular value. Equation (2), evaluated by SVD at $T = 6$, is the primary FTLE definition. A parallel full-basis QR iteration every ten steps supplies a conditioning diagnostic following standard Lyapunov-spectrum practice [3], [4]; it is not substituted for the finite-window singular maximum.

Three checks make the implementation falsifiable. The tangent map agrees with centered finite differences to 8.11×10^{-11}. Its one-step symplectic residual is 1.11×10^{-16}. Across the cohort, the maximum residual $\|M_T^\top \Omega M_T - \Omega\|_F$ is 6.3e-11, while the largest reciprocal-singular-pair error is 7.57×10^{-12}. Energy drift falls by a factor of four whenever h is halved, as expected for the second-order map.

C. Three exits require competing risks

Forecast time is divided into $(0, 10], (10, 20], (20, 40], (40, 60]$. For interval j, a multinomial logit estimates

$$h_{jc}(z) = P(T \in I_j, C = c \mid T > \inf I_j, z), \tag{3}$$

for causes $c = 1, 2, 3$, plus conditional survival $c = 0$. Person-interval rows stop after the first exit. Survival and cumulative incidence follow the recursion

$$S_j = S_{j-1} h_{j0}, \tag{4}$$
$$F_{jc} = F_{j-1,c} + S_{j-1} h_{jc}. \tag{5}$$

This preserves the fact that one exit prevents the other two and that survival through 60 is an observed administrative endpoint, rather than treating every non-event as missing [5].

D. What the models may see

Four fixed degree-two, L_2-regularized multinomial models share the same interval terms and penalty. They receive: (i) energy only; (ii) energy, eight noisy positions, fitted slopes, latest finite differences, history acceleration, path length, displacement, and final observed radius; (iii) the same history plus oracle $\lambda_{\max}(6)$; or (iv) energy plus exact landmark (x, y, p_x, p_y) as an upper reference. No outcome, escape time, channel, eligibility flag, or post-landmark state is a feature.

E. No neighborhood or energy leakage

Initial polar angle defines six regions $R_0, \ldots, R_5$ before any outcome is simulated. For the test cell (E_e, R_r), training is

$$\mathcal{D}_{-e,-r} = \{i : E_i \neq E_e \text{ and } R_i \neq R_r\}. \tag{6}$$

The 36 cross-classified folds predict every landmark survivor exactly once. Test labels never tune features, regularization, calibration, or support. Primary uncertainty resamples whole energy-region cells 2,000 times, retaining within-cell dependence and the paired model contrast.

ROC AUC evaluates escape-by-60 ranking. The Brier score evaluates probability error [6]; its mean over four horizons is the integrated Brier score (IBS). Final four-outcome Brier score and macro one-vs-rest channel AUC evaluate competing outcomes. Quantile reliability curves, calibration bias, a weakly ridge-stabilized slope, and expected calibration error (ECE) expose probability defects rather than hiding behind rank metrics [7].

III. Executed results

A. Noisy history transfers, but energy does much of the ranking

Figure 1(b) shows the main difficulty: escape incidence ranges from 0.07 to 0.95 across cold cells. Energy alone nevertheless reaches AUC 0.815. Noisy positions reach 0.816, only 0.001 higher, so most endpoint ranking comes from energy. The history matters elsewhere: IBS improves from 0.180 to 0.160, and macro exit-channel AUC rises from 0.622 to 0.738.

TABLE I
JOINT COLD-FOLD PERFORMANCE

Features	Escape AUC	IBS ↓	Cause AUC
Energy only	0.815	0.180	0.622
Noisy positions	0.816	0.160	0.738
Positions + max FTLE	0.814	0.160	0.742
Exact landmark state	0.835	0.152	0.785

A time series can add timing and destination information even when a one-number rank statistic barely moves.

Exact landmark state is useful but not magical under this transfer test: AUC 0.835, IBS 0.152, and macro cause AUC 0.785. Holding out an entire angular region asks polynomial logits to extrapolate, not to recover a familiar local basin boundary.

