Warm-Start Randomized SVD for Streaming Data
Hurricane Simulation Benchmark

Ahmer Nadeem Khan

Numerical Linear Algebra

1 May 2026

Outline

Outline

The Streaming Low-Rank Approximation Problem

- Many applications produce a **sequence of large matrices** $A_1, A_2, \ldots, A_T$ that evolve slowly over time.
- At each timestep t, we want a rank-k approximation

$$A_t \approx U_t \operatorname{diag}(\mathbf{s}_t) V_t^T, \qquad U_t \in \mathbb{R}^{m \times k}$$

- **Cold-start**: compute from scratch every step — ignores prior structure.
- **Warm-start**: reuse U_{t-1} to guide the sketch — exploits assumed temporal structure.

Central question: Does reusing U_{t-1} reduce approximation error and complexity?

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: $Q, \_ \leftarrow \mathrm{QR}(Y)$ [**dominant cost:** $\sim$83% of runtime]
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

Matmul count: AX once, $A^T X$ once $\Rightarrow$ **2 total**.

No memory of prior timesteps.

Halko et al. Error Bound

Expected approximation error (spectral norm)

For a rank-k approximation with oversampling $p \geq 2$:

$$\mathbb{E}\|A - QQ^*A\| \leq \left[1 + \frac{4\sqrt{k+p}}{p-1} \cdot \sqrt{\min\{m,n\}}\right] \sigma_{k+1}$$

- Q is the orthonormal basis from the randomized range finder.
- The bound shows the error is controlled by σ_{k+1}, the first neglected singular value.
- Oversampling p tightens the bound: as p grows, the prefactor $\to 1$.

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

Matmul count: AX twice, $A^T X$ twice $\Rightarrow$ **4 total**.

Key trade-off: more matmuls, but narrower QR — which effect dominates?

Wedin's Theorem (Perturbation of Singular Subspaces)

Let $\boldsymbol{M}$ and $\tilde{\boldsymbol{M}} \in \mathbb{R}^{m\times n}$ be two matrices with rank-r SVDs:

$$\boldsymbol{M} = [\boldsymbol{U}_1 \;\; \boldsymbol{U}_2] \begin{bmatrix} \boldsymbol{\Sigma}_1 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{V}_1^T \\ \boldsymbol{V}_2^T \end{bmatrix}, \quad \tilde{\boldsymbol{M}} = \boldsymbol{M} + \boldsymbol{\Delta} = [\tilde{\boldsymbol{U}}_1 \;\; \tilde{\boldsymbol{U}}_2] \begin{bmatrix} \tilde{\boldsymbol{\Sigma}}_1 & \mathbf{0} \\ \mathbf{0} & \tilde{\boldsymbol{\Sigma}}_2 \end{bmatrix} \begin{bmatrix} \tilde{\boldsymbol{V}}_1^T \\ \tilde{\boldsymbol{V}}_2^T \end{bmatrix}.$$

If $\delta = \min\{\min_{1\le i\le r,\, r+1\le j\le n} |\sigma_i - \tilde{\sigma}_j|,\, \min_{1\le i\le r} \sigma_i\} > 0$, then

$$\|\sin\boldsymbol{\theta}(\tilde{\boldsymbol{U}}_1, \boldsymbol{U}_1)\|_F^2 + \|\sin\boldsymbol{\theta}(\tilde{\boldsymbol{V}}_1, \boldsymbol{V}_1)\|_F^2 \le \frac{\|\boldsymbol{U}_1^T\boldsymbol{\Delta}\|_F^2 + \|\boldsymbol{\Delta}\boldsymbol{V}_1\|_F^2}{\delta^2}.$$

Why Warm-Start Can Help: Intuition

Cold-start sketch

$$Y = A \underbrace{\Omega}_{n \times (k+p_c)}$$

- Width $k + p_c = 30$
- Relies entirely on random luck to cover the dominant subspace

Warm-start sketch

$$Y = \left[AG \mid A\Omega \right]$$

- Width $k + p_w = 25$ (< 30)
- AG *deterministically* covers the previous subspace
- $A\Omega$ explores *new* directions with only $p_w = 5$ vectors

Net effect (when data is temporally correlated)

Warm-start needs fewer random vectors because prior structure is already in the sketch. Narrower QR $\Rightarrow$ cheaper dominant phase $\Rightarrow$ lower wall-clock time *and* lower error.

Outline

Approximation Quality Metrics

Relative Frobenius Error

$$\varepsilon_F = \frac{\|A - U\,\mathrm{diag}(\mathbf{s})\,V^T\|_F}{\|A\|_F} = \frac{\sqrt{\|A\|_F^2 - \|\mathbf{s}\|^2}}{\|A\|_F}$$

Fraction of energy *not* captured by the rank-k approximation. Computed without forming the full residual via the identity above (avoids catastrophic cancellation: accumulated in float64).

Optimal Rank-k Frobenius Error (Eckart–Young)

$$\varepsilon_F^* = \frac{\sqrt{\sum_{i>k} \sigma_i^2}}{\sqrt{\sum_i \sigma_i^2}} = \sqrt{\frac{\sum_{i>k} \lambda_i(A^T A)}{\mathrm{tr}(A^T A)}}$$

Theoretical minimum for any rank-k approximation. Computed via eigendecomposition of the small ($n \times n$) Gram matrix $A^T A$ (avoids full SVD of the 250,000 $\times$ 100 data matrix).

Approximation Quality Metrics (cont.)

Subspace Distance via Principal Angles (Frobenius)

Let $\sigma_1 \geq \cdots \geq \sigma_k$ be the singular values of $U_1^T U_2$ (cosines of principal angles θ_i). We use:

$$\|\sin\theta\|_F = \sqrt{\sum_i (1-\sigma_i^2)}, \qquad \text{frac\_aligned} = 1 - \frac{\|\sin\theta\|_F^2}{k} = \frac{\|U_1^T U_2\|_F^2}{k}$$

Note: $\sin\theta_{\max} = \sqrt{1-\sigma_k^2}$ saturates to ≈ 1 for any two rank-k subspaces with $k > 1$ and is *not used*. Three comparisons: (1) **warm drift** $U_{t-1} \to U_t^{\text{warm}}$, (2) **cold vs. warm** at the same t, (3) **warm prior quality** — $\|\sin\theta\|_F(U_{t-1}, U_t^{\text{cold}})$.

