A FREE KNOB: DECOUPLING CALIBRATION AND PREDICTIVE SKILL IN THRESHOLD-BASED EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Many dense-prediction benchmarks evaluate rare events by pooling prediction and target over spatial blocks, thresholding each, and scoring the resulting contingency table. At a fixed rare operating point, the max-pooled Critical Success Index (CSI) confounds spatial discrimination with amplitude calibration: sharp observations promote many blocks above threshold, while attenuated predictions from squared-error-type regression leave the same blocks below it. We repurpose classical monotone calibration as a symmetric audit: a post-hoc transform is fitted on held-out data and applied separately to each system. The transform requires no retraining and cannot reverse pixel ordering, so any contrast it reproduces cannot establish improved spatial ranking. On SEVIR, two released checkpoints of one architecture differ by -29.5% in extreme-threshold CSI before the control and by +5.3% after it. Across 450 pairwise contrasts among 6 systems, the difference in pooled frequency-bias deviation is associated with how far the CSI contrast moves under the control ($r = +0.796$ $[+0.753, +0.832]$), and 51 contrasts reverse sign. At CasCast's published extreme-event operating point, the cascade-over-backbone CSI gap falls from 0.1601 to 0.0339, a 78.8% reduction; the remaining gap stays positive. The effect persists when the transform is fitted on a window preceding the test period, and it acts in both directions: calibration also reveals advantages hidden by a better-calibrated baseline. A rarity sweep on geostationary infrared imagery shows the relative gain from the transform growing as events become rarer, crowd-density experiments reproduce the bias–gain relationship under patch-sum pooling, and semantic segmentation marks the boundary: where frequency bias is already near one, the average change is small. The confound therefore requires both a fixed operating point and a training regime that leaves the output miscalibrated there. We recommend reporting pooled frequency bias and a symmetric held-out *FreeKnob Audit*, released as an installable package, alongside rare-event pool-and-threshold scores.

1 INTRODUCTION

Many dense-prediction benchmarks score rare, high-magnitude events in three steps (Gong et al., 2024; Song et al., 2026; Nguyen et al., 2025): the predicted and observed fields are pooled over spatial blocks to tolerate small location errors, each pooled field is thresholded to isolate the events of interest, and the resulting hits, misses, and false alarms are counted. In precipitation nowcasting the usual score is the Critical Success Index (CSI), the number of hits divided by the number of hits, misses, and false alarms. However, CSI also depends on *frequency bias*, the ratio of predicted to observed event counts (Schaefer, 1990): a system that predicts too few events loses score through misses even when every event it predicts is correctly placed, and can raise its score by predicting more events without placing them better. The comparisons audited in this paper report pooled CSI without first equalising how many events each system predicts. We ask one question: *how much of a reported between-system CSI difference can be produced by amplitude calibration alone, without improving where events are placed?*

The question is particularly important on SEVIR (Veillette et al., 2020), a storm-event radar dataset widely used to benchmark deep precipitation nowcasting. At the highest commonly reported SEVIR threshold, only 0.081% of pixels are positive, or about one in every 1238. CasCast (Gong et al., 2024), a nowcasting model that adds a diffusion cascade to a deterministic backbone, reports its

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

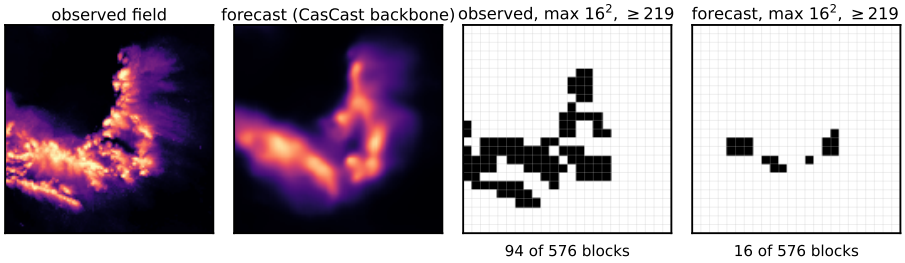

Figure 1: One SEVIR event. The forecast broadly follows the observed storm but attenuates its peaks. After 16×16 max pooling and thresholding at 219, the observation has 94 positive blocks and the forecast 16, illustrating the large deficit in forecast exceedances. Its block ratio, $16/94 \approx 0.17$, is the closest of any frame to the backbone's pooled frequency bias over the whole split (0.172), so the event shows the backbone's typical deficit.

main extreme-event comparison at this threshold using 16×16 max pooling. More recent generative nowcasting studies report the same score (Song et al., 2026; Nguyen et al., 2025). Max pooling, however, treats sharp observations and smoother predictions differently. A block is classified as positive if any pixel within it exceeds the threshold. Thus, a sharp observed event can produce many positive blocks, whereas a smoother prediction of the same event can leave some of those blocks below the threshold (Figure 1). These additional misses push the pooled frequency bias below one, even when the spatial ordering of the prediction remains unchanged (Section 5).

Correcting this frequency-bias deficit changes two comparisons of released models. First, two independently trained checkpoints of EarthFormer's cuboid-attention architecture (Gao et al., 2022) differ by -29.5% in CSI at this operating point. A gap of this size would ordinarily be read as an architectural improvement, yet the two checkpoints share the same architecture. After the same held-out monotone calibration is applied to both checkpoints, the difference becomes +5.3%, and the ranking reverses at 5 of the 6 thresholds (Section 6.2). Second, the CSI gain of the CasCast cascade over its deterministic backbone, evaluated using our common pipeline at CasCast's published operating point, decreases from 0.1601 to 0.0339, a 78.8% reduction, although the gain remains positive (Section 6). Both results show that amplitude calibration alone can substantially change CSI-based comparisons of released models. Such rankings guide which architectures the field adopts, so a calibration difference read as a modelling advance misdirects that choice.

We audit how much calibration alone changes a comparison, using a monotone post-hoc transform fitted on held-out data. The transform shifts or rescales amplitudes, or maps the forecast marginal onto the observed marginal; it requires no retraining and cannot reverse the ordering of two pixels. We call this adjustment a *free knob*: it can change how much forecast mass crosses the operating threshold, but it cannot improve spatial ranking. Under max pooling, the quantile-map version corresponds to matched-percentile thresholding (subject to empirical tie conventions), the classical remedy (Mittermaier et al., 2013). We use that remedy as a symmetric negative control: fit the same calibration procedure separately to every evaluated system (hereafter an *arm*), then ask how much of the original contrast survives.

The audit establishes a concrete identification failure. *Pool-and-threshold CSI at a fixed rare operating point does not separately identify spatial discrimination from amplitude calibration.* A contrast that disappears under this control does not, by itself, establish improved spatial ranking. A surviving contrast is the difference remaining after the specified marginal adjustment; it is not a direct measurement or a guaranteed bound on spatial skill. Pooled frequency bias is the diagnostic: if two arms are evaluated at very different frequency biases, the raw CSI contrast mixes the intended comparison with that operating-point difference.

Across 450 contrasts among 6 arms, pooled frequency-bias differences track contrast movement, with 51 sign reversals (Section 6.1). Calibration can reduce a claimed gain or reveal one hidden by a better-calibrated baseline. Further experiments examine pooling, rarity, and transfer across domains. Crowd counting reproduces the bias–gain relationship; segmentation shows small average movement where frequency bias is near one (Section 7).

Contributions. We (i) quantify how max pooling turns amplitude attenuation into missed events, and how this effect grows with rarity and pooling block size; (ii) decompose published comparisons on released checkpoints and show that pooled frequency-bias differences track contrast distortion across all 450 pairwise contrasts, 51 of which reverse sign; and (iii) show that the effect carries to crowd counting under sum pooling and is small on average in the segmentation control, where frequency bias is already near one. The resulting *FreeKnob Audit*, provided as an installable package in the supplement, reports pooled frequency bias and repeats every comparison after the same held-out monotone calibration procedure is fitted separately to each arm. To our knowledge, it is the first such audit of published deep-learning nowcasting comparisons: a test of how much a reported contrast changes under symmetric marginal calibration.

2 Related work

Frequency bias and verification. CSI (the threat score) has long been known to depend on frequency bias as well as discrimination (Schaefer, 1990; Hamill & Juras, 2006). Verification work also studies the resulting incentives to hedge (Stephenson, 2000; Hogan et al., 2010) and has used probability/frequency matching to compare forecasts at matched event frequencies (Ebert, 2001; Mittermaier et al., 2013), including bias equalisation before numerical models are compared (Hamill, 1999; Pyle & Brill, 2019). Frequency matching has also been applied to a deep-learning nowcast as bias correction (Zhang et al., 2022). We establish how strongly this known sensitivity changes comparisons of released deep nowcasting systems at rare, max-pooled operating points.

Pooling and alternative scores. Spatial pooling belongs to the neighbourhood-verification family (Ebert, 2008; Gilleland et al., 2009). Earlier work showed that fixed-threshold neighbourhood verification can confound intensity error with spatial or size error, and proposed percentile-based thresholds to reduce that effect (Zhu et al., 2015). Our setting differs in using a held-out, symmetric marginal control to audit released modern checkpoints while retaining the original CSI operating point. Related work also shows that neighbourhood scores such as FSS are themselves sensitive to frequency bias (Necker et al., 2024), and that rankings of AI and numerical precipitation forecasts depend on neighbourhood size (Loveday & Hertneky, 2025). Under max pooling, the quantile-map control corresponds to matched-percentile thresholding, subject to empirical tie and interpolation conventions (Section 3). Replacing CSI with another dichotomous score does not remove the problem: bias-adjusted, chance-corrected, and rare-event alternatives make different assumptions and can remain inequitable at low base rates (Mesinger & Brill, 2004; Hogan et al., 2010; Stephenson et al., 2008; Ferro & Stephenson, 2011). Reweighting hits, misses, and false alarms also cannot supply our control, because pooling and thresholding have already determined which samples enter the contingency table; we therefore transform the forecast and re-score the same metric. Appendix O reviews these alternatives, and Appendix K repeats the main decomposition under ETS.

Recalibration and benchmark audits. Monotone recalibration is standard in machine learning (Platt, 1999; Guo et al., 2017), and the broader distinction between calibration and resolution is classical (Murphy & Winkler, 1987; Gneiting et al., 2007). Recent forecast-verification work likewise argues for fair comparison by granting competing deterministic weather models the same statistical postprocessing and evaluating their potential CRPS (Gneiting et al., 2026). That construction uses isotonic distributional regression and yields a score invariant under strictly increasing transformations. The framework has since been extended to extremes using weighted potential CRPS (Biegert et al., 2026); there, model ordering is largely insensitive to which extreme-event thresholds receive emphasis. This is a different stability question from ours: stability of a post-processed potential score across threshold emphasis does not test whether an original fixed-threshold model contrast changes after symmetric marginal calibration. We fit an unpaired marginal map separately to each arm on held-out data and then re-evaluate the original pool-and-threshold CSI; the transform serves as a negative control, and we do not propose a replacement benchmark score. Relatedly, Ranjan et al. (2024) show that model-performance trends can reverse after post-hoc transformations and argue that model-selection conclusions should account for those transformations. This places the paper alongside benchmark audits showing that apparent progress can depend on unequal treatment of compared systems (Musgrave et al., 2020; Ferrari Dacrema et al., 2019; Henderson et al., 2018; Bello et al., 2021). Here the control requires no retraining or hyperparameter search, and its effect is diagnosed by pooled frequency bias.

