
State or Pattern? Do Zero-Shot Time-Series Foundation Models Use the Full State When Simulating Chaotic Systems?

Qinzheng Ma¹ Jialin Wu² Qingxi Meng¹ Shichen Tang¹
¹Rice University ²UC San Diego
qm18@rice.edu

Abstract

A forecaster continues a series; a simulator propagates a *state*. Zero-shot time-series foundation models (TSFMs) now accept multivariate input, but whether they treat the other channels as the state of a coupled system or as further patterns to continue has not been tested. We run a pre-registered, paired experiment on 67 chaotic ODEs from the DYSTS collection: every target coordinate is forecast 512 steps from a 512-point context under three arms — the coordinate alone (U), all d coordinates jointly (M), and the coordinate together with the other coordinates taken from a segment of the same trajectory tens of Lyapunov times away, which keeps them on the same attractor and at the same token count but destroys the coupling (C). Measured in valid prediction time, the full state buys Chronos-2 -0.05 Lyapunov times (95% CI $[-0.14, +0.03]$), TimesFM-3 $+0.22$ $[+0.07, +0.48]$ and Toto $+0.04$ $[+0.00, +0.10]$, while a next-generation reservoir computer fitted on the same 512 points gains $+1.94$ $[+1.16, +2.68]$ and outscores every foundation model in any arm. The models are not blind to the extra channels — their forecasts move by 0.24 – 0.47 context standard deviations, and the true state beats the decoupled one for all three — yet they recover at most 6.2% of NVAR’s coupling gain (37% of the weaker analog reference’s); on the secondary metric all three do gain, Chronos-2 included. On the road from forecasting to world modelling, attending across channels is not yet buying the predictability that the state contains.

1 Introduction

A forecaster continues a series; a simulator propagates a *state*. That distinction is the premise of treating temporal foundation models as world models rather than forecasters, and it is sharpest for deterministic dynamical systems: the future of one coordinate of a chaotic ODE is a function of the full state vector, not of that coordinate’s own past. Zero-shot time-series foundation models (TSFMs) now accept multivariate inputs: Chronos-2 attends across the series of an item [Ansari et al., 2025], TimesFM-3 is natively multivariate [Jain and Sen, 2026], Toto lets channels attend jointly [Cohen et al., 2025], and Moirai-1 flattens variates into one token sequence [Woo et al., 2024], though Moirai-2 has since dropped multivariate support [Liu et al., 2025b]. Whether they use the other coordinates as a *state*, an initial condition for coupled dynamics, or merely as more patterns to continue, has not been tested.

The closest prior study evaluated Chronos on the 135 chaotic systems of the DYSTS collection [Gilpin, 2021] one coordinate at a time, because Chronos is univariate; it reports that full-state input makes the task easier for its *trained* baselines, but never gives the full state to the foundation model [Zhang and Gilpin, 2025]. The same authors then showed that copying the best-matching context segment (*context parroting*) is hard for TSFMs to beat on chaotic systems [Zhang and Gilpin, 2026],

suggesting that univariate zero-shot forecasting of chaos is largely pattern continuation. Parroting has a full-state analogue that predates deep learning: analog forecasting [Lorenz, 1969] finds the past state closest to the present one and follows its successor. Together with a next-generation reservoir computer (NVAR; Gauthier et al., 2021) fitted on the same context, these classical methods measure how much predictability the full state *adds* to a 512-point context, so they lower-bound the gain a TSFM could reach and give the scale on which its use of the state can be read.

Design. On every (system, window, coordinate) we score the same target coordinate under three arms: U sees the coordinate alone; M sees all d coordinates of the window jointly; C pairs the target with the other $d - 1$ coordinates taken from a segment of the *same* trajectory tens of Lyapunov times away, so the extra channels are the same signals off the same attractor, with the same token count, but carry no information about the current state. $\Delta_1 = \text{VPT}(\text{M}) - \text{VPT}(\text{U})$ is the coupling gain in valid prediction time; $\Delta_2 = \text{VPT}(\text{M}) - \text{VPT}(\text{C})$ isolates the part that comes from the current state rather than from in-context exposure to more series of the same system.

Contributions.

- A pre-registered, paired U/M/C test of state use for three zero-shot multivariate TSFMs on 67 chaotic systems (536 windows, 1,784 scored forecasts per arm), with a decoupled-channel control separating state use from system-identity in-context learning.
- Classical context-only references showing the state does carry usable information: NVAR fitted on the same 512 points gains +1.94 Lyapunov times of valid prediction from it, and analog forecasting gains on 81% of systems.
- The verdict: the three models recover 0.6–6.2% of that gain, depending on model and metric (Δ_1 between -0.05 and $+0.22$ Lyapunov times), although their forecasts do change when the channels change and the true state beats a decoupled one ($\Delta_2 > 0$ throughout); whatever the three models do with the extra channels, it is not extracting the predictability the state contains.

Related work. General-purpose TSFMs [Ansari et al., 2024b,a, Das et al., 2024, Woo et al., 2024] are pretrained on real series and univariate synthetic generators [Oreshkin et al., 2026]; the multivariate generation [Ansari et al., 2025, Jain and Sen, 2026, Cohen et al., 2025, Podest et al., 2026] reports large gains, Chronos-2 and TiRex-2 on the covariate tasks of Shchur et al. [2025], attributed to in-context learning across series [Auer et al., 2025, Das et al., 2025]. Panda ablates the channel-attention layers inside its own architecture [Lai et al., 2025] and ChaosNexus builds variable attention in without isolating it [Liu et al., 2025a], but no study asks whether general-purpose zero-shot TSFMs exploit state coupling, nor uses a decoupled control. Durstewitz et al. [2026] call for evaluating time-series models as dynamical-systems reconstructions, Adler et al. [2026] study calibration under rollout, Perez-Diaz et al. [2025] note that most TSFMs emit marginals rather than trajectories, transformers on dynamical data have been reported to fail where reservoir computers succeed [Zhai et al., 2026], and probes find TSFM representations dominated by local motifs [Pandey et al., 2025, Bao et al., 2026].

2 Experimental setup

The design was frozen before any model was run; deviations are listed in Appendix D.

