Adaptive Error-Bounded Compression with Warm-Start Randomized SVD Hurricane Isabel, NYX, and Miranda

Ahmer Nadeem Khan

HDPS Lab

24 June 2026

Outline

Outline

Cold-Start rSVD

Algorithm Cold-Start Randomized SVD

Require: $A \in \mathbb{R}^{m \times n}$, rank k, oversampling p_c
Ensure: $U \in \mathbb{R}^{m \times k}$, $\mathbf{s} \in \mathbb{R}^{k}$, $V^T \in \mathbb{R}^{k \times n}$

1: Draw $\Omega \in \mathbb{R}^{n \times (k+p_c)} \sim \mathcal{N}(0,1)$ [sketch width $= k + p_c$]
2: $Y \leftarrow A\Omega$ [range approximation]
3: $Q, \_ \leftarrow \mathrm{QR}(Y)$
4: $B \leftarrow Q^T A$
5: $\hat{U}, \mathbf{s}, V^T \leftarrow \mathrm{SVD}(B)$
6: $U \leftarrow Q\hat{U}$; truncate to first k columns

Warm-Start rSVD

Algorithm Warm-Start Randomized SVD

Require: $A \in \mathbb{R}^{m \times n}$, **prior basis** $U_{\text{prev}} \in \mathbb{R}^{m \times k}$, rank k, p_w
Ensure: $U \in \mathbb{R}^{m \times k}$, $\mathbf{s} \in \mathbb{R}^{k}$, $V^T \in \mathbb{R}^{k \times n}$

1: $G \leftarrow A^T U_{\text{prev}}$ [warm projection]
2: $Y_1 \leftarrow AG$ [exploit prior subspace, width k]
3: Draw $\Omega \in \mathbb{R}^{n \times p_w} \sim \mathcal{N}(0, 1)$
4: $Y_2 \leftarrow A\Omega$ [fresh exploration, width p_w]
5: $Y \leftarrow [Y_1 \mid Y_2]$ [sketch width $= k + p_w < k + p_c$]
6: $Q, \_ \leftarrow \text{QR}(Y)$; $B \leftarrow Q^T A$; SVD; lift; truncate

The Three Datasets

	Isabel	**Miranda**	**NYX**
Domain	WRF hurricane	Turbulent mixing (RT)	Cosmology
Raw volume	$100 \times 500 \times 500$	$384 \times 384 \times 256$	$512 \times 512 \times 512$
Matrix	250000×100	147456×256	262144×512
Dtype	float32	float64 $\rightarrow$ f32	float32
Variables	13 (4 featured)	7	6
Time axis	**48 steps**	single	single
Warm-start	**yes**	cold only	cold only

- Matricised identically: outermost axis $\rightarrow$ columns, the two inner spatial axes flattened $\rightarrow$ rows (tall, low-rank-friendly).

Outline

Adaptive Two-Layer Decomposition: $L + S$

Given a user tolerance $\tau > 0$, each snapshot is stored as:

$$\hat{A}_t = \underbrace{U_{k^*}\,\mathrm{diag}(\mathbf{s}_{k^*})\,V_{k^*}^T}_{L\ (\text{adaptive rank } k^*)} + \underbrace{S}_{\text{sparse corrections}}$$

L — Low-rank layer. *One* rSVD (warm or cold), truncated at the cost-optimal rank k^*. Captures all exploitable smooth structure.

S — Sparse layer. Stores (i, j, R_{ij}) for every residual entry with $|R_{ij}| > \tau$, where $R = A - L$. At decode, these entries become exact.

The rank layer's goal is **storage optimisation**: each rank added has to pay for itself by deleting sparse entries.

Storage Cost Objective

One **cost objective function** (in bytes):

$$C(k;\,\tau) \;=\; \underbrace{k\,(m+n+1)\,c_{\text{float}}}_{\text{rank factors}} \;+\; \underbrace{s_k(\tau)\cdot c_{\text{entry}}}_{\text{sparse corrections}}, \qquad k^* = \arg\min_k \; C(k;\,\tau)$$

where, in bytes:

- $s_k(\tau) = \#\{(i,j) : |A_{ij} - (A_k)_{ij}| > \tau\}$ — the "violations" at rank k
- $c_{\text{float}} = 4$
- $c_{\text{entry}} = 12$ — one sparse correction (COO: row + col + value)

C is a *model*: $c_{\text{entry}} = 12$ B is an assumed encoding.

Marginal Economics of Optimizing Rank

Isabel: one rank $\approx$ 1.00 MB

A rank pays for itself iff it removes $> 83\,000$ violations. Example (U_f, $\tau = 1.0$ m/s, $t = 2$):

- $k = 12 \to 13$: removes 95 981 violations $= 1.15$ MB sparse, costs 1.00 MB $\Rightarrow$ **take it**
- $k = 13 \to 14$: removes only 42 447 $= 0.51$ MB $\Rightarrow$ **stop** ($k^* = 13$)

The cost curve is (empirically) U-shaped and *shallow near its minimum* — single-rank precision in the search is worth real storage.

Rank Search: Coarse-to-Fine Truncation Sweep

One factorization per snapshot — all candidates are truncations of it:

1. Compute rSVD at the window top k_{hi} (sketch always oversampled: $\geq k_{\text{hi}} + p$ columns)
2. **Coarse sweep** $k \in \{1, \ldots, 8, 12, 16, \ldots\}$: maintain a running reconstruction A_k, one rank-1 term per step [$O(mn)$ per candidate]
3. **Exact prune:** once $k \cdot \text{rank\_bytes} \geq$ best cost, no larger k can win — stop
4. **Fine sweep:** every rank within ± 3 of the coarse argmin
5. **Walk-down:** keep stepping -1 while the bottom rank improves [rank collapse is free: no new SVD]
6. **Boundary rule:** argmin on the window top $\Rightarrow$ widen window, recompute rSVD, repeat

Optimality Under Unimodality

Structure of the objective:

$$C(k) = \underbrace{k\,(m+n+1)\,c_{\text{float}}}_{\text{strictly increasing}} + \underbrace{s_k(\tau)\cdot c_{\text{entry}}}_{\text{non-increasing in } k}$$

Violations s_k can only fall as rank grows (every candidate truncates the *same* factorization) — empirically $C(k)$ is U-shaped and *shallow* near its minimum.

Bracket argument: why ± 3 is enough

The coarse grid has step ≤ 4. If C is **unimodal** and the coarse argmin is k_c, then $C(k_c) \leq C(k_c \pm 4)$ forces the true minimiser into the open interval $(k_c - 4,\, k_c + 4)$ — and the fine sweep evaluates *every* integer in $[k_c - 3,\, k_c + 3]$.

Conclusion: under unimodality, k^* is the *exact* argmin over all truncations of the computed factorization.

Streaming: Warm Tracking of Rank and Subspace

Warm state carried across timesteps: exactly the k^* stored columns $U_{\text{prev}} = U_{k^*_{t-1}}$ (= reconstructible from the compressed stream itself)

- **Sketch seeding:** $Y = [\, A(A^T U_{\text{prev}}) \mid A\,\Omega\,]$ — prior subspace *plus* one implicit power iteration on the new data
- **Oversampling preserved:** random width $p_{\text{needed}} = \max(p_w,\ k_{\text{hi}} - k^*_{t-1} + p_w) \Rightarrow$ sketch always $\geq k_{\text{hi}} + p_w$ columns
- **Window centering:** search $[k^*_{t-1} - \Delta,\ k^*_{t-1} + \Delta]$ instead of $[1, k_{\max}]$ — rank evolves slowly

Caveat: when the window expands past k^*_{t-1}, the sketch is padded with more random columns.

