EVEREST: AN EVIDENTIAL, TAIL-AWARE TRANSFORMER FOR RARE-EVENT TIME-SERIES FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Forecasting rare events in multivariate time-series data is challenging due to severe class imbalance, long-range dependencies, and distributional uncertainty. We introduce EVEREST, a transformer-based architecture for probabilistic rare-event forecasting that delivers calibrated predictions and tail-aware risk estimation, with auxiliary interpretability via attention-based signal attribution. EVEREST integrates four components: (i) a learnable attention bottleneck for soft aggregation of temporal dynamics; (ii) an evidential head for estimating aleatoric and epistemic uncertainty via a Normal–Inverse–Gamma distribution; (iii) an extreme-value head that models tail risk using a Generalized Pareto Distribution; and (iv) a lightweight precursor head for early-event detection. These modules are jointly optimized with a composite loss (focal loss, evidential NLL, and a tail-sensitive EVT penalty) and act only at training time; deployment uses a single classification head with no inference overhead (approximately 0.81M parameters). On a decade of space-weather data, EVEREST achieves state-of-the-art True Skill Statistic (TSS) of 0.973/0.970/0.966 at 24/48/72-hour horizons for C-class flares. The model is compact, efficient to train on commodity hardware, and applicable to high-stakes domains such as industrial monitoring, weather, and satellite diagnostics. Limitations include reliance on fixed-length inputs and exclusion of image-based modalities, motivating future extensions to streaming and multimodal forecasting.

1 INTRODUCTION

Problem and setting. Forecasting rare, high-impact events in multivariate time series is important in science and operations (e.g., space weather, industrial monitoring, and power systems). The setting is challenging due to extreme class imbalance, long-range temporal dependencies, and the need for calibrated probabilities and explicit tail-risk assessment. Thresholded, asymmetric decisions are the norm, so average losses are a poor proxy for operational utility; models should be accurate, calibrated, and efficient at inference.

What makes it hard. (1) Imbalance and long horizons: rare positives and long contexts dilute signal. Recent long-horizon approaches improve aggregation via frequency decompositions and patches, and pure-convolutional stacks rival attention for long dependencies, but calibration under class rarity remains nontrivial (Zhou et al., 2022; Nie et al., 2023; Luo & Wang, 2024). (2) Calibration and decision relevance: in high-stakes regimes, miscalibration directly degrades thresholded utility; reliability (ECE, Brier) and decomposition of uncertainty support rational cutoffs (Sensoy et al., 2018; Amini et al., 2020; van Amersfoort et al., 2020). (3) Tail behaviour: catastrophic outcomes live in the far tail, where standard objectives under-weight exceedances. Peaks-over-threshold methods from extreme value theory (EVT) provide a principled way to model exceedances beyond a high quantile (Coles, 2001; de Haan & Ferreira, 2006).

Approach. EVEREST co-optimizes discrimination, calibration, and tail-risk within a single encoder and schedule, producing material rare-event gains while keeping deployment identical to a standard classifier. A single-query attention bottleneck aggregates long-range temporal evidence

with minimal compute, acting as a lightweight, task-conditioned pooling mechanism (cf. global tokens/attention pooling; §2). Training adds evidential (NIG), EVT exceedance, and precursor auxiliaries to shape rare-event dynamics and reliability; deployment uses only the classification logit, with the auxiliaries optional for diagnostics, so inference cost is unchanged.

Contributions. (1) A practical recipe that co-optimizes discrimination, calibration, and tail-risk in one compact backbone (encoder $\rightarrow$ attention bottleneck $\rightarrow$ shallow shared MLP), yielding consistent TSS and reliability gains with unchanged inference. The model is compact (0.81M params) and retains single-head inference (no runtime overhead); auxiliaries are training-only regularizers that shape this shared backbone. (2) Solar-flare SOTA with fair comparisons: on SHARP–GOES (2010–2023), EVEREST reaches TSS 0.973/0.970/0.966 ($\geq$C, 24/48/72 h), 0.898/0.920/0.906 ($\geq$M), and 0.907/0.936/0.966 ($\geq$M5) with strong calibration (e.g., M5–72 h ECE $= 0.016$), on the same SHARP–GOES split, our reported scores are higher than the baseline values (e.g., +0.251 TSS for $\geq$C–48,h; +0.237 for $\geq$M5–72,h; Section 5). (3) Cross-domain transfer on SKAB (industrial valve forecasting) without architectural changes, achieving F1 $= 98.16\%$ and TSS $= 0.964$. (4) Ablations and diagnostics: we quantify the marginal contribution of the bottleneck, evidential, EVT, and precursor heads; report reliability diagrams and cost–loss analyses; and show that attention attributions align with known precursors. (5) Reproducibility and clarity: documented splits/preprocessing, fixed configurations, explicit evaluation protocol (point-wise), and full complexity reporting.

Roadmap. §2 situates the design among recent time-series Transformers, calibration methods, and EVT. §3 details the bottleneck and heads, the composite loss, and thresholding choices. §4 covers datasets, metrics, baselines, and efficiency reporting. §5 presents results, ablations, and diagnostics. §6 discusses limitations (fixed windows, unimodal inputs) and extensions.

2 Related Work

Rare-event time series and imbalance. Rare-event forecasting combines severe class imbalance with long temporal contexts. Early pipelines treated sequences as snapshots (hand-crafted per-window features + classifier), while modern sequence models (TCNs/Transformers) exploit temporal evolution. Cost-sensitive training such as focal loss mitigates skew without distorting the data distribution; by contrast, aggressive oversampling can harm physical consistency and cause temporal leakage, motivating a "train once, weight in loss" strategy for operational settings. recent supervised pipelines report strong solar-flare discrimination at 24–72,h horizons, e.g., CNN/RNN hybrids and task-specific Transformers (e.g., Liu et al., 2019; Sun et al., 2022; Abduallah et al., 2023). In this paper we report our results on the same SHARP–GOES dataset.

Transformers for time series. Transformers are competitive forecasters but naïve self-attention scales as $\mathcal{O}(T^2)$. Recent designs compress or restructure temporal information: patch/token reorganization with channel-first encoders (Nie et al., 2023), frequency/decomposition modules for long horizons (Zhou et al., 2022), and inverted architectures that summarize time before mixing channels (Liu et al., 2024). Pure-convolutional stacks rival attention at lower cost on long sequences (Luo & Wang, 2024). Our single-query attention bottleneck is a lightweight, task-conditioned aggregator in this space, closer to attention pooling/global tokens (Ilse et al., 2018; Lee et al., 2019) than to full self-attention over all steps.

Calibration and evidential learning. Operational thresholds make reliability as important as discrimination. Beyond TSS/AUPRC, reporting ECE and Brier score supports decision-quality assessment; temperature scaling can improve marginal fit but discards input-conditional epistemic cues (Guo et al., 2017). Deterministic OOD surrogates and deep ensembles address uncertainty at higher compute (van Amersfoort et al., 2020; Lakshminarayanan et al., 2017), whereas evidential methods learn distributional parameters in closed form (Dirichlet for classification; NIG for regression) enabling mean/variance without Monte Carlo (Sensoy et al., 2018; Amini et al., 2020). Recent conformal developments provide distribution-shift–robust error guarantees and selection control, complementing probabilistic calibration in time series (Ding et al., 2023). We adopt an evidential NIG head over the logit with explicit reliability reporting.

Tail risk and EVT in ML. Average losses under-weight catastrophic extremes. Peaks-over-threshold modeling with Generalized Pareto exceedances provides a principled account of distribution tails widely used in the sciences; incorporating EVT-inspired objectives focuses learning on high-quantile regions where decisions are costly (Coles, 2001; de Haan & Ferreira, 2006). We adopt a training-time EVT exceedance loss (no inference-time fitting) that complements calibrated probabilities by reallocating gradient mass to high-risk tails.

Auxiliary/precursor supervision and multi-task learning. Auxiliary heads can improve a primary task through a shared backbone and implicit regularization—even when those heads are unused at inference (Caruana, 1997; Standley et al., 2020). We include a lightweight precursor head that provides early-window supervision as a training-only auxiliary objective; in ablations it improves TSS and calibration, while deployment remains single-head.