Data. We use the DYSTS collection of chaotic ODEs [Gilpin, 2021] in its public multivariate form (Hugging Face dataset `williamgilpin/dysts`, fine granularity, CC BY 4.0): 117 of its 135 systems, one 10,000-step multivariate trajectory each. These are synthetic integrations of named ODEs and appear in no documented TSFM pretraining corpus, which sidesteps the leakage reported for standard benchmarks [Meyer et al., 2025, Qiao et al., 2026]. The granularity, calibrated once against a regenerated Lorenz trajectory, is $p = 100$ points per dominant period, so each system’s Lyapunov time in samples is $\tau = p/(\lambda_{\max}T)$ from the collection’s metadata. Inclusion, frozen before any model run: $\tau \in [4, 128]$ samples (so context and horizon each span at least four Lyapunov times), $d \leq 6$, no degenerate coordinate, no non-finite values. This leaves 67 systems of dimension 3 to 6 (mean 3.3), with τ between 4.1 and 126.6 samples (Appendix B). *Windows.* Per system, eight forecast origins $o_k = 512 + \lfloor k(3584 - 512)/7 \rfloor$, $k = 0, \dots, 7$, with context $[o_k - 512, o_k]$ and target $[o_k, o_k + 512]$ ($L = H = 512$). The “far” segments of arm C are steps $[5000 + 512k, 5512 + 512k]$ of the same trajectory, at least 31 Lyapunov times after the window they replace (median 63): points

Table 1: Valid prediction time in Lyapunov times (mean over 67 systems) and the paired contrasts with 95% bootstrap intervals over systems; bold marks intervals excluding zero. R : fraction of NVAR’s coupling gain recovered, over the 56 systems where NVAR gains (so $R > 0$ is possible with $\Delta_1 < 0$). “State?”: the pre-registered verdict; † marks a model satisfying both of its clauses at once (Toto; Holm-corrected $p = 0.32$).

Model	VPT by arm			paired contrast		R	State?
	U	M	C	Δ_1 [95% CI]	Δ_2 [95% CI]		
Chronos-2	2.60	2.56	2.43	−0.05 [−0.14, +0.03]	+ 0.12 [+0.07, +0.18]	0.6%	no
TimesFM-3	2.59	2.82	2.76	+ 0.22 [+0.07, +0.48]	+ 0.06 [+0.02, +0.09]	3.8%	yes
Toto	1.82	1.87	1.81	+ 0.04 [+0.00, +0.10]	+ 0.06 [+0.01, +0.14]	1.3%	yes†
Chronos-Bolt	1.06	—	—	—	—	—	—
Moirai-2	1.51	—	—	—	—	—	—
Persistence	0.57	—	—	—	—	—	—
Context mean	0.18	—	—	—	—	—	—
Analog	2.81	2.65	1.75	−0.15 [−0.60, +0.21]	+ 0.90 [+0.48, +1.29]	—	incon.
NVAR	2.34	4.28	—	+ 1.94 [+1.16, +2.68]	—	—	—

of the same attractor whose spectra match the window as closely as any other segment of the same run does (Appendix D), but whose dynamics are decoupled from it.

Arms (Table 2, Appendix A). In M one call per window returns all d coordinates, scored one by one; U and C use one call per (window, coordinate). Inputs are raw values (every model normalises per series); the scored output is the median forecast of the target coordinate over the 512 horizon steps.

Models. Primary (multivariate) TSFMs, zero-shot with package defaults: Chronos-2 [Ansari et al., 2025] (multivariate tensor interface), TimesFM-3.0 [Jain and Sen, 2026] (two-dimensional contexts, single-pass decoder) and Toto-Open-Base-1.0 [Cohen et al., 2025] (coordinates in the channel dimension, median of 64 samples). Moirai-2.0-R-small [Liu et al., 2025b] appears in arm U only, because Moirai-2 dropped cross-variate support by design, and Chronos-Bolt-base [Ansari et al., 2024a] is a univariate reference; see Appendix C.

Classical context-only baselines. All baselines see the 512 context points and nothing else; M-analog and NVAR-M are the positive controls, persistence and the context mean the floors. U-*analog* is context parroting [Zhang and Gilpin, 2026]: delay-embed the target coordinate (dimension 4, delay 1), copy the continuation of its nearest earlier embedded point, and re-search whenever the copied segment runs out. M-*analog* is Lorenz’s analog forecast [Lorenz, 1969]: the same with the neighbour taken in the full d -dimensional state; C-*analog* uses the far segment’s coordinates and must not help. NVAR [Gauthier et al., 2021] is a ridge regression from delayed states and their quadratic monomials to s_{t+1} , rolled out 512 steps (Appendix D); its U variant sees the target only, its M variant the full state.

Metric. For each (system, window, coordinate, arm, model) the normalised error is $e_t = |\hat{y}_t - y_t|/\sigma_{\text{ctx}}$, with σ_{ctx} the standard deviation of that coordinate over the 512 context points, smoothed by a centred 5-step moving average. The valid prediction time is the first t with $e_t > 0.5$, censored at 512, reported in Lyapunov times, $\text{VPT} = t/\tau$, and averaged over the eight windows and d coordinates to one number per (system, arm, model); *systems are the independent units*.

Statistics. For each primary model we report the paired per-system contrasts Δ_1 and Δ_2 as means over systems with 95% percentile-bootstrap confidence intervals over systems ($B = 10,000$), Wilcoxon signed-rank p -values, and Holm correction across the primary models. Decision rule, fixed in advance: a model *uses the state* iff both intervals exclude zero on the positive side; it *does not use the state* iff both point estimates are below 0.2 Lyapunov times in magnitude with intervals inside $[-0.4, 0.4]$; otherwise the result is inconclusive. The *recovery ratio* R is a model’s Δ_1 over a classical reference’s, on systems where the reference gains; the table uses NVAR-M, Appendix E also the pre-registered M-analog. Secondaries: sMAPE over the first Lyapunov time, the Spearman correlation between the classical coupling gain and each TSFM’s Δ_1 , and per-system and robustness tables (appendix).