B. The FTLE result is negative

The oracle FTLE is visibly structured and is larger for many trajectories that later escape (Fig. 3). Yet adding it changes pooled escape AUC from 0.816 to 0.814: $\Delta = -0.002$ with interval $[-0.008, +0.004]$. IBS changes from 0.160 to 0.160: $\Delta = +0.0006$ with $[-0.0013, +0.0025]$. Both intervals cross zero, and the point estimates are adverse. Macro channel AUC rises slightly from 0.738 to 0.742, but the final multiclass Brier contrast is -0.0006 with interval $[-0.0066, 0.0050]$. There is no resolved incremental benefit.

This does not mean the system is regular or the FTLE is wrong. It means the fastest local tangent growth over $[0, 6]$ contributes little beyond an energy setpoint and the observed path for this later, coarse event forecast. Description and prediction answer different questions.

C. Calibration is usable, not perfect

For noisy history, mean predicted risk differs from the event rate by 0.003 and ECE is 0.033, but the stabilized calibration slope is 0.69. The mean is nearly right while individual probabilities are too spread out. FTLE worsens ECE to 0.053 and slope to 0.66. Figure 2(b) also reveals a late-horizon channel error: the model keeps accumulating risk after empirical incidence has nearly flattened. AUC alone would miss both defects.

D. The support gate fails honestly

A label-free support score standardizes noisy-history features within each training fold and measures test distance to training energy-region centroids. Its threshold is the frozen 90th percentile of pseudo-cold training distances, where each training row is compared only with centroids sharing neither its energy nor region.

The rule accepts 89.7% of predictions, but distance has rank correlation -0.313 with absolute error. Supported predictions have IBS 0.165; rejected predictions have the better 0.110. High-distance cells are often high-energy cells with near-certain escape, so unfamiliar does not mean ambiguous. The score remains a transparent domain-shift flag, but abstaining