Sharpness as a reported virtue. Modern nowcasting papers often read high pooled CSI as evidence that predictions are sharper or capture local extremes better. This interaction has been observed directly: FACL reports that for sharp forecasts CSI increases with pooling size, whereas for blurry forecasts it changes little with pooling size (Yan et al., 2024). Their experiment changes the training loss; ours holds released predictions fixed and grants each the same held-out marginal control. Sharpness can be valuable, but fixed-threshold pooled CSI need not separate getting amplitudes across the threshold from putting those amplitudes in the right places. Reporting practice varies: RainPro-8 tabulates frequency bias by threshold and finds that EarthFormer and SimVP substantially under-predict (Pablos Sarabia et al., 2026), whereas the SEVIR comparisons audited here report pooled CSI without it.

3 METHOD: THE FREEKNOB AUDIT

Metric. For a forecast field $\hat{y}$, observation y, threshold τ, and pooling operator Π, let h, f, m be hits, false alarms, and misses after pooling both fields and thresholding:

$$\begin{aligned} h &= \sum \mathbf{1}[\Pi\hat{y} \geq \tau]\mathbf{1}[\Pi y \geq \tau], \\ f &= \sum \mathbf{1}[\Pi\hat{y} \geq \tau]\mathbf{1}[\Pi y < \tau], \\ m &= \sum \mathbf{1}[\Pi\hat{y} < \tau]\mathbf{1}[\Pi y \geq \tau]. \end{aligned} \tag{1}$$

We report $\mathrm{CSI} = h/(h+f+m)$ and pooled frequency bias $\mathrm{Bias} = (h+f)/(h+m)$, accumulating one contingency table over the split. A *cell* is one (Π, τ) pair; the *reference cell* is max 16×16 pooling at $\tau = 219$, where CasCast reports its extreme-event result.

The free knob. Let g be an order-preserving amplitude map applied to the raw forecast before pooling and thresholding. We consider shifts, positive gains, and marginal quantile mapping,

$$\mathcal{G} = \{x \mapsto x + \delta\} \cup \{x \mapsto \gamma x : \gamma > 0\} \cup \{x \mapsto F_y^{-1}(F_{\hat{y}}(x))\}. \tag{2}$$

Such a map cannot reverse the ordering of any two pixels, so it cannot improve spatial ranking; it only changes which amplitudes cross the operating threshold. We call the resulting change in CSI the *knob gain*. The observed marginal used to fit it carries information about amplitude but no paired information about location. Throughout, "calibration" means this marginal, operating-point notion, summarised by pooled frequency bias rather than probabilistic reliability. We use $\mathrm{dev} = |\log \mathrm{Bias}|$ as a symmetric measure of deviation from $\mathrm{Bias} = 1$. We use r for Pearson correlation and ρ for Spearman rank correlation.

For continuous, strictly increasing marginals, $g(\hat{y}) \geq \tau \iff \hat{y} \geq F_{\hat{y}}^{-1}(F_y(\tau))$, which is matched-percentile thresholding (Mittermaier et al., 2013). Our empirical map uses linear interpolation and explicit tie conventions (Appendix A). A monotone map commutes with max pooling but need not with average pooling. The *FreeKnob Audit* applies the same procedure separately to every arm on held-out data, reports the surviving contrast, and places pooled frequency bias beside the raw score.

Fitting protocol. We split events by index parity into calibration and evaluation halves. All main results use one *global* quantile map per arm, fitted once on 1001 quantiles and then held fixed across every threshold and pooling convention. The fit uses 33 evenly spaced calibration events (5.8×10^7 pixels), the largest sample under our fixed pixel cap; refitting on 128 events from an earlier window leaves the conclusions unchanged (Section 6). We also report a stronger *per-cell* variant, selected separately for each (Π, τ) to minimise $|\mathrm{Bias} - 1|$ on the calibration half; its role is to show how conclusions depend on the allowed calibration budget (Section 6 and Appendix E). Neither variant is selected on evaluation CSI. The global map is fitted to the raw marginal, so it does not force post-pooling $\mathrm{Bias} = 1$; Section 6 additionally fits it on a window preceding the test period.

4 EXPERIMENTAL SETUP

Data and models. We use SEVIR (Veillette et al., 2020): 1872 test events, 13 input frames, 12 output frames at 384×384, and VIL thresholds $\{16, 74, 133, 160, 181, 219\}$. The 5 primary arms are persistence (the last observed frame repeated at every lead time), pysteps Lucas–Kanade advection (Pulkkinen et al., 2019), the official EarthFormer checkpoint (Gao et al., 2022), and CasCast's

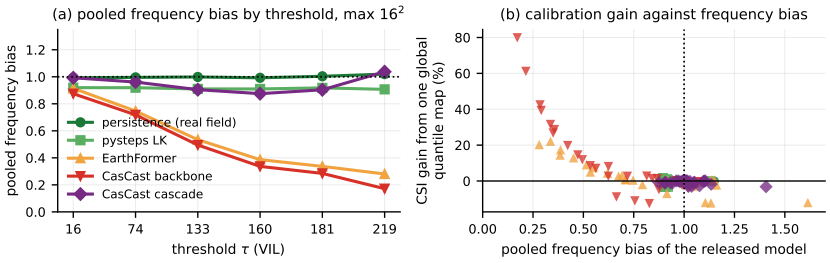

Figure 2: **(a)** Pooled frequency bias across thresholds under max 16×16 pooling. **(b)** The consequence: each of the 30 points is one arm at one threshold under max 16×16, the vertical axis is the CSI gain from a single global quantile map fitted on held-out events. The knob gain grows with the bias deficit and is generally small near $\mathrm{Bias} = 1$ in the evaluated arms. The event in Figure 1 is the frame whose block ratio is closest to the backbone's pooled bias over the whole split.

released deterministic backbone and diffusion cascade (Gong et al., 2024). All learned arms are released artefacts.

Reproduction fidelity. The central experiment decomposes a published number, so matching the released evaluation protocol matters. We run the released checkpoints through one common pipeline and compare our deterministic CasCast backbone directly with the authors' reported cells; the worst discrepancy is 2.8%. One implementation detail affects the scores: CasCast's released evaluation script samples with classifier-free guidance weight 2. Using the unguided path, although it is a natural reading of the model description, gives a different extreme-threshold score. We therefore use the released guided setting throughout and keep the unguided sample only as a sixth diagnostic arm. Appendix A.1 records the checks and Appendix I isolates the guidance effect.

5 The Mechanism and Its Effect Size

Consider a forecast that places precipitation perfectly but reproduces only a fraction of the observed local dynamic range, a characteristic failure mode of squared-error regression under uncertainty: its population minimiser is the conditional mean, which averages over sharp outcomes and can therefore be smooth. Under no pooling this costs the forecast the pixels between its attenuated peak and the threshold. Under max pooling with block size k it is compared against $\max_{k^2} y$, an order statistic that rises with k for the sharp observation and much less for the smooth forecast, so the gap widens with k and with τ and enters the contingency table as misses, driving Bias below one.

Persistence is a control because its forecasts are observed, sharp radar fields. If the pooled marginal is approximately stable over the forecast window, its bias should stay near one and the knob should move its score little; storm growth or decay can violate that condition.

As predicted, persistence's pooled bias stays within $[0.993, 1.020]$ across all six thresholds under max 16×16 pooling, and the global quantile map changes its CSI by at most 0.8%, in some cells slightly negative, as expected for a transform fitted on a different half. Figure 2(b) plots knob gain against bias for the 30 arm–threshold points the 5 primary arms give at max 16×16, the reference convention: the gain is ordered by how far each model's bias sits below one, $\rho = 0.675$, and $\rho = 0.599$ over all 150 cells once the other four pooling conventions are included as well. Both correlations are limited by ties: three of the 5 arms sit at $\mathrm{Bias} \approx 1$ with gains within a percentage point of zero, and their relative order is noise. The same holds for the strongest arm: CasCast's cascade has the highest CSI we score, its bias is already 1.037 at $\tau = 219$, and the map that lifts its own backbone by +79.8% moves it by -2.9%, so calibration adds little once sharpness is restored.

Rarity. The released SEVIR checkpoints are scored at thresholds they were not trained for, and retraining them at each threshold is not feasible. Rarity is therefore varied on a smaller task where the model is retrained at every operating point: downscaling ERA5 reanalysis predictors (Hersbach et al., 2020) to infrared brightness temperature from the geostationary Himawari-8 satellite (Bessho

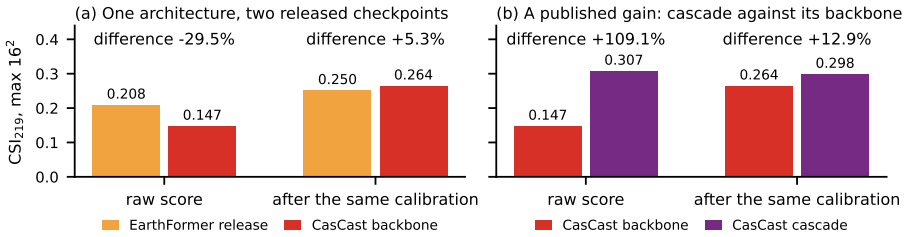

Figure 3: Two comparisons of released SEVIR models at $\tau = 219$ under max 16×16 pooling, before and after the same held-out monotone calibration is applied to each system. **(a)** Two independently trained checkpoints of one architecture. **(b)** A published cascade against its backbone. The calibration cannot reorder pixels, so the part of a raw difference that it removes cannot be attributed to improved spatial ranking.

et al., 2016), with the event defined as a temperature at or below τ (Appendix D). We retrain an attention U-Net at each of six operating points whose positive fraction ranges from $R = 20.270\%$ to $R = 0.264\%$. The knob gain rises from -2.3% at the most common level to +250.8% at the rarest, with monotone ordering ($\rho = -1.000$ against $\log_{10} R$). The absolute CSI change remains bounded within $[-0.0062, +0.0238]$ while CSI itself collapses 31-fold, so calibration becomes an increasingly large fraction of the reported score as events become rare (full sweep and permutation tests in Appendix C).

Pooling. The pooling convention also changes the knob gain. For CasCast's deterministic backbone at $\tau = 219$, it is -8.9% under average 16×16 but +79.8% under max 16×16, an 89-point range with the model held fixed. The deficit deepens with max-pooling block size: the backbone's bias falls from 0.291 unpooled to 0.214 at 4×4 and 0.172 at 16×16, and the gain rises from +39.3% to +61.1% and +79.8%. Under max 16×16 pooling, pysteps outranks EarthFormer at 5 of 6 thresholds; under every other convention it does so at 0 (full results in Appendix E).

6 Decomposing a published comparison

CasCast (Gong et al., 2024) reports that adding its diffusion cascade on top of its own deterministic backbone improves CSI_{219} under 16×16 max pooling by +91.8%, and quotes this as evidence for cascaded modelling of regional extremes. Both stages are released checkpoints, so the comparison can be decomposed directly: we run both through one pipeline on one split and report the 2×2 of $\{\text{backbone}, \text{cascade}\} \times \{\text{uncalibrated}, +\text{global quantile map}\}$. Our preprocessed SEVIR subset differs in sample and window construction from the original CasCast evaluation, so agreement with the published scores is a fidelity check, not an exact reproduction of that protocol. The calibration is the same procedure in both arms, one map per arm fitted on events neither is scored on. A single shared map would not be symmetric, because it favours the arm whose marginal it was fitted on.

At the reference cell, the backbone scores 0.1467 at pooled bias 0.172 and the cascade 0.3069 at 1.037. After symmetric global calibration, the corresponding CSI scores are 0.2639 and 0.2978 (Figure 3b). Thus, in our common-pipeline evaluation, the absolute cascade-over-backbone gap falls from 0.1601 to 0.0339, a 78.8% reduction. Expressed instead as a relative advantage over the backbone, the contrast falls from 109.1% to 12.9%, an 88.2% reduction. Each reduction is one minus the calibrated-to-raw ratio and is reported only where the raw advantage is positive.