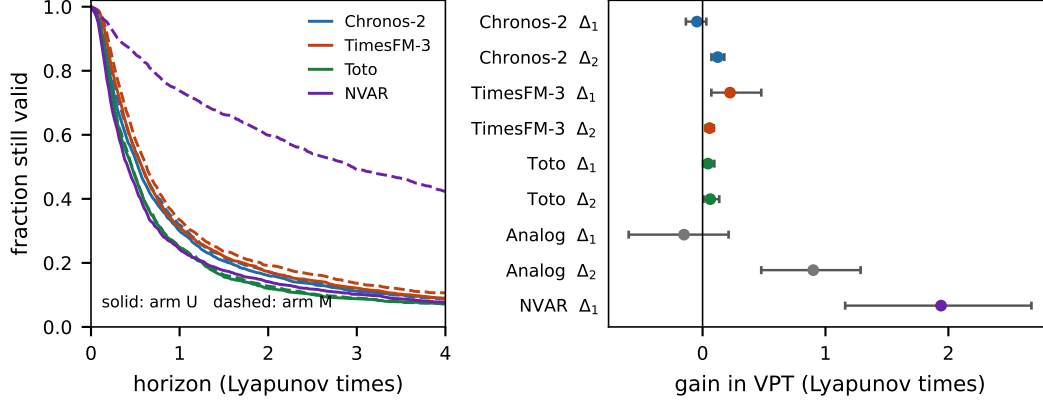


Figure 1: **Left:** fraction of forecasts still valid against horizon in Lyapunov times, arm U (solid) and arm M (dashed); the state lifts NVAR far above its univariate curve, while Chronos-2’s and Toto’s two curves coincide and TimesFM-3’s separate slightly. **Right:** the paired contrasts, 95% bootstrap intervals.

3 Results

The state does carry usable information. NVAR, fitted per window on the same 512 points, gains +1.94 Lyapunov times from the full state (95% CI [+1.16, +2.68], better on 84% of systems), reaching 4.28 in arm M — above every model’s score in any arm, and a measure of what the context contains rather than a zero-shot competitor. Analog forecasting gains on 81% of systems (median +0.18; its mean, -0.15 [$-0.60, +0.21$], is pulled down by a heavy tail), and the control behaves as intended: C-analog loses 1.05 Lyapunov times against U-analog [$-1.43, -0.74$].

The foundation models barely use it. Δ_1 is -0.05 [$-0.14, +0.03$] for Chronos-2, $+0.22$ [$+0.07, +0.48$] for TimesFM-3 and $+0.04$ [$+0.00, +0.10$] for Toto (Table 1). Under the pre-registered rule Chronos-2 *does not use the state* and TimesFM-3 *uses it*; Toto satisfies both clauses at once — intervals above zero, effects inside the negligible band — which is the result in miniature. All are one to two orders of magnitude below the classical reference: recovery ratios against NVAR of 0.6%, 3.8% and 1.3%. Only TimesFM-3 survives Holm correction ($p = 6 \times 10^{-5}$; Toto 0.32, Chronos-2 0.60). As in Zhang and Gilpin [2026], the best models stay below context parroting (2.81 against 2.60 and 2.59 in arm U).

They are not ignoring the extra channels. Arm M moves each model’s forecast of the same coordinate by 0.47 (Chronos-2), 0.31 (TimesFM-3) and 0.24 (Toto) context standard deviations, and no forecast is unchanged. The channels are read as more than noise: $\Delta_2 > 0$ for all three ($+0.12$ [$+0.07, +0.18$], $+0.06$ [$+0.02, +0.09$], $+0.06$ [$+0.01, +0.14$]), so the true state beats a stand-in drawn from the same attractor at the same token count.

Robustness. Under sMAPE over the first Lyapunov time all three improve with the state (Chronos-2 by 2.3 points [1.0, 3.6], TimesFM-3 by 3.1 [2.0, 4.2], Toto by 0.7 [0.3, 1.1]), so Chronos-2’s null is specific to valid prediction time; NVAR improves by 48.4 [37.9, 58.9], leaving recoveries of 4.6%, 6.2% and 1.5%. The censoring ceiling H/τ ranges from 4.0 to 124.7 Lyapunov times across systems, so we also capped every VPT at a common 4: the models move to $+0.02$, $+0.10$ and $+0.02$ against NVAR’s $+1.67$ [$+1.29, +2.04$], the same ratio. Each positive mean leans on one system: dropping the most influential leaves TimesFM-3 at $+0.11$ [$+0.06, +0.17$] and Toto at $+0.03$ [$-0.00, +0.07$]. At the strictest error threshold Chronos-2 turns significantly negative and Toto’s gain vanishes (Appendix E).

Limitations. This measures deployed zero-shot behaviour, not capability in principle: no TSFM corpus we can inspect documents coupled-ODE trajectories, so an architectural limit cannot be separated from missing pretraining data, and chaos specialists [Lai et al., 2025, Liu et al., 2025a] may differ. Arm C shows series from the same system, so Δ_2 separates state use from system-identity in-context learning but neither from pattern continuation, and multivariate interfaces differ, so a null for one is about that interface as released.

