TOPIQ: Statistical Error Propagation for Quantity-of-Interest Prediction under Lossy Compression

Youyuan Liu[1], Bo Jiang[1], Taolue Yang[1], Sheng Di[2], Robert Underwood[2], Sian Jin[1]

1. Temple University 2. Argonne National Laboratory

SC26 • Chicago, IL • November 2026

Error bounds protect the data, not the analysis

Lossy compression is no longer optional

- Climate: hundreds of TB every 16 s
- Cosmology (Nyx 4096^3): 2.8 TB per snapshot
- Fusion on exascale: exabytes per run

Error-bounded compressors (SZ3, ZFP, SPERR) guarantee

$$|\hat{x}_i - x_i| \leq \varepsilon \quad \text{for every element } i.$$

But scientists consume quantities of interest (QoIs)

- regional means, neural-network predictions, multi-field ratios, ...
- a per-element bound gives **no bound on the QoI**: errors get amplified, cancelled, or biased
- spatially correlated errors make a block mean **up to $10^3\times$ noisier** than the i.i.d. estimate
- PSNR / SSIM do not see any of this

Takeaway The question scientists ask is "how wrong is *my* QoI?", not "how wrong is each pixel?"

Existing approaches fix the QoI before compression

Dynamic, agent-driven workflows define QoIs after compression

Approach	Example	Limitation	Runtime QoI?
Quality-metric tuning	QoZ	PSNR-level, not QoI-level	No
Fixed-ratio control	FXRZ	targets storage budget	No
ML quality surrogate	QualityNet, DeepCQ	per-dataset training; predefined metrics	No
QoI preservation	MGARD, cpSZ, QPET	QoI & region fixed at compression; per-QoI derivation	No
Statistical propagation	**TOPIQ (ours)**	**statistical, not deterministic**	**Yes**

LLM / agent pipelines query compressed data at **arbitrary regions** with **QoIs composed on the fly** — nothing can be fixed at compression time.

TOPIQ in one slide

Predict QoI bias $\pm$ uncertainty from tiny compression metadata

Key idea

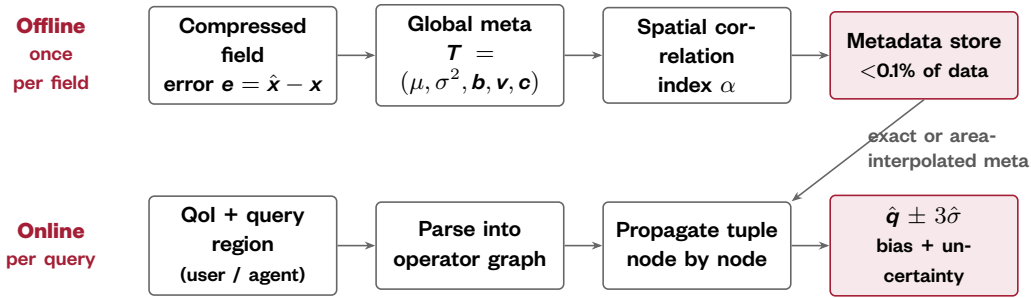
- Treat a QoI as a graph of **primitive operators**: linear, nonlinear, multi-field, aggregation
- Each operator has a **closed-form rule** that maps an error tuple $T_{\text{in}} \mapsto T_{\text{out}}$
- New QoI $=$ new composition: **no per-QoI derivation, no retraining**
- Models what naïve propagation misses: spatial error correlation (α) and data–error coupling ($\text{Cov}(x, e)$)

Results at a glance

93.1%	of 552 evaluations calibrated
$<$0.1%	metadata vs. original data
56–402$\times$	faster than direct computation
$\geq$99%	3σ coverage on arbitrary regions

4 datasets $\times$ 3 compressors $\times$ 4 QoI families $\times$ 8 error bounds

Overview: offline metadata, online propagation



Single-field $f(x)$
linear (exact), nonlinear (2nd-order expansion)

Multi-field $f(x, w)$
$\pm, \times, \div$ under cross-field independence

Aggregation $\sum_i w_i \delta_i$
spatially correlated errors via v_u / v_c split

The error tuple: one interface for every operator

$$\boldsymbol{T} = (\underbrace{\mu,\ \sigma^2}_{\text{signal}},\ \underbrace{\boldsymbol{b},\ \boldsymbol{v},\ \boldsymbol{c}}_{\text{error}})$$