Outline

Datasets: One Temporal, Two Static

Dataset	Physics	Variables	Matrix	Time
Isabel	WRF hurricane	4 featured†	250000×100	48 steps
NYX	cosmology	6 (all)	262144×512	single
Miranda	hydrodynamics	7 (all)	147456×256	single

† U_f (wind), TC_f (temperature), $QVAPOR_f$ (vapour), $QRAIN_f$ (rain — the sparse/unstable-rank control).

- NYX and Miranda are single-snapshot $\Rightarrow$ they exercise only the **cold bootstrap initializaion** of the algorithm.
- Isabel runs in **two arms** on identical data: `warm` (streaming, U_{prev} carried) vs `cold` (every timestep bootstrapped standalone)

Relative Tolerance

Per-snapshot value-range-relative tolerance

$$\tau_t \;=\; \varepsilon \cdot \big(\max(A_t) - \min(A_t)\big), \qquad \varepsilon \in \{10^{-2}, 10^{-3}, 10^{-4}\}$$

- Cross-dataset and warm-vs-cold comparisons share the same ε grid.
- τ changes with the data: e.g. U_f at $\varepsilon = 10^{-3}$: $\tau = 0.106 \to 0.125$ m/s over $t = 1 \ldots 3$.
- The guarantee is **per-snapshot relative**, not absolute over the sequence. A fixed physical tolerance (e.g. "1 m/s always") is the `abs` mode.

Isabel Experiment: Adaptive Warm vs Adaptive Cold

What the comparison isolates — the total value of temporal structure, which is two bundled effects:

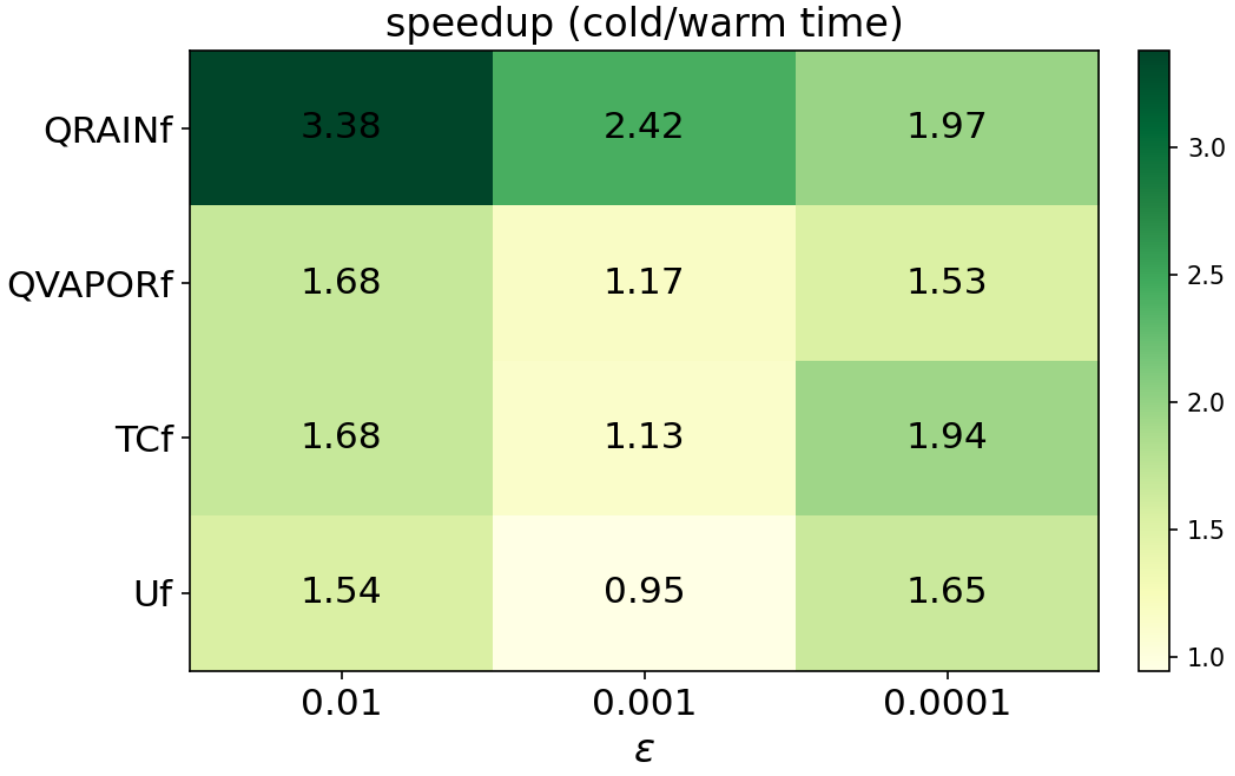
1. **Sketch seeding:** warm carries k^* informed columns from U_{prev} ($+\Delta$ window headroom) vs cold's $k_{\max}$ random columns, every step (oversampling implicit)
2. **Window centering:** $\sim$12 candidates around k^*_{t-1} vs re-searching from rank 1

- $t=1$ is bootstrap in both arms (same seed) $\Rightarrow$ rows are **identical by construction**.
- Arms may pick different k^* $\Rightarrow$ compare at the **system level**: total bytes, total time, fidelity at chosen storage
- Both arms satisfy $\|A - \hat{A}\|_{\max} \leq \tau$ always — fidelity differences appear in *storage* and *PSNR*.

Parameters: $k_{\max}=50$ (Isabel) / 32 (static), $\Delta=8$, expansion 16/32, fine radius 3, $p_c=10$, $p_w=5$, $q=0$, $c_{\text{entry}}=12$ B, seed 42. 4 vars $\times$ 3 ε $\times$ 2 arms $\times$ 48 steps = 1152 snapshots/arm.