References

- Coen Adler, Yuxin Chang, Felix Draxler, Samar Abdi, and Padhraic Smyth. Beyond Accuracy: Are Time Series Foundation Models Well-Calibrated? In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2510.16060.
- Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Oleksandr Shchur, et al. Chronos-bolt: fast and accurate zero-shot forecasting (model card). Hugging Face model card, <https://huggingface.co/amazon/chronos-bolt-base>, 2024a.
- Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Hao Wang, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Yuyang Wang. Chronos: Learning the Language of Time Series. *Transactions on Machine Learning Research*, 2024b. arXiv:2403.07815.
- Abdul Fatir Ansari, Oleksandr Shchur, Jaris Küken, Andreas Auer, Boran Han, Pedro Mercado, Syama Sundar Rangapuram, Huibin Shen, Lorenzo Stella, Xiyuan Zhang, Mononito Goswami, Shubham Kapoor, Danielle C. Maddix, Pablo Guerron, Tony Hu, Junming Yin, Nick Erickson, Prateek Mutalik Desai, Hao Wang, Huzefa Rangwala, George Karypis, Yuyang Wang, and Michael Bohlke-Schneider. Chronos-2: From Univariate to Universal Forecasting. arXiv technical report; model amazon/chronos-2, 2025. arXiv:2510.15821.
- Andreas Auer, Raghu Parthipan, Pedro Mercado, Abdul Fatir Ansari, Lorenzo Stella, Bernie Wang, Michael Bohlke-Schneider, and Syama Sundar Rangapuram. Zero-Shot Time Series Forecasting with Covariates via In-Context Learning. arXiv preprint, 2025. arXiv:2506.03128.
- Anthony Bao, Venkata Hasith Vattikuti, Jeffrey Lai, and William Gilpin. Universal Redundancies in Time Series Foundation Models. arXiv preprint, 2026. arXiv:2602.01605.
- Ben Cohen, Emaad Khwaja, Youssef Doubli, Salahidine Lemaachi, Chris Lettieri, Charles Masson, Hugo Miccinilli, Elise Ramé, Qiqi Ren, Afshin Rostamizadeh, Jean Ogier du Terrail, Anna-Monica Toon, Kan Wang, Stephan Xie, Zongzhe Xu, Viktoriya Zhukova, David Asker, Ameet Talwalkar, and Othmane Abou-Amal. This Time is Different: An Observability Perspective on Time Series Foundation Models. arXiv preprint; model Datadog/Toto-Open-Base-1.0, 2025. arXiv:2505.14766.
- Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024. arXiv:2310.10688.
- Abhimanyu Das, Matthew Faw, Rajat Sen, and Yichen Zhou. In-Context Fine-Tuning for Time-Series Foundation Models. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025. arXiv:2410.24087.
- Daniel Durstewitz, Christoph Jürgen Hemmer, Florian Hess, Charlotte Ricarda Doll, and Lukas Eisenmann. Position: A Dynamical Systems Perspective is Needed to Advance Time Series Modeling. arXiv preprint, 2026. arXiv:2602.16864.
- Daniel J. Gauthier, Erik Bollt, Aaron Griffith, and Wendson A. S. Barbosa. Next generation reservoir computing. *Nature Communications*, 12:5564, 2021. doi: 10.1038/s41467-021-25801-2. arXiv:2106.07688.
- William Gilpin. Chaos as an interpretable benchmark for forecasting and data-driven modelling. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2021. arXiv:2110.05266.
- Ayush Jain and Rajat Sen. TimesFM-3: A zero-shot foundation model for multivariate forecasting. Google Research blog (31 August 2026) and Hugging Face model card google/timesfm-3.0-pytorch, <https://research.google/blog/timesfm-3-a-zero-shot-foundation-model-for-multivariate-forecasting/>, 2026.
- Jeffrey Lai, Anthony Bao, and William Gilpin. Panda: A pretrained forecast model for chaotic dynamics. arXiv preprint, 2025. arXiv:2505.13755.

- Chang Liu, Bohao Zhao, Jingtao Ding, and Yong Li. ChaosNexus: A Foundation Model for ODE-based Chaotic System Forecasting with Hierarchical Multi-scale Awareness. arXiv preprint, 2025a. arXiv:2509.21802.
- Chenghao Liu, Taha Aksu, Juncheng Liu, Xu Liu, Hanshu Yan, Quang Pham, Silvio Savarese, Doyen Sahoo, Caiming Xiong, and Junnan Li. Moirai 2.0: When Less Is More for Time Series Forecasting. arXiv preprint; model Salesforce/moirai-2.0-R-small, 2025b. arXiv:2511.11698.
- Edward N. Lorenz. Atmospheric predictability as revealed by naturally occurring analogues. *Journal of the Atmospheric Sciences*, 26(4):636–646, 1969. doi: 10.1175/1520-0469(1969)26<636:APARBN>2.0.CO;2.
- Marcel Meyer, Sascha Kaltenpoth, Kevin Zalipski, and Oliver Müller. Rethinking Evaluation in the Era of Time Series Foundation Models: (Un)known Information Leakage Challenges. arXiv preprint, 2025. arXiv:2510.13654.
- Boris N. Oreshkin, Mayank Jauhari, Ravi Kiran Selvam, Malcolm Wolff, Wenhao Pan, Shankar Ramasubramanian, Kin G. Olivares, Tatiana Konstantinova, Andres Potapczynski, Mengfei Cao, Dmitry Efimov, Michael W. Mahoney, and Andrew G. Wilson. Zero-shot Forecasting by Simulation Alone. arXiv preprint, 2026. arXiv:2601.00970.
- Atharva Pandey, Abhilash Neog, and Gautam Jajoo. On the Internal Semantics of Time-Series Foundation Models. arXiv preprint, 2025. arXiv:2511.15324.
- Alvaro Perez-Diaz, James C. Loach, Danielle E. Toutoungi, and Lee Middleton. Foundation Model Forecasts: Form and Function. arXiv preprint, 2025. arXiv:2510.19345.
- Patrick Podest, Marco Pichler, Elias Bürger, Levente Zólyomi, Bernhard Voggenberger, Wilhelm Berghammer, Daniel Klotz, Sebastian Böck, Günter Klambauer, and Sepp Hochreiter. TiRex-2: Generalizing TiRex to Multivariate Data and Streaming. arXiv preprint, 2026. arXiv:2607.01204.
- Zhongzheng Qiao, Sheng Pan, Anni Wang, Viktoriya Zhukova, Yong Liu, Xudong Jiang, Qingsong Wen, Mingsheng Long, Ming Jin, and Chenghao Liu. It’s TIME: Towards the Next Generation of Time Series Forecasting Benchmarks. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026. arXiv:2602.12147.
- Oleksandr Shchur, Abdul Fatir Ansari, Caner Turkmen, Lorenzo Stella, Nick Erickson, Pablo Guerron, Michael Bohlke-Schneider, and Yuyang Wang. fev-bench: A Realistic Benchmark for Time Series Forecasting. arXiv preprint, 2025. arXiv:2509.26468.
- Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified Training of Universal Time Series Forecasting Transformers. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024. arXiv:2402.02592.
- Zheng-Meng Zhai, Celso Grebogi, and Ying-Cheng Lai. Can Transformers predict system collapse in dynamical systems? arXiv preprint, 2026. arXiv:2605.04024.
- Yuanzhao Zhang and William Gilpin. Zero-shot forecasting of chaotic systems. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2409.15771.
- Yuanzhao Zhang and William Gilpin. Context parroting: A simple but tough-to-beat baseline for foundation models in scientific machine learning. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2505.11349.

A The three arms

Delay embedding recovers the state of a deterministic system only asymptotically; within a 512-point context, coordinates that are poor observables (Lorenz z , blind to the $(x, y) \mapsto (-x, -y)$ symmetry) leave the state genuinely ambiguous, which is why the univariate arm is not simply a harder version of the multivariate one.

Table 2: The three arms. All arms are scored on the same target coordinate, window and system; only the context differs. d is the system dimension.

