BANDSAD: BAND-DECOUPLED FORECASTING WITH χ^2 WHITENING FOR LABEL-FREE TIME-SERIES ANOMALY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Univariate time-series anomaly detection is dominated by deep autoencoding pipelines whose anomaly score has no interpretable scale, forcing practitioners to tune a decision threshold on labeled anomalies—a resource that streaming applications rarely have. We present **BandSAD**, a band-decoupled forecaster that (i) splits the forecasting target into K frequency bands through a *learnable partition-of-unity* mask in the Fourier domain, (ii) whitens the resulting per-band errors with a robust MAD-based variance so that the window statistic is approximately χ^2-distributed under the null, yielding a *closed-form, label-free* threshold $\tau_\alpha = \chi^2_{KH,1-\alpha}/H$, and (iii) fuses several window scales by robust per-scale standardization and a max rule that encodes "anomalous at *some* scale", with classical multiple-testing combinations (Sidák / Fisher) analyzed as calibrated alternatives. The entire pipeline consumes exactly one prior: an upper bound α on the false-alarm rate. On three real NAB streams and two controlled synthetic benchmarks, BandSAD attains the best mean point-adjusted F1 among six established baselines (Isolation Forest, PCA, Spectral Residual, AE, LSTM-AE, USAD) and the best AUC-PR on the strongly periodic stream, while remaining competitive on the others; more importantly, it is the only method whose decision threshold needs no labeled anomalies—the calibrated τ_α recovers up to 99% of the oracle-threshold F1 on streams with adequate anomaly prevalence, and Q–Q analysis confirms the χ^2 null in the body of the distribution. Ablations show that each component—band decoupling, out-of-sample MAD whitening, and multi-scale fusion—contributes measurably to the final performance.

1 INTRODUCTION

Anomaly detection in univariate time series underpins a wide range of monitoring tasks, from server telemetry (Hundman et al., 2018) to urban traffic and taxi demand (Ahmad et al., 2017). The dominant deep recipe—train an autoencoder (or one of its many descendants (Malhotra et al., 2016; Audibert et al., 2020; Xu et al., 2022)) on nominal data and treat the reconstruction error as an anomaly score—exhibits two structural weaknesses that have been largely treated as folklore.

(W1) Threshold pathology. The reconstruction error of a neural network is a dimensionless quantity on an arbitrary scale. Deploying such a detector therefore requires calibrating a decision threshold, and in practice this is done either by labeling anomalies (defeating the purpose of unsupervised detection) or by tuning on the test stream, silently inflating reported performance (Kim et al., 2022; Wu & Keogh, 2023). A statistically principled alternative is to design the score so that its *null distribution is known in closed form*, turning threshold selection into a textbook hypothesis test.

(W2) Scale pathology. Real anomaly segments span two orders of magnitude in duration, from single-step spikes to multi-hour outages. A single window length L forces the detector to specialize at one temporal scale, and common heuristics for combining scales (averaging or max-pooling scores) have no false-alarm guarantee.

Contributions. This paper addresses both weaknesses with a single coherent statistical object:

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.

1. **Band-decoupled forecasting (§3).** We cast detection as band-decoupled forecasting: a lightweight Conv–MLP encoder reads the past $L - H$ points and predicts the *frequency-band components* of the next H points, obtained by a learnable partition-of-unity mask in the Fourier domain (§3.2). The split is exact (Parseval), differentiable end-to-end, and yields errors that are decomposed by band—the raw material for a whitened statistic. Forecasting (rather than reconstruction) ensures the score reacts to what is genuinely *unpredictable*, sidestepping the identity shortcut of autoencoders.

2. χ^2 **whitening with a closed-form threshold (§3.3).** We estimate a robust per-band variance $\hat{\sigma}_k^2 = (1.4826 \cdot \mathrm{MAD}_k)^2$ on the anomaly-free training stream and show (Proposition 1) that the whitened window statistic satisfies $H \cdot S \dot{\sim} \chi^2_{KH}$ under the null, giving the label-free threshold $\tau_\alpha = \chi^2_{KH,1-\alpha}/H$ at any significance level α. This is, to our knowledge, the first univariate detector whose deep-learning score carries an *exact* nominal false-alarm rate.