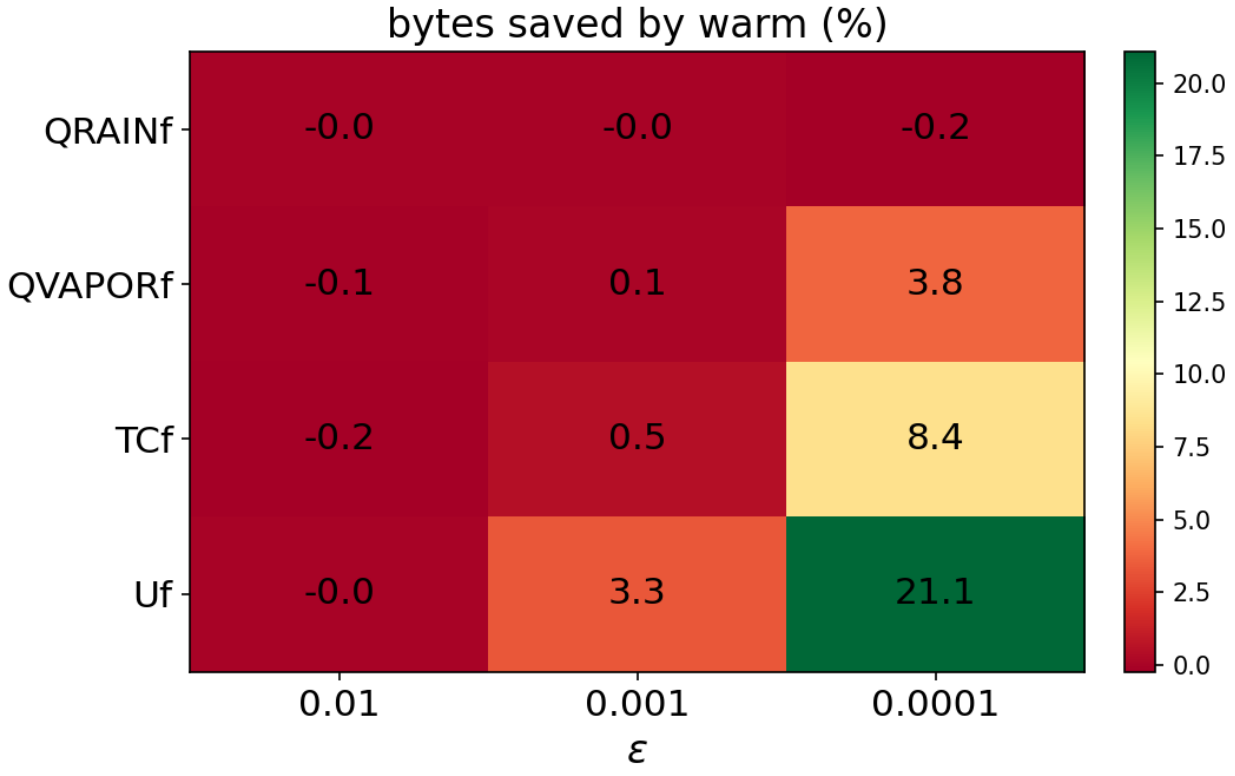
Outline

Warm-Start Advantage: Speedup



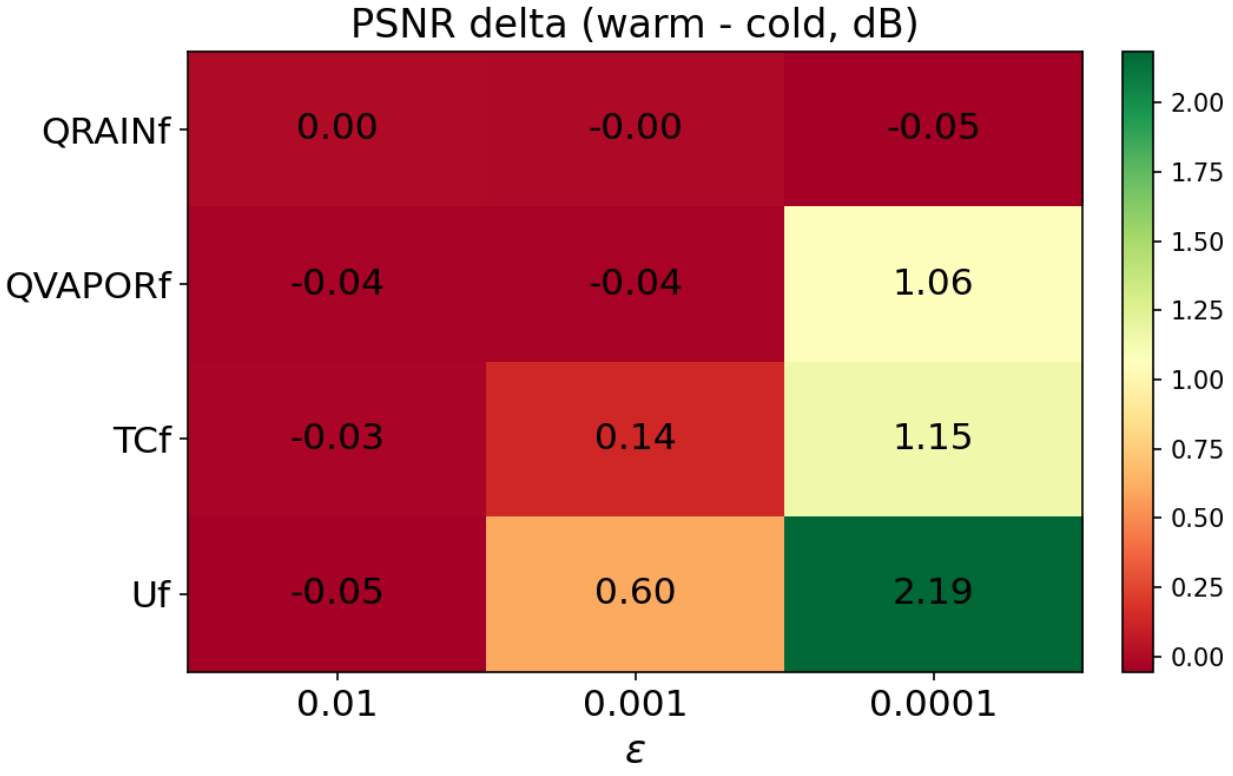
Warm is faster at *every* (var, ε) — 1.0–3.4×. Largest where k^* is small (loose ε, $QRAIN_f$): a tight search window is cheap.

Warm-Start Advantage: Bytes Saved



≈ 0 at loose/mid ε (both arms pick the same k^*); jumps at tight $\varepsilon = 10^{-4}$ on dense fields — U_f stores **21% fewer bytes**.

Warm-Start Advantage: PSNR Gain



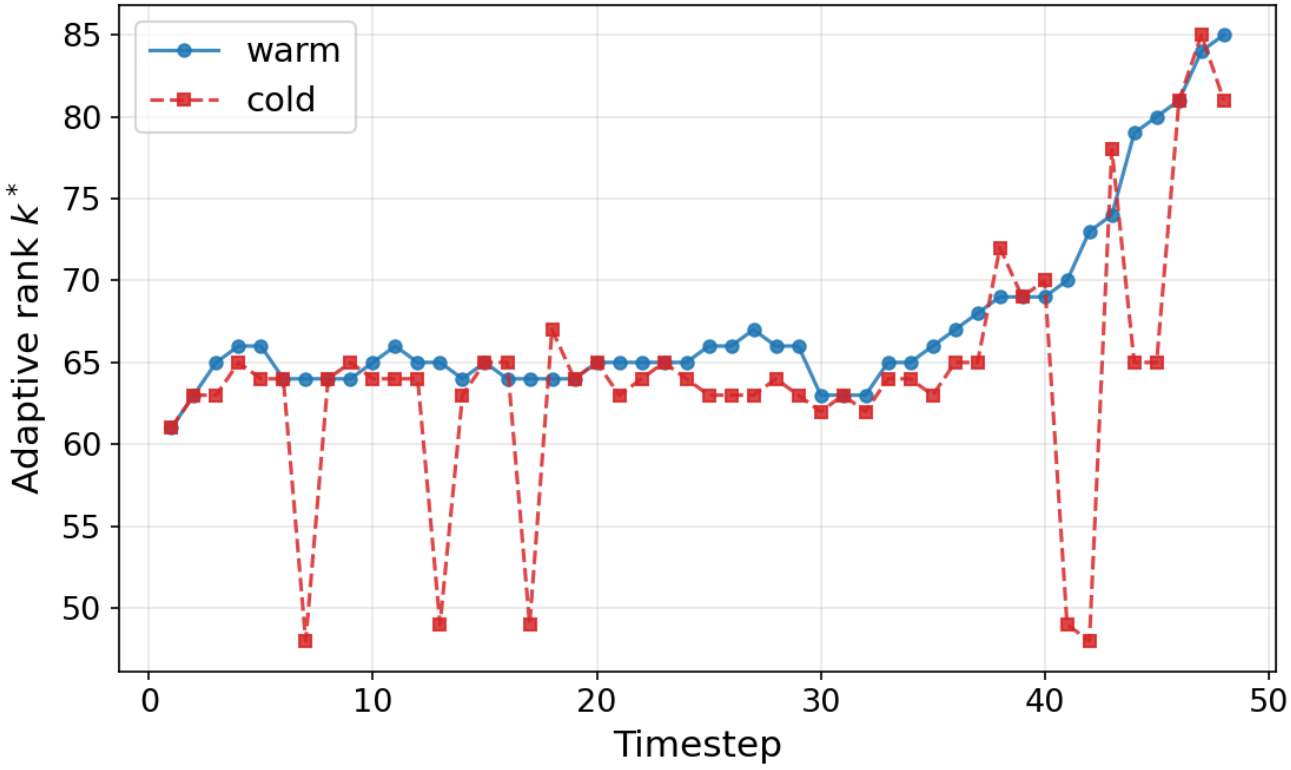
A tie at loose ε; warm gains $+1$ **to** $+2.2$ **dB** at tight $\varepsilon = 10^{-4}$, where its high-rank subspace is the more accurate one.