Arm	Context given to the model	Series	What it tests
U (univariate)	target coordinate of the window	1	pattern continuation of one observable
M (multivariate)	all d coordinates of the window as one item	d	U + current state
C (control)	target coordinate of the window + the other $d - 1$ coordinates from the far segment	d	U + same system, wrong state

B Included systems

The Hugging Face release carries 117 of the 135 DYSTS systems at fine granularity. Of these, 7 are excluded for $d = 10 > 6$ and 43 for a Lyapunov time outside $[4, 128]$ samples (all of them too slow, $\tau > 128$, so a 512-step horizon would not resolve error growth), leaving the 67 systems of Table 3: dimension 3 to 6 (mean 3.3; 52 systems have $d = 3$, 11 have $d = 4$, one $d = 5$ and three $d = 6$), τ from 4.1 to 126.6 samples. Every system contributes 8 windows and d scored coordinates, i.e. 1,784 forecasts per arm for the univariate arms and 536 multivariate calls in arm M.

Table 3: The 67 included systems: dimension d , largest Lyapunov exponent $\lambda_{\max}$, dominant period T and Lyapunov time in samples $\tau = p/(\lambda_{\max}T)$ at $p = 100$ points per period.

System	d	$\lambda_{\max}$	T	τ	System	d	$\lambda_{\max}$	T	τ
ArnoldBeltramiChildress	3	0.367	2.85	95.7	ArnoldWeb	5	0.065	67.25	22.8
AtmosphericRegime	3	0.011	89.59	105.9	BeerRNN	3	0.039	181.58	14.2
BelousovZhabotinsky	3	132.088	0.04	19.8	Blasius	3	0.126	6.60	120.2
BlinkingRotlet	3	2.120	2.95	16.0	BlinkingVortex	3	2.863	2.76	12.6
BurkeShaw	3	2.325	0.78	54.9	CellCycle	6	0.010	204.71	48.0
CellularNeuralNetwork	3	2.326	3.20	13.4	Chen	3	2.006	0.59	84.9
Chua	3	1.230	1.42	57.1	CoevolvingPredatorPrey	3	0.007	122.77	111.2
Colpitts	3	0.374	4.27	62.7	Coulet	3	0.137	5.77	126.4
Dadras	3	0.406	3.29	74.9	DoubleGyre	3	0.122	13.95	58.8
DoublePendulum	4	3.985	4.02	6.2	Duffing	3	0.140	7.45	95.7
Finance	3	0.094	9.89	108.1	Halvorsen	3	0.696	1.49	96.4
HastingsPowell	3	0.013	135.19	59.0	Hopfield	6	0.568	10.87	16.2
HyperCai	4	1.678	0.61	97.5	HyperPang	4	1.447	0.84	81.9
HyperQi	4	12.787	0.06	126.4	HyperWang	4	1.293	0.65	119.3
HyperXu	4	0.982	1.13	89.9	HyperYan	4	1.395	0.91	78.6
InteriorSquirmers	3	0.845	28.82	4.1	ItikBanksTumor	3	0.036	49.74	55.5
JerkCircuit	3	0.292	19.45	17.6	LidDrivenCavityFlow	3	2.173	2.73	16.8
LiuChen	3	0.246	8.02	50.8	Lorenz	3	0.892	1.50	74.7
Lorenz96	4	1.336	2.20	34.1	LorenzBounded	3	0.712	1.37	102.3
LorenzCoupled	6	1.028	1.62	60.0	LorenzStenflo	4	0.381	3.89	67.6
MooreSpiegel	3	0.698	1.29	111.2	MultiChua	3	1.462	2.26	30.3
NewtonLiepnik	3	0.149	8.51	78.6	NuclearQuadrupole	4	0.296	6.02	56.2
PanXuZhou	3	0.613	1.78	91.5	Qi	4	2.192	0.41	111.8
QiChen	3	4.028	0.39	63.4	RayleighBenard	3	1.814	0.92	59.9
Rosler	3	0.151	5.91	112.4	Rucklidge	3	0.194	6.07	85.1
Sakarya	3	0.265	4.98	75.7	SaltonSea	3	1.070	0.97	96.6
SanUmSrisuchinwong	3	0.261	7.47	51.4	SprottB	3	0.207	11.05	43.8
SprottC	3	0.158	8.89	71.2	SprottF	3	0.118	19.24	43.9
SprottG	3	0.091	8.71	126.6	SprottH	3	0.131	20.23	37.7
SprottJerk	3	0.088	11.92	95.5	SprottMore	3	1.095	6.62	13.8
SprottS	3	0.274	3.35	109.0	SprottTorus	3	0.106	13.61	69.5
Thomas	3	0.632	4.97	31.8	ThomasLabyrinth	3	1.391	6.05	11.9
VallisElNino	3	0.548	2.21	82.7	WindmiReduced	3	0.404	4.67	53.0
YuWang	3	1.853	1.34	40.4					

C Per-model input conventions and decoding

All models are run zero-shot with package defaults and inputs in raw units; every model applies its own per-series normalisation. Contexts have $L = 512$ points and forecasts $H = 512$ points. No arm truncates the context: Chronos-2 allows 8192 points per variate, TimesFM-3 and Toto keep the sequence length at 512 and place the d variates on a separate axis, and the one model that flattens variates into a single token sequence (Moirai-2) is run only in arm U. Versions: chronos-forecasting 2.3.2, timesfm 3.0.2, toto-ts 0.2.0, uni2ts 2.0.0, transformers 5.17.0, torch 2.11.0+cu128.

- **Chronos-2.** Arms M and C pass a tensor of shape $[B, d, L]$ (variates along the second axis), arm U one of shape $[B, 1, L]$, with up to $B = 64$ independent windows per call. The scored series is the 0.5 quantile of the target variate. The horizon of 512 lies within the model’s direct multi-patch output (output patch 16, up to 64 patches), so no autoregressive rollout is needed. Cross-item learning is left at its package default (`cross_learning=False`), so the items batched into a call do not attend to one another.
- **TimesFM-3.0.** Arms M and C pass a two-dimensional context of shape $[d, L]$ as one multivariate series; arm U passes a one-dimensional context. The point forecast is the scored output; the horizon is decoded in a single non-autoregressive pass.
- **Toto-Open-Base-1.0.** Variates are placed in the channel dimension, $[1, d, L]$, so that they attend jointly; independent series are never batched into one item. The scored series is the median of 64 samples; the horizon is produced in patches of 64 with the model’s internal autoregressive loop. The checkpoint is loaded with `snapshot_download` followed by `load_from_checkpoint`, because `from_pretrained` is incompatible with `huggingface_hub` 1.x.
- **Moirai-2.0-R-small.** Arm U uses context length 512, prediction length 512 and target dimension 1. Moirai-2 has no cross-variate path: its authors dropped multivariate and covariate support and treat multivariate forecasting as independent univariate tasks [Liu et al., 2025b]; correspondingly the `target_dim>1` route of `uni2ts` 2.0.0 raises a shape error at prediction length 512, and we did not patch it. Moirai-2 therefore appears in arm U only.
- **Chronos-Bolt-base** (arm U only). Univariate context $[1, L]$; the 0.5 quantile is scored; beyond its native 64-step output the horizon is filled by the package’s autoregressive continuation.