Outline

Hurricane Isabel: Dataset

- **Scope:** 13 physical variables $\times$ 48 hourly snapshots.
- **Variables:**
  - Wind: U_f, V_f, W_f
  - Temperature: TC_f; Pressure: P_f
  - Cloud/moisture: $CLOUD_f$, $QVAPOR_f$, $QCLOUD_f$, $QICE_f$, $QRAIN_f$, $QSNOW_f$, $QGRAUP_f$
  - Precipitation: $PRECIP_f$

- Structured variables (wind, temperature, vapor) have a dominant low-rank subspace that evolves smoothly over time — ideal for warm-start exploitation.
- Sparse near-zero variables (QGRAUP, QRAIN) lack a stable subspace.

Hurricane Isabel: Data Format

Raw format

Each snapshot is stored as a $500, 500, 100)$ float32 binary array ($\approx$95 MB). Total dataset: $13 \times 48 = 624$ files.

Matrix unfolding

Reshape each snapshot to $A_t \in \mathbb{R}^{250{,}000 \times 100}$:

- **Rows:** 500×500 horizontal spatial grid (x, y)
- **Columns:** 100 vertical levels (z)

Low-rank structure arises because the 100 vertical profiles are highly correlated across space.

Hurricane Isabel: Streaming Setup

- Each of the 13 variables is processed **independently** as a stream of 48 matrices.
- $t = 1$: **cold start only** — no prior basis available; U_1 is computed and stored.
- $t \geq 2$: both cold and warm rSVD run on the **same** A_t:
  - Cold draws a fresh random sketch each step.
  - Warm reuses U_{t-1} (from cold at $t-1$) as the warm component.
- All timing and error metrics are averaged over the **47 non-initial** timesteps.
- Timings measured in C++ with `std::chrono::steady_clock` wrapping individual BLAS phases — no Python/PyTorch overhead.

Data Characterization: Why Some Variables Are Hard

Variable	Sparsity	Max $\lvert v \rvert$	Energy in top-20	Stable rank
TC_f	1.8%	78.6	99.999%	1.00
P_f	18.3%	4757	99.9998%	1.03
U_f	7.3%	69.5	99.97%	1.86
V_f	7.2%	67.9	99.98%	1.68
W_f	85.0%	16.1	99.99%	3.00
$CLOUD_f$	95.3%	0.002	99.81%	2.14
$QVAPOR_f$	51.2%	0.023	99.999%	1.05
$QICE_f$	94.8%	0.001	99.995%	1.32
$QSNOW_f$	95.6%	0.001	99.996%	1.38
$PRECIP_f$	97.2%	0.012	99.997%	1.19
$QCLOUD_f$	98.2%	0.002	99.948%	1.90
$QGRAUP_f$	99.0%	0.011	99.9996%	1.12
$QRAIN_f$	98.6%	0.007	99.9999%	1.05

- **Sparsity**: fraction of values below 1% of variable's own maximum. **Stable rank** $= \|A\|_F^2/\|A\|_2^2 = \sum \sigma_i^2/\sigma_1^2$.
- All variables appear low-rank (top-20 captures >99.8% of energy) and σ_1 dominates (stable rank $\approx$ 1–3).
- Key differentiator is **sparsity**: sparse fields ($QGRAUP_f$, $QRAIN_f$, >98% near-zero) have a shifting active region each hour $\Rightarrow$ U_{t-1} carries no useful signal for the next step.

Experiment Design: C++ Benchmark

Parameter	**Value**
Target rank k	20
Cold oversampling p_c	10
Warm oversampling p_w	5
Sketch width (cold)	30
Sketch width (warm)	25
Power iterations q	0
Timesteps per variable	48
Data type	float32

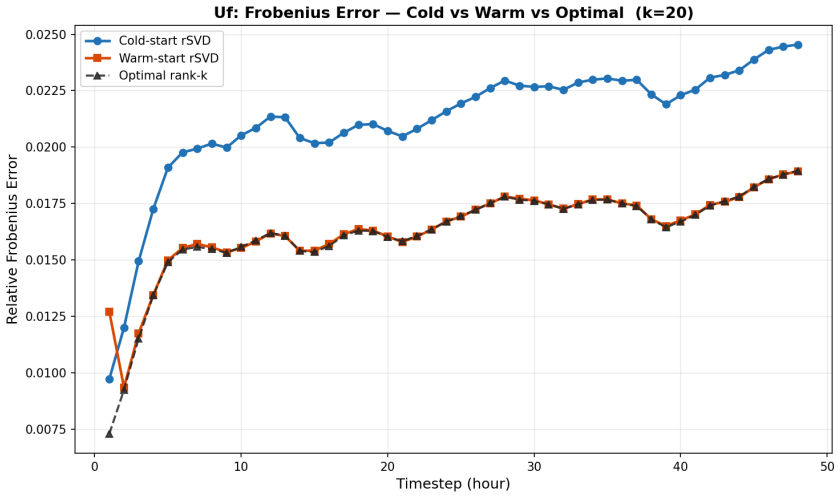
Per-Phase Timing Breakdown

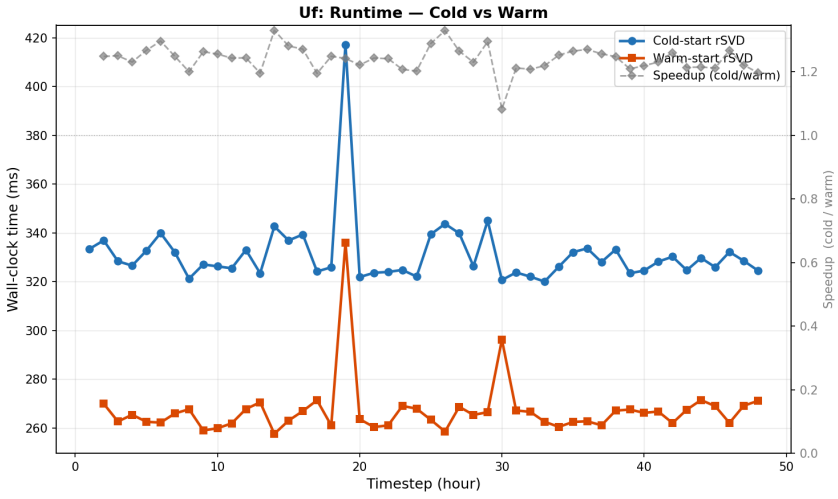
Why does warm-start end up faster despite more matmuls?