Warm vs Cold: Paired Deltas ($t \geq 2$)

Var	ε	Speedup	Bytes saved	Viol. reduced	ΔPSNR	Same k^*
U_f	10^{-2}	1.54×	−0.0%	−2.1%	−0.05 dB	96%
U_f	10^{-3}	0.95×	3.3%	15.3%	+0.60 dB	87%
U_f	10^{-4}	**1.66×**	**21.1%**	**70.2%**	**+2.19 dB**	21%
TC_f	10^{-2}	1.68×	−0.2%	−2.4%	−0.03 dB	94%
TC_f	10^{-3}	1.13×	0.5%	−3.9%	+0.14 dB	68%
TC_f	10^{-4}	**1.94×**	**8.4%**	**39.2%**	**+1.15 dB**	47%
$QVAPOR_f$	10^{-2}	1.68×	−0.1%	−2.1%	−0.04 dB	98%
$QVAPOR_f$	10^{-3}	1.17×	0.1%	−2.4%	−0.04 dB	89%
$QVAPOR_f$	10^{-4}	**1.53×**	**3.8%**	**34.6%**	**+1.06 dB**	53%
$QRAIN_f$	10^{-2}	**3.38×**	−0.0%	−0.0%	0.00 dB	100%
$QRAIN_f$	10^{-3}	**2.42×**	−0.0%	−0.0%	−0.00 dB	100%
$QRAIN_f$	10^{-4}	**1.97×**	−0.2%	−1.7%	−0.05 dB	98%

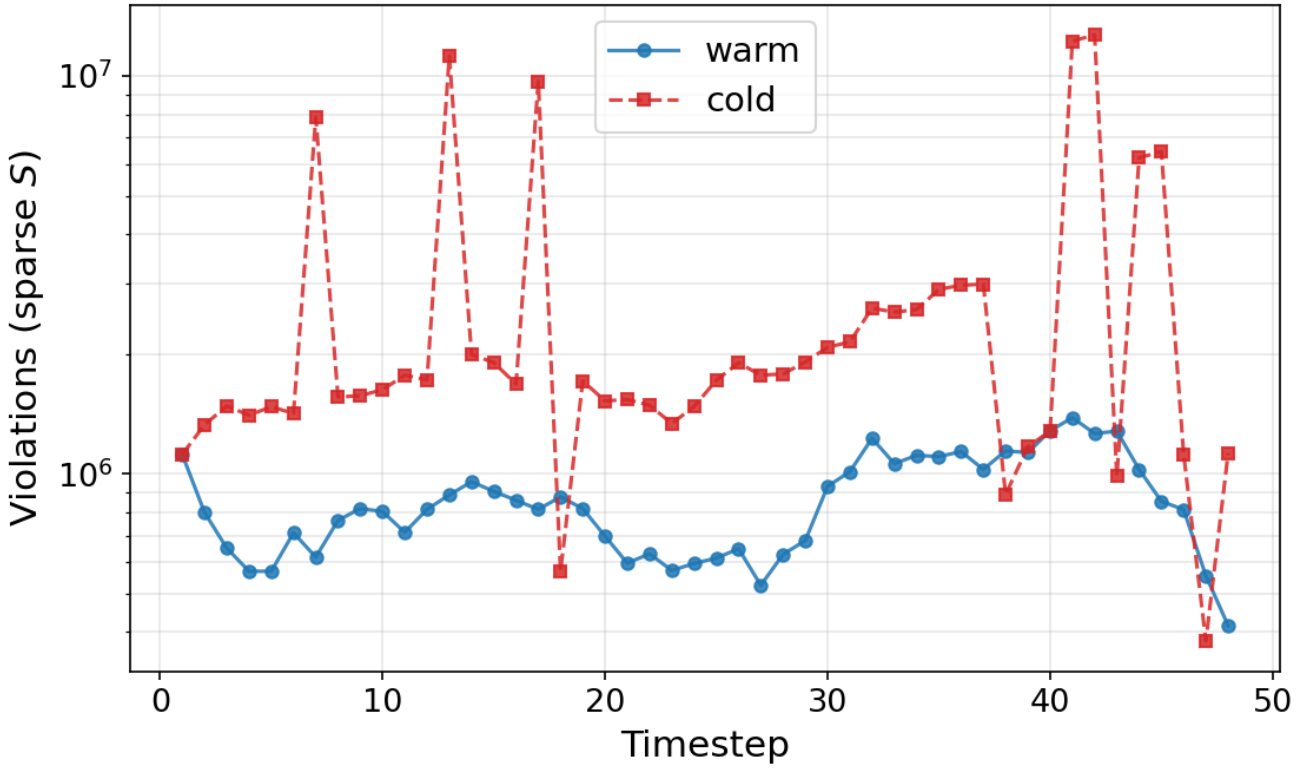
The τ guarantee holds on **all 2304 rows**, both arms.

Tight-Tolerance U_f ($\varepsilon = 10^{-4}$): Rank



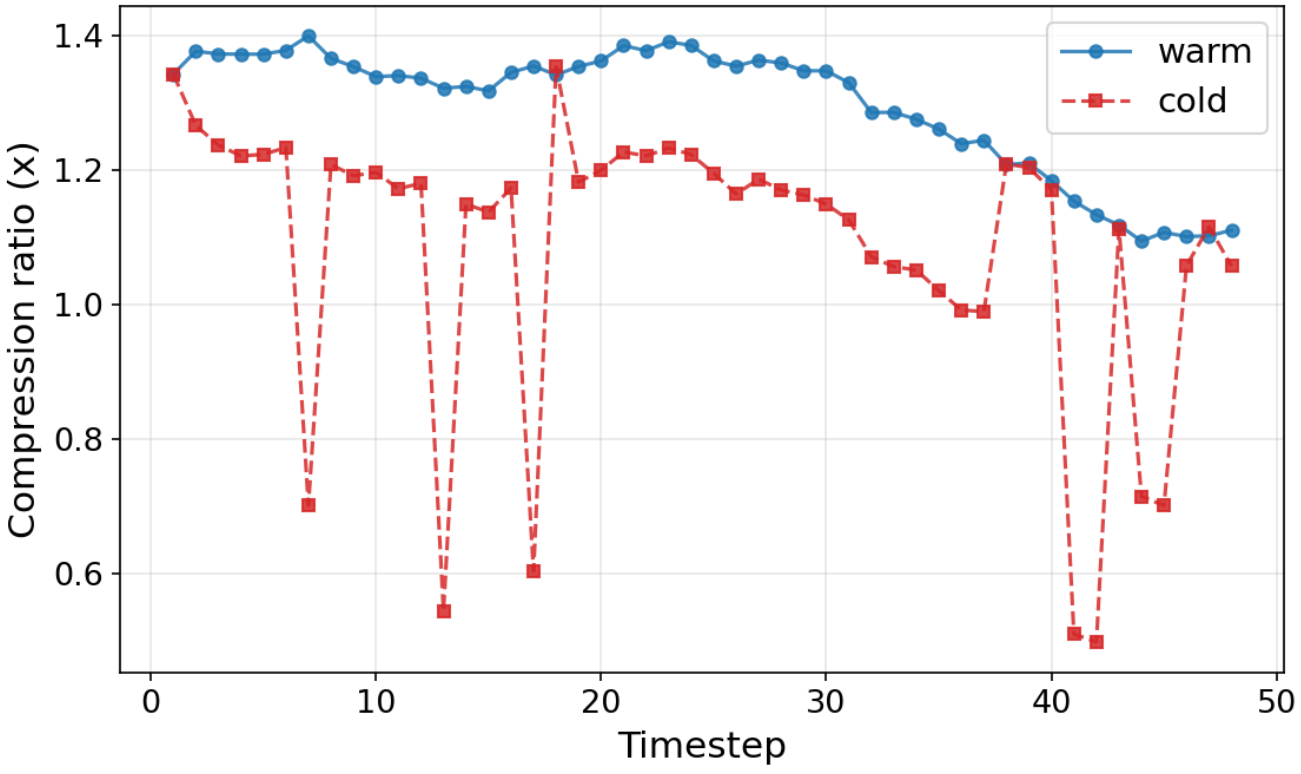
Warm's k^* tracks smoothly ($\sim$60 $\rightarrow$ 85); cold's drops to $\sim$48 at several steps — it occasionally fails to recover the high rank; Warm-start is more robust.