Our raw relative gain exceeds the published +91.8% because our cascade reproduction scores above the published value (0.3069 against 0.2841, +8.0%); we therefore treat the published percentage as context rather than as the denominator of the calibrated decomposition.

Beyond the reference cell, the relative-gain reduction increases with threshold, from 36% at $\tau = 74$ to 88% at $\tau = 219$, a threshold dependence consistent with the rarity pattern of Section 5 without identifying rarity separately from intensity. The residual remains positive under the global map; thus marginal matching does not account for the whole gap. The result is also specific to the reporting

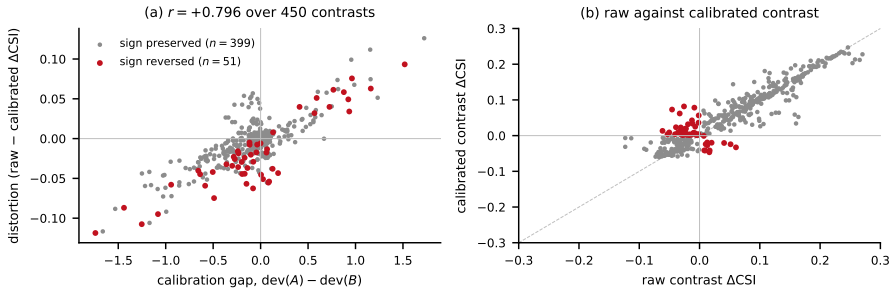

Figure 4: All 450 pairwise contrasts among the 6 SEVIR arms. **(a)** Distortion (raw minus calibrated CSI contrast) against the pair's calibration gap. **(b)** Calibrated against raw contrast; points off the diagonal moved under the control, and points in the upper-left and lower-right quadrants reversed sign. Red marks the 51 reversals in both panels.

convention: under 16×16 average pooling the cascade leads at 0 of 6 thresholds, and at $\tau = 219$ scores 0.2276 against the backbone's 0.2347.

Fitting before the evaluation period. The global map is fitted on the disjoint calibration half of the test split because that is the symmetric control for a between-arm comparison. To test a more deployment-like separation, we refit each arm on 128 events from 1 January to 30 April 2019, which precedes the test period. The relative-gain reduction is then 96% rather than 88%, and the calibration-gap correlation remains $r = +0.766$ rather than $r = +0.796$ (Appendix H).

The surviving contrast also depends on the admissible calibration budget. The global map is restrictive by design: one transform per arm is fixed across all thresholds and pooling conventions. A per-cell control, still fitted only on held-out data and granted to both arms, leaves a residual of -5.5%, which places the cascade behind its backbone. We therefore report the global map as the primary result and the range between the two controls, $[-5.5\%, +12.9\%]$, as the sensitivity range; guidance and alternative knob-family analyses are in Appendices I and N.

6.1 All pairwise contrasts

To extend the decomposition beyond one paper, we run the same 2×2 over every pair of the 6 arms, 15 unordered pairs at 30 cells each, 450 in all. These are repeated comparisons among the same 6 evaluated arms, not independent model replications. For a pair (A, B) the *raw* contrast $\mathrm{CSI}_B - \mathrm{CSI}_A$ is what a results table prints, the *calibrated* contrast is the same difference after one global quantile map fitted per arm on the calibration half is applied to both, and their difference is the distortion. That distortion equals the difference of the two arms' knob gains, so it is not independent evidence for the mechanism. It tests the predictive form of the claim: distortion should be ordered by the pair's *calibration gap*, $\mathrm{dev}(A) - \mathrm{dev}(B)$, in the deviation of Section 3. This test could fail even if the per-arm mechanism held, since two arms miscalibrated in the same direction cancel and leave the contrast between them unchanged. The prediction holds (Figure 4a): $r = +0.796$ over all 450 cells, $r = +0.884$ at the reference pooling, and $r = +0.869$ at the extreme thresholds. In these comparisons, two bias numbers track how far the contrast moves (Spearman $\rho = +0.787$ at the reference pooling, against $\rho = 0.675$ for the single-arm relationship of Section 5).

These cells are not independent, so we resample the independent unit, the event, 2000 times, recomputing every pooled CSI and bias, all 450 contrasts and the correlation inside each replicate, with one event-index vector shared by all arms so the contrasts stay paired. The primary correlation is $+0.884$ $[+0.874, +0.893]$ and over all cells $+0.796$ $[+0.753, +0.832]$; the reversal count is 51 $[45, 56]$. Averaging within each pair before computing the correlation gives $+0.847$ $[+0.801, +0.884]$ (Appendix L).

Table 1 shows, at the most extreme threshold, the 3 pairs that published papers themselves compare: CasCast's cascade against its backbone; the cascade against the official EarthFormer checkpoint, which Gong et al. (2024) evaluate as a baseline for regional extremes; and EarthFormer against

Table 1: Our evaluation of the three model pairs compared in prior work, at max 16×16 pooling and $\tau = 219$. "Raw" is the relative CSI gain reported without calibration; "calibrated" is the same relative gain after the global held-out quantile map is fitted separately to both arms. The final column is the reduction in that relative gain; ↑ marks a gain that grows. Bias columns give the uncalibrated pooled frequency bias of each arm.

comparison	bias_{ref}	bias_{new}	raw	calibrated	rel.-gain reduction
cascade vs. backbone	0.172	1.037	+109.1%	+12.9%	88%
cascade vs. EarthFormer	0.281	1.037	+47.4%	+18.9%	60%
EarthFormer vs. persistence	1.020	0.281	+8.2%	+31.3%	↑

persistence (Gao et al., 2022). The second is a further audited comparison: in our evaluation the margin falls from +47.4% to +18.9%, a 60% relative-gain reduction. Over the 90 cells of these pairs, $r = +0.791$ and 17 contrasts reverse sign. The distortion acts in both directions: calibration shrinks the cascade's gains over its backbone and over EarthFormer, but it reveals a larger EarthFormer advantage over persistence. Across all 450 cells, 51 contrasts reverse sign (Figure 4b), and 23 of those exceed 0.01 CSI on both sides. The reversing contrasts are small in absolute CSI but comparable to published increments in relative terms: their median raw relative gain is 5.7%, and 30 exceed 5%. Excluding the unguided diagnostic arm leaves 40 reversals. Full threshold-by-threshold results are in Appendix L, Table 11.

6.2 Two released checkpoints of the same architecture

A more direct check holds architecture fixed. CasCast's deterministic backbone uses EarthFormer's cuboid-attention architecture, retrained independently on the same benchmark. At the reference cell the two released checkpoints differ by -29.5% without calibration (Figure 3a); after both receive the global map the difference becomes +5.3%, with sign reversals at 5 of 6 thresholds. Their pooled biases are 0.281 and 0.172 (Appendix G). A gap large enough to read as architectural can therefore arise between two trainings of one architecture whose operating-point biases differ. The two groups' recipes differ in more than the seed, so we do not attribute the gap to seed variance; we conclude only that architecture cannot explain it.

7 Generality beyond one benchmark

The claim is not specific to radar. On ShanghaiTech Part A we deliberately use 32×32 patch sums, a reduction that does not create the SEVIR max-pooling asymmetry. Across trained CSRNet arms, the bias deficit still orders the absolute knob gain ($\rho = 0.568$ over 320 cells), although the modality, architecture, and pooling rule all differ from SEVIR.

With the recipe held fixed, seed variation alone produces the effect. Five CSRNet seeds trained with one recipe span pooled bias 0.002–1.146 and CSI 0.0017–0.3194, and pooled bias orders their CSI at $\rho = 0.90$. Their dense-patch count errors differ by a factor of 1.7: the much larger CSI spread is not accompanied by a comparable spread in dense-patch count error (Appendix F). The identification problem therefore appears across published systems and within one training recipe.

Separating placement from attenuation. A target-corruption control separates the two components. Spatial displacement is strongly penalised by CSI, but leaves pooled bias in $[0.926, 0.979]$ and a knob gain of at most 0.0023. A mass-conserving blur likewise remains close to calibration under patch sums. By contrast, compressing each pixel toward its local mean creates the deficit (bias 0.761) and a knob gain of +0.0809. The metric therefore reacts to both placement and amplitude, and the monotone control measures sensitivity to amplitude adjustment (full results in Appendix F).

Semantic segmentation as the boundary case. Per-class intersection over union has the same contingency-table arithmetic,

$$\text{IoU} = \frac{TP}{TP + FP + FN} = \frac{h}{h + f + m} = \text{CSI}. \tag{3}$$

We evaluate each class independently at a probability threshold of 0.5; this is a binary IoU control rather than standard multiclass argmax evaluation. We score deeplabv3-resnet50 on COCO `val2017` without retraining and apply one held-out monotone logit shift per class. Here our rarity prediction fails: over 20 classes, $\rho = -0.019$ between event frequency and knob gain, and the mean movement is only -0.0020 IoU, because 11 of 20 classes already lie within 0.06 of $\mathrm{Bias} = 1$ across a wide range of class frequencies and leave the knob little deficit to remove.

Where segmentation classes are miscalibrated, the relationship reappears: knob gain tracks the bias deficit at $\rho = +0.663$, rising to +0.836 under max 8×8 pooling. A short-bias class such as boat gains +0.0272, while an over-forecasting class such as diningtable is hurt by -0.0351 when pulled toward one. The confound therefore requires two conditions: a fixed operating point and a training regime that leaves the model miscalibrated there. Rarity makes the problem severe for the squared-error-type regression systems we study, but rarity alone is not the diagnostic. Class-level results and the argmax reproduction check are in Appendix B.

8 Discussion and recommendations

Recommended use. The practical recommendation is to report pooled frequency bias beside every rare pool-and-threshold score and, when compared arms sit at different biases, repeat the comparison after the same held-out monotone calibration procedure is fitted separately to each arm, stating the calibration split and the transform family. In reference form, using the `freeknob` package in the supplementary material:

```
report = free_knob_audit(pred[val], obs[val],
            pred[test], obs[test],
            threshold=tau, pooling=pool)
```

The raw-to-calibrated movement is the part of the contrast that can be generated without improving spatial ranking, and the audit is a control for that comparison; it does not produce a corrected leaderboard. In the CasCast decomposition the residual under the global map is +12.9%, and its paired bootstrap interval excludes zero at every threshold (Appendix L). The cascade retains skill; the much larger raw number combines that skill with an amplitude-calibration advantage.

Amplitude as skill and as confound. Generative nowcasters are designed in part to restore the sharp amplitudes that smoother regression models lose, and getting magnitude right at a physical threshold is real predictive skill. A pool-and-threshold metric is therefore not at fault for rewarding amplitude. The identification problem arises when the same score is read as evidence that an architecture places extremes better. If an unpaired, order-preserving marginal transform reproduces most of a contrast, the score alone cannot separate improved placement from improved amplitude.

Scope and limitations. The relative-gain reductions are measured on two published comparisons among three released SEVIR systems, not estimated as a literature-wide average; the effect size has not been established across independent forecasting families and their native benchmarks. Our conclusions also apply specifically to contingency-table scores at fixed operating points after pooling; we did not test CRPS, LPIPS, or SSIM, and we do not transfer that numerical relative-gain reduction to other metrics. The magnitude of the audit depends on the admissible transform family as well: per-cell and per-horizon controls can leave different residuals (Appendices M and N). The audit is an identification device. It does not imply that amplitude is irrelevant or that every residual is spatial skill: a monotone marginal map rules out only what such a map can explain.

Reading and reporting the audit. Agreement indicates stability under the specified control; shrinkage or growth measures its effect on the contrast. A sign reversal prevents reading the raw ordering as an ordering of spatial discrimination. None of these readings treats $\mathrm{Bias} = 1$ as application-optimal. Near-unit pooled bias usually accompanies small movement; exceptions warrant checking the fitted map and the raw-versus-pooled marginal distinction (Appendices B and H). The attenuation hypothesis predicts larger positive knob gains for larger under-forecasting deficits; failure of this association on held-out data would weaken that explanation.