Phase	Cold (ms)	Warm (ms)	Delta
Warm proj ($A^T U_{\text{prev}}$)	—	14.3	+14.3
Warm matmul (AG)	—	16.4	+16.4
Random matmul ($A\Omega$)	20.1	10.4	−9.7
QR decomposition	**276.6**	**188.4**	**−88.2**
Projection ($Q^T A$)	12.1	11.4	−0.7
Small SVD	0.27	0.20	−0.07
Lift ($Q\hat{U}$)	14.6	11.9	−2.7
Total	**331 ms**	**267 ms**	**−64 ms**

- **QR dominates**: 83.5% of cold time, 70.5% of warm time.
- Narrower sketch ($25 < 30$) saves ≈88 ms on QR.
- Two warm-unique matmuls add ≈31 ms overhead.
- **Net: ≈24% speedup.** (U_f; representative of all non-sparse vars.)

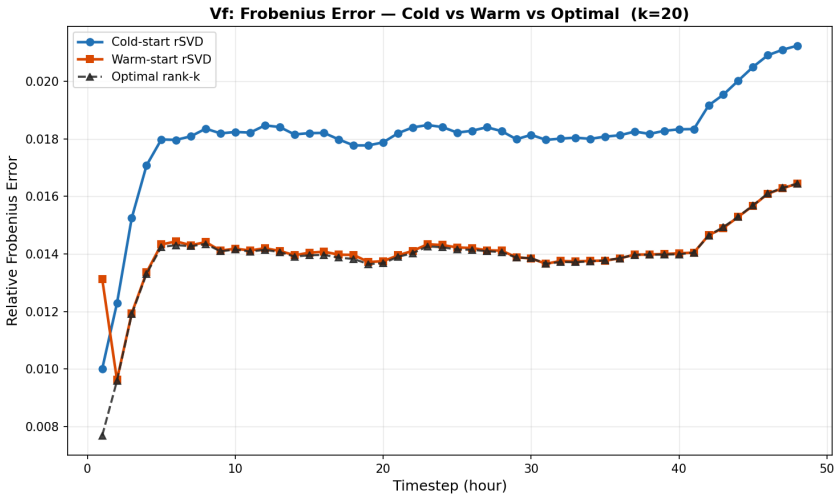
Per-Variable: U_f (East–West Wind)

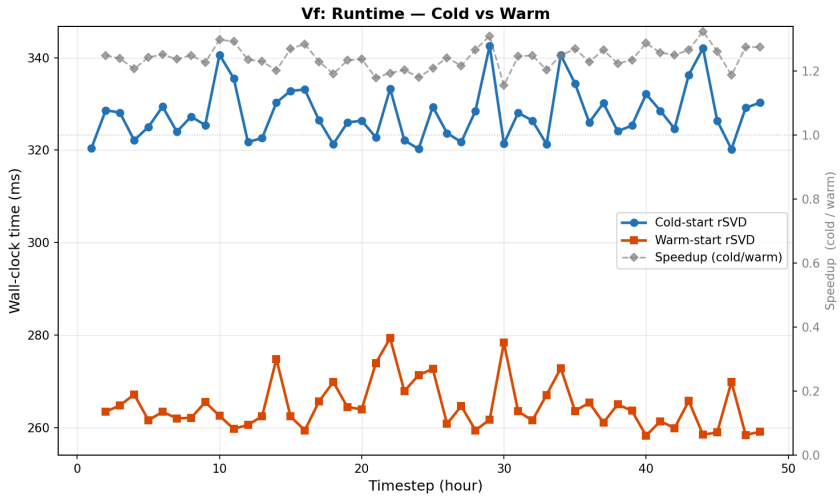




Left: Warm error tracks the optimal bound; cold is consistently above.
Right: Warm is faster at every non-initial step (∼24% median speedup).

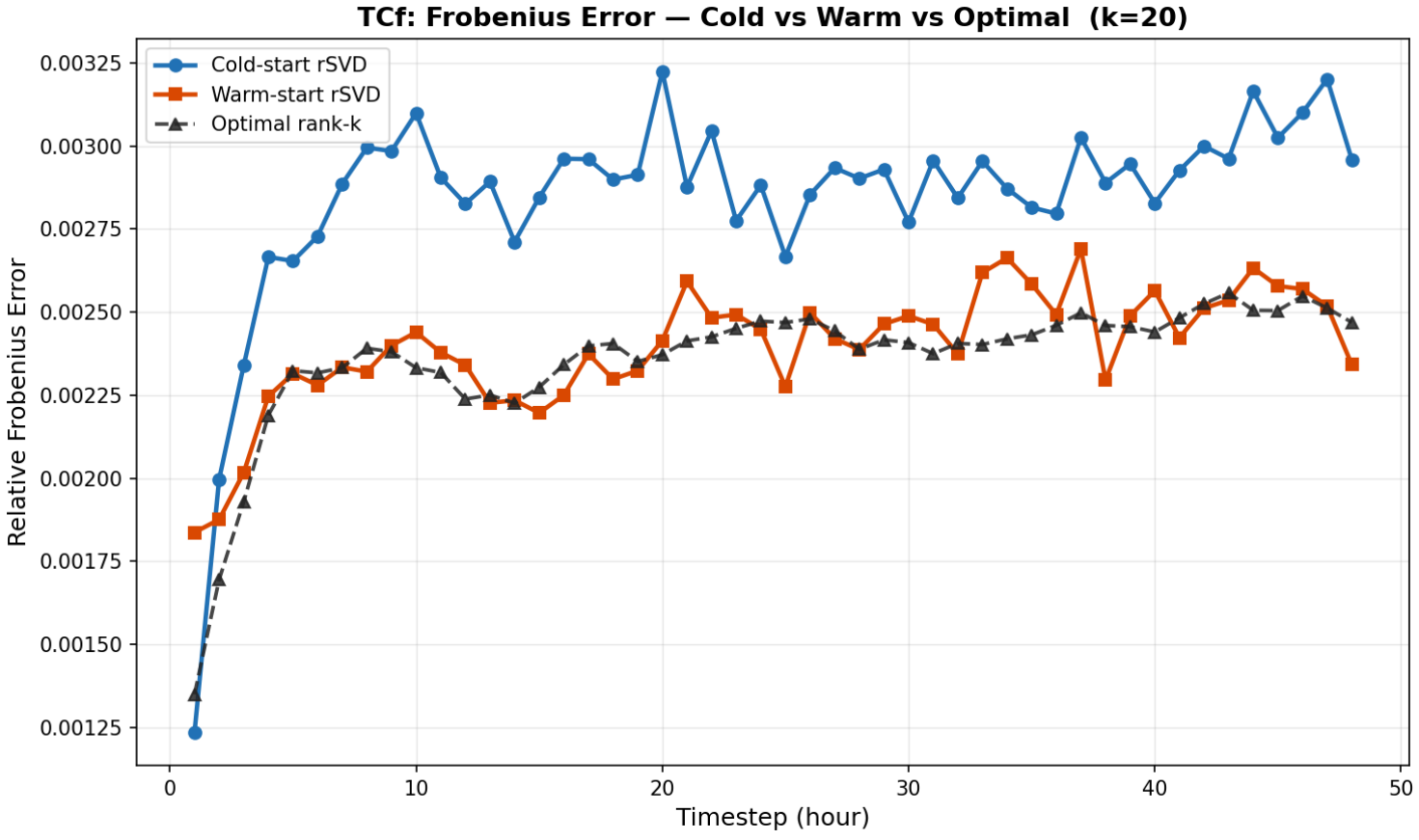
Per-Variable: V_f (North–South Wind)