$\mu,\ \sigma^2$	mean, variance of original data x
$\boldsymbol{b} = \mathbb{E}[\boldsymbol{e}]$	bias of compression error
$\boldsymbol{v} = \mathrm{Var}(\boldsymbol{e})$	error variance
$\boldsymbol{c} = \mathrm{Cov}(\boldsymbol{x}, \boldsymbol{e})$	data–error coupling

Every operator: tuple in → tuple out

- Signal and error stay separated, so both remain identifiable after long chains
- Linear operators propagate **exactly**:

$$\boldsymbol{T_{x+a}} = (\mu+\boldsymbol{a},\ \sigma^2,\ \boldsymbol{b},\ \boldsymbol{v},\ \boldsymbol{c})$$

$$\boldsymbol{T_{kx}} = (\boldsymbol{k}\mu,\ \boldsymbol{k}^2\sigma^2,\ \boldsymbol{kb},\ \boldsymbol{k}^2\boldsymbol{v},\ \boldsymbol{kc})$$

- Bias b: the measured global mean error is always propagated

Takeaway Five numbers per field are enough to drive the whole propagation graph.

Nonlinear operators: second-order expansion

Expand $f(\hat{x})$ around μ, with $\hat{x} = \mu + (x - \mu) + e$:

$$b_f = \underbrace{f'(\mu)\, b}_{\text{input bias}} + \underbrace{\tfrac{1}{2} f''(\mu)\, v}_{\text{Jensen}} + \underbrace{f''(\mu)\, c}_{\text{coupling}}$$

$$v_f = [f'(\mu)]^2\, v, \qquad c_f = f'(\mu)\, c$$

Three bias sources — the coupling term c is what naïve propagation drops.

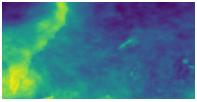
Example: $f(x) = x^2$

$$v_{x^2} = 4\,\mathbb{E}[x^2]\, v = 4(\mu^2 + \sigma^2)\, v$$

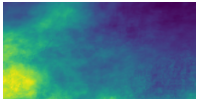
- **Both factors share the same error at each pixel**
- **Treating $x \cdot x$ as two independent fields gives $2(\mu^2+\sigma^2)v$ — half the true variance**
- **Hence single-field and multi-field need different rules**

Takeaway Same-field inputs share errors; different fields do not.

Multi-field: data correlated, errors not

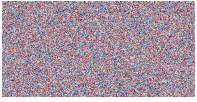


CLDTOT

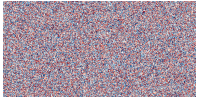


CLDHGH

data: Pearson $r = 0.43$



error CLDTOT



error CLDHGH

errors: Pearson r = **-0.002**

Cross-field independence: separately compressed fields share no prediction context or coefficients.

Add/sub: $v_{x\pm w} = v_x + v_w$

Multiply:

$$v_{xw} = (\mu_w^2+\sigma_w^2)\, v_x + (\mu_x^2+\sigma_x^2)\, v_w$$

Divide: reciprocal (nonlinear rule) $\times$ multiply

Within-field $c = \mathrm{Cov}(x, e)$ is still tracked; only cross-field terms are dropped — no $O(k^2)$ pairwise bookkeeping.

CESM-ATM, SZ3, $\varepsilon_{\mathrm{rel}} = 10^{-3}$.

Aggregation: correlated errors don't average out

Block sum of n errors needs $O(n^2)$ covariances:

$$\mathrm{Var}\left(\sum_i e_i\right) = nv + n(n-1)\,\bar{c}$$

Split $v = v_u + v_c$:

- v_u: pixel noise, averages away (scales n)
- v_c: common-mode offset, persists (scales n^2)

$$\mathrm{Var}(S) = n\,v_u + n^2\,v_c$$

Weighted sum, $s_1 = \sum w_i$, $s_2 = \sum w_i^2$:
$s_2\,v_u + s_1^2\,v_c$

Calibrate with one number α

$$\alpha = \frac{\mathrm{Var}\left(\sum_{\text{block}} e_i\right)}{n\,\mathrm{Var}(e_i)} \qquad \begin{cases} \alpha = 1 & \text{i.i.d.} \\ \alpha \gg 1 & \text{correlated} \end{cases}$$

$$v_c = \frac{(\alpha - 1)\,v}{n - 1}, \qquad v_u = v - v_c$$

- $r = v_c/v$ is approx. **scale-invariant**: measure α once, reuse for any query block size
- Pixel-wise operators scale v_u and v_c equally

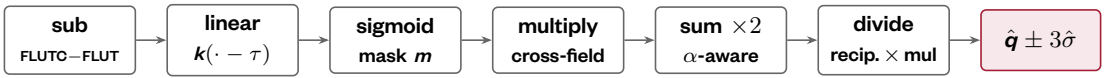
Takeaway One scalar α per field replaces $O(n^2)$ pairwise covariances.