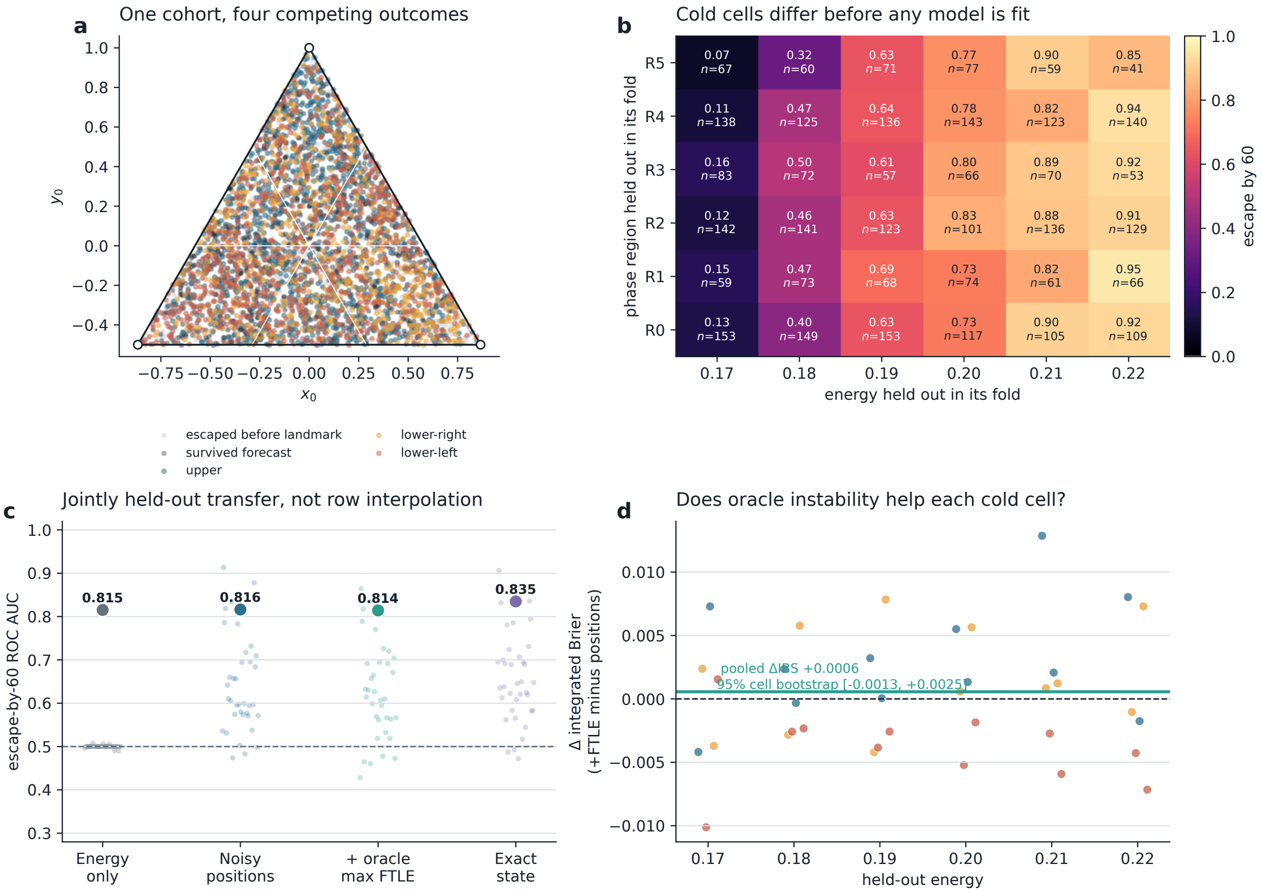

Fig. 1. Primary evidence. (a) Cohort and mutually exclusive outcomes; white rays mark frozen regions. (b) Incidence across 36 cold cells. (c) Fold and pooled AUC. (d) Paired FTLE increments; positive ΔBrier is worse.

on it would make the retained evaluation worse. No threshold is retuned after seeing outcomes.

E. Noise is not the bottleneck tested here

With clean positions, history AUC/IBS are 0.817/0.159. At noise standard deviation 0.04, they are 0.819/0.159. These small, non-monotone differences do not show that noise helps; they show that within this declared range, cold energy-region shift dominates measurement noise. Larger errors, missing observations, biased sensors, or unknown energy could behave differently.

IV. Second-cycle forecastability atlas

A pooled score can hide where prediction works. We therefore returned every cold prediction to its held-out energy-region cell and asked three separate questions at horizons 10, 20, 40, and 60: how often does the cell escape, how much does position history improve on energy alone, and how much does oracle FTLE improve on position history?

At the primary position-noise standard deviation of 0.015, the equal-cell Brier gain of history over energy is 0.0566 at horizon 10, with a 95% two-way product interval of $[0.0311, 0.0834]$. It falls to 0.0294 $[0.0078, 0.0544]$ at 20, 0.0128 $[-0.0043, 0.0350]$ at 40, and 0.0066 $[-0.0078, 0.0249]$ at 60. The short observed path is most useful for the near forecast.

Adding oracle FTLE does not show the same pattern. Its Brier gain is 0.0018 $[-0.0009, 0.0057]$ at 10, 0.0005 $[-0.0024, 0.0034]$ at 20, -0.0014 $[-0.0054, 0.0022]$ at 40, and -0.0019 $[-0.0078, 0.0033]$ at 60. Every interval includes zero.

The unadjusted cell association looks strong: median FTLE versus escape prevalence has Spearman correlation 0.895 $[0.450, 0.949]$ at horizon 10 and 0.863 $[0.398, 0.939]$ at 60. This pattern shares an energy gradient. At horizon 10, FTLE-energy correlation is 0.892 and risk-energy correlation is 0.963. After residualizing FTLE and risk ranks on energy rank, partial Spearman is 0.291 $[-0.569, 0.812]$ at horizon 10, 0.024 $[-0.734, 0.733]$ at 20, -0.067 $[-0.804, 0.652]$ at 40, and -0.089 $[-0.776, 0.677]$ at 60. This descriptive adjustment does not identify a within-energy mechanism. It agrees with the conditional prediction result: adding a trajectory's oracle FTLE after energy and its observed path yields no resolved gain.