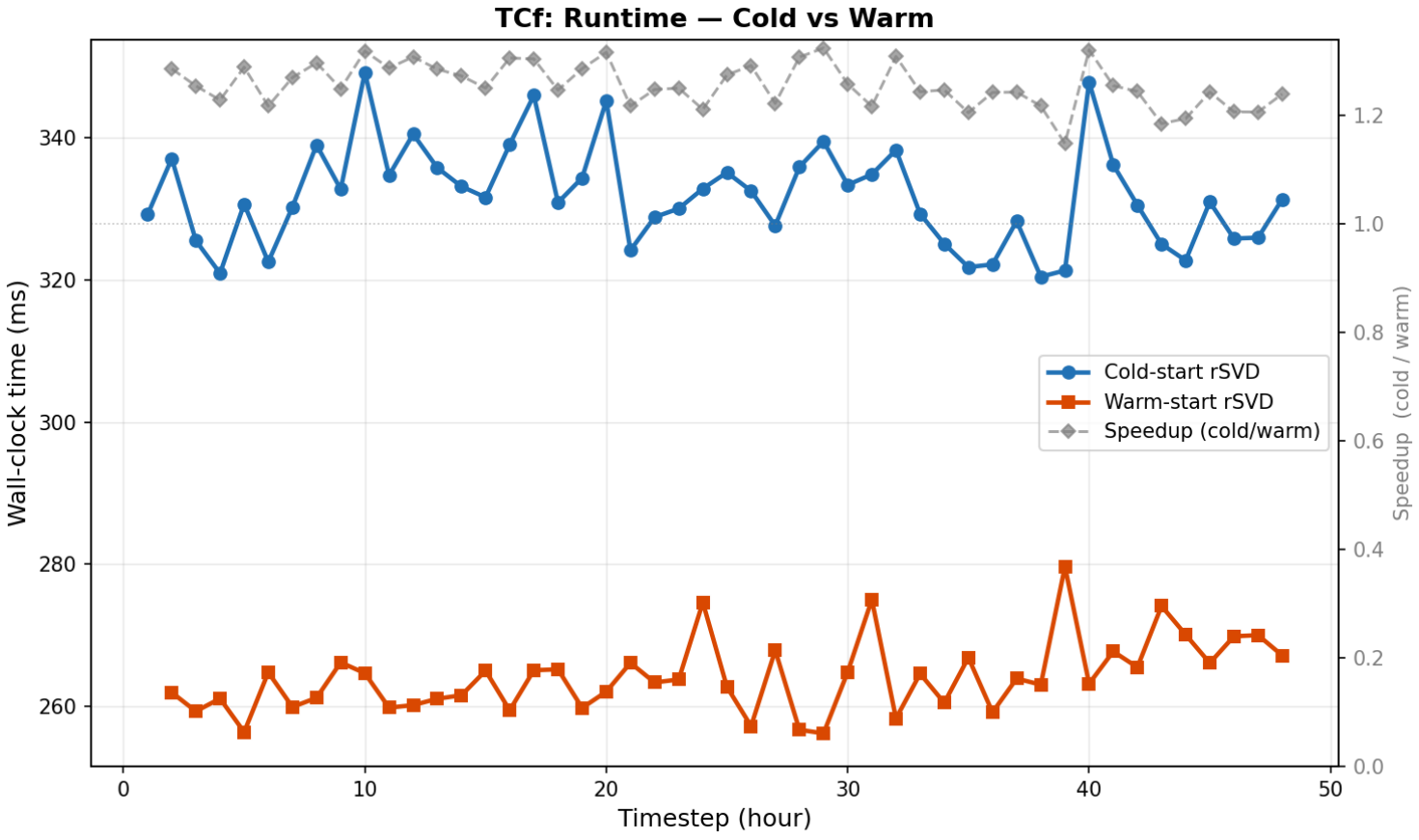
Left: Near-identical profile to U_f (ratio $= 0.784$, warm better 97.9% of steps).
Right: Consistent $1.24\times$ speedup; wind variables share the same subspace structure.