REPRODUCIBILITY STATEMENT

All reported results are generated programmatically from stored experimental artefacts rather than entered manually. We evaluate released checkpoints through a common pipeline, use deterministic calibration/evaluation splits where applicable, and provide the `freeknob` implementation of the proposed audit in the supplementary material. Appendix A documents the evaluation protocol, statistical conventions, checkpoint settings, reproduction checks, and model-inclusion decisions needed to reproduce the experiments.

AI USE STATEMENT

The research question and the identification claim originated with the authors. An AI assistant helped sharpen that framing and helped design the matched-control protocol and robustness checks in the appendices; these were then stress-tested with an AI assistant acting as an adversarial reader. It implemented the `freeknob` package and evaluation pipeline, wrote the data-assembly scripts for the ERA5/Himawari-8 record, and served as a sounding board while we interpreted intermediate statistics and settled on our readings. It also drafted and edited prose, produced figures and tables, and helped search for and summarise related work.

We did not use AI to generate synthetic data, formulate mathematical claims, or assist with translation. The paper contains no proofs, so proof-related categories do not apply.

The authors reviewed all AI-assisted work. Code was checked against a reference implementation that reproduces all thirty SEVIR cells to 10^{-7} (Appendix A); dataset descriptors and references were checked by hand against their sources; and AI-drafted prose was revised throughout. We take responsibility for the final content, including text, claims, and artefacts produced with AI assistance.

REFERENCES

Irwan Bello, William Fedus, Xianzhi Du, Ekin D Cubuk, Aravind Srinivas, Tsung-Yi Lin, Jonathon Shlens, and Barret Zoph. Revisiting ResNets: Improved training and scaling strategies. In *Advances in Neural Information Processing Systems*, 2021.

Kotaro Bessho, Kenji Date, Masahiro Hayashi, Akio Ikeda, Takahito Imai, Hidekazu Inoue, Yukihiro Kumagai, Takuya Miyakawa, Hidehiko Murata, Tomoo Ohno, Arata Okuyama, Ryo Oyama, Yukio Sasaki, Yoshio Shimazu, Kazuki Shimoji, Yasuhiko Sumida, Masuo Suzuki, Hidetaka Taniguchi, Hiroaki Tsuchiyama, Daisaku Uesawa, Hironobu Yokota, and Ryo Yoshida. An introduction to Himawari-8/9: Japan's new-generation geostationary meteorological satellites. *Journal of the Meteorological Society of Japan. Ser. II*, 94(2):151–183, 2016. doi: 10.2151/jmsj.2016-009.

Tobias Biegert, Sam Allen, Annika Alber, and Sebastian Lerch. Towards fair comparisons of AI- and physics-based weather models for extreme events via the weighted potential CRPS. *arXiv preprint arXiv:2606.21170*, 2026.

Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

Christopher Davis, Barbara Brown, and Randy Bullock. Object-based verification of precipitation forecasts. Part I: Methodology and application to mesoscale rain areas. *Monthly Weather Review*, 134(7):1772–1784, 2006.

Elizabeth E Ebert. Ability of a poor man's ensemble to predict the probability and distribution of precipitation. *Monthly Weather Review*, 129(10):2461–2480, 2001.

Elizabeth E Ebert. Fuzzy verification of high-resolution gridded forecasts: a review and proposed framework. *Meteorological Applications*, 15(1):51–64, 2008.

Maurizio Ferrari Dacrema, Paolo Cremonesi, and Dietmar Jannach. Are we really making much progress? A worrying analysis of recent neural recommendation approaches. In *ACM Conference on Recommender Systems*, 2019.

Christopher A. T. Ferro and David B. Stephenson. Extremal dependence indices: Improved verification measures for deterministic forecasts of rare binary events. *Weather and Forecasting*, 26(5): 699–713, 2011.

Zhihan Gao, Xingjian Shi, Hao Wang, Yi Zhu, Yuyang Wang, Mu Li, and Dit-Yan Yeung. Earthformer: Exploring space-time transformers for Earth system forecasting. In *Advances in Neural Information Processing Systems*, 2022.

Zhihan Gao, Xingjian Shi, Boran Han, Hao Wang, Xiaoyong Jin, Danielle Maddix, Yi Zhu, Mu Li, and Yuyang Wang. PreDiff: Precipitation nowcasting with latent diffusion models. In *Advances in Neural Information Processing Systems*, 2023.

Eric Gilleland, David Ahijevych, Barbara G Brown, Barbara Casati, and Elizabeth E Ebert. Intercomparison of spatial forecast verification methods. *Weather and Forecasting*, 24(5):1416–1430, 2009.

Tilmann Gneiting, Fadoua Balabdaoui, and Adrian E. Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B*, 69(2):243–268, 2007.

Tilmann Gneiting, Tobias Biegert, Kristof Kraus, Eva-Maria Walz, Alexander I. Jordan, and Sebastian Lerch. Probabilistic measures afford fair comparisons of AIWP and NWP model output. *Artificial Intelligence for the Earth Systems*, 5(1):e250054, 2026. doi: 10.1175/AIES-D-25-0054.1.

Junchao Gong, Lei Bai, Peng Ye, Wanghan Xu, Na Liu, Jianhua Dai, Xiaokang Yang, and Wanli Ouyang. CasCast: Skillful high-resolution precipitation nowcasting via cascaded modelling. In *International Conference on Machine Learning*, 2024.

Junchao Gong, Siwei Tu, Weidong Yang, Ben Fei, Kun Chen, Wenlong Zhang, Xiaokang Yang, Wanli Ouyang, and Lei Bai. PostCast: Generalizable postprocessing for precipitation nowcasting via unsupervised blurriness modeling. In *International Conference on Learning Representations (ICLR)*, 2025.

Bin Guan and Duane E. Waliser. Detection of atmospheric rivers: Evaluation and application of an algorithm for global studies. *Journal of Geophysical Research: Atmospheres*, 120(24):12514–12535, 2015. doi: 10.1002/2015JD024257.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, 2017.

Thomas M Hamill. Hypothesis tests for evaluating numerical precipitation forecasts. *Weather and Forecasting*, 14(2):155–167, 1999.

Thomas M Hamill and Josip Juras. Measuring forecast skill: is it real skill or is it the varying climatology? *Quarterly Journal of the Royal Meteorological Society*, 132(621C):2905–2923, 2006.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *AAAI Conference on Artificial Intelligence*, 2018.

Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, Adrian Simmons, Cornel Soci, Saleh Abdalla, Xavier Abellan, Gianpaolo Balsamo, Peter Bechtold, Gionata Biavati, Jean Bidlot, Massimo Bonavita, Giovanna De Chiara, Per Dahlgren, Dick Dee, Michail Diamantakis, Rossana Dragani, Johannes Flemming, Richard Forbes, Manuel Fuentes, Alan Geer, Leo Haimberger, Sean Healy, Robin J. Hogan, Elías Hólm, Marta Janisková, Sarah Keeley, Patrick Laloyaux, Philippe Lopez, Cristina Lupu, Gabor Radnoti, Patricia de Rosnay, Iryna Rozum, Freja Vamborg, Sebastien Villaume, and Jean-Noël Thépaut. The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020. doi: 10.1002/qj.3803.

Robin J. Hogan, Christopher A. T. Ferro, Ian T. Jolliffe, and David B. Stephenson. Equitability revisited: Why the "equitable threat score" is not equitable. *Weather and Forecasting*, 25(2): 710–726, 2010.

Yuhong Li, Xiaofan Zhang, and Deming Chen. CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft COCO: Common objects in context. *arXiv preprint arXiv:1405.0312*, 2014.

Nicholas Loveday and Tracy Hertneky. Evaluating extreme precipitation forecasts: A threshold-weighted, spatial verification approach for comparing an AI weather prediction model against a high-resolution NWP model. *arXiv preprint arXiv:2510.25045*, 2025.

Fedor Mesinger. Bias adjusted precipitation threat scores. *Advances in Geosciences*, 16:137–142, 2008.

Fedor Mesinger and Keith Brill. Bias normalized precipitation scores. In *17th Conference on Probability and Statistics in the Atmospheric Sciences, American Meteorological Society*, 2004.

Marion Mittermaier, Nigel Roberts, and Simon A Thompson. A long-term assessment of precipitation forecast skill using the fractions skill score. *Meteorological Applications*, 20(2):176–186, 2013.

Allan H. Murphy and Robert L. Winkler. A general framework for forecast verification. *Monthly Weather Review*, 115(7):1330–1338, 1987.

Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. A metric learning reality check. In *European Conference on Computer Vision*, 2020.

Tobias Necker, Ludwig Wolfgruber, Lukas Kugler, Martin Weissmann, Manfred Dorninger, and Stefano Serafin. The fractions skill score for ensemble forecast verification. *Quarterly Journal of the Royal Meteorological Society*, 150(764):4457–4477, 2024.

Thao Nguyen, Jiaqi Ma, Fahad Shahbaz Khan, Souhaib Ben Taieb, and Salman Khan. RainDiff: End-to-end precipitation nowcasting via token-wise attention diffusion. *arXiv preprint arXiv:2510.14962*, 2025.

Rafael Pablos Sarabia, Joachim Nyborg, Morten Birk, Jeppe Liborius Sjørup, Anders Lillevang Vesterholt, and Ira Assent. RainPro-8: An efficient deep learning model to estimate rainfall probabilities over 8 hours. In *International Conference on Learning Representations*, 2026.

John Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*. MIT Press, 1999.

Seppo Pulkkinen, Daniele Nerini, Andres Perez Hortal, Carlos Velasco-Forero, Alan Seed, Urs Germann, and Loris Foresti. Pysteps: an open-source Python library for probabilistic precipitation nowcasting (v1.0). *Geoscientific Model Development*, 12(10):4185–4219, 2019.

Matthew E Pyle and Keith F Brill. A comparison of two methods for bias correcting precipitation skill scores. *Weather and Forecasting*, 34(1):3–13, 2019.

Rishabh Ranjan, Saurabh Garg, Mrigank Raman, Carlos Guestrin, and Zachary Lipton. Post-hoc reversal: Are we selecting models prematurely? In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Suman Ravuri, Karel Lenc, Matthew Willson, Dmitry Kangin, Remi Lam, Piotr Mirowski, Megan Fitzsimons, Maria Athanassiadou, Sheleem Kashem, Sam Madge, et al. Skilful precipitation nowcasting using deep generative models of radar. *Nature*, 597:672–677, 2021.

Nigel M Roberts and Humphrey W Lean. Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Monthly Weather Review*, 136(1):78–97, 2008.

Joseph T Schaefer. The critical success index as an indicator of warning skill. *Weather and Forecasting*, 5(4):570–575, 1990.

Changhoon Song, Teng-Yuan Chang, and Youngjoon Hong. Extreme weather nowcasting via local precipitation pattern prediction. In *International Conference on Learning Representations (ICLR)*, 2026.

David B. Stephenson. Use of the "odds ratio" for diagnosing forecast skill. *Weather and Forecasting*, 15(2):221–232, 2000.

David B. Stephenson, Barbara Casati, Christopher A. T. Ferro, and Clive A. Wilson. The extreme dependency score: a non-vanishing measure for forecasts of rare events. *Meteorological Applications*, 15(1):41–50, 2008.

Mark Veillette, Siddharth Samsi, and Chris Mattioli. SEVIR: A storm event imagery dataset for deep learning applications in radar and satellite meteorology. In *Advances in Neural Information Processing Systems*, 2020.