D Hyperparameters, pre-registered constants and deviations

- Windows: $L = 512, H = 512$; eight origins $o_k = 512 + \lfloor k(3584 - 512)/7 \rfloor, k = 0, \dots, 7$; far segments $[5000 + 512k, 5512 + 512k)$.
- Inclusion: $\tau \in [4, 128]$ samples, $d \leq 6$, coordinate standard deviation $> 10^{-6} \times$ the largest coordinate’s, finite values only.
- Analog forecasters: embedding dimension 4 and delay 1 for U-analog; the last two points of all d coordinates for M-analog and C-analog; Euclidean distance; re-search from the current predicted state when a copied segment reaches the end of the context.
- NVAR: features $[1, s_t, s_{t-1}, s_{t-2}]$ plus all quadratic monomials of (s_t, s_{t-1}) ; ridge parameter selected on the last 64 context points from $\{10^{-6}, 10^{-5}, \dots, 10^0\}$; 512-step rollout with clipping at ± 10 context standard deviations.
- Metric: normalised error with respect to the context standard deviation; centred 5-step moving average; threshold 0.5; censoring at 512 samples; VPT in Lyapunov times.
- Statistics: 95% percentile bootstrap over systems with $B = 10,000$ resamples (seed 20260916); Wilcoxon signed-rank test; Holm correction across the three primary models.
- Toto sampling: 64 samples, median, package default seed.

How well the arm-C segment matches the window it replaces. The far segment is not a distributional copy of the window, and we do not claim it is; the relevant benchmark is how much two segments of the *same* run differ anyway. Averaged over all systems, windows and coordinates, the far segment differs from its window by 0.52 context standard deviations in mean and 24% in standard deviation, against 0.31 and 20% for a different window of the same trajectory; the normalised power spectra differ by a total-variation distance of 0.40 (other window: 0.39) with a log-spectrum correlation of 0.937 (other window: 0.939). Spectrally the control is therefore as good a match as any other segment of the same run, while its marginals drift somewhat further; since every model normalises per series, and since C-analog collapses exactly as a decoupled control should, we read Δ_2 as a coupling effect, but a residual distribution-shift component cannot be excluded.

Deviations from the frozen design. (i) The pre-registered recovery ratio is defined against M-analog, whose mean coupling gain turned out to be negative (-0.15) although its median is positive and it wins on 81% of systems; a ratio against a near-zero denominator is unstable, so the main table reports R against NVAR-M and Table 4 gives both. (ii) Moirai-2 was pre-registered as a primary model conditional on its multivariate path running; it has none by design (Appendix C), so it is reported for arm U only and three primary models remain. (iii) No other deviation: the system list, windows, arms, metric, decision rule and statistics are as frozen before the first model call.

E Additional tables

Table 4: Secondary statistics. Win rates and medians are over the 67 systems; p is Holm-corrected for the three primary models and raw for the baselines. R_{analog} and R_{NVAR} are the recovery ratios against the two classical references; ρ is the Spearman correlation between the analog coupling gain and the model’s Δ_1 (p in brackets).

Model	win Δ_1	win Δ_2	med Δ_1	med Δ_2	$p(\Delta_1)$	R_{analog}	R_{NVAR}	ρ
Chronos-2	58%	79%	+0.016	+0.075	6.0e-01	9.6%	0.6%	0.24 (0.05)
TimesFM-3	75%	67%	+0.044	+0.032	5.6e-05	37.2%	3.8%	0.21 (0.09)
Toto	60%	57%	+0.004	+0.002	3.2e-01	6.7%	1.3%	0.17 (0.18)
Analog	81%	94%	+0.175	+0.768	1.0e-04	–	–	–
NVAR	84%	–	+1.754	–	1.5e-07	–	–	–

Table 5: Robustness of the primary contrasts to the metric constants. Rows vary the error threshold and the smoothing window; the pre-registered setting is 0.5 and 5.

thr.	smooth	Chronos-2		TimesFM-3		Toto	
		Δ_1	Δ_2	Δ_1	Δ_2	Δ_1	Δ_2
0.3	5	−0.20 [−0.47, −0.01]	+0.06 [−0.01, +0.12]	+0.34 [+0.03, +0.92]	+0.06 [+0.02, +0.10]	−0.00 [−0.02, +0.01]	+0.03 [−0.02, +0.09]
0.5	1	−0.12 [−0.35, +0.04]	+0.11 [+0.06, +0.17]	+0.34 [+0.06, +0.85]	+0.07 [+0.03, +0.11]	+0.05 [+0.00, +0.10]	+0.05 [+0.01, +0.09]
0.5	5	−0.05 [−0.14, +0.03]	+0.12 [+0.07, +0.18]	+0.22 [+0.07, +0.48]	+0.06 [+0.02, +0.09]	+0.04 [+0.00, +0.10]	+0.06 [+0.01, +0.14]
0.5	9	−0.05 [−0.15, +0.03]	+0.12 [+0.07, +0.17]	+0.23 [+0.06, +0.51]	+0.06 [+0.03, +0.09]	+0.04 [+0.01, +0.09]	+0.05 [+0.00, +0.12]
0.7	5	−0.02 [−0.11, +0.05]	+0.14 [+0.07, +0.23]	+0.15 [+0.07, +0.26]	+0.04 [−0.00, +0.08]	+0.03 [+0.00, +0.05]	+0.03 [−0.01, +0.08]

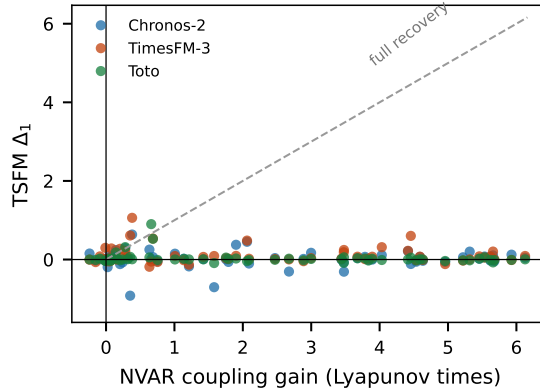


Figure 2: Per-system coupling gain of NVAR ($\text{VPT}(\text{NVAR-M}) - \text{VPT}(\text{NVAR-U})$) against each primary model’s Δ_1 , one marker per system and model; the dashed line is what full recovery of the classically available gain would look like. The models stay on the zero line across the whole range. The axis is clipped at the 97th percentile (6.2 Lyapunov times); two systems lie beyond it and also have model gains near zero.