Per-Variable: TC_f — Frobenius Error



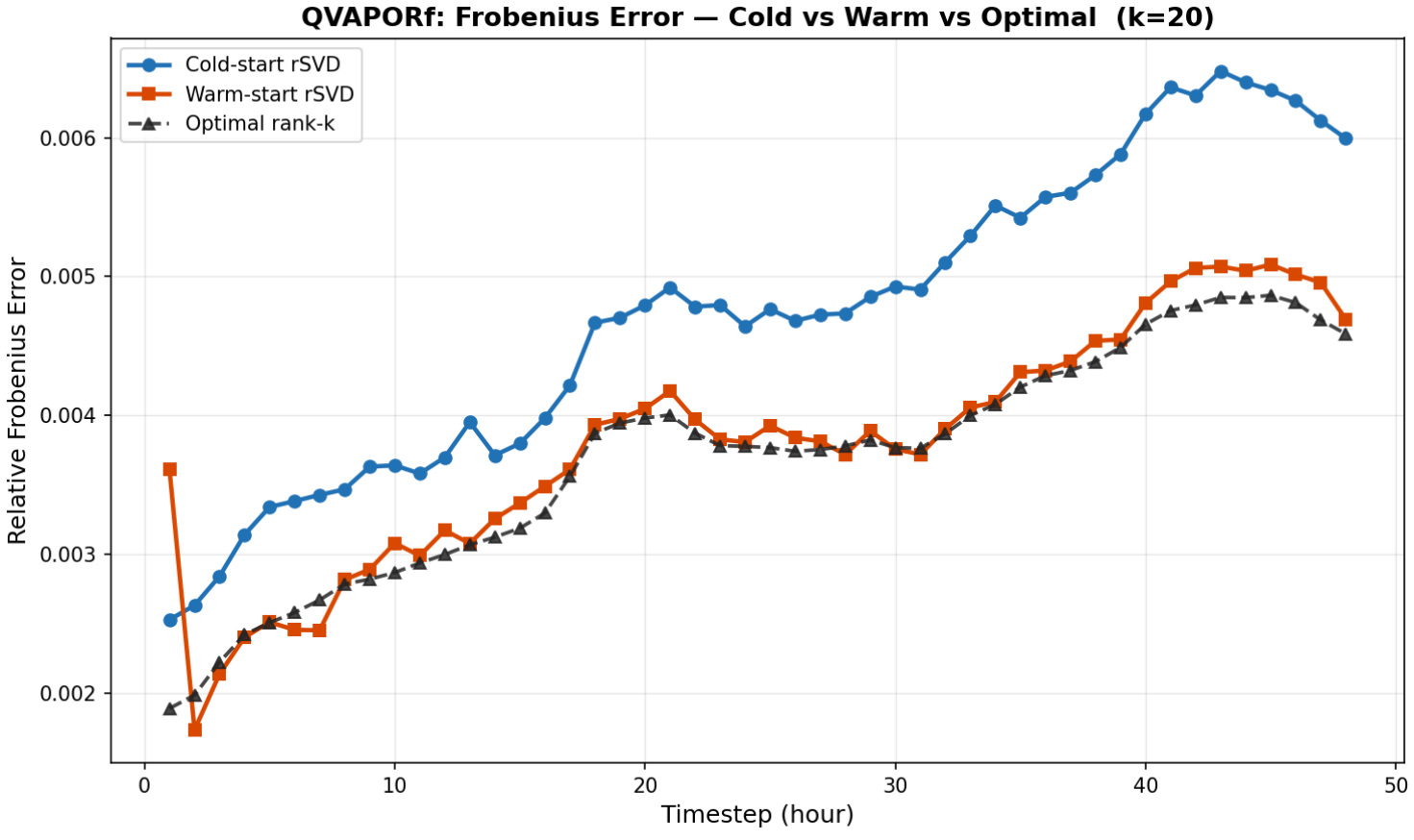
Most stable subspace of all 13 variables (frac_aligned = 0.71); warm nearly matches the Eckart–Young optimal bound throughout all 47 steps.

Per-Variable: TC_f — Runtime



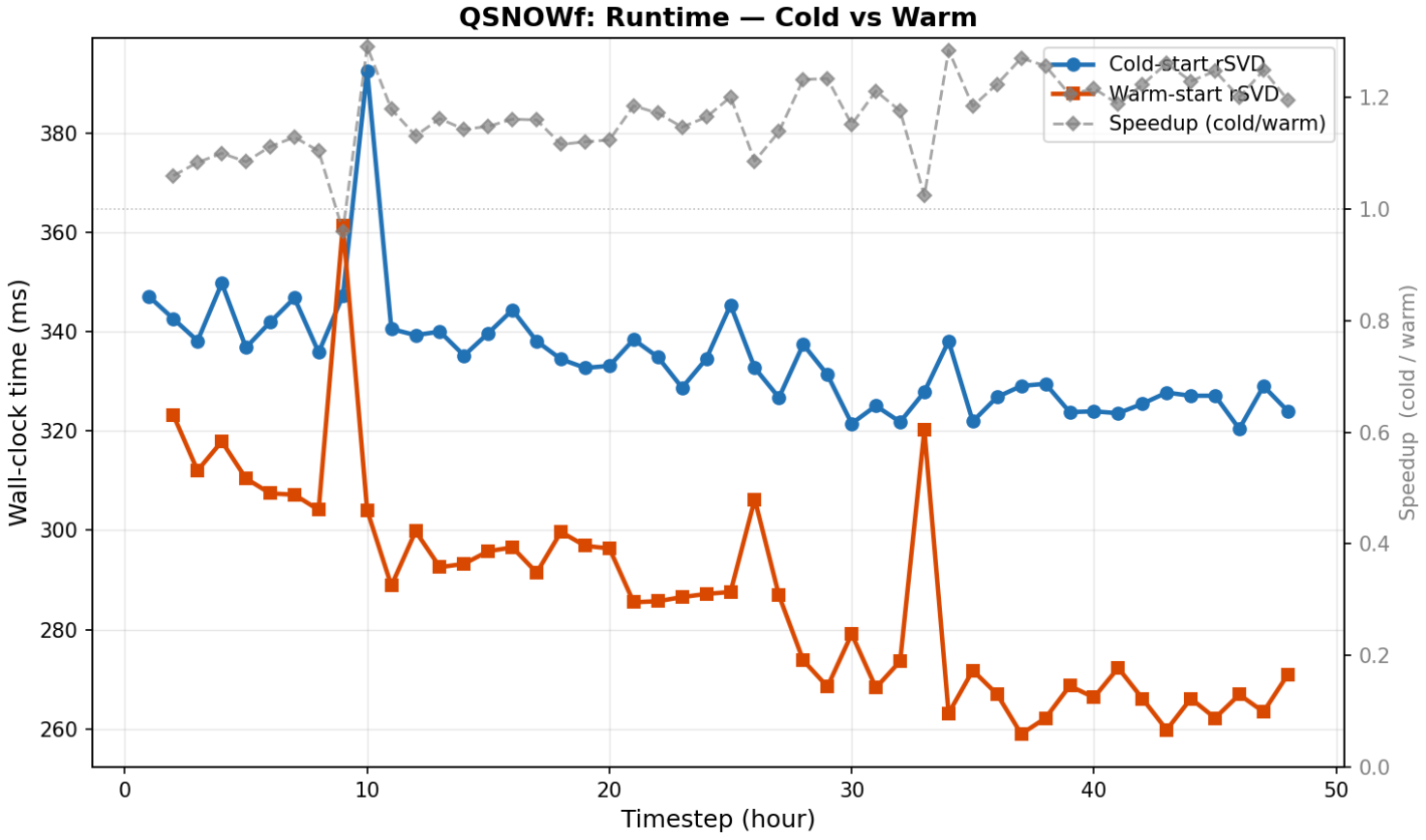
Highest speedup of all 13 variables (1.26×); the highly stable subspace means U_{t-1} consistently covers the current dominant direction, maximising QR compression.