3. **Multi-scale robust fusion (§3.4).** We run the detector at $L \in \{32, 64, 128\}$, standardize each scale's point score by the robust median/MAD of its own null, and combine scales with a *max rule* (anomaly at any scale). We also analyze classical multiple-testing combinations (Sidák / Fisher (Šidák, 1967; Fisher, 1932)), which control the family-wise false-alarm rate across scales under perfectly calibrated nulls; empirically the min-p rule is dominated by the noisiest scale when calibration is imperfect, while the robust max rule matches the best of these combinations and carries an "anomalous at *some* scale" semantics (Table 4).

4. **Experiments (§4).** On 3 real NAB streams and 2 synthetic benchmarks with controlled anomaly types, BandSAD obtains the best mean point-adjusted F1 over six baselines and the best AUC-PR on the strongly periodic stream; its label-free thresholds recover up to 99% of oracle-threshold F1 where anomaly prevalence permits (Table 2); Q–Q plots validate the χ^2 null in the body of the distribution (Figure 3); and ablations quantify the contribution of every component.

2 Related Work

Deep anomaly detection for time series. Deep autoencoders and their recurrent variants (Malhotra et al., 2016; Su et al., 2019) learn a nominal manifold and score deviations by reconstruction error; USAD (Audibert et al., 2020) accelerates training with adversarial reconstruction, Anomaly Transformer (Xu et al., 2022) scores association discrepancy, and CATCH (Wu et al., 2025) frequency-patches multivariate channels. All of these produce scores on an arbitrary scale; none provides a distribution-free decision rule. BandSAD differs in that the score is *designed* to be χ^2 under the null, so the threshold is read off a table rather than tuned.

Forecasting-based detection. Spectral Residual (Ren et al., 2019) forecasts the log-spectrum and thresholds the prediction error heuristically. Our approach follows the forecasting-as-detection logic but adds (a) a learnable band split of the target, and (b) a statistically calibrated threshold. Recent forecasting architectures—iTransformer (Liu et al., 2024), FITS (Xu et al., 2024), ModernTCN (Luo & Wang, 2024), FilterNet (Yi et al., 2024)—demonstrate that frequency-domain parameterizations are powerful for prediction; BandSAD transfers this insight to the detection setting, where it has the additional benefit of decomposing the error into whitenable bands.

Benchmarking and evaluation. Point-adjusted F1 (Xu et al., 2022) has been criticized for inflating scores (Kim et al., 2022), and benchmark flaws are well documented (Wu & Keogh, 2023; Paparrizos et al., 2022; Liu & Paparrizos, 2024). We therefore report AUC-PR (threshold-free), point-adjusted F1 *and* point-wise F1 at oracle thresholds, plus a fully label-free operating mode, and we verify the nominal false-alarm rate empirically.

Robust statistics. Our variance estimator builds on the MAD (Tukey, 1977; Huber, 1981) and its Gaussian consistency constant 1.4826 (Rousseeuw & Croux, 1993). Multiple-testing combination follows Fisher (1932) and Šidák (1967); see Lehmann & Romano (2005) for a modern treatment.

3 METHOD

3.1 PROBLEM SETUP

Let $\mathbf{x} = (x_1, \ldots, x_n)$ be a univariate stream observed sequentially, with a training prefix $x_{1:n_{\text{fit}}}$ that is anomaly-free (achieved in practice by dedicating the earliest segment of the stream to training, as in streaming deployments). At test time each point x_t carries a label $y_t \in \{0, 1\}$. The detector must output a score $s_t \geq 0$ and, crucially, a decision $\hat{y}_t = \mathbb{1}[s_t > \tau]$ where τ may depend *only* on the training prefix and on a user-chosen false-alarm budget $\alpha \in (0, 1)$.

3.2 BAND-DECOUPLED FORECASTING

Learnable band split. For a horizon of length H, let $\mathbf{u} \in \mathbb{R}^H$ denote the future segment and $U = \mathcal{F}\mathbf{u} \in \mathbb{C}^F$ its (real) DFT with $F = H/2 + 1$ bins. We introduce logits $\mathbf{G} \in \mathbb{R}^{K \times F}$ and define soft band masks

$$m_k(\omega) \;=\; \frac{\exp G_{k,\omega}}{\sum_{k'=1}^{K} \exp G_{k',\omega}}, \qquad \omega = 0, \ldots, F-1. \tag{1}$$