Industrial anomaly benchmarks. SKAB provides multivariate valve traces widely used for time-series anomaly detection (Filonov et al., 2020); among strong baselines, TranAD reports leading F1 on several valves (Tuli et al., 2022). We retain the same protocol as published research for comparison.

Gap and positioning. Most prior work optimizes sequence encoders *or* calibration in isolation; tail risk is rarely addressed jointly with reliability in a compact model. EVEREST contributes a practical recipe that (i) focuses long contexts via a single-query attention bottleneck, (ii) learns calibrated, closed-form uncertainty (evidential NIG on the logit), and (iii) emphasizes extremes through an EVT exceedance penalty—trained jointly yet deployed with a *single* classification head at test time.

3 METHOD

3.1 PROBLEM FORMULATION AND NOTATION

We consider binary rare–event forecasting on multivariate time series. Each example is a window $X \in \mathbb{R}^{T\times F}$ with label $y \in \{0,1\}$ indicating whether an event occurs within a fixed forecast horizon. The model outputs a probability $\hat{p} \in [0,1]$ used with a decision threshold τ to produce an alert. We report skill with the True Skill Statistic (TSS) and assess reliability with Brier score and Expected Calibration Error (ECE).

3.2 ARCHITECTURE OVERVIEW

The network comprises: (i) an input embedding with scaled positional encoding, (ii) a $6\times$ Transformer encoder, (iii) a single-query attention bottleneck that aggregates the sequence into one latent vector z, and (iv) a shallow shared MLP (128-d) from which four parallel heads branch: a primary binary classification logit (used at inference) and three training-only auxiliaries—evidential (NIG), EVT (GPD) exceedance, and a lightweight precursor head.

Deployment path. Unless explicitly stated otherwise, inference uses only the classification head (single forward pass). Evidential/EVT/precursor heads are training-time auxiliaries; they can be evaluated offline for diagnostics but are never required for test-time decisions.

3.3 EMBEDDING AND TRANSFORMER BACKBONE

Raw inputs X are projected to d-dimensional tokens and combined with sinusoidal positional codes scaled by a learnable global factor α:

$$h_0 = \mathrm{LN}(W_{\mathrm{emb}}X + b_{\mathrm{emb}})\,, \qquad H^{(0)} = \mathrm{Drop}(h_0 + \alpha \cdot \mathrm{PE})\,.$$

We apply L=6 encoder blocks with multi-head self-attention and position-wise feed-forward networks:

$$\tilde{H}^{(l)} = \mathrm{LN}\big(H^{(l-1)} + \mathrm{Drop}[\mathrm{MHA}(H^{(l-1)})]\big), \quad H^{(l)} = \mathrm{LN}\big(\tilde{H}^{(l)} + \mathrm{Drop}[\mathrm{FFN}(\tilde{H}^{(l)})]\big),$$

for $l = 1,\ldots,6$. The reference setting (§4) uses d=128, L=6, H=4 attention heads, FFN width 256, and dropout p=0.20.

3.4 ATTENTION BOTTLENECK (TEMPORAL FOCUSING)

Let $\mathbf{H} = [h_1, \ldots, h_T] \in \mathbb{R}^{d \times T}$ denote the final encoder states and $w \in \mathbb{R}^d$ a learned scorer. We compute a single soft attention distribution over time and the pooled vector

$$\alpha_t = \mathrm{softmax}_t\big(w^\top h_t\big), \qquad z = \sum_{t=1}^{T} \alpha_t \, h_t, \quad w \in \mathbb{R}^d.$$

This *single-query* bottleneck adds only $+d$ parameters and $\mathcal{O}(Td)$ flops, yet concentrates capacity on weak, distributed precursors that global average pooling (GAP) tends to dilute. In ablations (§5), replacing the bottleneck with mean pooling substantially reduces skill (e.g., $\Delta\mathrm{TSS} = +0.427$ on the hardest M5–72 h task).

3.5 HEADS AND PROBABILISTIC TARGETS

The pooled representation z feeds four parallel linear heads:

- **Classification (logit):** $l = W_{\text{clf}} z + b_{\text{clf}}$, with $\hat{p} = \sigma(l)$.
- **Evidential (NIG) head:** predicts (μ, v, α, β) and minimises a *closed-form evidential objective* over the *logit*, yielding analytic predictive mean/variance without Monte Carlo sampling. In ablations it primarily improves *discrimination* (e.g., $\Delta\mathrm{TSS} = +0.064$ on M5–72 h; §5).
- **EVT (GPD) head:** predicts Generalized Pareto parameters (ξ, σ) for logit exceedances above a high batchwise quantile (90% by default), with a stability regularizer; this shifts gradient mass to the risky upper tail and improves rare-event skill.
- **Precursor (auxiliary) head:** trained with the *same binary label* as an auxiliary objective (anticipatory supervision) via binary cross-entropy. It is *not* used at inference. Removing it degrades M5–72 h TSS by -0.650 (§5).

3.6 COMPOSITE LOSS AND TRAINING SCHEDULE

The training objective unifies four complementary criteria—discrimination, calibration, tail awareness, and anticipatory supervision—within a single composite loss. Formally, we optimise

$$\mathcal{L} = \lambda_f \, \mathcal{L}_{\text{focal}} + \lambda_e \, \mathcal{L}_{\text{evid}} + \lambda_t \, \mathcal{L}_{\text{evt}} + \lambda_p \, \mathcal{L}_{\text{prec}}.$$

This structure can be interpreted through the lens of the Information Bottleneck (IB) principle (Tishby et al., 2000): the encoder compresses inputs X into a latent Z while maximising mutual information $I(Z;Y)$ with the event label. Each loss term targets a distinct component of this balance: $\mathcal{L}_{\text{focal}}$ improves separation under extreme rarity, $\mathcal{L}_{\text{evid}}$ regularises predictive entropy (reducing $H(Y|Z)$), $\mathcal{L}_{\text{evt}}$ reallocates bits toward tail exceedances, and $\mathcal{L}_{\text{prec}}$ enriches $I(Z;Y_{\text{early}})$ to capture anticipatory structure.

Focal discrimination. The focal term $\mathcal{L}_{\text{focal}}$ addresses extreme class imbalance by re-weighting misclassified positives. With focusing parameter γ, it emphasises hard rare-event examples:

$$\mathcal{L}_{\text{focal}} = -\tfrac{1}{N} \sum_i (1-\hat{p}_i)^\gamma y_i \log \hat{p}_i + \hat{p}_i^\gamma (1-y_i) \log(1-\hat{p}_i).$$

We anneal $\gamma : 0 \to 2$ linearly over the first 50 epochs, initially allowing broad exploration and later sharpening emphasis on difficult rare-event instances.

Evidential calibration. The evidential term $\mathcal{L}_{\text{evid}}$ learns Normal–Inverse–Gamma (NIG) parameters over the logit, yielding closed-form predictive mean and variance without Monte Carlo sampling. This acts as a Bayesian surrogate: rather than only predicting $\hat{p}$, the model quantifies epistemic and aleatoric uncertainty. In practice, ablations show small effects on ECE but consistent gains in discrimination on the hardest tasks (§5).

Tail emphasis via EVT. The EVT term $\mathcal{L}_{\text{evt}}$ fits a Generalized Pareto Distribution (GPD) to logit exceedances above a high quantile u. For a batch of logits $\{l_i\}$, exceedances $\{l_i - u : l_i > u\}$ are modelled with

$$\Pr(L > u + x \mid L > u) \approx \left(1 + \frac{\xi x}{\sigma}\right)^{-1/\xi},$$

where (ξ, σ) are learned tail parameters. This reallocates gradient mass to rare high-risk predictions, aligning optimization with extreme-value theory and improving sensitivity to catastrophic outcomes.

Precursor supervision. The precursor term $\mathcal{L}_{\text{prec}}$ reuses the binary event label as an auxiliary signal, optimised with binary cross-entropy. It serves as *anticipatory supervision*, encouraging Z to encode early discriminative cues rather than only near-term signals. In the IB view, this enriches $I(Z; Y)$ by regularising toward features predictive of both early and late outcomes.