Heini Wernli, Marcus Paulat, Martin Hagen, and Christoph Frei. SAL—A novel quality measure for the verification of quantitative precipitation forecasts. *Monthly Weather Review*, 136(11): 4470–4487, 2008.

Chiu-Wai Yan, Shi Quan Foo, Van Hoan Trinh, Dit-Yan Yeung, Ka-Hing Wong, and Wai-Kin Wong. Fourier amplitude and correlation loss: Beyond using L2 loss for skillful precipitation nowcasting. In *Advances in Neural Information Processing Systems*, volume 37, pp. 100007–100041, 2024.

Ming Zhang, Yanhui Yin, Haonan Chen, and Lei Han. A frequency-matching method for bias correction of deep learning nowcasts. In *IGARSS 2022 – IEEE International Geoscience and Remote Sensing Symposium*, pp. 6710–6713, 2022. doi: 10.1109/IGARSS46834.2022.9884009.

Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. Single-image crowd counting via multi-column convolutional neural network. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 589–597, 2016.

Yuchen Zhang, Mingsheng Long, Kaiyuan Chen, Lanxiang Xing, Ronghua Jin, Michael I Jordan, and Jianmin Wang. Skilful nowcasting of extreme precipitation with NowcastNet. *Nature*, 619: 526–532, 2023.

Kefeng Zhu, Yi Yang, and Ming Xue. Percentile-based neighborhood precipitation verification and its application to a landfalling tropical storm case with radar data assimilation. *Advances in Atmospheric Sciences*, 32(11):1449–1459, 2015. doi: 10.1007/s00376-015-5023-9.

A Reproducibility details

Code, artefacts, and evaluation pipeline. Every SEVIR result in this paper is produced from a released checkpoint evaluated through one common pipeline on the official preprocessed data used in our experiments. Tables, figures, and the statistics quoted in the text are generated programmatically from the resulting JSON artefacts. Dataset descriptors, such as the split dates and base rates of Appendix D, are transcribed from the released data audit. The supplementary material includes the installable `freeknob` package, whose `free_knob_audit` function implements the matched-control procedure used in the paper and returns the corresponding diagnostics and pairwise contrasts.

Empirical quantile-map conventions. The implementation linearly interpolates paired forecast and observed quantile knots and extrapolates using the endpoint segments. A left-sided search selects the segment at a knot; a repeated source knot uses the lower endpoint when the segment has zero width. The same deterministic rule is applied to equal inputs. These conventions resolve empirical ties without introducing pixel-specific information.

The supplied parity record covers all thirty SEVIR pooling–threshold cells for the CasCast deterministic backbone and agrees with the reference computation to an absolute tolerance of 10^{-7}. The supplementary verifier checks this stored record; rerunning the complete parity harness additionally requires the external data and prediction files.

Splits and statistical conventions. For the main SEVIR experiments, calibration and evaluation events are separated by index parity, so the split requires no random seed. Unless stated otherwise, rank correlations are Spearman's ρ with tied ranks assigned their average rank, and the associated p-values are computed by exact permutation tests. This convention matters because several evaluated quantities contain exact ties.

SEVIR and CasCast settings. Predictions are handled in VIL units using the released repository's $x/255$ convention. CasCast's diffusion cascade is sampled with classifier-free guidance weight 2 and one ensemble member, i.e. one sample per event, matching the authors' released evaluation script. These choices materially affect the extreme-threshold results. Our re-run of the released deterministic backbone agrees with the published cells to within 2.8%; the cell-level comparison is reported below.

A.1 Reproduction of the CasCast backbone

Table 2 compares our re-run of the released deterministic checkpoint with the four cells reported by Gong et al. (2024). This reproduction check is what allows the analysis in Section 6 to study a released model rather than a re-implementation.

Table 2: Reproduction of the CasCast deterministic backbone against the values published in the authors' Table 4. CSI-M is the mean over the six thresholds.

metric	pooling	published	ours	deviation
CSI-M	none	0.4310	0.4299	-0.2%
CSI-M	max 16^2	0.4351	0.4327	-0.5%
CSI-219	none	0.1448	0.1488	+2.8%
CSI-219	max 16^2	0.1481	0.1467	-0.9%

Model inclusion and exclusion. We restrict the primary comparison to systems for which the relevant released artefacts can be evaluated faithfully. DGMR (Ravuri et al., 2021) and NowcastNet (Zhang et al., 2023) provide released checkpoints, but those checkpoints were trained on different radar products, resolutions, and temporal sampling from SEVIR. Evaluating either on SEVIR would therefore require retraining, turning the comparison into one against our re-implementation rather than the released model.

We also attempted to include PreDiff (Gao et al., 2023). Its original released weight URLs no longer resolve. A third-party mirror could be loaded at the parameter-name, spatial, and channel levels, but its temporal dimensions are inconsistent with the released SEVIR-LR configurations: the denoising UNet and relative-position parameters imply a temporal extent of 5, whereas the released configuration uses 7 context and 6 target frames, and the alignment module has temporal extent 1 against a configured 6. We therefore could not establish that the mirrored weights were the checkpoints underlying the published results and did not substitute them for the original artefact.

These exclusions affect the breadth of the empirical comparison but not the definition of the audit: any additional checkpoint can be evaluated with the same pooled-frequency-bias diagnostic and symmetric held-out calibration procedure.

B Per-class segmentation results

Table 3 gives the per-class results behind Section 7, using DeepLabv3 (Chen et al., 2017) on COCO (Lin et al., 2014). Every class with enough calibration support to fit a knob is shown, ordered by positive-pixel fraction; the gain column does not follow that ordering.

One class shows that near-unit bias does not guarantee an inert knob. tvmonitor sits at bias 0.963, inside the 0.06 band, and still loses -0.0248 IoU. It is a 0.414% class, so the bias estimated on the calibration half is noisy, and the shift that zeroes it there is the wrong shift on the evaluation half. Re-selecting the same family on validation IoU rather than on $|\mathrm{Bias} - 1|$ recovers most of it, -0.0101

instead of -0.0248. Half-to-half fitting noise is one possible explanation for the residual harm at low support, but this comparison does not establish it as the unique cause.

Across the 11 classes inside the near-unit-bias band, 6 move by less than 0.001 IoU, 4 lose, 1 gain, and the mean change is -0.0048. Thus the average movement is small in this band, although individual classes can move appreciably.

Our mIoU is -5.3% relative to the published figure. The gap reflects two protocol differences: (i) all val2017 images, not the VOC-containing subset; and (ii) iscrowd regions kept in the target. We keep the first because restricting to that subset would select images on the classes we are measuring, and the second because deleting those regions would remove exactly the crowded scenes where a rare class is hardest to place. Both choices lower our mIoU relative to the published number, and neither affects the comparison made here, which is each class against itself before and after one monotone shift.

Table 3: COCO `val2017`, deeplabv3-resnet50, released checkpoint. Each class is evaluated independently as a binary mask at probability threshold 0.5; per-class IoU $\equiv$ CSI is reported at that operating point and after one monotone logit shift δ fitted on the disjoint calibration half by $|\mathrm{Bias} - 1|$. This audit is distinct from the separate multiclass argmax reproduction check. Classes are ordered by positive-pixel fraction; those without enough calibration support to fit a knob are omitted rather than fitted on noise.

class	R	Bias	IoU	Δ	class	R	Bias	IoU	Δ
person	9.012%	1.022	0.8155	-0.0008	tvmonitor	0.414%	0.963	0.6992	-0.0248
diningtable	3.002%	1.732	0.3802	-0.0351	horse	0.336%	0.997	0.6969	+0.0010
bus	0.868%	1.005	0.7802	-0.0123	cow	0.299%	0.964	0.7718	+0.0000
train	0.788%	0.975	0.7764	+0.0000	sheep	0.258%	0.988	0.7033	-0.0143
chair	0.729%	0.643	0.3384	+0.0204	aeroplane	0.227%	0.999	0.8153	+0.0000
sofa	0.702%	0.825	0.5066	+0.0226	bottle	0.216%	0.780	0.5125	+0.0207
dog	0.615%	1.003	0.7203	+0.0000	bicycle	0.201%	0.942	0.6093	-0.0019
cat	0.583%	0.945	0.8360	+0.0000	boat	0.171%	0.771	0.4487	+0.0272
motorbike	0.552%	0.899	0.7270	+0.0051	bird	0.163%	0.861	0.6220	-0.0172
car	0.551%	0.846	0.5242	+0.0000	pottedplant	0.156%	1.093	0.3490	-0.0311

C FULL RESULTS OF THE RARITY SWEEP

This appendix gives the per-threshold results behind Section 5. The model is an attention U-Net trained with a weighted mean squared error on geostationary infrared band B08, with the event defined as $y \leq \tau$ kelvin, so that rarity rises down the table. The model is retrained at every threshold, so every row is scored at the operating point its own model was trained for; R and the physical intensity of the event still move together down the table, as they must on real fields. Each row averages 10 runs, five seeds crossed with the two single-decoder loss variants, *graded and tiered*: graded weights each pixel continuously by how far it lies past the threshold, tiered places hard breakpoints on the target and weights the resulting bands. Averaging over a factor can manufacture an ordering, so the sweep is also computed inside each variant: $\rho = -0.943$ for graded and $\rho = -0.829$ for tiered against $\log_{10} R$, against -1.000 pooled, with exact two-sided permutation p of 0.0167, 0.0583 and 0.0028 respectively. We attach no significance to the pooled correlation being the strongest of the three, and the test has limited resolution: six levels admit $6! = 720$ orderings, so the smallest two-sided p available is 0.0028 and only a near-perfect ordering reaches conventional significance. The tiered variant does not. Every departure from a perfect ordering falls at the three most common thresholds, where the gain is within a few percent of zero in both variants and its sign is set by out-of-sample fitting noise rather than by rarity; across the three rarest thresholds, where the claim is made, both variants are separately monotone, and the absolute and median columns of Table 4 support the claim without a rank test. Here R is the all-scene fraction on the valid evaluation footprint, while the coverage-filtered score excludes scenes with observed coverage at or below 0.1%. Because the retained cohort shrinks with τ, we separately recompute pooled contingency scores over all 150 scenes in Appendix D.

Table 4: Gain from a free monotone recalibration against event rarity. R is the all-scene test fraction on the valid evaluation footprint; scores retain the observed-coverage $> 0.1\%$ frame filter. The relative gain is undefined for a run whose uncalibrated CSI is zero, so n gives the runs entering that column against the 10 trained at that τ. Columns three to six are means over those n runs, which is why ΔCSI and ΔCSI (all) differ at the two rarest levels; the last two columns are computed over all 10. The 4 runs the relative column omits score exactly zero uncalibrated and are lifted to between 0.0258 and 0.0315 CSI by the knob, so their true relative gain is unbounded and dropping them can only understate the trend.

τ	R (test)	CSI identity	CSI recal	ΔCSI	rel. gain	n	ΔCSI (all)	median (all)
220	20.270%	0.2672	0.2610	-0.0062	-2.3%	10/10	-0.0062	-0.0048
215	11.537%	0.1992	0.1956	-0.0036	-1.8%	10/10	-0.0036	-0.0011
210	6.378%	0.1381	0.1368	-0.0013	-0.9%	10/10	-0.0013	-0.0039
205	3.124%	0.0578	0.0739	+0.0160	+27.7%	10/10	+0.0160	+0.0143
200	0.884%	0.0117	0.0352	+0.0236	+202.1%	9/10	+0.0238	+0.0245
197	0.264%	0.0085	0.0299	+0.0214	+250.8%	7/10	+0.0237	+0.0265

D THE INFRARED DOWNSCALING TASK

Data and split. We construct this setting rather than adopting a packaged benchmark. A six-hourly scene enters the record when the Guan–Waliser global catalogue (Guan & Waliser, 2015) marks any atmospheric-river object inside 20–27°N, 88–93°E. The usable intersection with geostationary observations contains 1,500 scenes from 2015-07-17 06:00 through 2023-12-17 06:00. The target is Himawari B08 brightness temperature (Bessho et al., 2016) on a $0.02°$ grid. The stored ERA5 (Hersbach et al., 2020) predictors are integrated-vapour-transport magnitude, temperature at 500 and 850 hPa, relative humidity at 700 hPa, and vertical velocity at 500 hPa, linearly interpolated to that grid; the scored models use the latter four, with integrated vapour transport stored but excluded. Predictor means and population standard deviations are fitted per channel on training pixels only.