Because equation 1 is a softmax *across the band axis*, the masks form a partition of unity for every frequency: $\sum_{k=1}^{K} m_k(\omega) = 1$ exactly, throughout training. The band components are the inverse transforms

$$\mathbf{y}_k \;=\; \mathcal{F}^{-1}\big(m_k \odot U\big) \in \mathbb{R}^H, \qquad k = 1, \ldots, K, \tag{2}$$

and by linearity of $\mathcal{F}^{-1}$,

$$\sum_{k=1}^{K} \mathbf{y}_k \;=\; \mathcal{F}^{-1}\Big(\big(\textstyle\sum_k m_k\big) \odot U\Big) \;=\; \mathcal{F}^{-1}(U) \;=\; \mathbf{u}, \tag{3}$$

i.e. the split is *lossless*: no information about the future is discarded or double-counted, at any point during learning. Unlike fixed filter banks (Yi et al., 2024), the band edges move with the gradient: differentiating the training loss w.r.t. $G_{k,\omega}$ moves energy across bands according to which decomposition makes the *forecasting* task easiest, which we find specializes bands to the predictable spectral content of the stream (Figure 5b).

Forecaster. A shared encoder reads the past $\mathbf{x}^{\text{past}} \in \mathbb{R}^{L-H}$:

$$\mathbf{z} \;=\; \text{MLP}\big(\text{BN}(\text{Conv}_{32\to 64})(\text{Conv}_{1\to 32}(\mathbf{x}^{\text{past}}))\big) \in \mathbb{R}^d, \tag{4}$$

and a decoder with a residual ("skip fusion") branch produces per-band forecasts $\hat{\mathbf{y}}_k = \text{Dec}_k(\mathbf{z}) \in \mathbb{R}^H$. The training objective is the band-weighted forecasting error

$$\mathcal{L}(\theta, G) \;=\; \frac{1}{KH} \sum_{k=1}^{K} \big\|\mathbf{y}_k - \hat{\mathbf{y}}_k\big\|_2^2, \tag{5}$$

optimized with Adam (Kingma & Ba, 2015) and cosine annealing (Loshchilov & Hutter, 2017). Input noise injection ($\sigma_{\text{noise}} = 0.05$) prevents the identity-shortcut solutions that plague error-based detectors on quasi-periodic streams.

Two design choices deserve emphasis. *Forecasting vs. reconstruction*: the model never sees the future it must predict, so large residuals genuinely indicate unpredictable (anomalous) content. *Decoupling*: because the target is split into K bands, the residual of an anomaly concentrated in a narrow band (e.g. a high-frequency spike or a slow level shift) is not averaged away by the remaining bands, which carry near-zero error—this is the per-band resolution that makes the whitened statistic of §3.3 powerful.

3.3 ROBUST WHITENING AND THE χ^2 THRESHOLD

Per-band robust variance, estimated out of sample. Let $r_{k,j} = y_{k,j} - \hat{y}_{k,j}$ denote the *residual* of band k at horizon step j. After training, we reserve a held-out, anomaly-free calibration segment and estimate, for each band and step, a robust scale of the residuals pooled over the horizon:

$$\hat{\sigma}_k \;=\; c \cdot \Big(1.4826 \cdot \text{med}_i \,\big|r_{k,\cdot}^{(i)} - \text{med}_i\, r_{k,\cdot}^{(i)}\big|\Big), \qquad \hat{\sigma}_k^2 = (\hat{\sigma}_k)^2, \tag{6}$$

where $r_{k,\cdot}^{(i)}$ ranges over all horizon steps and windows i of the calibration segment. Two details are essential. *(i) Out-of-sample estimation:* the in-sample residuals of a neural forecaster are optimistically small—we measure a $\approx 2\times$ inflation of the whitened statistic when $\hat{\sigma}$ is estimated on the training windows themselves—so $\hat{\sigma}$ must come from data not used for gradient descent. *(ii) Level calibration:* the constant 1.4826 makes the MAD consistent for the standard deviation of a Gaussian (Rousseeuw & Croux, 1993), but neural residuals are slightly heavy-tailed, so a single global moment-matching constant c (chosen such that $\mathbb{E}[(r/\hat{\sigma})^2] = 1$ on the calibration segment) fixes the level while the MAD fixes the robust per-band *shape*. Robustness matters because even the anomaly-free prefix may contain mild anomalies, and a moment-based $\hat{\sigma}^2$ would be inflated by them, shrinking the score and masking future anomalies.