Weighting and robustness. We set $(\lambda_f, \lambda_e, \lambda_t, \lambda_p) = (0.8,\ 0.1,\ 0.1,\ 0.05)$ by small-grid search. Ablations confirm stability to $\pm 20\%$ perturbations. This reflects the relative dominance of discrimination, with auxiliary heads providing calibrated and tail-sensitive regularization.

Overall, the composite objective can be viewed as enforcing a multi-view consistency: $\mathcal{L}_{\text{focal}}$ drives separation, $\mathcal{L}_{\text{evid}}$ calibrates predictive entropy, $\mathcal{L}_{\text{evt}}$ shapes the heavy tail, and $\mathcal{L}_{\text{prec}}$ enforces temporal anticipation. Together, they yield an encoder that balances predictive skill with uncertainty fidelity under extreme rarity.

optimization. We train with AdamW (β_1=0.9, β_2=0.999), learning rate 3×10^{-4}, weight decay 10^{-4}, and automatic mixed precision on CUDA. The focal parameter γ is annealed linearly from 0 to 2 over the first 50 epochs. One mini-batch update computes all four losses in a single back-propagation pass (no extra memory pass for auxiliaries).

Training vs. inference. All four losses act only at training; deployment uses the classification head $\hat{p} = \sigma(l)$, with uncertainty/tail diagnostics evaluated offline if desired.

3.7 Complexity and efficiency

At the reference shape, the model has $\sim$ **8.14** $\times 10^5$ parameters and $\sim$ **1.66** $\times 10^7$ FLOPs per window; the six-layer backbone accounts for $> 97\%$ of both, while the bottleneck adds only $+d$ parameters. A full per-module budget and a comparison to *SolarFlareNet* (Abduallah et al., 2023) are provided in Appendix A.

4 Experimental Setup

4.1 Datasets and Splits

Solar flares (SHARP–GOES). We adopt the SHARP–GOES protocol and splits consistent with prior work (Abduallah et al., 2023): SHARP vector-magnetogram parameters aligned to GOES flare labels across Solar Cycle 24–25, with standard quality masks (`QUALITY`=0, $|\text{CMD}| \leq 70°$, observer radial-velocity filter) applied before windowing. We use the same nine SHARP parameters and the same window construction for 24/48/72 h horizons. To prevent leakage, we use the identical `HARPNUM`-stratified train/validation/test split; the resulting per-horizon, per-class counts are consolidated in Appendix B (Table 5). All preprocessing (normalization, cadence handling, label alignment) follows that setup to ensure 1:1 comparability.

SKAB (industrial transfer). We evaluate cross-domain transfer on the Skoltech Anomaly Benchmark (SKAB) (Filonov et al., 2020) using fixed-length windows (stride two), stacked raw+diff channels, chronological 70/15/15 splits, and standardization fitted on train only. We do not apply oversampling or task-specific loss reweighting. TranAD is the strongest published reference (Tuli et al., 2022). Full data-processing protocol, model configuration, and the complete results/comparisons are provided in Appendix C (Tables 6, 7).

4.2 Metrics and Evaluation Protocol

Primary and secondary metrics. Our primary discrimination metric is the *True Skill Statistic* (TSS),

$$\mathrm{TSS} = \frac{\mathrm{TP}}{\mathrm{TP+FN}} - \frac{\mathrm{FP}}{\mathrm{FP+TN}},$$

reported at the task-specific operating threshold $\tau^\star$ (below). We also report Precision/Recall/F1, AUROC and PR-AUC for ranking quality, and the *Brier score* for probabilistic accuracy. Reliability is quantified via *Expected Calibration Error* (ECE) with equal-frequency binning (15 bins).

Operating thresholds and cost sensitivity. Decision thresholds are selected by grid search over $\tau \in \{0.10, 0.11, \ldots, 0.90\}$ using the balanced score (40% TSS, 20% F1, 15% Precision, 15% Recall, 10% Specificity). For sensitivity to asymmetric costs, we complement this with a cost–loss sweep (e.g., C_{FN}:C_{FP}=20:1) and report the minimum-cost threshold in §5 alongside the balanced operating point.

Statistical rigor and leakage control. Each model is trained/evaluated over five random seeds; we report means and 95% CIs via 10^4-draw bootstrap on the held-out test set. The solar data splits are `HARPNUM`-stratified to preclude temporal leakage across active-region instances. All threshold selection and early stopping are performed on the validation split only.

4.3 Training details and HPO

All models are trained in PyTorch with automatic mixed precision (AMP), AdamW (β_1=0.9, β_2=0.999), cosine-decayed learning rate, gradient-norm clipping (1.0), and the composite objective from §3 with $\boldsymbol{\lambda}$=$(0.8, 0.1, 0.1, 0.05)$ and focal γ annealed 0→2 over the first 50 epochs. Hyper-parameter optimization follows the three-stage protocol (Sobol scan → Optuna refinement → confirmation), limited to the six knobs that explained the bulk of validation-TSS variance: embedding width d, encoder depth L, dropout p, focal γ, peak LR $\eta_{\max}$, and batch size B. The search priors and the final chosen configuration are in Appendix D; per-scenario optima are tabulated in Appendix D.4.

Statistical protocol. For each threshold–horizon task we train five seeds and report means with 95% CIs via 10^4-draw bootstrap on the held-out test set, stratified by NOAA active-region identifier to preclude temporal leakage. Operating thresholds are selected by a grid over $\tau \in \{0.10, \ldots, 0.90\}$ (step 0.01) using a balanced score (40% TSS, 20% F1, 15% Precision, 15% Recall, 10% Specificity). Unless stated, headline metrics use the task-specific $\tau^\star$ from this procedure. We also filter obviously failed runs and report minimum detectable effects (e.g., $\Delta\mathrm{TSS} \geq 0.02$) alongside p-values from the bootstrap test.

4.4 Figures and Tables for Reproducibility

To keep the setup self-contained within the page budget, we reuse the same artefacts and protocol as our released implementation:

- **SHARP feature list and motivations** (Table 4): the nine input parameters with brief physical rationale.
- **Dataset distribution** (Table 5): counts per horizon, class, and split under `HARPNUM` stratification.
- **CMD filtering diagram** (Fig. 2): effect of the $|\mathrm{CMD}| \leq 70^\circ$ mask on the usable sequence pool during solar data pre-processing.

5 Results

5.1 Headline Performance

Compared to strong baselines on the solar flare dataset (Liu et al., 2019; Sun et al., 2022; Abduallah et al., 2023), EVEREST shows large TSS gains across horizons, with especially strong improve-

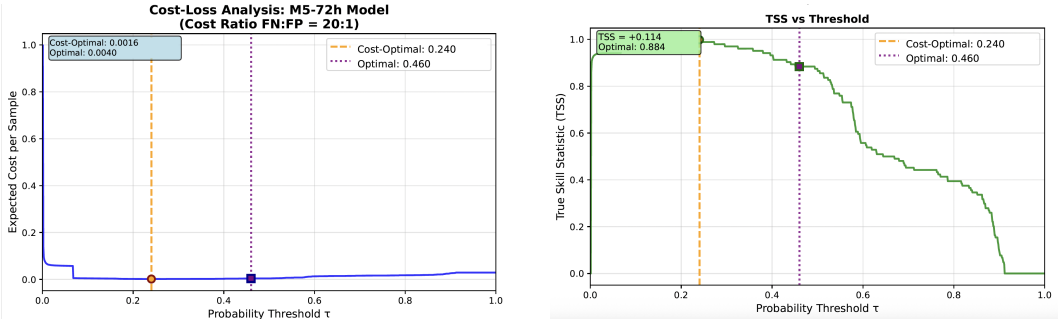

Figure 1: Cost–loss analysis for the M5–72 h model under asymmetric costs (C_{FN}:C_{FP} = 20:1). The left panel shows the cost curve; the right panel highlights the minimum-cost threshold $\tau^\star = 0.240$ versus the balanced-score threshold $\tau = 0.460$.

ments for rare M5 events. All nine tasks exceed the reported baseline TSS values (Table 1). Table 11 reports bootstrapped metrics; EVEREST delivers consistently high discrimination for common C-class events (TSS ≥ 0.966 at all horizons) and strong performance for rarer M and M5 classes. See §5.2 for calibration diagnostics, and Appendix E for full per-task results and operating thresholds.