Per-Variable: $QVAPOR_f$ — Frobenius Error



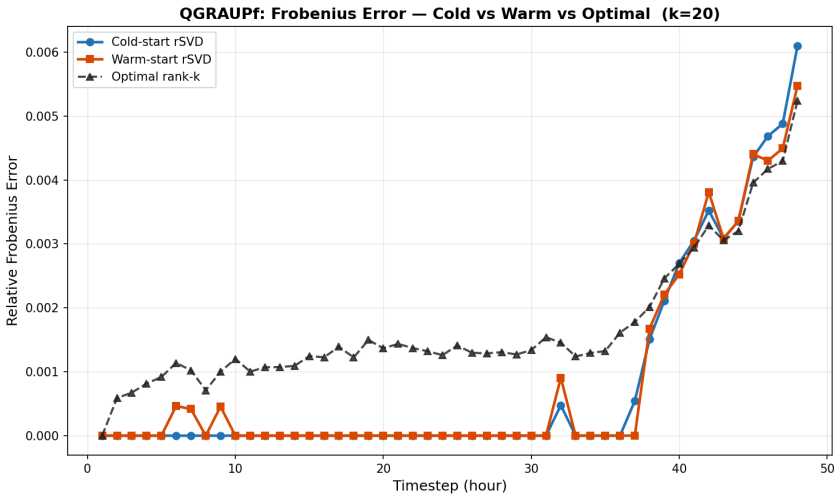
Strong and consistent error reduction (ratio = 0.811, warm better 97.9% of steps); water vapor has high spatial coherence across the 48-hour simulation window.

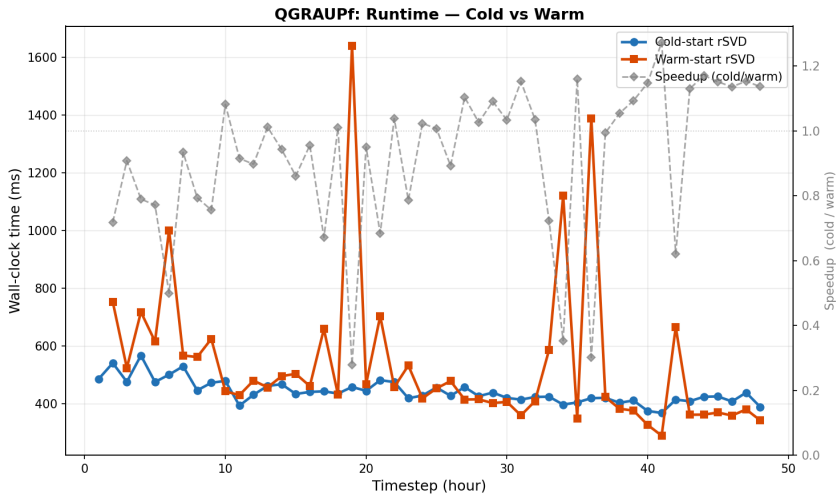
Per-Variable: $QSNOW_f$ — Runtime



Moderate speedup (1.17×); subspace stability (frac_aligned = 0.158) is lower than wind or temperature variables, but the warm basis still delivers consistent QR compression.

Per-Variable: $QGRAUP_f$ (Graupel — Sparse Failure Case)





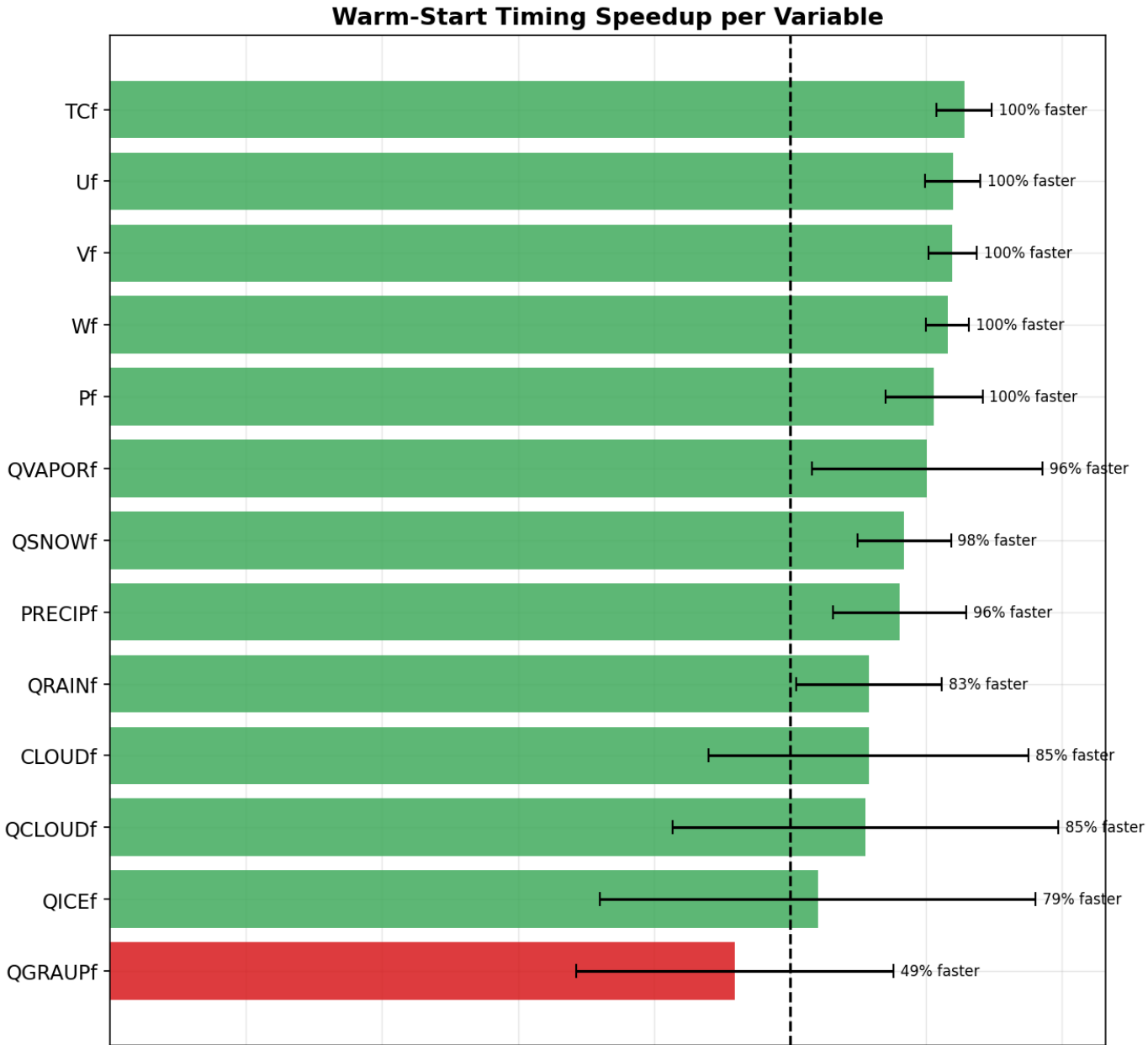
Left: No reliable error reduction; 99% near-zero sparsity means the active region shifts each hour — U_{t-1} carries no useful signal.
Right: Warm is on average *slower* (0.92×): the mismatched basis adds matmul overhead with no QR savings.