Composing QoIs: no per-QoI derivation

(a) Block-input neural network $q = W_2\,\sigma(W_1\mathbf{x} + b_1) + b_2$, $\mathbf{x} \in \mathbb{R}^n$, 1.28M params



(b) Cloudy radiative ratio $q = \dfrac{\sum_i m_i(\mathrm{FLUTC}_i - \mathrm{FLUT}_i)}{\sum_i m_i + \epsilon}$, $m_i = \sigma\big(k(\mathrm{CLDTOT}_i - \tau)\big)$



Scope: any DAG of **smooth** operators. Non-smooth ops (ReLU, max, counts) are future work.

Setup: 552 evaluations

Datasets (float32)

Dataset	Shape	Fields
CESM-ATM	1800×3600	CLDTOT, CLDHGH, FLUT, FLUTC
NYX	512^3	dark matter, baryon, vel_x
SCALE-LETKF	$98 \times 1200 \times 1200$	PRES, T
Hurricane	$100 \times 500 \times 500$	TC, U

Compressors: SZ3, SPERR, ZFP (abs. error bound)
Error bounds: 8 values, $\varepsilon_{\text{rel}} \in [10^{-5}, 5 \times 10^{-3}]$
Blocks: 200×200 (2D), 32^3 (3D)

QoI families

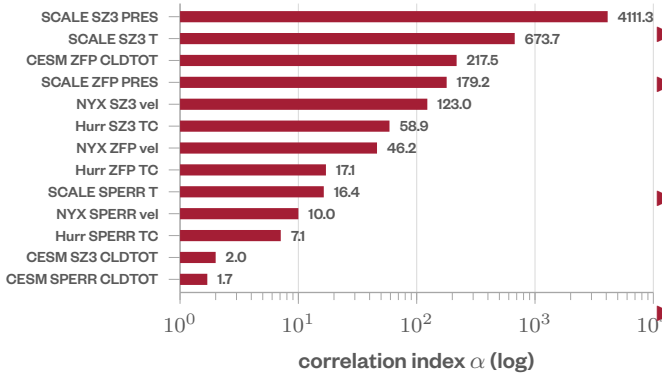
1. Mean square $\frac{1}{n} \sum x_i^2$
2. Block-input neural network
3. Weighted sum $\sum x_i w_i$ (2 fields)
4. Cloudy ratio (3 fields)

Metric: z-score $z = (q^* - \hat{q})/\hat{\sigma}$

- bias $\mu_z \to 0$, calibration $\sigma_z \to 1$
- calibrated: $\sigma_z \in [0.7, 1.3]$; 3σ cov. $\to 99.7\%$

69 configurations $\times$ 8 error bounds $=$ 552 evaluations. Dual Intel Xeon E5-2695 v4 (36 cores), 128 GB.

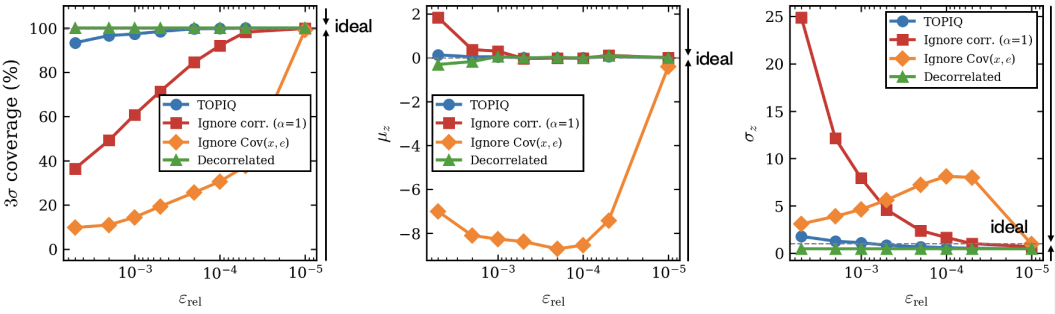
Error correlation α spans 1 to 4111



- α ranges from ≈ 1 (i.i.d.) to **4111**
- Randomly permuting error pixels gives $\alpha_{\text{shuf}} \approx 1$ everywhere $\Rightarrow$ the effect is **spatial**
- Ignoring it (set $\alpha{=}1$) on ZFP/CLDTOT ($\alpha{=}218$): calibration σ_z **1.0** $\rightarrow$ **14.7**
- When α is small (SPERR/CLDTOT, 1.7) the model reduces to the i.i.d. case

Takeaway Assuming independent pixel errors underestimates aggregated QoI error by up to $14.7\times$.