F Reproducibility

Everything the study consumes is public and pinned: the trajectories are the fine split of the Hugging Face dataset williamgilpin/dysts; the Lyapunov exponents and dominant periods are the dysts package metadata; the checkpoints are

Table 6: Per-coordinate valid prediction time (Lyapunov times) for four named systems, first three coordinates, averaged over the eight windows. The last column is the mean coupling gain over those coordinates.

Model	U	M	C	mean Δ_1
<i>Lorenz ($d = 3$, $\tau = 74.7$ samples)</i>				
Chronos-2	0.29, 0.36, 0.67	0.40, 0.31, 0.52	0.32, 0.27, 0.62	-0.03
TimesFM-3	0.31, 0.23, 0.72	0.34, 0.30, 0.86	0.29, 0.32, 0.59	0.08
Toto	0.28, 0.20, 0.66	0.29, 0.21, 0.60	0.21, 0.20, 0.59	-0.01
Analog	0.46, 0.50, 0.80	0.80, 0.69, 0.85	0.23, 0.19, 0.51	0.19
NVAR	0.20, 0.17, 0.51	6.36, 5.73, 5.64	—	5.62
<i>Rosler ($d = 3$, $\tau = 112.4$ samples)</i>				
Chronos-2	1.82, 1.70, 0.70	1.20, 1.41, 0.68	1.13, 1.03, 0.68	-0.31
TimesFM-3	1.33, 1.80, 0.57	1.64, 2.11, 0.68	2.18, 2.18, 0.67	0.24
Toto	1.02, 1.04, 0.45	1.02, 1.21, 0.46	1.05, 1.14, 0.46	0.06
Analog	2.86, 3.07, 0.69	3.46, 3.65, 0.93	2.29, 2.11, 0.78	0.47
NVAR	0.83, 0.99, 0.12	4.15, 4.15, 4.07	—	3.48
<i>Chua ($d = 3$, $\tau = 57.1$ samples)</i>				
Chronos-2	1.22, 0.68, 0.60	1.27, 0.88, 1.07	1.14, 0.64, 0.87	0.24
TimesFM-3	1.07, 1.46, 0.97	0.91, 1.28, 1.34	0.92, 1.33, 1.26	0.01
Toto	1.07, 0.58, 0.96	1.26, 0.97, 0.92	0.97, 0.89, 0.91	0.18
Analog	1.02, 1.44, 1.28	1.44, 1.31, 1.53	0.38, 0.22, 0.40	0.18
NVAR	0.88, 1.11, 1.02	1.08, 1.08, 1.28	—	0.14
<i>Halvorsen ($d = 3$, $\tau = 96.4$ samples)</i>				
Chronos-2	0.22, 0.24, 0.26	0.24, 0.30, 0.18	0.34, 0.36, 0.21	0.00
TimesFM-3	0.27, 0.28, 0.29	0.33, 0.28, 0.29	0.36, 0.29, 0.29	0.02
Toto	0.34, 0.24, 0.44	0.35, 0.24, 0.46	0.36, 0.26, 0.45	0.01
Analog	0.22, 0.20, 0.20	0.48, 0.41, 0.44	0.21, 0.19, 0.10	0.23
NVAR	0.19, 0.24, 0.19	3.66, 3.65, 3.63	—	3.44

amazon/chronos-2, google/timesfm-3.0-pytorch, Datadog/Toto-Open-Base-1.0, Salesforce/moirai-2.0-R-small and amazon/chronos-bolt-base at their released weights. The inclusion rule, window origins, far-segment offsets, arm construction, baseline hyperparameters, metric constants and bootstrap seed in Appendices B–D determine every input exactly; no model is trained or fine-tuned and all decoding is deterministic apart from Toto’s 64 samples, whose median we score. Code will be released with the camera-ready version.

G Influence and censoring

VPT is censored at $H = 512$ samples, i.e. at H/τ Lyapunov times, which ranges from 4.0 (SprottG) to 124.7 (InteriorSquirm) with median 8.1; a mean over systems therefore averages quantities with very different ceilings. Censoring is common only for the strongest baseline: it affects 1.6–5.0% of each foundation model’s forecasts and 4.2% of NVAR-U’s, but 20.2% of NVAR-M’s, so NVAR’s coupling gain — the denominator of every recovery ratio — is if anything understated. Capping every VPT at a common 4 Lyapunov times before taking the contrast gives Δ_1 of +0.015 [−0.037, +0.065] (Chronos-2), +0.104 [+0.059, +0.156] (TimesFM-3), +0.017 [−0.000, +0.039] (Toto), +0.161 [+0.061, +0.250] (M-analog) and +1.668 [+1.293, +2.042] (NVAR): every model moves to a small positive value, the classical reference stays an order of magnitude above, and the conclusion is unchanged. Single systems matter: 54% of Chronos-2’s Δ_1 comes from BlinkingVortex (−1.69), 50% of TimesFM-3’s from InteriorSquirm (+7.53) and 39% of Toto’s from BlinkingVortex (+1.15). Dropping each model’s most influential system leaves −0.02 [−0.09, +0.04], +0.11 [+0.06, +0.17] and +0.03 [−0.00, +0.07]: TimesFM-3’s gain survives at half its size, Toto’s interval then touches zero, and Chronos-2 stays null.