5.2 Calibration and Reliability

We report calibration with Brier score and Expected Calibration Error (ECE; 15 equal-frequency bins) alongside TSS. On the most imbalanced task (M5–72 h) we obtain ECE $= 0.016$ with a near-diagonal reliability curve; similar trends hold for C–72 h and M–72 h. Diagnostics use the same seeds, splits, and binning as the headline metrics in Table 11. Full reliability diagrams are provided in Appendix F.

5.3 Decision Analysis under Asymmetric Costs

Operational use often values missed-event costs far above false alarms. For M5–72 h, a cost–loss sweep with C_{FN}:C_{FP}=20:1 yields a minimum-cost threshold of $\tau^\star = 0.240$, distinct from the balanced-score $\tau = 0.460$. Figure 1 illustrates the trade-off; the corresponding confusion matrices are in Appendix G.

By threshold class. **$\geq$C:** TSS remains within $0.973/0.970/0.966$ (24/48/72 h), with precision $0.994/0.993/0.992$ and minor horizon decay (ΔTSS$= 0.007$ from 24 h to 72 h). **$\geq$M:** Despite stronger imbalance, TSS reaches $0.898/0.920/0.906$ with recall ≥ 0.908; precision gains with horizon ($0.728 \rightarrow 0.834$). **$\geq$M5:** For the rarest events, TSS is $0.907/0.936/0.966$ with tight CIs and the best ECE (e.g., 0.016 at 72 h).

Comparison to prior work. Table 1 summarizes TSS versus reported baselines. Our reported scores are higher than published baseline values (e.g., +0.251 TSS for $\geq$C–48 h and +0.237 for $\geq$M5–72 h). Significance testing is applied within our models.

This explicit operating-point choice addresses decision relevance under asymmetric costs without retraining, and the full confusion matrices are provided in Appendix G.

5.4 Ablations

A leave-one-component-out suite (five seeds each) quantifies the marginal utility of each module on the hardest task (M5–72 h). Headline effect sizes are:

- **Attention bottleneck:** $+0.427$ TSS over mean pooling.
- **EVT head:** $+0.285$ TSS with major extreme-Brier gains.
- **Evidential NIG head:** $+0.064$ TSS with lower ECE.
- **Composite schedule:** $+0.045$ TSS from γ annealing and stable joint training.

Table 1: TSS performance across flare thresholds and horizons. Bold indicates the best performance within each horizon. Reported values for EVEREST are mean (standard deviation) over 5 seeds.

Method	Horizon	≥C	≥M	≥M5.0
Liu et al. (2019)	24h	0.612	0.792	0.881
Sun et al. (2022)	24h	0.756	0.826	–
Abduallah et al. (2023)	24h	0.835	0.839	0.818
	48h	0.719	0.728	0.736
	72h	0.702	0.714	0.729
EVEREST	24h	**0.973** (0.001)	**0.898** (0.011)	**0.907** (0.025)
	48h	**0.970** (0.001)	**0.920** (0.007)	**0.936** (0.021)
	72h	**0.966** (0.001)	**0.906** (0.012)	**0.966** (0.024)

Removing the precursor auxiliary degrades performance by -0.650 TSS, showing that anticipatory supervision materially shapes the backbone even though it is discarded at inference. Mixed-precision (AMP) was also indispensable: FP32 runs diverged or underperformed. Full per-variant metrics, bootstrap significance tests, and calibration effects are consolidated in Appendix H.

5.5 Interpretability

Saliency analysis highlights how EVEREST differentiates between prediction outcomes. True positives show coordinated increases in `USFLUX` and `MEANGAM` in the final hours before the forecast horizon, consistent with flux emergence and field-inclination steepening. True negatives and false positives exhibit flatter or noisier signatures. Confidence-stratified TP cases show that gradients are strongest when predictive confidence is high. Full gradient visualisations are provided in Appendix I.

5.6 Prospective case study

We evaluate EVEREST on the unseen 6 Sep 2017 X9.3 flare (NOAA AR 12673), the largest event of Solar Cycle 24. Data from 3–7 September 2017 were excluded from training and threshold calibration. The probability trace and lead-time statistics are provided in Appendix J (Figure 7 and Table 15).

5.7 Cross-domain transfer: SKAB

With the architecture unchanged, EVEREST achieves mean TSS = 0.964 and F1 = 98.16% on SKAB (Filonov et al., 2020). We include SKAB because it is multivariate, rare-event–oriented, and widely used in anomaly detection; baseline results (e.g., TranAD)(Tuli et al., 2022). Full valve-level metrics and calibration diagnostics are in Appendix C.

5.8 Efficiency Snapshot

Training uses AMP and the composite schedule from §3. The model is compact (814k params) yet compute-dense (16.6M FLOPs/reference shape), with mean epoch times $\sim$24 s on RTX A6000 and $\sim$69 s on M2 Pro; full energy and carbon accounting appears in the supplement; results remain within typical "Green AI" norms for this model scale.

Summary. Across nine tasks, EVEREST reports higher TSS than the baselines with strong calibration, clear module-level attributions for its gains, and actionable threshold analyses. The same backbone generalises to SKAB without architectural changes.

6 Conclusion

We presented EVEREST, a compact, domain-agnostic Transformer and unified training recipe for rare-event time series that jointly targets discrimination, calibration, and tail-risk. From an infor-

mation–bottleneck perspective (Tishby et al., 2000), the model shapes a latent representation Z that preserves maximal mutual information with the event label Y while discarding nuisance variability. Each auxiliary term enforces a distinct view of this principle: focal loss drives separation under rarity, the evidential head regularises predictive entropy, the EVT penalty reallocates gradient mass to tail exceedances, and the precursor head biases compression toward anticipatory signals. Deployment remains single-head and incurs no inference overhead.

Across nine solar-flare tasks, EVEREST achieves strong TSS (e.g., C: $0.973/0.970/0.966$ at 24/48/72 h; M5: $0.907/0.936/0.966$), with well-calibrated probabilities (e.g., M5–72 h ECE $= 0.016$). The same backbone transfers *unchanged* to SKAB with F1=98.16%, TSS=0.964, surpassing published baselines (Filonov et al., 2020; Tuli et al., 2022). Ablations attribute gains to temporal focusing (+0.427 TSS), EVT tail emphasis (+0.285), and evidential calibration (+0.064). Interpretability analyses show attention concentrating on physically meaningful precursors, and a prospective X9.3 case study demonstrates early, well-calibrated alerts. Training is efficient (814k params, AMP-enabled), supporting practical deployment.

Limitations. Our study inherits several constraints: (i) a fixed context window, which may miss very slow precursor dynamics; (ii) data gaps and quality filters that reduce effective coverage; (iii) potential cycle-dependent drift between training and deployment periods; (iv) extreme scarcity of the highest-magnitude events (e.g., X-class), limiting tail fitting and evaluation; and (v) unimodal inputs—image and radio modalities are not considered here.

Future work. Promising directions include (i) streaming/state-space memory or compressive transformers for indefinite context; (ii) multimodal fusion (e.g., SHARP + EUV/radio) with cadence-aware alignment; (iii) federated or continual training to mitigate cross-cycle drift and institutional data silos; (iv) model compression (quantisation/distillation) and hardware-aware compilation for edge/ops deployment; and (v) richer time-series XAI (counterfactuals, TS-IG) to strengthen operational trust and post-hoc auditing.

Broader impacts. Reliable, calibrated, and tail-aware rare-event forecasts can improve risk communication and decision-making in high-stakes domains (e.g., space weather, industrial monitoring, power systems). EVEREST emphasises small-model efficiency and mixed-precision training, maintaining a "Green AI" footprint while providing actionable probabilities and threshold analyses. We provide an anonymized artifact (code and splits) to support transparent benchmarking and reproducible research.