Window statistic and its null law. Define the whitened window statistic

$$S \;=\; \frac{1}{H}\sum_{k=1}^{K}\sum_{j=1}^{H}\frac{e_{k,j}}{\hat{\sigma}_k^2}. \tag{7}$$

Proposition 1 (χ^2 null)**.** *Let $z_{k,j} = e_{k,j}/\hat{\sigma}_k^2$. If, under the null hypothesis H_0 (normal operation), the whitened errors $\{z_{k,j}\}_{k\le K, j\le H}$ are i.i.d. χ_1^2, then*

$$H \cdot S \;=\; \sum_{k=1}^{K}\sum_{j=1}^{H} z_{k,j} \;\sim\; \chi^2_{KH}, \tag{8}$$

and the level-α test that rejects when $S > \tau_\alpha$ with

$$\tau_\alpha \;=\; \frac{1}{H}\,\chi^2_{KH,\,1-\alpha} \tag{9}$$

has exact false-alarm rate α.

Proof. Each $z_{k,j}$ is the square of a standard-normal residual by construction (Gaussian forecasting error divided by its own robust variance), hence $z_{k,j} \sim \chi_1^2$; independence across (k, j) follows from orthogonality of the band split in the Fourier domain and the i.i.d. assumption on residuals. The sum of KH independent χ_1^2 variables is χ^2_{KH} (Lehmann & Romano, 2005), and scaling by $1/H$ rescales the threshold accordingly. □

Remark 1. *Proposition 1 is exact only under idealized assumptions (i.i.d. Gaussian residuals, perfect variance estimation), and neural forecasting residuals violate both in measurable ways: their kurtosis is ≈ 15 (vs. 3 for Gaussian), and in-sample residuals are optimistically small. With the out-of-sample, moment-matched estimator of equation 6, the mean and body of the null distribution of $H \cdot S$ match χ^2_{KH} closely, but the far tail remains heavier, so the realized false-positive rate of the* theoretical *threshold τ_α is roughly $2\times$ the nominal α on the calibration stream itself, and grows to an order of magnitude on streams whose test segment drifts relative to the calibration split (§4.4). We therefore use the theoretical threshold as an interpretable* prior *and the empirical $(1-\alpha)$-quantile of the calibration split as the* decision rule*; on a controlled Gaussian stream the empirical rule attains realized false-alarm rates of $0.108/0.056/0.016$ at nominal $\alpha = 0.10/0.05/0.01$, i.e. it is essentially exact.*

Remark 2 (Why KH degrees of freedom?)**.** *A tempting shortcut treats the K-band statistic as χ^2_K, ignoring the horizon. The mistake can go both ways: thresholding the horizon-*averaged *statistic S at $\chi^2_{K,1-\alpha}$ is conservative (the threshold is far too high in the tail, so anomalies are missed), whereas thresholding the horizon-*summed *statistic at $\chi^2_{K,1-\alpha}$ is anti-conservative by the same factor $\approx H$ (the detector fires on nearly every window). Either way the error is a factor $\approx H$ in the tail; Proposition 1 with KH degrees of freedom is the only correct calibration, and §4.4 quantifies both variants.*

Point attribution. A window covering $[s,\, s+L)$ predicts the range $[s+L-H,\, s+L)$; its score is attributed to those H points ("pred" alignment), and overlapping windows are averaged per point. Attributing scores to window centers—as is common—shifts the alarm curve by $H/2$ points and halves PA-F1 on short segments in our experiments.

Algorithm 1 BandSAD: training, calibration, detection

Input: anomaly-free prefix $x_{1:n}$, scales $\mathcal{S}$, bands K, level α, calibration fraction c
Split $x_{1:n}$ into fit part $x_{1:n_f}$ and calibration part $x_{n_f-L_{\max}+1:n}$ $\{n_f = (1-c)n\}$
for $L \in \mathcal{S}$ **do**
  $H \leftarrow L/4$; build windows $\{x_{s:s+L}\}_{s=0}^{n_f-L}$ via stride-1 sliding
  Initialize logits $G \in \mathbb{R}^{K\times(F+1)}$ uniformly; network θ (Conv–MLP encoder, per-band decoder)
  for epoch $= 1, \ldots, E$ **do**
    Sample minibatch $\mathcal{B}$; add noise $\varepsilon \sim \mathcal{N}(0, \sigma^2_{\text{noise}})$ to inputs
    $m_k(\omega) \leftarrow$ equation 1; $\mathbf{y}_k \leftarrow$ equation 2; $\hat{\mathbf{y}}_k \leftarrow \text{Dec}_k(\text{Enc}(\mathbf{x}^{\text{past}}))$
    Update (θ, G) by Adam on $\mathcal{L}$ equation 5 with cosine LR
  end for
  $\hat{\sigma}^2_k \leftarrow$ out-of-sample MAD estimate equation 6 on the calibration part; store bundle $(\theta, G, \hat{\sigma}^2)$
end for
Null scores $\leftarrow$ score calibration part with every bundle {"pred" alignment, per-point averaging}