We split selected scenes chronologically: 1,200 train (2015-07-17 06:00–2021-11-14 18:00), 150 validation (2021-11-15 06:00–2023-03-19 06:00), and 150 test (2023-03-19 12:00–2023-12-17 06:00). This avoids random-split leakage between nearby selected times, but it creates temporal shift: at 220 K the full-grid positive fractions are 18.75%, 13.95%, and 19.62%, respectively. Thus the validation window is the quietest, and the operating point selected there is applied across that shift. Moreover, six-hourly scenes are autocorrelated, so the effective sample size is below the nominal scene count. We disclose both as limitations and report the no-frame-filter pooled sensitivity below.

Geometry, cohort and base rate. Stored targets are 351×251. Evaluation follows the training pipeline's centre crop to a valid 256×251 footprint inside its nominal 256×256 tensor. In the tables, R always means the fraction of valid test pixels with $y \leq \tau$ over all 150 test scenes on that footprint; at 220 K this is 20.270%, against 19.622% on the full grid. The coverage-filtered CSI protocol then scores only scenes whose observed coverage exceeds 0.1%. The event-conditioned training climatology on this footprint is 26.007% at 220 K. These quantities are intentionally different and are emitted together by the released base-rate audit.

As a sensitivity check, we also retain every validation and test scene and form one pooled contingency table per run. The rarity ordering becomes less exact ($\rho = -0.829$, $p = 0.0583$), but the knob gain still grows from +11.7% at 220 K to +92.2% at 197 K. The rarity association therefore does not depend on the threshold-dependent frame cohort.

E FULL RESULTS BY POOLING CONVENTION

Table 5 is the detail behind Section 5: one arm at one threshold under every pooling convention, with the ranking of the 2019 advection scheme against the 2022 transformer counted beside it.

Table 5: Pooling convention determines the size of the calibration effect and the ranking. Bias and CSI are for the CasCast deterministic backbone at $\tau = 219$. The last three columns count, out of six thresholds, how often pysteps (2019 advection) outranks EarthFormer (2022) when both are left uncalibrated, both receive the global quantile map, and both receive the per-cell knob.

pooling	Bias	CSI	CSI+qmap	gain	pysteps > EarthFormer		
					uncal.	+global	+per-cell
none	0.291	0.1488	0.2074	+39.3%	0	0	0
avg 4^2	0.350	0.1724	0.2185	+26.7%	0	0	0
max 4^2	0.214	0.1426	0.2297	+61.1%	0	0	0
avg 16^2	0.664	0.2347	0.2138	-8.9%	0	0	0
max 16^2	0.172	0.1467	0.2639	+79.8%	5	5	0

F Full crowd-density results under the patch-sum protocol

This appendix gives the CSRNet (Li et al., 2018) results behind Section 7. The pool-and-threshold score is our construction; ShanghaiTech (Zhang et al., 2016) itself reports MAE and RMSE on image counts. The domain contributes a sum reduction, the convention that does not induce a deficit on SEVIR. Table 6 gives the trained arms and the constructed controls together, and Table 7 the per-seed spread behind the same section.

Section 7 quotes the bias–gain rank correlation over the trained arms alone, $\rho = 0.568$ over 320 cells, because those are the analogue of the SEVIR statistic, which ranges over systems. Over all 1760 cells, controls included, it reads 0.629. We do not report that figure in the main text: the controls make up most of the cells and form a large tie at exactly zero gain, so the statistic depends on how that tie is ranked.

Table 6: Crowd density under the patch-sum protocol of Section 7 (32×32 patch sums), averaged over 5 seeds. ΔCSI is the absolute change from one monotone knob, selected on a held-out split by $|\mathrm{Bias} - 1|$ and never on CSI. The last two rows are targets we corrupted deliberately, to move placement and sharpness one at a time. Selecting on $|\mathrm{Bias} - 1|$ cannot tune for the score we report, but it is not neutral toward an arm whose bias is long; re-selecting the same family on validation CSI removes the harm and leaves every conclusion unchanged (Appendix J).

arm	pooled Bias			ΔCSI from the knob		
	τ=3	τ=7	τ=10	τ=3	τ=7	τ=10
CSRNet	1.122	0.682	0.289	+0.0072	+0.0352	+0.0911
CNN baseline (MSE)	1.000	1.128	1.426	+0.0021	-0.0339	-0.0442
target, displaced 2 blocks	0.938	0.954	0.969	-0.0009	+0.0000	+0.0000
target, blurred (mass conserved)	0.995	0.975	0.976	+0.0000	+0.0000	+0.0000
target, peaks compressed (mass conserved)	0.942	0.831	0.761	+0.0047	+0.0000	+0.0809

Table 7 is the seed spread quoted in Section 7, with a count-level error beside the score. The seed at $\mathrm{Bias} = 0.002$ has an error on the dense patches of 8.1 persons against 4.7 for the best seed, a factor of 1.7, while its CSI is 188 times smaller. Thus, the much larger CSI spread is not accompanied by a comparable spread in dense-patch count error. This comparison is consistent with the attenuation mechanism of Section 5, but does not by itself establish that every training run converged adequately. The patch error orders the score perfectly across the seeds, $\rho = -1.000$, whereas pooled bias orders it at $\rho = 0.90$ with a single adjacent inversion. This does not show that the score tracks placement. Patch MAE on the dense patches is itself amplitude-sensitive: a flattened density map under-counts where the count is large, so patch MAE and pooled bias both measure the same attenuation, the former at finer resolution. In addition, a patch counts as positive only if it holds at least $\tau = 10$ people, so a predictor that output zero everywhere would score a patch error of at least 10 on exactly these patches; the worst seed's 8.1 is below that bound, and the best seed's 4.7 well below. Every seed performs better than the corresponding all-zero-predictor bound on this dense-patch error, while the score separates them 188-fold. We report patch-level error because the artefacts retain it per cell;

image-level count MAE, this benchmark's principal metric, was not cached per seed and we do not reconstruct it.

Table 7: CSRNet on ShanghaiTech Part A, 5 seeds, one recipe, at $\tau = 10$ under the patch-sum protocol. "patch MAE" is the mean absolute count error over patches above the threshold and "bg MAE" over the rest. Only the seed changes across rows.

seed	pooled Bias	CSI	patch MAE	bg MAE
s42	0.002	0.0017	8.12	0.794
s43	0.107	0.0777	6.23	0.806
s44	0.136	0.0687	6.38	0.885
s45	1.146	0.3194	4.72	0.798
s46	0.053	0.0368	6.61	1.017

G TWO RELEASED CHECKPOINTS OF ONE ARCHITECTURE

Both arms of Table 8 are the same cuboid transformer. CasCast's deterministic backbone reuses EarthFormer's attention code, in the file `networks/earthformer_xy.py`, and was trained by the CasCast authors. This is not a published head-to-head comparison, and we do not present it as one. It measures how large an apparent architectural gap can arise between two trainings of one architecture.

Table 8: One architecture, two independent trainings, max 16×16 pooling. "raw" columns are what a results table would print; "calibrated" columns follow one global quantile map fitted per arm on the calibration half. The sign of the contrast reverses at 5 of the 6 thresholds.

τ	bias EF	bias CasCast	raw ΔCSI	cal. ΔCSI	raw	calibrated
16	0.915	0.875	-0.0109	+0.0065	-1.4%	+0.9%
74	0.747	0.718	-0.0149	-0.0007	-2.2%	-0.1%
133	0.534	0.495	-0.0236	+0.0055	-5.1%	+1.1%
160	0.387	0.337	-0.0315	+0.0054	-9.5%	+1.4%
181	0.337	0.285	-0.0324	+0.0118	-11.3%	+3.4%
219	0.281	0.172	-0.0614	+0.0133	-29.5%	+5.3%

H FITTING THE TRANSFORM ON AN EARLIER VALIDATION WINDOW

Every knob in the main text is fitted on the calibration half of the test split and evaluated on the other half. That control addresses the paper's question, whether a comparison between two arms survives giving both the same free transform, and a symmetric split is appropriate for it. It leaves open whether the transform remains effective when it must be fitted in advance and applied across temporal drift.

SEVIR's preprocessed nowcast release ships a training file and a testing file and no validation file, so the split has to be reconstructed. EarthFormer's loader takes `train_val_split_date` = 2019-01-01 and CasCast's file lists name their validation events `2019_0101_0430`, so both hold out the same window. The test file begins on 2019-06-11. That window is therefore held out from both models we decompose and lies entirely before the evaluation period. We run all six arms over 128 events sampled evenly across it, fit one global quantile map per arm there, and apply it unchanged to the same test evaluation half the rest of the paper scores. Table 9 gives the result at the reference cell.

The conclusion does not depend on which fit is used. The main decomposition reports +109.1% raw and retains +12.9% under the test-half fit against +4.6% under the val-fit, so the relative-gain reduction is 88% against 96%. Over all 450 pairwise contrasts the calibration gap is associated with the distortion at $r = +0.796$ under the test-half fit and $r = +0.766$ under the val-fit, with 51 and 49 sign reversals respectively.

Table 9: The same arms at max 16×16 pooling and $\tau = 219$, scored on the test evaluation half, with the global quantile map fitted two ways. "test-fit" is the paper's default, on the disjoint calibration half of the test split; "val-fit" is on 128 events from 2019-01-01 to 2019-04-30, which precede the test period and were held out of both models' training. The backbone's val-fit score comes from the earlier inference run described in Table 12.

arm	Bias	CSI	CSI test-fit	CSI val-fit	gain test-fit	gain val-fit
persistence	1.020	0.1924	0.1908	0.1936	-0.8%	+0.6%
pysteps	0.907	0.2700	0.2704	0.2719	+0.2%	+0.7%
EarthFormer	0.281	0.2082	0.2505	0.2877	+20.3%	+38.2%
CasCast backbone	0.172	0.1467	0.2639	0.2928	+79.8%	+99.6%
CasCast cascade	1.037	0.3069	0.2978	0.3063	-2.9%	-0.2%
CasCast cascade (no guidance)	0.548	0.2475	0.2899	0.2662	+17.1%	+7.6%

Fitting out of period does not reduce the knob gain. Across the 3 arms that gain from the knob, the val-fitted map is worth between 44% and 188% of what the test-half fit is worth, least for CasCast cascade (no guidance) and most for EarthFormer; and on persistence and pysteps it moves CSI by at most 0.8% under either fit, which is the persistence control of Section 5 repeated with a different fitting set. Those two are named rather than described as the arms near $\mathrm{Bias} = 1$, because pooled bias is not what selects them: the cascade sits at pooled bias 1.037, closer to one than pysteps at 0.907 by a factor of 2.7 in $|\log \mathrm{Bias}|$, and the knob still leaves pysteps unchanged while moving the cascade -2.9% test-fit and -0.2% val-fit. As noted in Section 3, the map is fitted on the raw, unpooled marginal while the metric is computed after pooling, so a small pooled bias deficit does not by itself imply an inert knob. We do not interpret the ratios above one as evidence that an earlier, smaller fitting sample is better; a single window cannot support that. The claim they support is the narrow one, that the knob gain is not an artefact of fitting inside the split it is scored on.