Ablation: both dependencies are necessary



Ignore α: σ_z inflates to 25 at loose bounds, coverage drops to 36%

Ignore Cov(x, e): bias μ_z shifts to -8.7, coverage $\approx$ 10–40%

TOPIQ corrects both; the decorrelated control restores $\sigma_z \approx 1$.

Both effects fade at tight bounds as SPERR's α drops $53 \to 2$.

mean(x^2) on NYX dark-matter density, SPERR.

Prediction accuracy: 94.2% calibrated at $\varepsilon_{\text{rel}} = 10^{-3}$

Data	Comp.	QoI / Field	μ_z	σ_z	cov$_{3\sigma}$
CESM	SZ3	mean(x^2) / CLDTOT	+0.00	1.01	98.8
CESM	SPERR	neural net / CLDTOT	+0.08	0.97	100.0
CESM	SZ3	cloudy ratio / 3 fields	+0.05	0.92	100.0
CESM	ZFP	cloudy ratio‡ / 3 fields	−0.00	0.48	100.0
Hurr.	SPERR	mean(x^2) / TC	+0.02	1.01	98.7
Hurr.	SZ3	weighted sum / TC, U	+0.15	1.13	97.8
NYX	SZ3	neural net / dark matter	+0.00	1.00	99.7
NYX	ZFP	neural net / baryon	−0.02	1.01	99.7
NYX	SZ3	weighted sum / dm, vel$_x$	−0.00	0.89	99.7
SCALE	ZFP	mean(x^2) / PRES	+0.01	1.01	100.0
SCALE	SPERR	neural net / PRES	+0.07	1.03	99.6
SCALE	SPERR	mean(x^2) / PRES†	+0.05	1.76	92.5
SCALE	SPERR	weighted sum† / PRES, T	+0.06	1.89	91.5

excellent | good | marginal | poor | failure † under-calibrated ‡ over-conservative

Representative subset of the 69 configurations at $\varepsilon_{\text{rel}} = 10^{-3}$; full table in the paper.

65 / 69

configs with $\sigma_z \in [0.7, 1.3]$

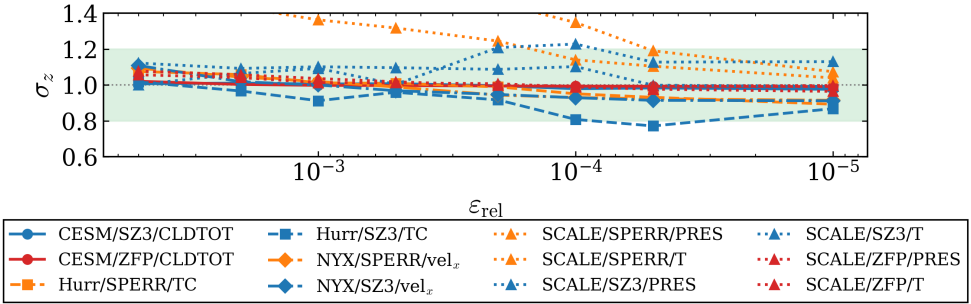
99.6%

median 3σ coverage

All four QoI families — including the 1.28M-param network and the 3-field ratio — calibrate as well as single-operator QoIs.

Failures: 3 SPERR/SCALE (†) + 1 safe over-estimate (‡).

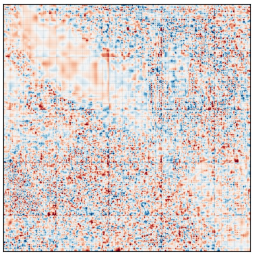
Calibration is stable across error bounds



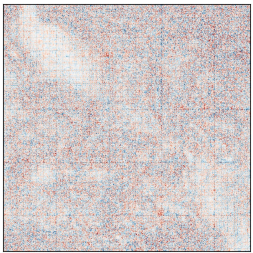
- **Across all 552 evaluations, 514 (93.1%) have $\sigma_z \in [0.7, 1.3]$**
- **Failures concentrate in SPERR on SCALE-LETKF (orange dotted, top): non-stationary errors**

mean(x^2), 12 dataset/compressor/field combinations; shaded band $= [0.8, 1.2]$.