Tight-Tolerance U_f ($\varepsilon = 10^{-4}$): Violations



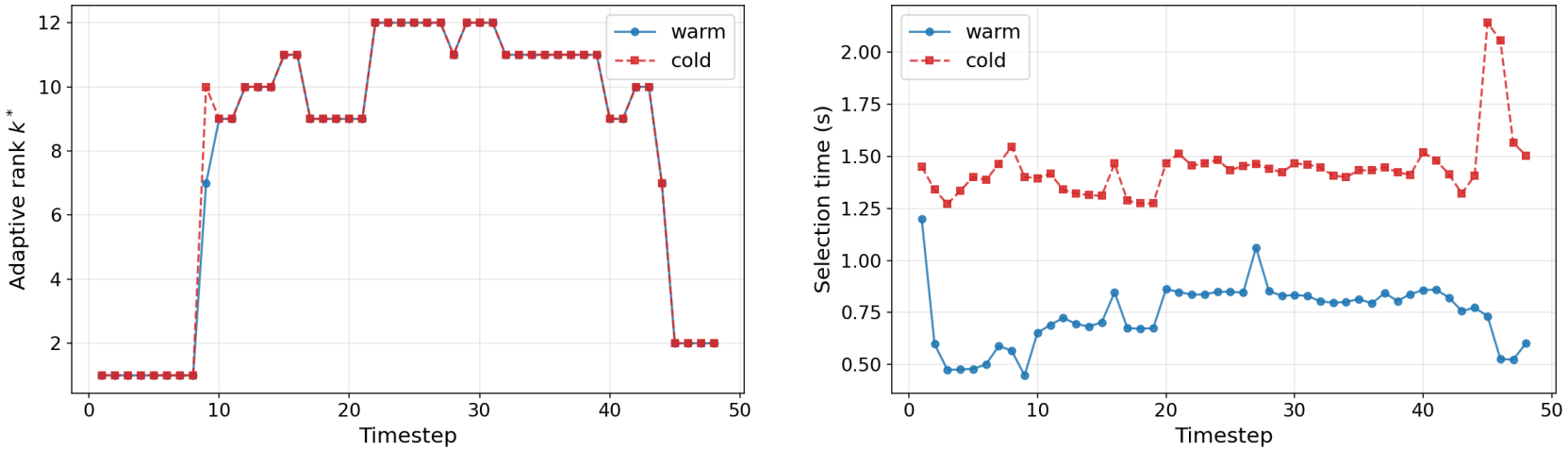
The sparse layer S: warm leaves consistently fewer entries above τ — **70% fewer violations** on average than cold.

Tight-Tolerance U_f ($\varepsilon = 10^{-4}$): Compression Ratio



Warm holds ratio ∼1.3–1.4 throughout; cold sits lower and *crashes below 1* (worse than raw) at the same steps its rank collapsed — mean **1.30 vs 1.09**, i.e. the 21% byte saving.

$QRAIN_f$: Warm speed-advantage ($\varepsilon = 10^{-4}$)



The sparse/unstable-rank control: warm and cold pick the *same* k^* every step (left, curves coincide) — *no* quality gain — yet warm runs **2–3**× **faster** (right). The speed benefit survives with no coherent subspace to reuse.

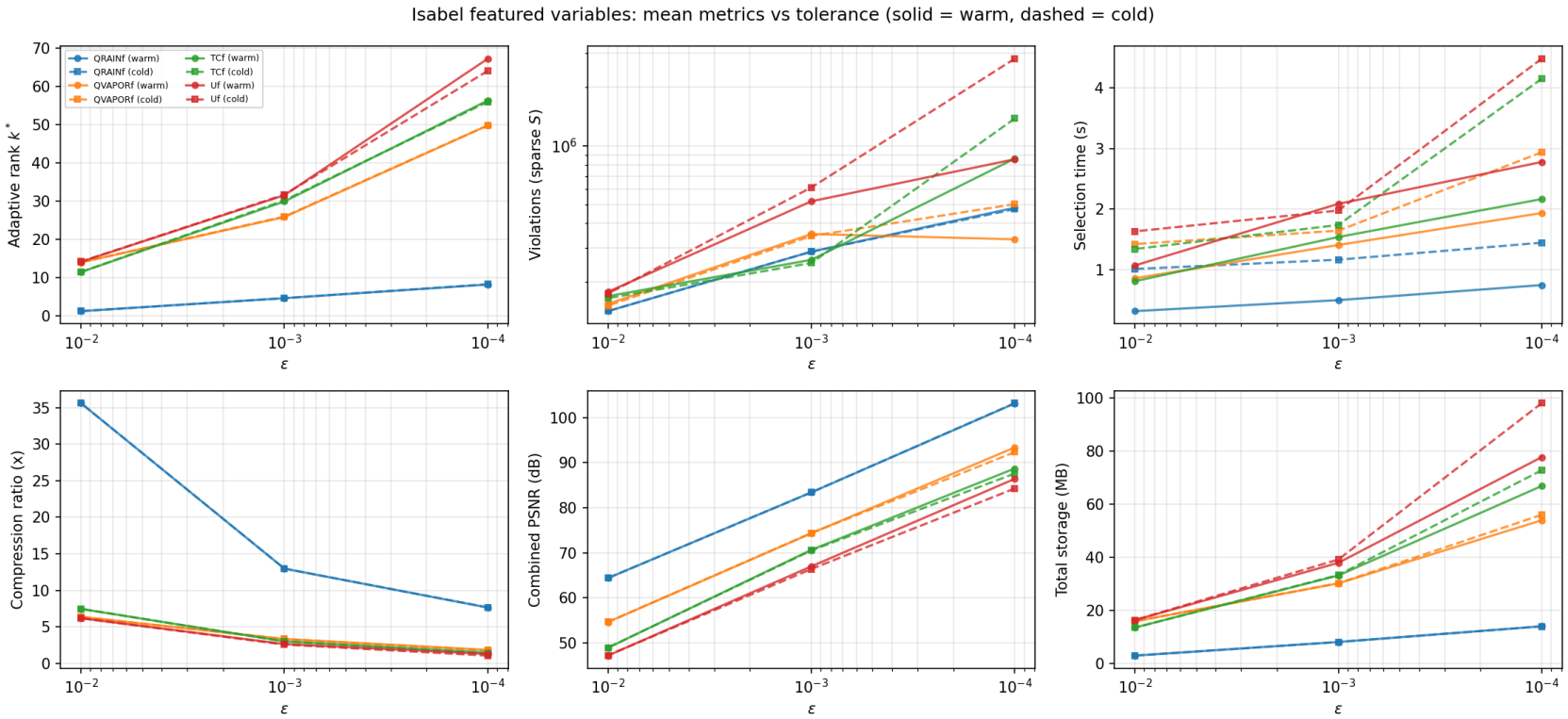
Why Warm Wins at Tight Tolerance

At $\varepsilon = 10^{-4}$ the optimal ranks ($\sim$50–67) exceed the cold bootstrap cap ($k_{\max} = 50$):