H Compute

All inference ran on one 24 GB consumer laptop GPU. The eleven (model, arm) runs took 11 minutes of wall clock in total: Chronos-2 2 s, 2 s, 6 s for arms U, M, C; TimesFM-3 20 s, 12 s, 37 s; Toto 77 s, 78 s, 267 s; Chronos-Bolt 44 s and Moirai-2 26 s in arm U (inference only, excluding model loading), at a peak of 7.0 GB. The classical baselines take 10 s on 8 CPU cores and the analysis, including

10,000 bootstrap resamples for every contrast, under a minute. No forecast contained a non-finite value and no output series was constant.

Table 7: Per-system coupling gain $\Delta_1 = \text{VPT(M)} - \text{VPT(U)}$ in Lyapunov times for the three primary models and the two classical references.

System	d	τ	Chronos-2	TimesFM-3	Toto	Analog	NVAR
ArnoldBeltramiChildress	3	95.7	-0.02	-0.07	-0.01	-0.22	-0.15
ArnoldWeb	5	22.8	-0.18	-0.13	0.00	-1.44	1.21
AtmosphericRegime	3	105.9	0.02	0.01	-0.04	-0.36	-0.01
BeerRNN	3	14.2	-0.19	-0.04	-0.00	0.29	0.02
BelousovZhabotinsky	3	19.8	0.64	1.06	0.01	0.34	0.38
Blasius	3	120.2	-0.00	0.07	0.02	0.34	3.89
BlinkingRotlet	3	16.0	-0.54	0.32	0.00	-2.16	-4.95
BlinkingVortex	3	12.6	-1.69	0.71	1.15	-5.06	-10.67
BurkeShaw	3	54.9	0.12	-0.01	-0.01	0.54	5.93
CellCycle	6	48.0	-0.12	0.27	-0.02	-0.05	0.20
CellularNeuralNetwork	3	13.4	0.06	0.53	0.53	0.26	0.69
Chen	3	84.9	0.06	-0.03	-0.02	0.01	5.22
Chua	3	57.1	0.24	0.01	0.18	0.18	0.14
CoevolvingPredatorPrey	3	111.2	0.02	0.06	0.01	0.02	0.32
Colpitts	3	62.7	0.38	0.09	0.03	0.07	1.90
Coullet	3	126.4	-0.07	0.04	0.00	0.31	0.24
Dadras	3	74.9	0.02	-0.02	0.02	0.06	1.14
DoubleGyre	3	58.8	-0.02	-0.01	0.01	-2.11	-0.62
DoublePendulum	4	6.2	0.29	0.22	0.32	1.58	0.28
Duffing	3	95.7	-0.25	0.66	0.02	0.18	-0.91
Finance	3	108.1	-0.06	0.02	0.01	0.24	1.79
Halvorsen	3	96.4	0.00	0.02	0.01	0.23	3.44
HastingsPowell	3	59.0	0.45	0.48	-0.04	0.45	2.06
Hopfield	6	16.2	-0.92	0.61	0.07	1.87	0.35
HyperCai	4	97.5	0.07	0.07	0.01	-0.11	4.59
HyperPang	4	81.9	0.08	0.05	0.05	0.16	5.54
HyperQi	4	126.4	0.01	0.03	0.05	0.03	1.75
HyperWang	4	119.3	0.07	0.01	0.03	-0.02	3.79
HyperXu	4	89.9	-0.03	-0.01	0.00	0.02	4.63
HyperYan	4	78.6	0.20	0.00	-0.02	0.04	5.32
InteriorSquirmer	3	4.1	-0.37	7.53	0.00	-11.02	-6.51
ItikBanksTumor	3	55.5	0.15	0.09	0.00	0.13	1.01
JerkCircuit	3	17.6	0.08	0.28	-0.02	2.24	0.08
LidDrivenCavityFlow	3	16.8	0.01	-0.05	0.90	-5.73	0.66
LiuChen	3	50.8	0.15	-0.00	0.00	0.15	-0.24
Lorenz	3	74.7	-0.03	0.08	-0.01	0.19	5.62
Lorenz96	4	34.1	0.02	-0.04	-0.02	0.13	2.89
LorenzBounded	3	102.3	0.01	0.07	0.00	0.16	1.42
LorenzCoupled	6	60.0	0.11	0.31	-0.01	0.18	4.03
LorenzStenflo	4	67.6	-0.06	-0.02	-0.02	0.01	4.54
MooreSpiegel	3	111.2	0.04	0.01	-0.04	0.46	0.06
MultiChua	3	30.3	0.25	-0.18	0.05	1.26	0.63
NewtonLiepnik	3	78.6	-0.01	0.02	-0.03	0.06	3.84
NuclearQuadrupole	4	56.2	0.01	0.08	0.02	0.52	-0.09
PanXuZhou	3	91.5	-0.70	0.09	-0.09	-0.03	1.58
Qi	4	111.8	-0.10	0.02	-0.01	0.12	2.09
QiChen	3	63.4	0.02	0.08	0.02	0.19	5.46
RayleighBenard	3	59.9	0.05	0.09	0.00	0.64	6.13
Rossler	3	112.4	-0.31	0.24	0.06	0.47	3.48
Rucklidge	3	85.1	-0.04	0.00	-0.01	0.23	5.68
Sakarya	3	75.7	0.04	0.01	-0.01	0.02	2.47
SaltonSea	3	96.6	-0.11	0.60	0.05	0.54	4.46
SanUmSrisuchinwong	3	51.4	-0.04	-0.06	-0.01	0.32	0.75
SprottB	3	43.8	0.10	0.20	-0.07	0.09	5.66
SprottC	3	71.2	0.21	0.22	0.00	0.25	4.41
SprottF	3	43.9	0.19	0.17	0.04	0.22	7.83
SprottG	3	126.6	-0.31	0.00	0.02	0.54	2.67
SprottH	3	37.7	0.02	0.17	0.05	0.13	9.02
SprottJerk	3	95.5	0.17	0.03	0.01	0.28	3.00
SprottMore	3	13.8	-1.37	-0.57	-0.21	-0.05	-2.49
SprottS	3	109.0	0.04	0.07	0.03	0.08	3.68
SprottTorus	3	69.5	-0.07	0.19	-0.08	0.21	3.48
Thomas	3	31.8	0.11	0.04	0.00	0.42	0.05
ThomasLabyrinth	3	11.9	0.04	-0.05	0.02	0.07	0.06
VallisElNino	3	82.7	-0.06	-0.12	-0.04	0.20	4.96
WindmiReduced	3	53.0	-0.02	0.30	-0.03	0.21	-0.01
YuWang	3	40.4	0.07	0.00	0.01	0.28	0.16