Recent work uses Poincaré geometry and finite-time Lyapunov exponents to test whether a learned Hamiltonian extrap-

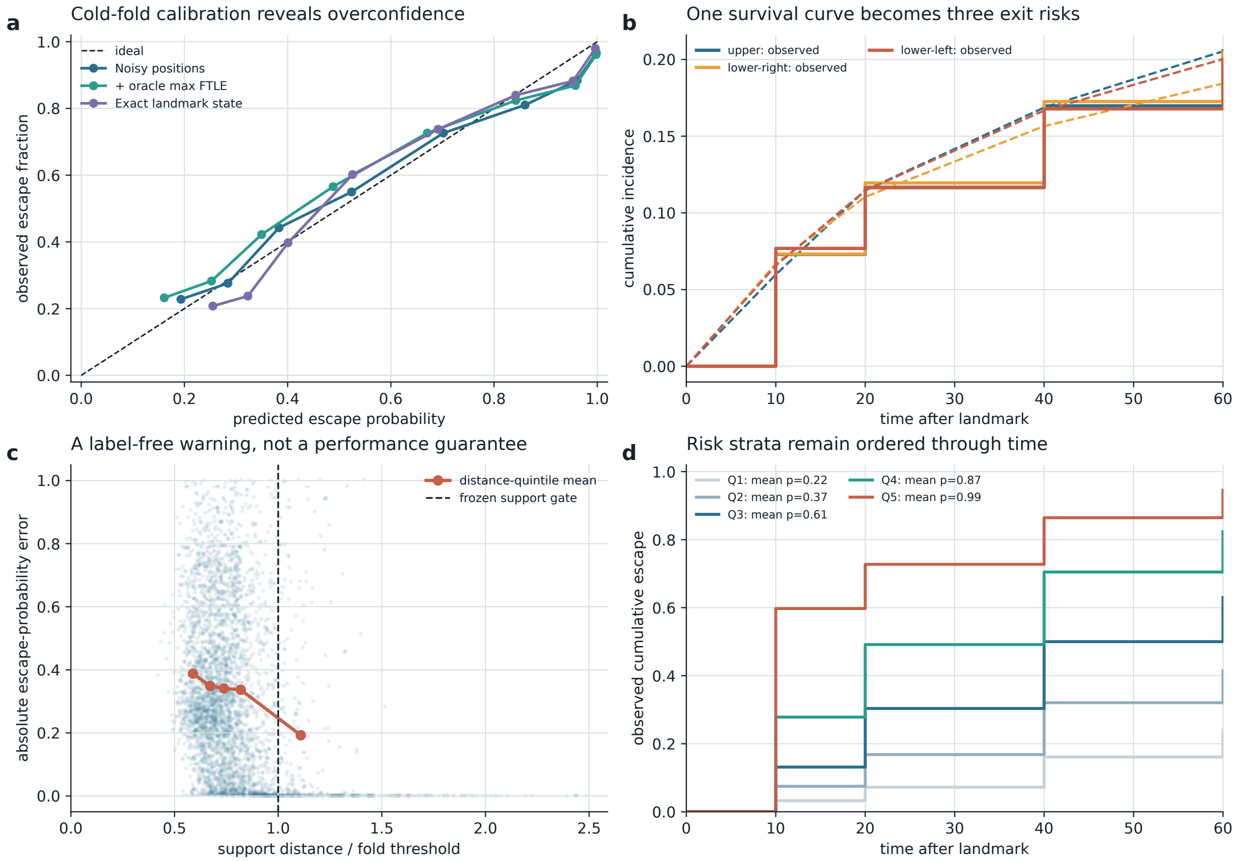

Fig. 2. Probability diagnostics. (a) Cold-fold reliability. (b) Predicted (dashed) and empirical (solid) exit incidence. (c) Frozen support distance is inversely related to error. (d) Risk quintiles remain ordered.