The val-fit column of Table 9 is a point estimate. The interval we quote on the residual, in Appendix L, is fitted on the test calibration half, which is the control the paper's claim is stated against. We attach no interval to the val-fit residual and rest nothing on its significance.

I THE GUIDANCE ARM

Classifier-free guidance weight is a sampling-time scalar that is set after training and changes no model parameter. On this benchmark it nonetheless moves pooled frequency bias. At max 16×16 and $\tau = 219$ the ladder runs backbone, unguided cascade, guided cascade: pooled bias 0.172, 0.548, 1.037, and CSI 0.1467, 0.2475, 0.3069. The free knob gain falls along the same ladder, +79.8%, +17.1%, -2.9%, which is the bias–gain relationship of Section 5 inside one system's own decoding settings. The ladder splits into two legs. Diffusion without guidance covers 67% of the backbone-to-guided distance in $|\log \mathrm{Bias}|$ and 43% of it on bias directly; the guidance weight covers the remaining 33% and 57% respectively. We quote $|\log \mathrm{Bias}|$ first because it is the deviation the pairwise contrasts of Section 6.1 are ordered by; the two scales disagree about which leg is larger, so we state the measure explicitly. A comparison that reports the cascade against the backbone at a fixed guidance weight is reporting a decoding choice alongside an architectural one, and on the measure this paper uses throughout the decoding choice is the smaller of the two. More generally, at a rare threshold under max pooling, any change that restores amplitude is scored as though it had improved placement.

J SELECTION RULES FOR THE TRANSFORM

The measured knob gain depends on both the transform family and its selection rule. A richer family or selection by validation CSI can change either arm's score, but need not reduce the between-arm residual or improve evaluation CSI. It is also not neutral between arms. Selecting on $|\mathrm{Bias} - 1|$ cannot tune for the score we report, but it is not neutral toward an arm whose bias is long: pulling such an arm to $\mathrm{Bias} = 1$ removes hits faster than false alarms at a rare threshold, and costs it CSI. The MSE baseline in Table 6 is that case, and under this rule the knob takes as much as -0.1286 CSI from it. Re-selecting the same family to maximise validation CSI, which gives each arm its best shot

while still never fitting on test, consumes paired labels rather than the observed marginal alone, so it is outside the "unpaired statistics only" description the rest of the paper claims and is reported as a robustness check rather than as the control. It all but removes the harm (worst case -0.0151, from -0.1286) and leaves the conclusion unchanged: CSRNet still recovers up to +0.2394, more than the +0.2162 the conservative rule grants it. We report the conservative rule throughout for that reason.

On SEVIR the two rules differ even less. At the reference cell the two rules select the same member, shift=32, so the backbone reaches 0.3201 either way. Across the cells of Appendix N the validation-CSI rule is not uniformly better: at max 16^2, $\tau = 181$ it selects a member scoring 0.3977 on the evaluation half against the bias rule's 0.4030, having fitted the calibration half's noise. Selecting on the score can lose out of sample what it gains in sample, which is a second reason to report the rule that cannot tune for the score at all.

K Decomposition under the equitable threat score

Table 10 is the decomposition of Section 6 recomputed under the equitable threat score, $\mathrm{ETS} = (h - h_r)/(h + f + m - h_r)$ with $h_r = (h + f)(h + m)/N$ and N the number of pooled blocks over the evaluation half. It reuses the per-event count cache, so the two columns are the same hits, false alarms and misses read by two scores rather than two experiments.

Chance correction does not remove the confound. Under ETS the calibration transform erases between 41% and 88% of the cascade's gain across the 5 thresholds at which there is a gain to erase, and at no threshold does the relative-gain reduction differ from the CSI figure by more than 5 points. The share is left blank at $\tau = 16$, where the cascade does not lead the backbone in the first place and a fraction of the gain is not defined; that row is undefined under both scores. This result is expected: h_r is computed at the forecast's own frequency, so a sharper forecast is charged a larger chance term, but the charge is calibrated to a random forecast and not to the geometry by which pooling converts sharpness into hits. Equitability is a statement about $\mathbb{E}[\text{score}]$ under a random forecast; it is not a statement about bias sensitivity for skilled ones (Hamill & Juras, 2006).

Table 10: The cascade-over-backbone gain of Section 6 at max 16×16, raw and after both arms receive the same monotone knob, scored under CSI and under ETS. "rel.-gain reduction" is the fraction of the raw relative gain removed by the knob.

	CSI			ETS		
τ	raw	calibrated	rel.-gain reduction	raw	calibrated	rel.-gain reduction
16	-0.2%	+4.9%	—	-1.8%	+5.0%	—
74	+7.4%	+4.7%	36%	+7.4%	+4.4%	41%
133	+26.9%	+12.2%	55%	+27.0%	+11.9%	56%
160	+52.4%	+14.8%	72%	+52.0%	+14.3%	73%
181	+64.7%	+14.5%	78%	+64.1%	+14.0%	78%
219	+109.1%	+12.9%	88%	+108.2%	+12.5%	88%

L Bootstrap intervals

Table 12 gives event-level paired bootstrap intervals for the four arms of the decomposition, 2000 replicates over the 936 evaluation events. The resampling is paired: one event-index vector is shared by all four arms within a replicate, so the interval on the contrast reflects the between-arm correlation rather than treating the arms as independent samples. Only the evaluation half is resampled; the calibration half fixes the transform once, so knob selection cannot leak into the interval.

These intervals condition on the fitted calibration maps and exclude calibration-fit uncertainty. The earlier-window refit provides a separate robustness check, rather than an estimate of that uncertainty.

The first contrast of interest is the knob gain of the backbone alone. At the reference cell it is +0.1171 CSI $[+0.1095, +0.1246]$. That is roughly three quarters of the whole cascade-over-backbone difference, and it grows monotonically with the threshold (column six) while the residual does not (column seven), which is where the rising relative-gain reduction in Table 11 comes from. In abso-

Table 11: Full threshold-by-threshold version of the three contrasts summarised in the main text, at max 16×16 pooling. The first block runs from $\tau = 74$ because the relative-gain reduction quoted in the text starts there; the others start at 133. "raw" is the relative gain as a results table would print it, "calibrated" the same gain after one global quantile map is fitted per arm on the calibration half and applied to both, and "rel.-gain reduction" is $1 - \text{calibrated}/\text{raw}$, with ↑ marking rows where calibration raises the gain so that ratio has no stable base. In each label "X vs. Y", Y is the reference arm and X the one claimed better, and the bias columns are their uncalibrated pooled biases in that order.

comparison	τ	bias_{ref}	bias_{new}	raw	calibrated	rel.-gain reduction
cascade vs. own backbone	74	0.718	0.961	+7.4%	+4.7%	36%
	133	0.495	0.905	+26.9%	+12.2%	55%
	160	0.337	0.875	+52.4%	+14.8%	72%
	181	0.285	0.904	+64.7%	+14.5%	78%
	219	0.172	1.037	+109.1%	+12.9%	88%
cascade vs. EarthFormer	133	0.534	0.905	+20.5%	+13.5%	34%
	160	0.387	0.875	+38.0%	+16.4%	57%
	181	0.337	0.904	+46.0%	+18.3%	60%
	219	0.281	1.037	+47.4%	+18.9%	60%
EarthFormer vs. persistence	133	0.999	0.534	+7.3%	+13.1%	↑
	160	0.993	0.387	-0.4%	+16.9%	↑
	181	1.004	0.337	+2.4%	+25.0%	↑
	219	1.020	0.281	+8.2%	+31.3%	↑

lute terms, at $\tau = 219$ a free monotone transform is worth +0.1171 CSI to the backbone, and the architecture it is being credited against is worth +0.0339.

The second contrast, in the last column of Table 12, is the part of the cascade's advantage that remains once both arms carry the same transform. At the reference cell it is +0.0339 $[+0.0234, +0.0439]$, $p < 0.001$. It is positive and its interval excludes zero at every threshold, which is the statistical form of the claim in Section 6 that the cascade retains real skill under the transform we report as primary. The interval is conditional on that transform and not on the knob family: under the per-cell member of $\mathcal{G}$ the same residual reads -5.5%, so the range between these two controls is $[-5.5\%, +12.9\%]$. The quoted interval applies only to the global-map result, not to extrema over $\mathcal{G}$. The intervals are narrow because the resampling is paired: the two arms see the same events, so the between-arm correlation is differenced out rather than added.

Table 12: Paired bootstrap over the 936 evaluation events, 2000 replicates, max 16×16 pooling. Brackets are 95% intervals on the paired contrast; the four CSI columns are point estimates. Column six is what the free knob is worth to the backbone; column seven is what remains of the cascade's advantage once both arms carry that same transform. Both contrasts have $p < 0.001$ at every threshold. Point estimates are those of Section 6. The intervals resample per-event counts cached from an earlier inference run of the deterministic backbone, whose CSI differs from the current run by at most 1.1×10^{-4} in any cell; the counts of every other arm are identical in the two runs.

τ	CSI (point estimate)				knob gain	residual
	backbone	+cal	cascade	+cal	backbone+cal − backbone	cascade+cal − backbone+cal
16	0.7850	0.7471	0.7833	0.7833	-0.0379 [-0.041, -0.035]	+0.0363 [+0.031, +0.041]
74	0.6654	0.6829	0.7146	0.7151	+0.0175 [+0.016, +0.018]	+0.0323 [+0.028, +0.036]
133	0.4442	0.4968	0.5638	0.5574	+0.0525 [+0.049, +0.055]	+0.0606 [+0.055, +0.066]
160	0.3013	0.3962	0.4592	0.4550	+0.0949 [+0.090, +0.100]	+0.0587 [+0.053, +0.065]
181	0.2536	0.3612	0.4176	0.4134	+0.1076 [+0.103, +0.112]	+0.0522 [+0.046, +0.058]
219	0.1467	0.2639	0.3069	0.2978	+0.1171 [+0.109, +0.125]	+0.0339 [+0.023, +0.044]

Table 13: Pooled frequency bias and CSI by lead time at max 16×16, $\tau = 219$, for CasCast's deterministic backbone. "id" is uncalibrated, "G" the single global quantile map, "L" the map fitted per lead time at the equal-pixel budget. The global map overshoots at the first lead and under-corrects at the last; the per-lead map is flatter in bias and scores higher at long leads, but the tables are summed across leads and the early leads carry the hits.

	pooled bias			CSI		
lead	id	G	L	id	G	L
1	0.572	1.158	0.703	0.5000	0.5952	0.5605
2	0.397	0.882	0.539	0.3466	0.5208	0.4344
3	0.262	0.652	0.435	0.2282	0.4167	0.3324
4	0.187	0.508	0.406	0.1604	0.3279	0.2894
5	0.144	0.382	0.367	0.1209	0.2559	0.2485
6	0.117	0.314	0.356	0.0947	0.2084	0.2257
7	0.096	0.265	0.352	0.0752	0.1668	0.2043
8	0.075	0.212	0.301	0.0572	0.1385	0.1722
9	0.064	0.184	0.305	0.0477	0.1159	0.1672
10	0.057	0.155	0.305	0.0426	0.0963	0.1547
11	0.045	0.128	0.303	0.0349	0.0791	0.1496
12	0.032	0.103	0.316	0.0228	0.0677	0.1396

M FITTING THE MAP PER LEAD TIME

A stricter version of the control fits the map separately at each lead time. Contrary to our expectation, this raises the residual instead of lowering it.