Reproducibility Statement

Code to reproduce all experiments is provided in the Supplementary Material, including an anonymized repository with `README.md`, `requirements.txt`, and ready-to-run scripts for solar flares (`models/train.py`, `models/evaluate_solar.py`) and SKAB (`models/train_skab.py`, `models/evaluate_skab.py`). The archive includes the exact processed train/validation/test splits, configuration files, and evaluation routines used to report results. Runs use five fixed seeds, mixed precision (AMP), AdamW, cosine learning-rate decay, gradient clipping, and deterministic cuDNN settings; thresholds are selected via a grid sweep and metrics include TSS, Brier score, and ECE with 15 equal-frequency bins. Environment versions are pinned in `requirements.txt`, enabling end-to-end replication.

REFERENCES

Y. Abduallah, X. Wang, W. Xu, B. Zhang, Y. Zheng, and S. E. Gibson. Operational prediction of solar flares using a transformer-based framework. *Scientific Reports*, 13(1):13665, 2023. doi: 10.1038/s41598-023-40884-1. URL https://www.nature.com/articles/s41598-023-40884-1.

Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. Deep evidential regression. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 14927–14937. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/aab085461de182608ee9f607f3f7d18f-Paper.pdf.

Jean Paul A. Audibert, Pietro Michiardi, Frédéric Guyard, Stéphane Marti, and Maria A. Zuluaga. USAD: UnSupervised anomaly detection on multivariate time series. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '20)*, pp. 3395–3404, New York, NY, USA, 2020. ACM. doi: 10.1145/3394486.3403392.

Md Abul Bashar and Richi Nayak. Tanogan: Time series anomaly detection with generative adversarial networks. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 1778–1785, 2020. doi: 10.1109/SSCI47803.2020.9308512. URL https://arxiv.org/abs/2008.09567. Also available as arXiv:2008.09567.

Rich Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997.

Stuart Coles. *An Introduction to Statistical Modeling of Extreme Values*. Springer, 2001.

Laurens de Haan and Ana Ferreira. *Extreme Value Theory: An Introduction*. Springer, 2006.

Tiffany Ding, Anastasios Angelopoulos, Stephen Bates, Michael Jordan, and Ryan J Tibshirani. Class-conditional conformal prediction with many classes. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 64555–64576. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/cb931eddd563f8d473c355518ce8601c-Paper-Conference.pdf.

Min Du, Feifei Li, Guineng Zheng, and Vivek Srikumar. Deeplog: Anomaly detection and diagnosis from system logs through deep learning. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (CCS '17)*, pp. 1285–1298, New York, NY, USA, 2017. ACM. doi: 10.1145/3133956.3134015.

Pavel Filonov, Andrey Lavrentyev, and Andrey Vorontsov. Skab: Skoltech anomaly benchmark. https://github.com/waico/SKAB, 2020. GitHub repository.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. 2017. URL https://arxiv.org/abs/1706.04599.

Maximilian Ilse, Jakub M. Tomczak, and Max Welling. Attention-based deep multiple instance learning. 2018. URL https://arxiv.org/abs/1802.04712.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/9ef2ed4b7fd2c810847ffa5fa85bce38-Paper.pdf.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam R. Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. 2019. URL https://arxiv.org/abs/1810.00825.

Hui Liu, Chang Liu, J. T. L. Wang, and Haimin Wang. Predicting solar flares using a long short-term memory network. *The Astrophysical Journal*, 877(2):121, 2019. doi: 10.3847/1538-4357/ab1b3c.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=JePfAI8fah.

Donghao Luo and Xue Wang. Moderntcn: A modern pure convolution structure for general time series analysis. In *ICLR*, 2024. URL https://openreview.net/forum?id=vpJMJerXHU.

Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Jbdc0vTOcol.

Daehyung Park, Yejin Hoshi, and Charles C. Kemp. A multimodal anomaly detector for robot-assisted feeding using an LSTM-based variational autoencoder. *IEEE Robotics and Automation Letters*, 3(3):1544–1551, 2018. doi: 10.1109/LRA.2018.2801475.

Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. 2018. URL https://arxiv.org/abs/1806.01768.

Trevor Standley, Amir R. Zamir, Dawn Chen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Which tasks should be learned together in multi-task learning? 2020. URL https://arxiv.org/abs/1905.07553.

Ya Su, Youjian Zhao, Chenhao Niu, Rong Liu, Wei Sun, and Dan Pei. Robust anomaly detection for multivariate time series through stochastic recurrent neural network. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '19)*, pp. 1907–1915, New York, NY, USA, 2019. ACM. doi: 10.1145/3292500.3330672.

Pengchao Sun, Wei Dai, Weiqi Ding, Song Feng, Yanmei Cui, Bo Liang, Zeyin Dong, and Yunfei Yang. Solar flare forecast using 3d convolutional neural networks. *The Astrophysical Journal*, 941(1):1, 2022. doi: 10.3847/1538-4357/ac9e53.

Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000. URL https://arxiv.org/abs/physics/0004057.

Shreshth Tuli, Giuliano Casale, and Nicholas R. Jennings. Tranad: Deep transformer networks for anomaly detection in multivariate time series data. *Proceedings of the VLDB Endowment*, 15 (6):1201–1214, 2022. doi: 10.14778/3514061.3514067. URL https://vldb.org/pvldb/vol15/p1201-tuli.pdf.

Joost van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. Uncertainty estimation using a single deep deterministic neural network. 2020. URL https://arxiv.org/abs/2003.02037.

Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. FEDformer: Frequency enhanced decomposed transformer for long-term series forecasting. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 27268–27286. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/zhou22g.html.

Table 2: Per-module parameter and FLOP budget for EVEREST (FP32 multiply–adds; T=10, F=9, batch = 1).

Module	Params (k)	FLOPs (M)
Embedding + positional encoding	1.54	0.03
Transformer encoder $\times 6$	794.88	16.24
Attention bottleneck	0.13	0.00
Classification head	16.64	0.34
Evidential (NIG) head	0.52	0.01
EVT (GPD) head	0.26	0.01
Precursor head	0.13	0.00
Total	**814.10**	**16.63**

Table 3: Complexity comparison with *SolarFlareNet* (T=10, F=9, batch = 1).

Model	Params (k)	FLOPs (M)	FLOPs / Param
SolarFlareNet (Abduallah et al., 2023)	6 120	0.62	0.10
EVEREST	**814**	**16.6**	**20.4**

A Complexity Profile

All numbers refer to a *single* forward pass with T=10 time steps, F=9 SHARP features, and batch size 1.

Per-module budget. The six-layer Transformer backbone accounts for the vast majority of parameters and computation, with 794.9k of 814.1k trainable weights (**97.6%**) and 16.24M of 16.63M FLOPs (**97.7%**). Each backbone weight is thus used about 20.4 times per inference. The auxiliary heads (evidential, EVT, precursor) together contribute only 0.91k parameters (0.11%) and 0.02M FLOPs (0.12%).

Cross-model comparison (SolarFlareNet). We compare EVEREST against *SolarFlareNet* (Abduallah et al., 2023) under the same input shape and profiling settings.

The reference architecture above underpins all reported experiments; hyper-parameter ranges, ablations, and evaluation protocols align with the modules and objectives in Section 3.

B Dataset and Pre-processing

Pipeline. Our data pipeline builds on Abduallah et al. (2023), enhancing temporal fidelity (12-minute cadence), enforcing stricter quality masks, and version-controlling all outputs. SHARP vector magnetograms (SDO/HMI) are merged with GOES flare data (NOAA/SWPC), programmatically harvested (JSOC, SunPy HEK), and segmented into supervised, HARPNUM-stratified windows.

Features. Nine SHARP parameters were retained from the original 25, following physical interpretability and prior studies (Abduallah et al., 2023). Table 4 lists the features.

Split strategy. The mission window spans May 2010–May 2025. We create datasets for nine tasks (three flare thresholds $\times$ three horizons). Each HARPNUM appears in exactly one split. Table 5 gives the per-class distribution.