Results: Cross-Variable Error and Timing

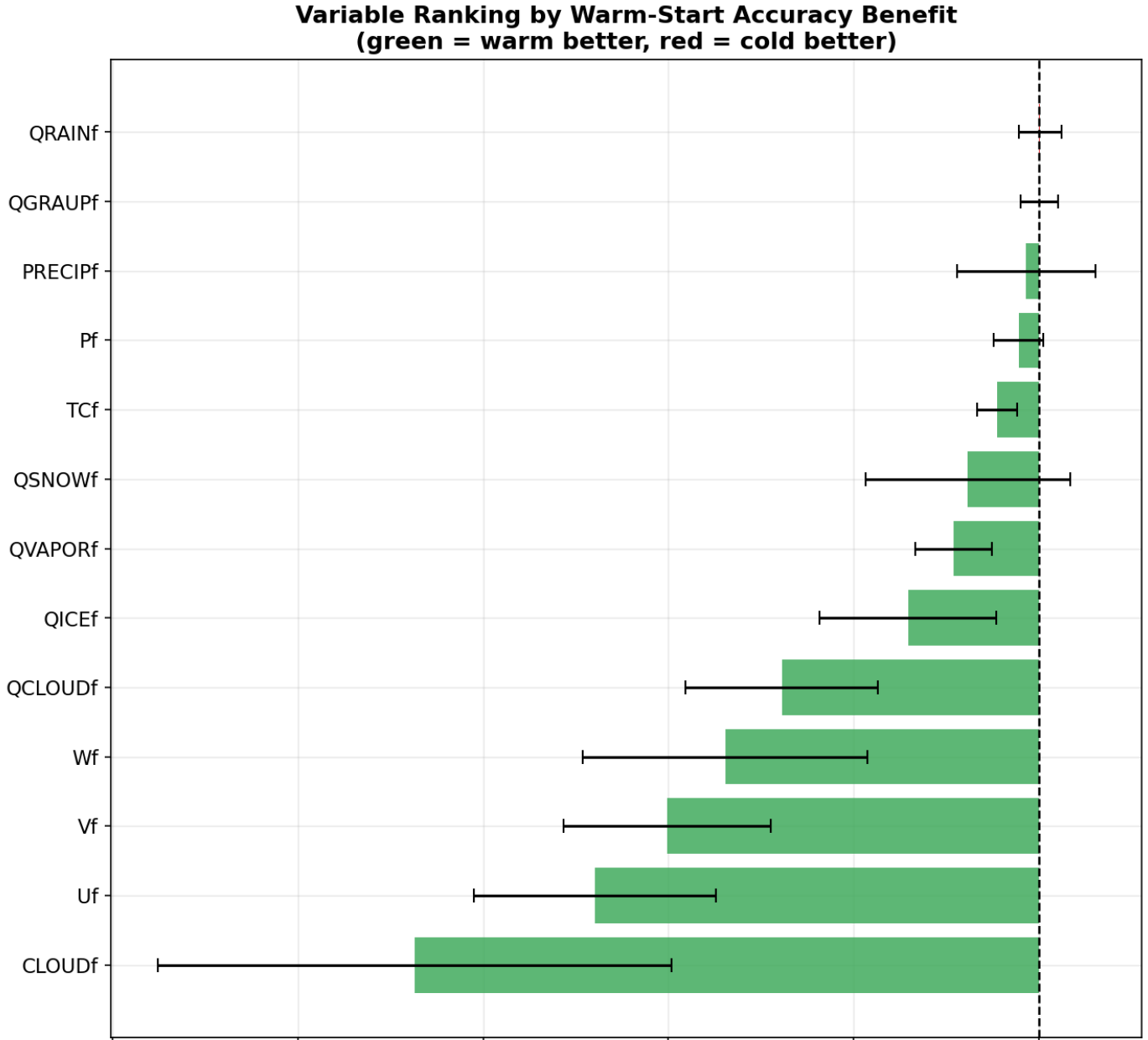
Var	Cold ε_F	Warm ε_F	Opt. ε_F^*	Ratio	Warm%	Speedup
U_f	0.0212	0.0164	0.0162	0.780	97.9%	1.24×
V_f	0.0181	0.0141	0.0139	0.784	97.9%	1.24×
W_f	0.0133	0.0099	0.0095	0.750	97.9%	1.23×
TC_f	0.0028	0.0024	0.0024	0.851	97.9%	1.26×
P_f	0.0016	0.0014	0.0014	0.877	87.5%	1.21×
$CLOUD_f$	0.0520	0.0453	0.0428	0.875	97.9%	1.12×
$QVAPOR_f$	0.0047	0.0038	0.0037	0.811	97.9%	1.20×
$QCLOUD_f$	0.0253	0.0225	0.0218	0.904	97.9%	1.11×
$QICE_f$	0.0081	0.0067	0.0066	0.845	97.9%	1.04×
$QSNOW_f$	0.0062	0.0055	0.0057	0.939	81.2%	1.17×
$PRECIP_f$	0.0040	0.0039	0.0045	1.040	56.2%	1.16×
$QGRAUP_f$	0.0008	0.0008	0.0017	0.985	14.6%	0.92×
$QRAIN_f$	0.0002	0.0002	0.0008	0.801	8.3%	1.12×

Warm%: fraction of 47 non-initial timesteps where warm error $<$ cold error.
Speedup: median per-timestep ratio $t_{\text{cold}}/t_{\text{warm}}$ (robust to OS timing spikes).
Sparse/near-zero fields ($QGRAUP_f$, $QRAIN_f$) have no dominant subspace to exploit. $QGRAUP_f$ also shows warm *slower* on average (0.92×): the mismatched warm basis adds overhead without QR savings.