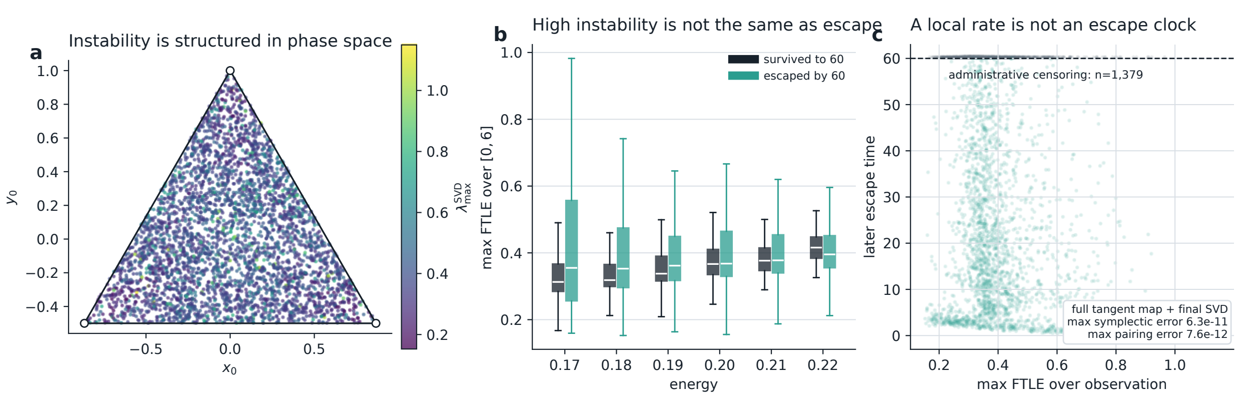

Fig. 3. Mechanism without overclaim. (a) Maximal SVD FTLE structure. (b) Escaped and surviving trajectories overlap. (c) Similar FTLEs lead to immediate, delayed, or censored outcomes.

olates into an unseen chaotic regime [8]. That is a different question. Here the Hamiltonian is known, the sensor is partial, and the target is a later escape event in a held-out energy-region cell. Reproducing chaos and forecasting an individual escape need not succeed together.

Intervals resample the six energy levels and six phase regions independently, then take their product. They represent this crossed design, not new trajectories, alternative Hamiltonians, or a population of physical systems.

V. Limits and Conclusion

This is a conditional landmark study. Early exits are counted, but the model answers a later question only for trajectories surviving to $t = 6$. All remaining subjects are followed until escape or the fixed 60-unit administrative horizon, so there is no random loss to follow-up. Region boundaries are