Window expansions (of 48 steps)	**Cold**	**Warm**
U_f	43	1
TC_f	44	0
$QVAPOR_f$	22	0

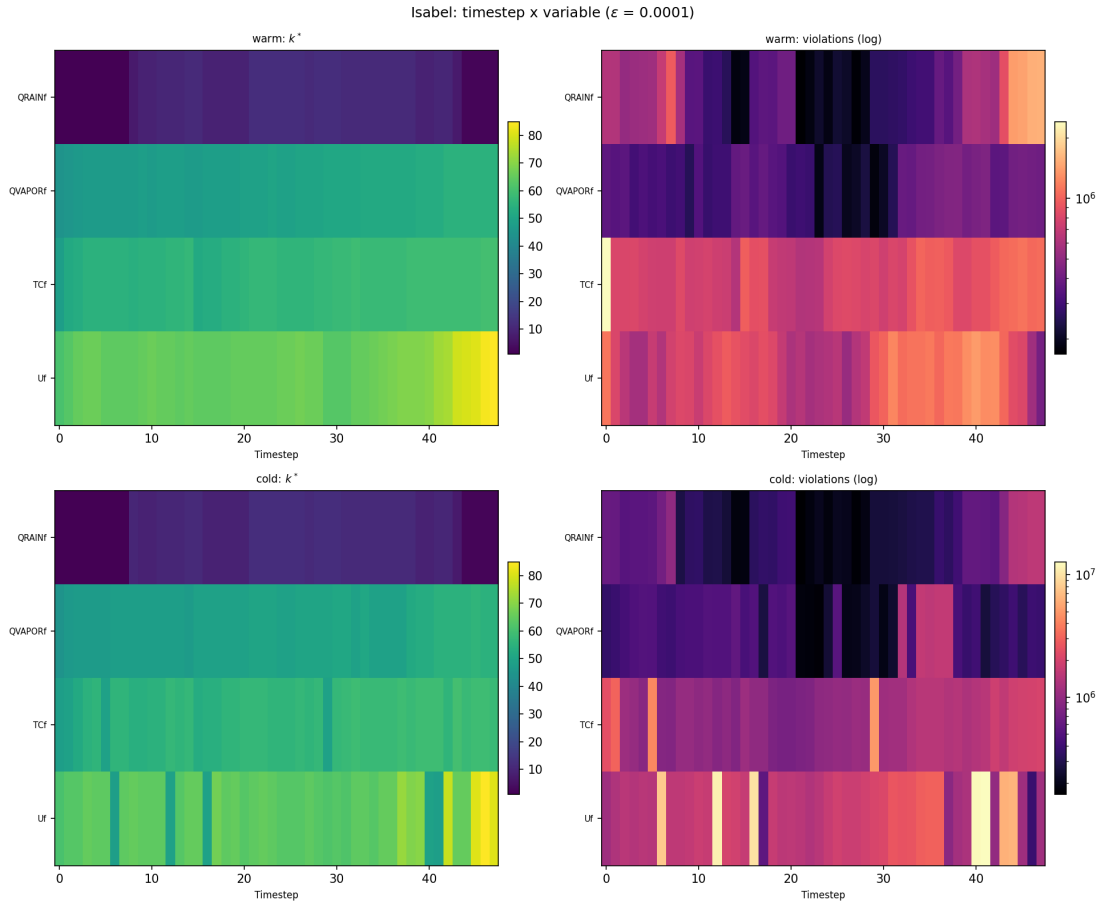
- **Cold** re-discovers the high rank from scratch every step: bootstrap, hit the cap, expand, recompute — and its trailing sketch directions are the least accurate
- **Warm** carries the rank forward: the window arrives already centred at $\sim$60, the sketch already spans the subspace
- Most striking (U_f): warm picks a *higher* mean rank (67.3 vs 64.2) yet stores **21% fewer bytes** — a better subspace deletes violations faster than the extra rank costs

Overview: All Metrics vs Tolerance



Solid = warm, dashed = cold. The arms separate only as ε tightens (rightward); selection time separates everywhere.

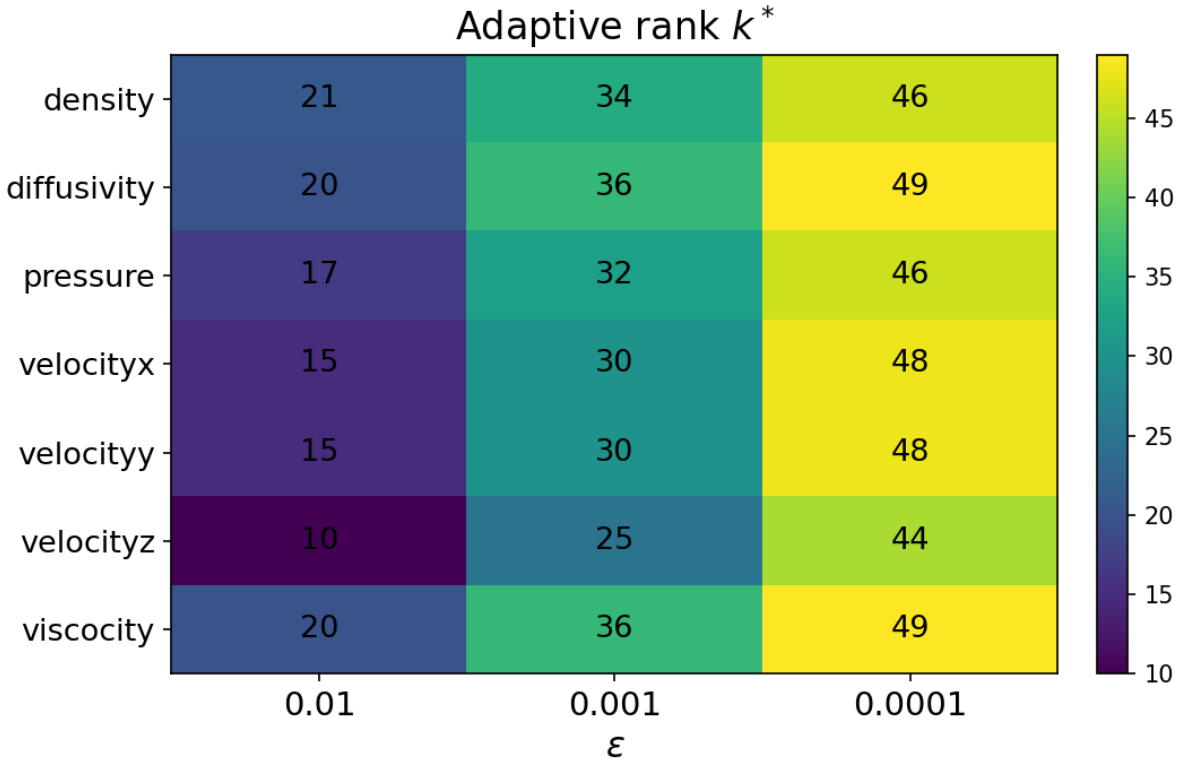
Rank and Violation Dynamics ($\varepsilon = 10^{-4}$)



Timestep $\times$ variable view: warm rank tracks are smoother; $QRAIN_f$'s rank collapses are handled identically in both arms (walk-down makes rank reduction free).