C SKAB Industrial Anomaly Benchmark

To assess cross-domain transfer, we evaluate EVEREST on the Skoltech Anomaly Benchmark (SKAB) (Filonov et al., 2020), a suite of multivariate valve-sensor traces with rare fault events.

Table 4: Selected SHARP features and their physical motivations.

Feature	Description	Physical motivation
TOTUSJH	Total unsigned current helicity	Magnetic twist; non-potentiality
TOTPOT	Total magnetic free energy density	Energy reservoir for reconnection
USFLUX	Total unsigned flux	AR size / activity
MEANGBT	Gradient of total field	Localised magnetic complexity
MEANSHR	Mean shear angle	Shearing near PIL
MEANGAM	Mean angle from radial	Loop inclination
MEANALP	Twist parameter α	Field line torsion
TOTBSQ	Total field strength squared	Energetic capacity
R_VALUE	PIL integral	Complexity near polarity inversion

Table 5: Number of positive and negative examples per flare class and horizon.

Flare	Horizon	Split	Positives	Negatives
C	24h	Train	244,968	218,217
		Test	31,897	15,878
	48h	Train	316,149	301,714
		Test	40,987	21,573
	72h	Train	356,219	350,853
		Test	46,066	25,663
M	24h	Train	13,989	449,196
		Test	1,368	46,407
	48h	Train	16,709	601,154
		Test	1,775	60,785
	72h	Train	18,505	688,567
		Test	2,131	69,598
M5	24h	Train	2,125	461,060
		Test	104	47,671
	48h	Train	2,255	615,608
		Test	104	62,456
	72h	Train	2,375	704,697
		Test	104	71,625

We adopt the standard windowing (24 steps, stride two), stacked raw+diff channels, chronological 70/15/15 splits, and standardisation fitted on train only. No oversampling or loss reweighting is used. Architecture and loss weights are unchanged except for a reduced width (d=96) and depth (L=4) to match the smaller dataset scale.

Results. Table 6 reports mean performance across all eight valve scenarios. EVEREST achieves strong discrimination (TSS 0.964 ± 0.028) and calibration, with F1 exceeding 98%.

Comparison with baselines. Table 7 situates our results against prior published methods. EVEREST surpasses the strongest reported baseline (TranAD (Tuli et al., 2022)) by roughly two F1 points, without task-specific tuning.

D Hyper-parameter Optimisation

Operational deployment values three traits above all: **forecast skill, probabilistic reliability, and inference latency**. We therefore tune only the hyper-parameters that collectively maximise skill $\times$ latency^{-1}.

Method synopsis. We run a three-stage Bayesian study (**Optuna v3.6** + Ray Tune) over six knobs: embedding width d, encoder depth L, dropout p, focal exponent γ, peak learning rate $\eta_{\max}$, and

Table 6: EVEREST averaged across all SKAB valves.

Metric	Precision (%)	Recall (%)	F1 (%)	TSS
EVEREST	97.7 ± 2.9	98.6 ± 3.2	98.2 ± 1.7	0.964 ± 0.028

Table 7: F1 comparison on SKAB valve anomalies.

Model	Reference	F1 (%)
Isolation F, LOF, etc.	Filonov et al. (2020)	65–75
Autoencoder	Filonov et al. (2020)	70–80
CNN/LSTM hybrids	Filonov et al. (2020)	75–85
TAnoGAN	Bashar & Nayak (2020)	79–92
DeepLog	Du et al. (2017)	87–91
LSTM-VAE	Park et al. (2018)	86–93
OmniAnomaly	Su et al. (2019)	88–94
USAD	Audibert et al. (2020)	89–95
TranAD	Tuli et al. (2022)	91–96
EVEREST	—	**98.2 ± 1.7**

batch size B. Median-stopping pruning halves the number of full trainings needed. A Sobol sensitivity scan (Appendix D.1) confirmed that these six knobs explain 91 of the variance in validation TSS.

Search logistics. Each flare-class/lead-time pair receives $\sim$165 trials split into *exploration*, *refinement*, and *confirmation* phases; exact budgets and early-stop criteria are in Appendix D.2.

Final tuning space and winner. Table 8 summarises the priors and the final configuration adopted for all production models. Full per-scenario optima are in Appendix D.4.

The selected tuple (d=128, L=6, γ=2, p=0.20, $\eta_{\max} = 4\times10^{-4}$, B=512) achieves **TSS** $= 0.795 \pm 0.005$ and inference latency of 4 ± 0.6 **s** on an NVIDIA RTX 6000. These values are frozen for all ablations and the compute-budget audit.

D.1 Sobol sensitivity scan

A 64-trial Sobol sweep assessed first-order and total-order effects; the six retained knobs jointly explain 91 of variance in validation TSS. Full indices and code are in the repository.

D.2 Search protocol

Each study followed the three-stage schedule in Table 9. Trials were pruned with Optuna's median rule after five epochs.

D.3 Hyper-parameter rationale

1. **Capacity:** d and L govern receptive field and FLOPs.
2. **Regularisation:** p mitigates over-fit.
3. **Imbalance:** γ addresses the 1:297 positive/negative ratio.
4. **Optimiser dynamics:** $\eta_{\max}$ sets AdamW step size.
5. **Throughput:** B trades GPU utilisation for generalisation.

Table 8: Search priors and final hyper-parameters used in production

Hyperparam	Prior	Rationale	Best
Embedding d	$\{64, 128, 192, 256\}$	capacity vs. latency	**128**
Encoder depth L	$\{4, 6, 8\}$	receptive field	**6**
Dropout p	$\mathcal{U}[0.05, 0.40]$	over-fit control	**0.20**
Focal γ	$\mathcal{U}[1, 4]$	minority gradient	**2.0**
Peak LR $\eta_{\max}$	Log-$\mathcal{U}[2\times10^{-4},\ 8\times10^{-4}]$	step size	4×10^{-4}
Batch size B	$\{256, 512, 768, 1024\}$	throughput vs. generalisation	**512**

Table 9: Trial budget per stage for each flare-class/lead-time study

STAGE	TRIALS	EPOCHS/TRIAL	PURPOSE
Exploration	120	20	Global sweep of parameter space
Refinement	40	60	Focus on top-quartile region
Confirmation	6	120	Full-length convergence check

D.4 Per-scenario optima

Table 10 lists the best trial for each of the nine studies.

A clear pattern emerges: C and M classes share a single optimum across all windows, while M5 requires larger capacity for short horizons and deeper, narrower networks for 72h forecasts.

D.5 Additional data pre-processing visuals

- **CMD filtering diagram** (Fig. 2): effect of the $|\mathrm{CMD}| \leq 70^\circ$ mask on the usable sequence pool during solar data pre-processing.

Table 10: Best hyper-parameters per flare class and forecast window

FLARE	WINDOW	d	L	p	γ	$\eta_{\max}$ (10^{-4})	B	TIME (s)
C	24h	128	4	0.353	2.803	5.337	512	3323
C	48h	128	4	0.353	2.803	5.337	512	4621
C	72h	128	4	0.353	2.803	5.337	512	4856
M	24h	128	4	0.353	2.803	5.337	512	3705
M	48h	128	4	0.353	2.803	5.337	512	5105
M	72h	128	4	0.353	2.803	5.337	512	5871
M5	24h	192	4	0.300	3.282	4.355	256	3778
M5	48h	192	4	0.300	3.282	4.355	256	4977
M5	72h	64	8	0.239	3.422	6.927	1024	5587

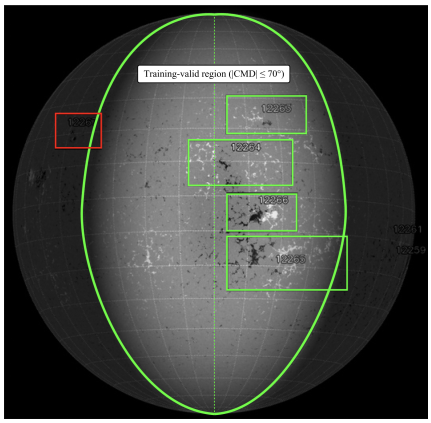

Figure 2: Central–meridian–distance (CMD) quality mask applied to an HMI synoptic magnetogram. The bright-green curve marks the acceptance limit $|\mathrm{CMD}| = 70°$; grey wedges beyond this boundary are discarded. Active-region boxes are color-coded by the centroid rule: green outlines (e.g., AR 12263, 12266) fall inside the limit and are retained, whereas red outlines (e.g., AR 12267) lie outside and are excluded. The mask removes limb data affected by foreshortening and line-of-sight artifacts while preserving the central disk used for training and evaluation.