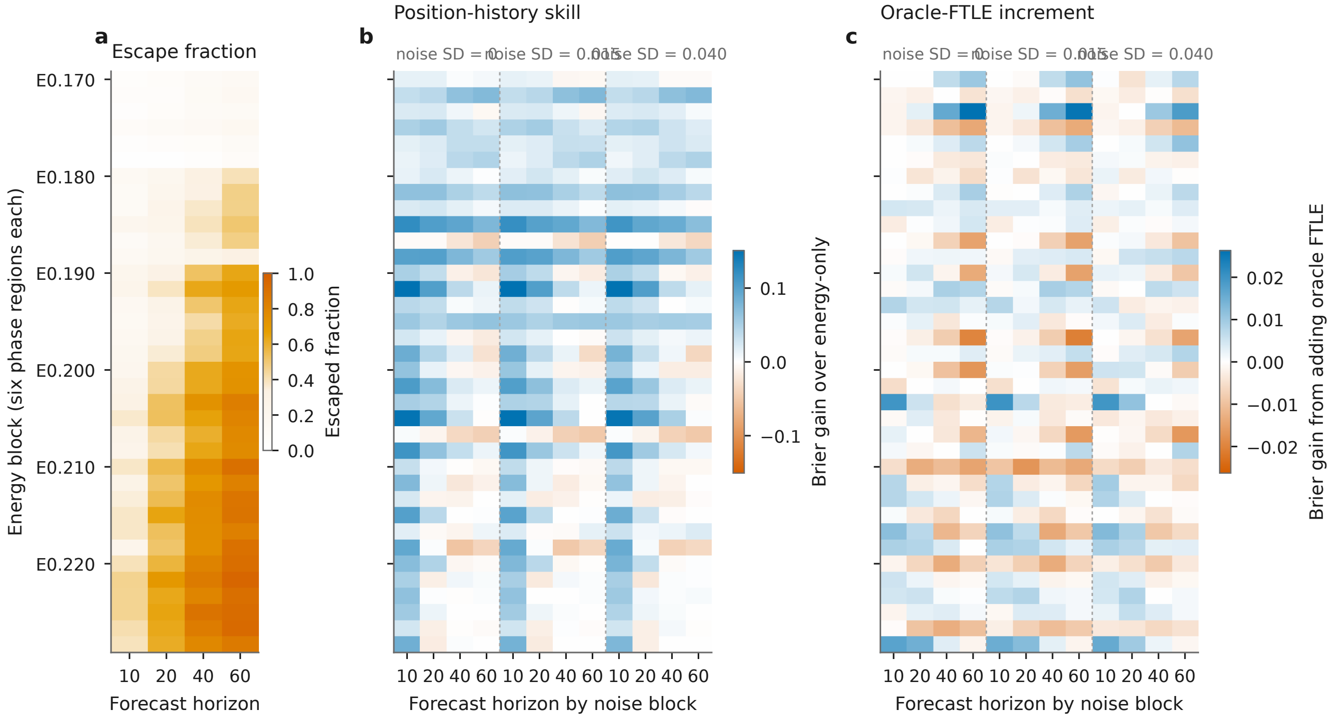

Fig. 4. Forecastability atlas for all 36 held-out energy-region cells. (a) Escape prevalence from the six-unit landmark; the annotation contrasts raw and energy-adjusted FTLE-risk rank associations. (b) Brier gain of noisy position history over energy alone. (c) Brier gain from adding oracle FTLE to history. Columns repeat the four horizons within clean, primary-noise, and high-noise blocks. Coral is adverse, cream is near zero, and teal is favorable.

frozen angular wedges, not a unique partition of phase space. The 36-cell bootstrap represents this finite cross-classified design; it does not cover alternative samplers, landmarks, radii, integrators, features, or model families.

The simulator is deterministic, energy is known exactly, and Gaussian observation error is idealized. The exact-state and FTLE models are oracles, not deployable sensors. Transfer is only to held-out cells within six energies of the same Hamiltonian. No celestial, molecular, or experimental claim follows.

The central result is layered. Under strict cold transfer, noisy short history gives a resolved Brier gain over energy alone at horizons 10 and 20, but that gain decays and is unresolved by 40 and 60. The strong raw cell-median FTLE association with regional escape prevalence is largely a common energy gradient: energy-adjusted rank correlations are small and unresolved, and the conditional FTLE increment after energy and history includes zero at every horizon. Geometric support distance also fails to identify the hard cells. The practical lesson is simple. Before calling chaos predictive, specify what is observed, which event and horizon are forecast, what domain is new, and whether a regional association adds information for an individual trajectory. A beautiful instability field is not yet a forecast.

References

[1] M. Hénon and C. Heiles, "The applicability of the third integral of motion: Some numerical experiments," *Astron. J.*, vol. 69, pp. 73–79, 1964, doi: 10.1086/109234.

[2] J. Aguirre, J. C. Vallejo, and M. A. F. Sanjuán, "Wada basins and chaotic invariant sets in the Hénon–Heiles system," *Phys. Rev. E*, vol. 64, 066208, 2001, doi: 10.1103/PhysRevE.64.066208.

[3] G. Benettin, L. Galgani, A. Giorgilli, and J.-M. Strelcyn, "Lyapunov characteristic exponents for smooth dynamical systems and for Hamiltonian systems; a method for computing all of them," *Meccanica*, vol. 15, pp. 9–30, 1980, doi: 10.1007/BF02128236.

[4] C. Skokos, "The Lyapunov characteristic exponents and their computation," in *Dynamics of Small Solar System Bodies and Exoplanets*, Lecture Notes in Physics 790, Springer, 2010, pp. 63–135, doi: 10.1007/978-3-642-04458-8_2.

[5] T. Meir and M. Gorfine, "Discrete-time competing-risks regression with or without penalization," *Biometrics*, vol. 81, no. 2, ujaf040, 2025, doi: 10.1093/biomtc/ujaf040.

[6] G. W. Brier, "Verification of forecasts expressed in terms of probability," *Mon. Weather Rev.*, vol. 78, pp. 1–3, 1950, doi: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

[7] B. Van Calster, D. Nieboer, Y. Vergouwe, B. De Cock, M. J. Pencina, and E. W. Steyerberg, "A calibration hierarchy for risk models was defined: from utopia to empirical data," *J. Clin. Epidemiol.*, vol. 74, pp. 167–176, 2016, doi: 10.1016/j.jclinepi.2015.12.005.

[8] J. Choi, *Extrapolating the emergence of Hamiltonian chaos with random-feature Hamiltonian neural networks*, arXiv:2607.28977, submitted July 31, 2026, doi: 10.48550/arXiv.2607.28977.