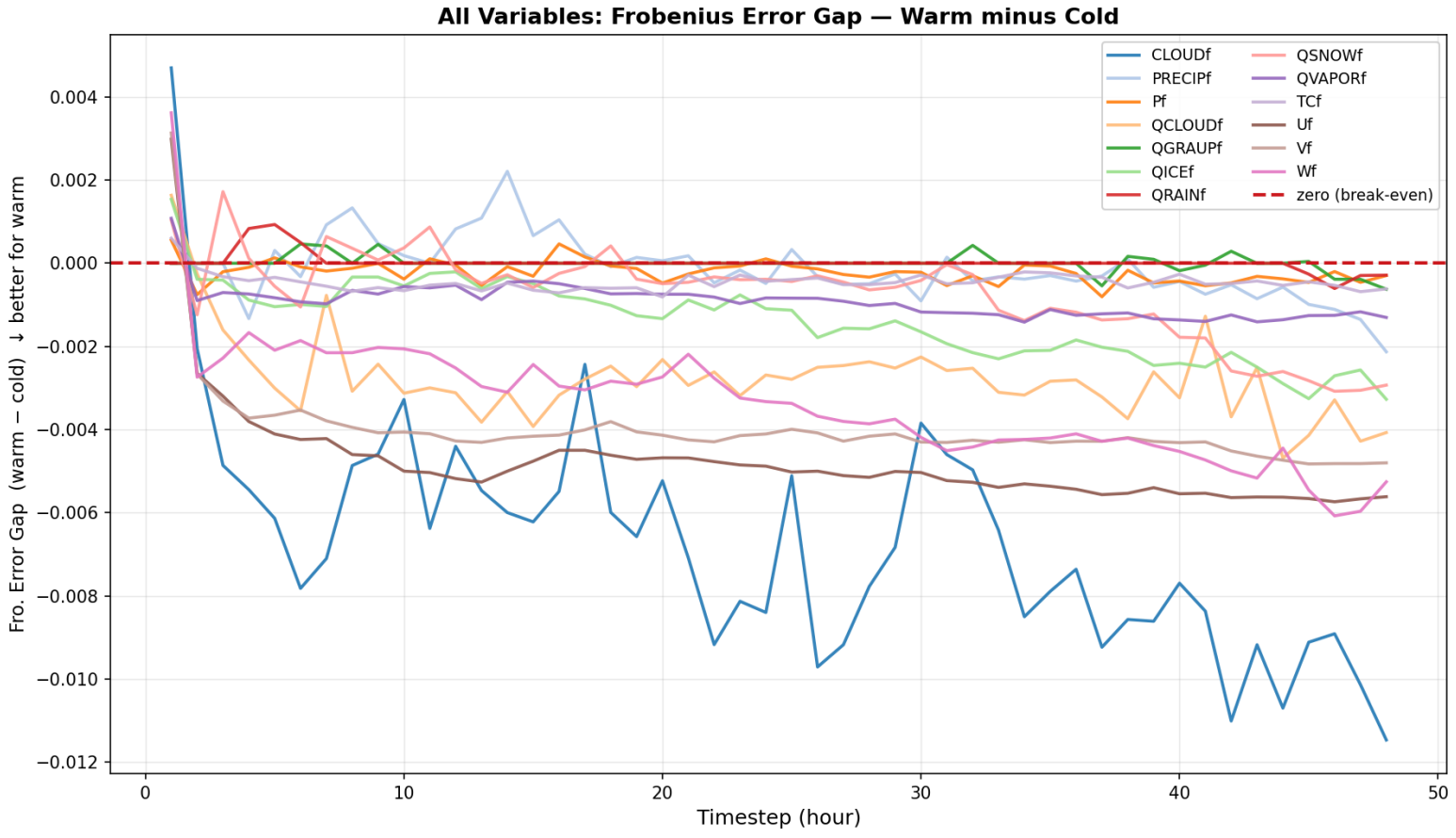
Cross-Variable: Speedup and Error Reduction



Cross-Variable: Warm Error Improvement Ranking



Cross-Variable: Error Gap Over Time



Negative gap = warm better. Most variables consistently negative; sparse fields near zero.

Outline

Key Findings

1. **Warm-start reduces approximation error** on correlated streaming data.
   - Synthetic: 13–25% lower relative Frobenius error vs. cold-start.
   - Hurricane: 6–25% error reduction on 11/13 variables, 97.9% of timesteps.
   - Sparse near-zero variables (QGRAUP, QRAIN) show no benefit.
2. **Warm-start is faster, not slower.**
   - QR dominates ($\sim$83% of cold runtime, $\sim$69% of warm).
   - Narrower sketch (25 vs 30 cols) saves $\sim$88 ms on QR.
   - Two extra matmuls cost only $\sim$31 ms overhead.
   - Median speedup: $\sim$17% across 12/13 variables; $QGRAUP_f$ (sparse) is the exception (0.92$\times$).
3. p_w **and** p_c **are the most important hyperparameters.**
   - Larger p_w broadens warm exploration; smaller p_c keeps cold weaker.

Future Directions

- **Power iterations:** Apply $q > 0$ rounds of $Y \leftarrow A(A^T Y)$ to sharpen the singular-value decay; may amplify benefit on flat-spectrum fields.
- **Adaptive p_w:** Increase oversampling when subspace drift $\sin\theta(U_{t-1}, U_t)$ is detected to be large; decrease when stable (?)
- **Incremental QR updates**
- **GPU implementation**
- **Compression pipeline**