Where TOPIQ fails, and why



PRES / SPERR: $\sigma_z \approx 1.8$



T / SPERR: $\sigma_z \approx 1.0$

(1) Non-stationary errors (3 configs)

SPERR's wavelets give tiny errors in smooth regions and large structured errors near fronts; block std varies $939\times$, per-block α by $9\times$ — no single global α fits.

Diagnose & act

- Flag: high coefficient of variation of per-block error variance
- Switch to a compressor with more uniform errors (SZ3 / ZFP calibrate on this field)
- Or estimate α per region (more metadata)

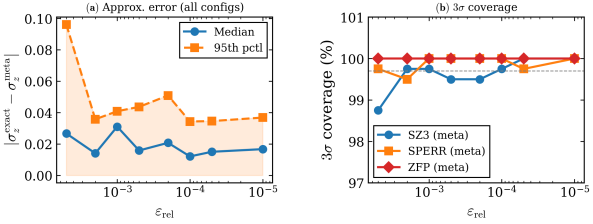
(2) Over-conservative (1 config)

- ZFP cloudy ratio: $\sigma_z \approx 0.5$
- Per-field biases partially cancel in the ratio, which independence does not model
- Wider interval than needed — **safe**

SCALE-LETKF slice $z{=}95$, $\varepsilon_{\text{rel}} = 10^{-3}$, $32{\times}32$ block grid; each panel has its own color scale.

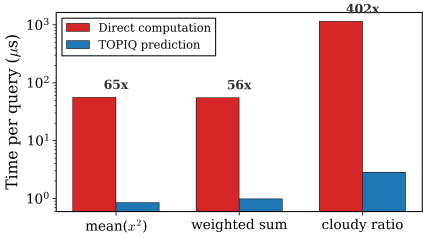
Arbitrary query regions, at a fraction of the cost

Post-hoc spatial queries



Coarse 150×150 meta grid (≈9 KB/field), area-interpolated to 200 random regions: median $|\Delta\sigma_z| < 0.03$, 3σ coverage ≥98.7% (mostly ≥99.5%).

Prediction overhead



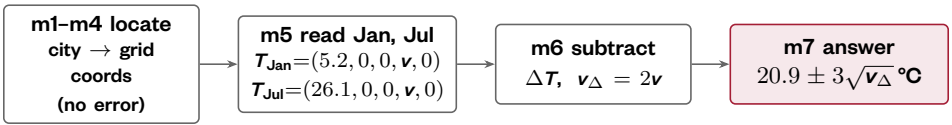
$O(1)$ per query vs. $O(n)$: 56–402× faster.

Direct computation needs the original data — impossible once it is discarded.

CESM-ATM, 200×200 query regions; overhead with SZ3, $\varepsilon_{\text{rel}} = 10^{-3}$; one-time pre-computation ≈33 ms/field.

Case study: LLM-composed climate queries

Q: *"How has average 2m temperature changed between January and July in Atlanta?"*



Each pipeline step is a TOPIQ operator

- Answer carries a 99.7% interval from metadata alone — no re-decompression
- Longer queries (regional means, multi-field comparisons) reuse the same validated operators

vs. QoI preservation (QPET)

- QPET tuned for mean(x^2) on a 200×200 grid: up to 11% of random 100×100 queries exceed its grid bound (max $4\times$), with no warning
- SZ3 + TOPIQ on the same queries: 99.9% coverage, $\sigma_z = 0.97$

Takeaway QPET suits fixed pipelines; TOPIQ suits dynamic exploration.

Conclusion and future work

TOPIQ

- Predicts QoI bias $\pm$ uncertainty from compact metadata ($<$0.1% of data)
- Composable closed-form operators — new QoIs need **no derivation, no retraining**
- Captures spatial correlation (α) and data–error coupling (Cov(x, e))
- **93.1%** of 552 evaluations calibrated; arbitrary regions; **56–402**$\times$ faster than direct computation

Open directions

- Non-smooth operators (ReLU, max, min): piecewise rules or distribution-shape stats (skewness, kurtosis)
- Discrete QoIs (feature counts) via a margin-based view
- Non-stationary errors: per-region α
- In-the-loop: pick the error bound that meets a QoI uncertainty budget

Takeaway Compress first, ask any (smooth) QoI later — and know how much to trust the answer.

Thank you!

Any questions are welcome!

Contact: Youyuan Liu youyuan.liu@temple.edu