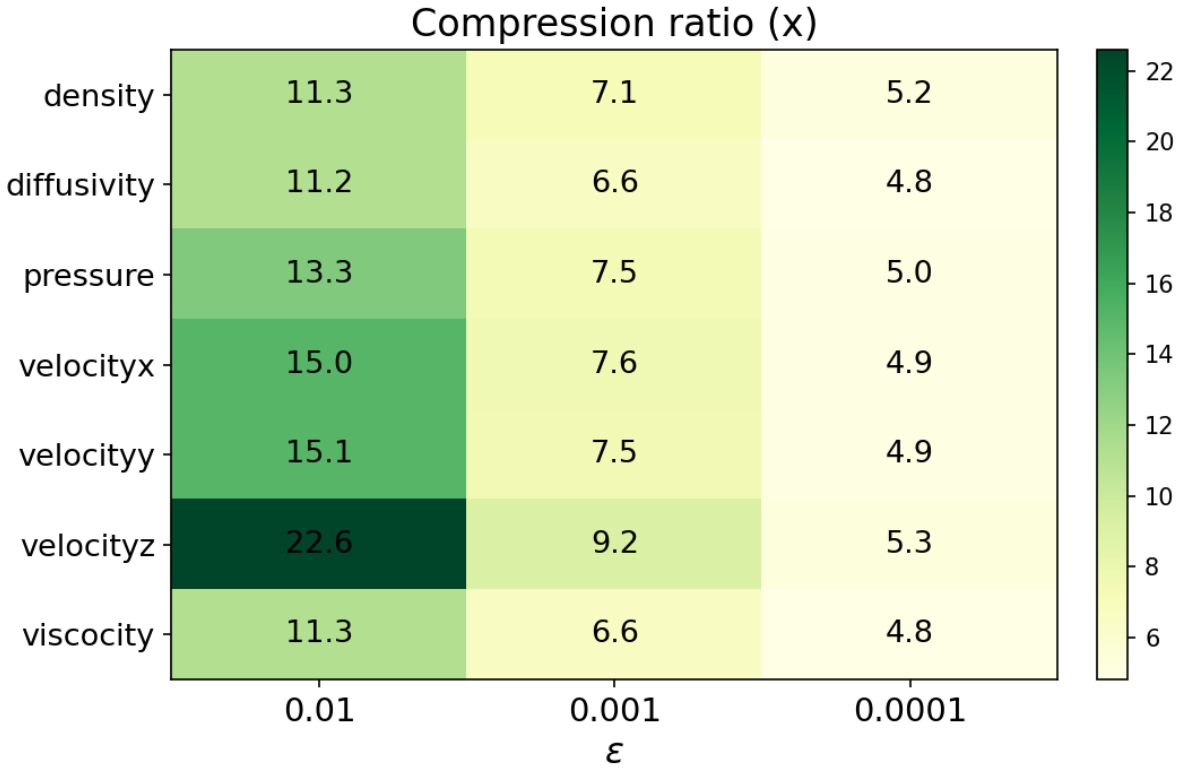
Outline

Miranda: Rank Scaling



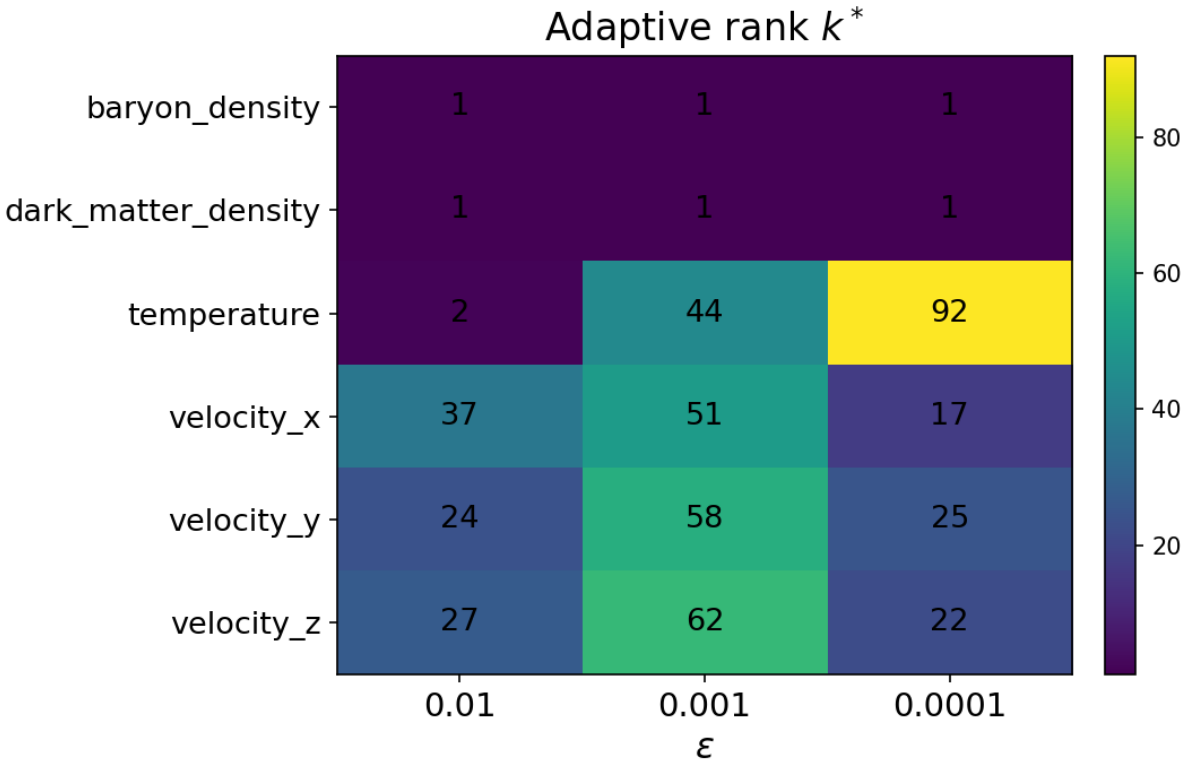
All 7 variables behave the same way: k^* rises *smoothly and monotonically* as τ tightens — 10–21 $(10^{-2}) \rightarrow$ 25–36 $(10^{-3}) \rightarrow$ 44–49 (10^{-4}).

Miranda: Compression Ratio



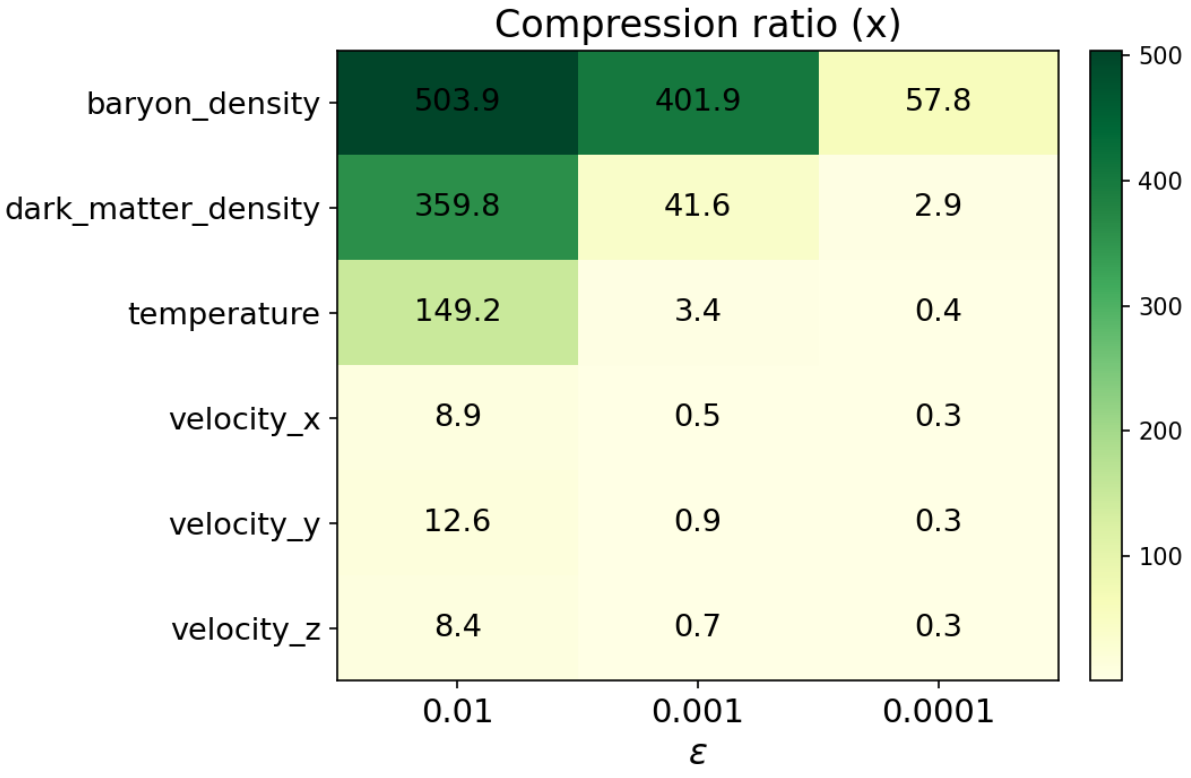
Compressible at *every* ε: ratios 22.6× down to 4.8× — never below 1. Stratified RT mixing has low-rank structure along the unfolded axis; the τ guarantee holds on all rows.