Status of the per-lead family. $\mathcal{G}$ is a set of maps of a scalar, so a horizon-indexed family $\{g_t\}$ is not a member of it: applied to the whole field it can send a pixel at lead 3 above one at lead 9 that it had been below. Its use here rests on a narrower property. Pooling is spatial, thresholding is pointwise, and the contingency tables are summed across frames, so every quantity we score is a function of each frame separately and the cross-frame reordering is invisible to all of them. Within a frame the family reorders nothing, and each g_t is fitted on the observed marginal at lead t alone, so it is unpaired in the same sense as the rest of the paper. At these cells it is therefore as free as a member of $\mathcal{G}$, although it is not one.

Two budgets. Each g_t is estimated from a twelfth as many pixels as the global map, and the upper percentiles that decide a $\tau = 219$ cell from a twelfth as many exceedances, so a loss could be estimation noise rather than the family. We separate the two. Under the *matched* budget every map is fitted from the same calibration events, which is the like-for-like comparison: same data, larger family. Under the *equal* budget each g_t is given as many pixels as the global map by drawing more calibration events. The global map is fitted from a fixed event sample in both, so its column is identical across the two and the per-lead columns are comparable.

Result. On the deterministic backbone at the reference cell the knob gain falls from +79.8% under the global map to +74.1% per lead at the matched budget and +62.1% at the equal budget. Decomposing the cascade contrast, whose raw gain is +111.3%, the relative-gain reduction falls from 87.3% to 84.8% and then to 76.4%. The smaller reduction persists under the equal-pixel-budget fit; this comparison does not separately identify family effects and estimation error. Both cascade arms here are the authors' released checkpoints run through their released sampler at the same guidance weight and step count; what differs is the latent-noise realisation, which this pipeline cannot fix exactly because the conditioning latent is a draw from the VAE posterior through an unseeded generator. One reproduces the published sampling scheme, one draws noise independently per event. Their raw gains, +111.3% and +107.6%, straddle the +109.1% of the arm scored in Section 6, and their relative-gain reductions under the global map, 87.3% and 89.1%, straddle its 88%. Per lead they read 84.8% and 86.3% at the matched budget and 76.4% and 76.7% at the equal budget, so neither the direction nor its size turns on which sample we score.

Source of the global map's higher score. Table 13 shows the tradeoff across horizons. At the first lead, the global map produces pooled bias above one and higher CSI than the per-lead map. At the last lead, the per-lead map roughly doubles CSI relative to the global map. Summing contingency tables across horizons weights these changes by their event counts; the resulting pooled CSI favours the global map. Thus horizon aggregation, like spatial pooling, affects the reported comparison.

Implications for the main result. The main relative-gain reduction uses the global map and is unaffected. Because per-horizon marginal matching raises the residual, the global result is not conservative along this axis. Appendix N finds the opposite direction for a bounded shift/gain grid under both selection rules. These are sensitivity results for the tested procedures. The range in Section 6 remains the range between its two reported controls, not a bound over $\mathcal{G}$ or its horizon-specific extensions.

N JOINT SENSITIVITY TO TRANSFORM FAMILY AND SELECTION RULE

Appendix J examines how the selection rule affects the knob gain, mainly on crowd density, and also discusses SEVIR. Appendix M found a per-horizon transform producing a smaller relative-gain reduction than its pooled-horizon counterpart. Both point to the same gap: the knob gain is a function of the family the knob is drawn from and of the rule that picks a member, and the paper reports one of each. This crosses them at the cell the main results are quoted at.

Terminology. "Global" is reserved throughout the paper for the single quantile map held fixed across every cell, which brings the backbone to 0.2639; that map does not appear in this appendix. The rows here are the discrete family of shifts and gains of Section 3, selected per cell: *pooled-horizon* selects one member per cell from a contingency table summed over lead times, which is the paper's per-cell knob, and *per-horizon* selects one member per (cell, lead).

Agreement with the main audit. Under pooled-horizon selection by bias, which is the per-cell knob of Section 6, the backbone goes from 0.1467 uncalibrated to 0.3201 at the reference cell. This matches the main audit's per-cell value, 0.3201, to four figures, although this harness accumulates a separate, lead-resolved contingency table.

Table 14: The knob family crossed against the selection rule, at max 16×16, $\tau = 219$, decomposing the cascade contrast whose raw gain is +111.3%. Both rules see the calibration half only. "per-horizon" selects a member per (cell, lead) and is not a member of $\mathcal{G}$; it is reported as an additional sensitivity result. The range in Section 6 compares its two specified controls only.

family	rule	backbone	cascade	residual	rel.-gain reduction
pooled-horizon	bias	0.3201	0.3059	-4.4%	104.0%
pooled-horizon	val-CSI	0.3201	0.3123	-2.4%	102.2%
per-horizon	bias	0.3366	0.2990	-11.2%	110.0%
per-horizon	val-CSI	0.3413	0.3086	-9.6%	108.6%

Effect of the family and of the rule. Read down Table 14, the two rules differ by less than the two families do. At the reference cell both rules select the same member, shift=32, so the pooled-horizon rows differ only through the cascade arm, and selecting on CSI is not uniformly better: Appendix J gives the cell where it loses out of sample. Crossing to per-horizon moves the backbone from 0.3201 to 0.3413 and the residual from -4.4% to -9.6%. The direction differs from Appendix M under both selection rules. The per-horizon quantile map and the bounded shift/gain grid impose different calibration constraints and select different operating points. This does not establish an inherent inability of shifts or gains to reproduce the quantile map's exceedances: at a single threshold under max pooling, an unrestricted shift can reproduce the same upper set, subject to boundary and tie conventions. The distinction here concerns the finite candidate grid and the fitting procedures tested.

Interpretation. A per-horizon member can send a pixel at one lead above a pixel at another that it had been below, so like the per-horizon map of Appendix M it is not a member of $\mathcal{G}$ as Section 3

defines it. It is as free as one at these cells, for the same reason: pooling is spatial, thresholding is pointwise, and the tables are summed across frames, so no scored quantity sees the cross-frame reordering. The second cascade sample of Appendix M agrees and is marginally stronger: against its raw +107.6% the pooled-horizon residual reads -5.8% and the per-horizon one -10.7%, corresponding to relative-gain reductions of 105.4% and 109.9%. The two samples straddle the -5.5% pooled-horizon residual reported in Section 6, as they straddle 88% under the global map, so the crossing is not a property of which sample of the cascade is scored.

We retain $[-5.5\%, +12.9\%]$ as the range between the two controls in Section 6, and report the per-horizon results as additional sensitivity checks. Both per-horizon rows leave a negative residual. The bias-selected row uses unpaired marginal information; the validation-CSI row uses paired labels and is a separate robustness check. Neither provides a certified bound over the transform family. The abstract's reduction remains tied to the global quantile map.

O ALTERNATIVE VERIFICATION SCORES

Section 2 claims that changing the score does not remove the confound studied here. Because that claim motivates our choice to transform the forecast, we examine the alternatives in turn. Throughout, B denotes frequency bias and s the base rate of the thresholded event; our reference cell sits at small s.

A published post-processor moving the same cell. PostCast (Gong et al., 2025) is the clearest external evidence that this operating point is movable after the forecaster is frozen. It is an unsupervised deblurring stage applied to a finished nowcast, and on SEVIR at the extreme threshold it takes TAU from $\mathrm{CSI} = 0.008$ to 0.043 unpooled, 0.014 to 0.074 under 4×4 max pooling, and 0.028 to 0.163 under 16×16. The absolute movement grows with the pooling scale, $+0.035$ unpooled against $+0.135$ at 16×16, consistent with the bias–gain relationship of Section 5 in independently published results; the relative gain is roughly flat near $5\times$, so it is the pooled cell that has the most room to move, not the raw one. No frequency bias is reported for any arm.

The PostCast transform, however, is a diffusion model conditioned on the blurry input, so it reorders pixels and sits outside $\mathcal{G}$: we cannot say its gain is free, and some of it may reflect placement. Its operating cell nevertheless matches ours: under the standard SEVIR VIL encoding $R(x) = \exp((x - 83.9)/38.9)$ for $x > 18$, their $\tau = 32.24\,\mathrm{kg\,m^{-2}}$ is byte 219 to three figures, and they pool 16×16 with a max. What differs is the grid (their SEVIR is resized to 256×256, so a block covers more ground than ours), and that, with checkpoints we could not confirm as released, is why their artefacts could not enter the pool of Section 6.1 without retraining. These results nonetheless indicate the scale of the effect: a post-hoc operation on a frozen forecast multiplies a reported extreme-threshold pooled CSI by about six, and the comparison is published without a frequency-bias diagnostic or a symmetric marginal control to assess that contribution.

Bias-adjusted threat score. Mesinger & Brill (2004); Mesinger (2008) estimate the threat score at unit bias using an assumed relationship between hits and forecast frequency. In the 2008 formulation, the increase in hits per additional false alarm is proportional to the remaining unhit area. The adjustment estimates a counterfactual hit count from the original contingency table; our audit instead applies a held-out transform to the forecast values and measures the resulting table. Two forecasts with identical original contingency counts can produce different counts after calibration, because those counts do not retain the ordering of values within the thresholded categories. A bias-adjusted score therefore cannot, in general, recover the contrast under our specified transform. It answers a complementary question; we do not claim that its assumptions fail on the fields evaluated here.

Chance-corrected scores. ETS subtracts the hits a random forecast of the same frequency would obtain. That corrects for chance, not for bias, and the two are distinct: a forecast can be perfectly chance-corrected and still be scored at a bias far from one. Hogan et al. (2010) sharpen the point, showing that ETS satisfies neither of Gandin and Murphy's requirements for equitability, and satisfies the first only in the limit of infinite sample size. Appendix K reports the decomposition of Section 6 re-run under ETS. HSS is a strictly increasing function of ETS, so every contrast sign and every reversal is identical under the two by identity rather than by experiment, and we do not report it separately.

Rare-event scores. The extreme dependency score (Stephenson et al., 2008) and the odds-ratio skill score (Stephenson, 2000) were built for the regime we operate in, where linear measures degenerate toward a constant as $s \to 0$; the extremal dependence indices (Ferro & Stephenson, 2011) repair EDS's own hedgeability and base-rate dependence. These scores improve on CSI for rare events, and reporting SEDI alongside CSI would benefit nowcasting evaluation. However, they do not remove the sensitivity examined here. Hogan et al. (2010) show that ORSS and the symmetric extreme dependency score are more strongly inequitable than ETS for rare events (at a base rate of 2% with 1000 samples, an unbiased random forecast scores about -0.5), so at small s the substitution moves the bias sensitivity around rather than removing it. Two further points are specific to our setting. First, none of these scores is what the benchmark reports: the claims we audit are stated in pooled CSI, and an audit conducted with a different score would address a different question. Second, and more basic, the mechanism we document is upstream of the choice of contingency-table score, because it operates on which cells enter the table at all. Pooling is applied to the field before any score sees it.

Neighbourhood and object-based alternatives. FSS (Roberts & Lean, 2008) is the neighbourhood literature's own answer, and it carries the same vulnerability: its large-neighbourhood asymptote is determined by frequency bias rather than by placement (Necker et al., 2024). Feature- and object-based methods provide separate diagnostics of structure, amplitude, and location. MODE uses object matching (Davis et al., 2006), whereas SAL summarises identified objects without one-to-one matching (Wernli et al., 2008). These methods require choices about object identification and, for MODE, matching; their sensitivity to amplitude is not evaluated here. They complement an audit of results already stated in pooled CSI, but answer a different question.

A control on the forecast. Every alternative above modifies the score. Our control modifies the forecast, using a map that is monotone, fitted without reference to the test targets under the protocol of Section 3, and available to every arm at no training cost. Its interpretation is correspondingly narrow: it does not say what the better model is, only how much of a reported gap survives when the free part of the gap is given to everyone. This narrower claim is the one that the released checkpoints can support.