E Extended Results and Protocols

E.1 Experimental Protocol

We evaluate nine benchmark tasks (three flare thresholds: C, M, M5; three horizons: 24h, 48h, 72h). Performance statistics are computed via 10,000-fold bootstrap resampling with splits stratified by NOAA active-region identifier to avoid temporal leakage. Each task is trained and evaluated with 5 random seeds; metrics are aggregated as mean (standard deviation) unless otherwise noted. Thresholds are selected by a balanced scoring rule over a grid of 81 values in $[0.1, 0.9]$ (step 0.01), with a fallback of 0.5 if no improvement is found. Statistical significance is assessed at $p < 0.05$ with a minimum effect size threshold $\Delta\mathrm{TSS} \geq 0.02$.

E.2 Bootstrapped Metrics (Full)

Table 11 reports bootstrapped performance on the held-out test set for all nine tasks (higher is better for TSS/Precision/Recall; lower is better for Brier/ECE).

Table 11: Bootstrapped performance (mean ± 95% CI) of EVEREST on the held-out test set. Thresholds are the task-specific optima from the balanced scoring rule.

Task	TSS	Precision	Recall	Brier	ECE
C-24h	0.973 ± 0.001	0.994 ± 0.000	0.986 ± 0.001	0.015 ± 0.000	0.049 ± 0.000
C-48h	0.970 ± 0.001	0.993 ± 0.000	0.984 ± 0.001	0.017 ± 0.000	0.054 ± 0.000
C-72h	0.966 ± 0.001	0.992 ± 0.000	0.982 ± 0.001	0.018 ± 0.000	0.052 ± 0.000
M-24h	0.898 ± 0.011	0.728 ± 0.016	0.908 ± 0.011	0.011 ± 0.000	0.037 ± 0.001
M-48h	0.920 ± 0.007	0.772 ± 0.010	0.928 ± 0.007	0.009 ± 0.000	0.029 ± 0.000
M-72h	0.906 ± 0.012	0.834 ± 0.015	0.911 ± 0.012	0.010 ± 0.000	0.033 ± 0.001
M5-24h	0.907 ± 0.025	0.686 ± 0.033	0.908 ± 0.025	0.003 ± 0.000	0.031 ± 0.000
M5-48h	0.936 ± 0.021	0.713 ± 0.035	0.937 ± 0.021	0.002 ± 0.000	0.020 ± 0.000
M5-72h	0.966 ± 0.024	0.727 ± 0.053	0.966 ± 0.024	0.002 ± 0.000	0.016 ± 0.000

Table 12: Statistical significance of TSS improvements over the strongest baseline (Abduallah et al. 2023). EVEREST values are mean (95% CI) from 10,000 bootstrap resamples stratified by HARPNUM. Asterisks denote bootstrap p-values for the null $H_0: \Delta\mathrm{TSS} \leq 0$: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Task	Baseline TSS	EVEREST TSS	Effect size ΔTSS
C-24h	0.835	0.973 (0.001)	+0.138***
M-24h	0.839	0.898 (0.011)	+0.059***
M5-24h	0.818	0.907 (0.025)	+0.089***
C-48h	0.719	0.970 (0.001)	+0.251***
M-48h	0.728	0.920 (0.007)	+0.192***
M5-48h	0.736	0.936 (0.021)	+0.200***
C-72h	0.702	0.966 (0.001)	+0.264***
M-72h	0.714	0.906 (0.012)	+0.192***
M5-72h	0.729	0.966 (0.024)	+0.237***

E.3 Significance vs. Baseline

We compare against the strongest baseline (Abduallah et al., 2023). Improvements are significant at $p < 0.01$ for all nine tasks (Table 12; bootstrap hypothesis testing).

E.4 Calibration and Operating Points

Reliability diagrams (15 equal-frequency bins) and cost–loss analyses are provided for representative tasks (figures referenced in the main text). Operating thresholds τ^* for each task are the grid-search optima under the balanced scoring rule; values are available in the code repository and summary tables.

F Additional Calibration Plot

G Confusion matrix analysis under asymmetric costs

The confusion matrices below quantify the effect of selecting different operating thresholds on the M5–72 h task. At the balanced-score threshold ($\tau = 0.460$), the model achieves strong overall discrimination but incurs some false negatives. At the cost-minimising threshold ($\tau^\star = 0.240$), all false negatives are eliminated at the expense of more false positives.

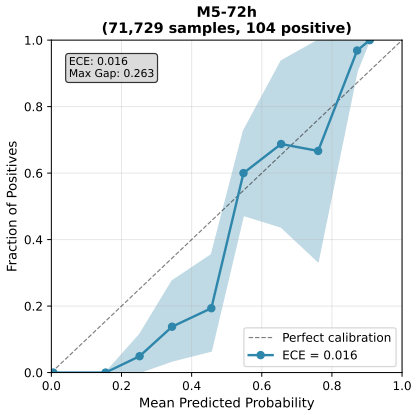

Figure 3: Reliability diagram for the M5–72 h task. Shaded region shows 95% bootstrap confidence intervals; the dashed line indicates perfect calibration. ECE $= 0.016$ with maximum bin gap 0.263.

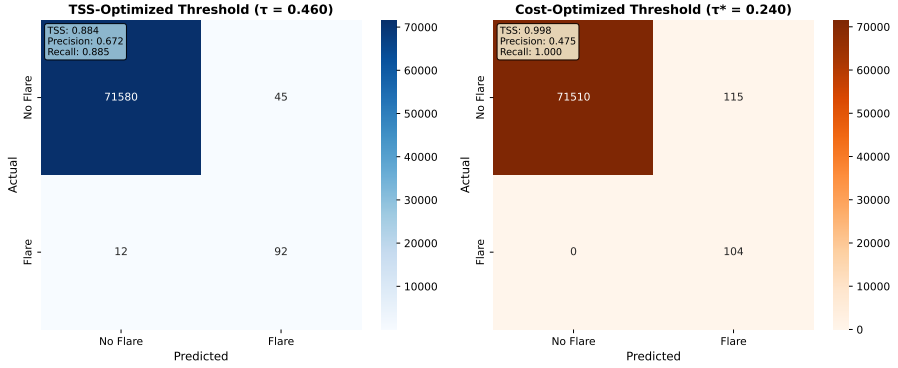

Figure 4: Confusion matrices for the M5–72 h model. Left: balanced-score threshold $\tau = 0.460$ (92 TP, 45 FP, 71,580 TN, 12 FN). Right: cost-minimising threshold $\tau^\star = 0.240$ (104 TP, 115 FP, 71,510 TN, 0 FN).

H Ablation Study Suite

We ran a systematic leave-one-component-out protocol with five seeds per variant to quantify the contribution of each EVEREST module. All runs targeted M5-class flares at 72 h horizon (the hardest task), with identical data splits, early stopping (120 epochs), and bootstrap evaluation (10^4 replicates). Tables 13 and 14 report mean metrics, effect sizes, and significance relative to the full model.

Interpretation. Four findings stand out: (i) mixed precision is numerically indispensable (FP32 diverged); (ii) the precursor auxiliary is the strongest regulariser, preventing collapse under extreme rarity; (iii) the attention bottleneck far outperforms mean pooling; (iv) evidential and EVT heads play complementary roles, with the former reducing calibration error and the latter improving tail-sensitive discrimination. These results support the design hypothesis that each module addresses a distinct failure mode.

Table 13: EVEREST ablation results on M5–72 h (mean $\pm$ s.d. over 5 seeds).