NYX: Adaptive Rank — Three Regimes



Density: rank-1 at every ε (one dominant mode). **Temperature:** rank climbs hard, $2 \to 44 \to 92$. **Velocity:** mid rank that never collapses — near-full-rank turbulence.

NYX: Compression Ratio



baryon_density reaches **504**×; the velocities fall *below 1* at tight ε (worse than raw). *Why* < 1: next slide.

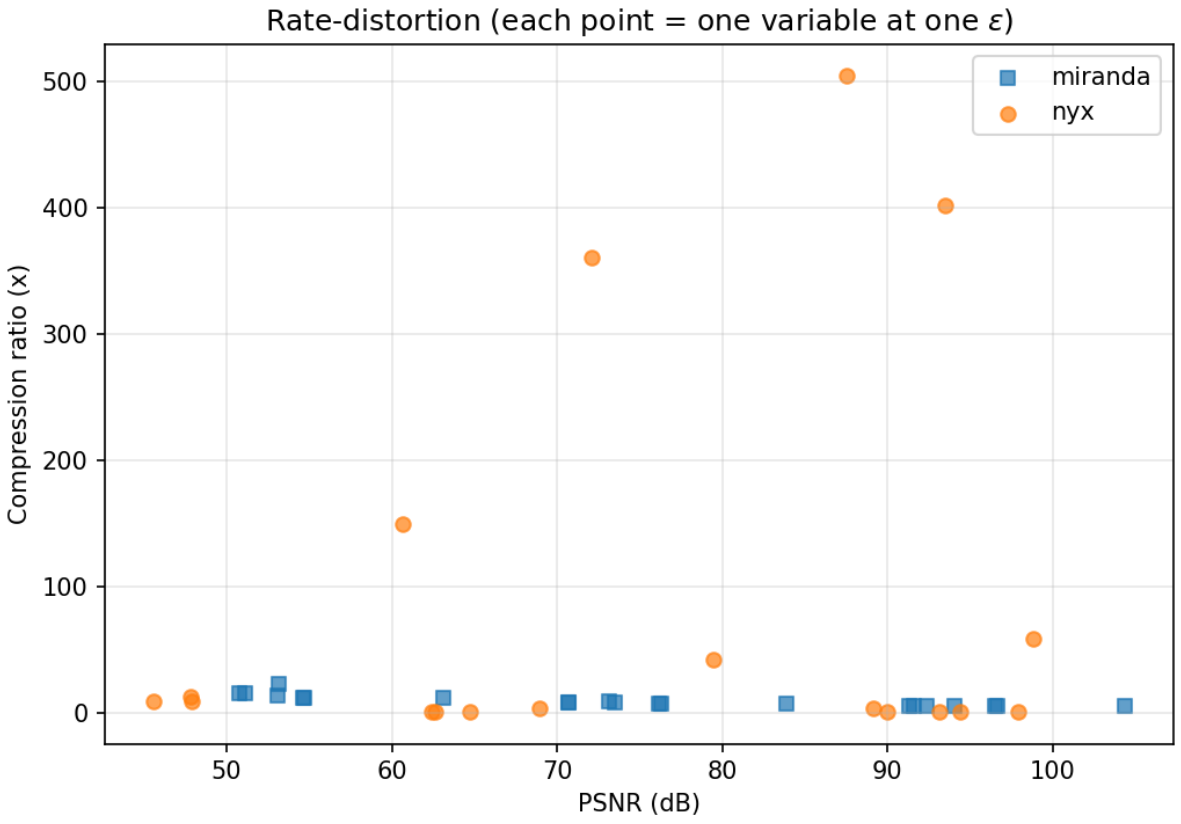
NYX Failure: Ratio < 1

At $\varepsilon = 10^{-4}$, velocity and temperature fall **below ratio 1** ($L + S$ larger than raw) — but for *opposite* reasons:

Field	**Spectrum**	k^*	**Cost sits in**
velocity x/y/z	flat / full-rank	17–25	almost all in S
temperature	steep head + long tail	**92**	both L (18%) and S

- **Floor** $= c_{\text{float}}/c_{\text{entry}} = 4/12 \approx 0.33$**:** the velocities land exactly there (0.34) — *every* entry in S.
- **Temperature fails only at tight** τ**:** $149\times \to 3.4\times \to 0.45\times$ over $\varepsilon = 10^{-2}, 10^{-3}, 10^{-4}$ — a rank-2 head, but an incompressible fine-scale tail.
- The optimiser is **correct** — the cost *minimum* exceeds raw $\Rightarrow$ a production codec needs a store-raw fallback and a cheaper sparse coder than 12 B/entry.

Cross-Dataset Rate–Distortion



Each point = one (variable, ε), highlighting the distinct physical regimes.

Outline

Adaptive Key Findings

1. **Hard guarantee everywhere.** $\|A - \hat{A}\|_{\max} \leq \tau$ holds by construction on all 2304 Isabel rows + 39 static runs — in both arms, at every tolerance.
2. **One algorithm tested on three datasets.** The adaptive compressor handles a temporal hurricane, a cosmology cube, and turbulent mixing with the same ε grid.
3. **Warm-start value has a regime structure.** Loose tolerance: same answers, 1.1–3.4× faster. Tight tolerance: **21% bytes saved, 70% fewer violations, +2.2 dB, 1.7× faster** (U_f) — warm tracking carries the high rank forward; cold re-expands its window at 43/48 steps.
4. **The cost model is now the frontier.** NYX velocities at $\varepsilon = 10^{-4}$ show ratio < 1: the 12 B/entry sparse model is the limiting factor.
5. **Single stage by design.** A second-stage residual rSVD ($L_1 + L_2 + S$) was explored, but dropped — it fired rarely and saved $< 0.3\%$ bytes for 8–32% extra time.

Next Steps

- **Sketch quality lever:** $q = 1$ power-iteration sweep under the adaptive cost objective
- **Warm-state policies:** carry k_{hi} columns (anti-dilution at expansions); V_{prev} seeding (one matmul instead of two); rank-dip diagnostic
- **Compression backend:** CSR/clustered sparse formats + entropy coding $\rightarrow$ true $c_{\text{entry}} \ll 12$ B; store-raw fallback; benchmark against SZ3
- **Scale out:** remaining 9 Isabel variables; a second temporal dataset to test warm-start generality