VARIANT	TSS	F1	BRIER	ECE	p
Full model	0.746 ± 0.146	0.747	0.0013	0.0110	—
No Evidential head	0.682 ± 0.193	0.626	0.0015	0.0111	< 0.01
No EVT head	0.461 ± 0.369	0.438	0.0039	0.0336	< 0.01
Mean pooling	0.319 ± 0.319	0.304	0.0229	0.1158	< 0.001
Cross-entropy loss	0.209 ± 0.332	0.195	0.0013	0.0023	< 0.001
No Precursor head	0.096 ± 0.174	0.095	0.0194	0.1105	< 0.001
FP32 training	0.000 ± 0.000	0.000	0.0520	0.2248	< 0.001

Table 14: Component ablation on M5–72 h. Paired bootstrap (10^4 replicates) vs. full model.

COMPONENT REMOVED	ΔTSS	REL. CHANGE (%)	p-VALUE
Mixed Precision (AMP)	-0.746	-100	< 0.001
Precursor head	-0.650	-87	< 0.001
Focal loss	-0.537	-72	< 0.001
Attention bottleneck	-0.427	-57	< 0.001
EVT head	-0.285	-38	< 0.001
Evidential head	-0.064	-9	0.004

I GRADIENT-BASED INTERPRETABILITY

We visualise feature–saliency gradients for representative tasks to probe the signals driving EVEREST predictions. Figure 5 summarises average gradient evolution across true positives (TP), true negatives (TN), and false positives (FP). Distinct temporal morphologies emerge: TP cases show sustained positive gradients in `USFLUX` and `MEANGAM`, while TN and FP cases lack such coherent rises.

To examine how predictive confidence aligns with saliency signals, Figure 6 shows TP cases stratified by model confidence. High-confidence TPs exhibit the strongest multi-feature gradients, whereas low-confidence TPs show weaker but still consistent rises.

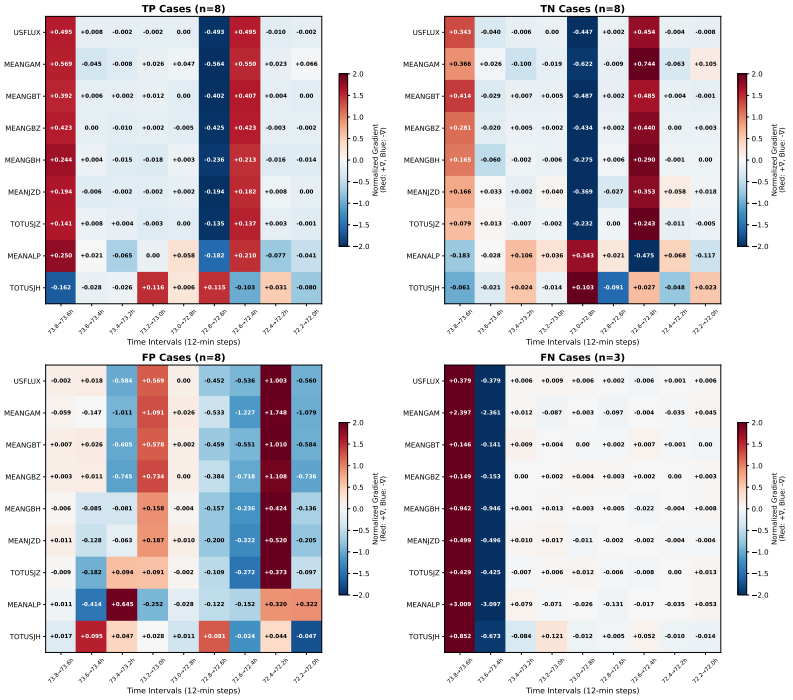

Figure 5: Feature evolution heatmaps across prediction outcomes (True Positive, True Negative, False Positive). Coordinated increases in `USFLUX` and `MEANGAM` appear in TPs, while TNs and FPs show flatter or noisier profiles.

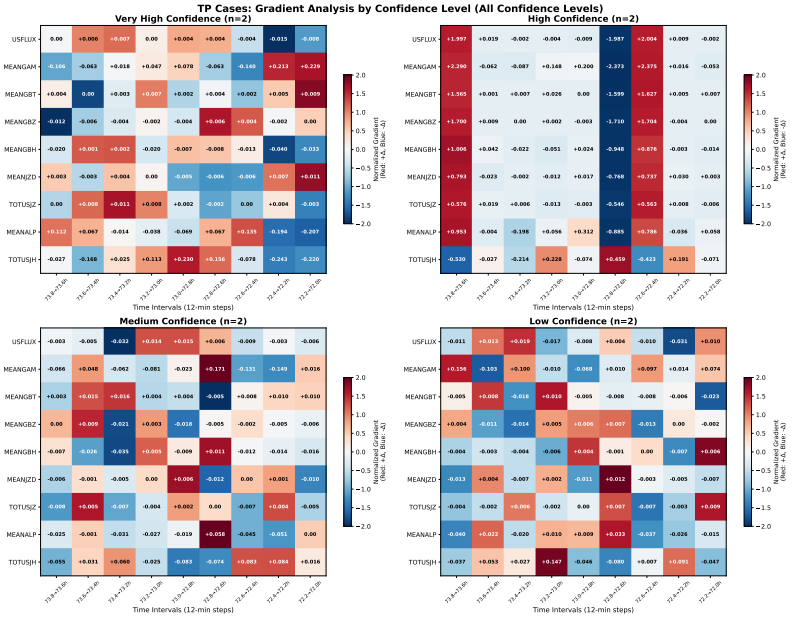

Figure 6: Gradient evolution for True Positive (TP) M5–72h predictions stratified by model confidence. Strongest gradients appear in high-confidence cases.

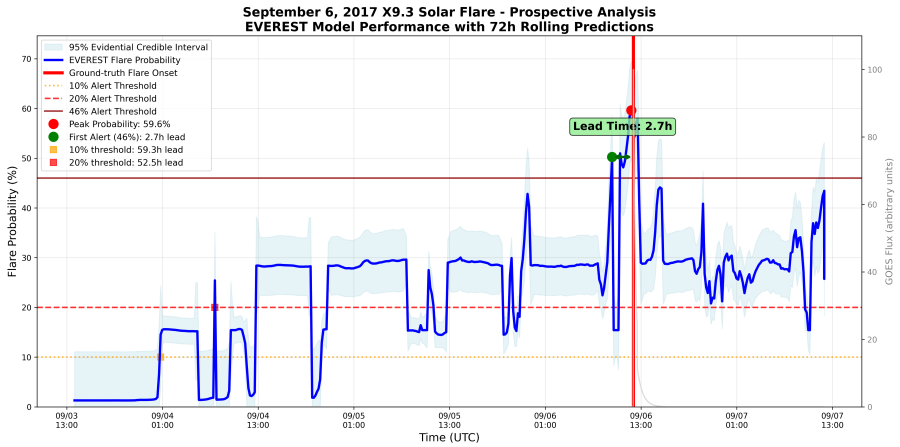

Figure 7: Prospective replay of the 6 September 2017 X9.3 flare. Blue: EVEREST M5–72 h probability (with 95% interval, if shown). Dashed lines mark alert thresholds (10%, 20%, 46%); grey shows GOES soft X-ray flux.

Table 15: Lead-time statistics for EVEREST (M5–72 h) on the 6 Sep 2017 X9.3 flare.

Threshold (τ)	**First crossing (UTC)**	**Lead time**	**Continuous alert length**
10%	04 Sep 00:57	59.3 h	60.8 h
20%	04 Sep 14:01	52.5 h	53.6 h
46%	06 Sep 09:19	2.7 h	2.3 h

J PROSPECTIVE REPLAY: 6 SEPTEMBER 2017 X9.3 FLARE

The X9.3 flare of 6 September 2017 (NOAA AR 12673, peak at 12:02 UT) was held out from training and threshold calibration (3–7 Sep 2017) to provide a true out-of-sample test. Figure 7 shows the M5–72 h probability trace; Table 15 lists the associated lead times for several alert thresholds.