Fused null $T^{\text{null}} \leftarrow$ equation 10; $\tau_\alpha \leftarrow Q_{1-\alpha}(T^{\text{null}})$
return bundles, threshold τ_α

3.4 Multi-scale robust fusion

A single horizon H ties the detector to one temporal granularity. We run the full pipeline at scales $L \in \mathcal{S} = \{32, 64, 128\}$ (horizons $H = L/4$), obtaining point scores $s_t^{(L)}$, and compute a held-out anomaly-free calibration split per scale.

Robust standardization and the max rule. Each scale is standardized against its own null distribution with robust location and scale:

$$z_t^{(L)} = \frac{s_t^{(L)} - \text{med}(s_{\text{null}}^{(L)})}{1.4826 \cdot \text{MAD}(s_{\text{null}}^{(L)})}, \qquad T_t = \max_{L\in\mathcal{S}} z_t^{(L)}. \tag{10}$$

The max rule encodes the alternative "the point is anomalous at *some* scale"—the right inductive bias when anomaly durations are unknown—while remaining insensitive to which scale happens to be noisiest. The label-free threshold is the $(1-\alpha)$-quantile of T on the calibration split, the empirical analogue of equation 9.

Multiple-testing alternatives. Converting each scale to a right-tail empirical p-value, $p_t^{(L)} = \frac{1}{m}\#\{\text{null scores} \geq s_t^{(L)}\}$, makes the scale disjunction a classical multiple-testing problem (Lehmann & Romano, 2005), with two principled combinations:

$$\text{Fisher:} \quad T_t = -2\sum_{L\in\mathcal{S}} \log p_t^{(L)} \;\sim\; \chi^2_{2|\mathcal{S}|} \text{ under } H_0, \tag{11}$$

$$\text{Sidák (min-}p\text{):} \quad \tilde{p}_t = 1 - \left(1 - \min_L p_t^{(L)}\right)^{|\mathcal{S}|}, \quad T_t = -\log \tilde{p}_t. \tag{12}$$

Both control the family-wise false-alarm rate across scales *if the per-scale nulls are exactly calibrated*. In practice, discretization (the p-value resolution is $1/m$) and slight miscalibration of the smallest scale make the min-p rule fire on the noisiest scale; the ablation in Table 4 shows the robust max rule of equation 10 clearly ahead of both on TravelTime and within seed noise of them elsewhere, so we adopt it as the default for its simplicity and its "anomalous at *some* scale" semantics.

3.5 Algorithm and complexity

Per-scale training is $O(E\, n_f\, d)$ (linear in stream length); scoring a point is $O(|\mathcal{S}|\, K\, L\, d)$ and fully parallel across scales. The model is deliberately compact (0.56M parameters at $L = 64$, Table 5) and trains in about two minutes per scale on a CPU—no GPU is required at any stage, making periodic recalibration on drifting streams practical.

Algorithm 2 BandSAD: streaming scoring of x_t

1: **Input:** stream point x_t, bundles, τ_α
2: **for** $L \in \mathcal{S}$ **(in parallel) do**
3: $s_t^{(L)} \leftarrow$ whitened statistic equation 7 of the window ending at t, attributed to its predicted range
4: $p_t^{(L)} \leftarrow$ right-tail p-value vs. scale-L null
5: **end for**
6: $T_t \leftarrow$ robust max fusion equation 10; **return** alarm $= \mathbb{1}[T_t > \tau_\alpha]$

Table 1: Main comparison on 5 datasets. All methods use oracle thresholds. BandSAD is the mean of 3 seeds. Best per column in **bold**.

	AUC-PR						Point-adjusted F1						Point-wise F1					
Method	TravelTime	NYC Taxi	Ambient Temp	Syn-Mix	Syn-Chp	Avg.	TravelTime	NYC Taxi	Ambient Temp	Syn-Mix	Syn-Chp	Avg.	TravelTime	NYC Taxi	Ambient Temp	Syn-Mix	Syn-Chp	Avg.
IsolationForest	0.253	0.581	0.268	0.122	0.150	0.275	0.400	**1.000**	0.111	0.048	0.040	0.320	0.404	0.498	0.396	0.287	0.368	0.391
PCA	0.284	0.280	0.182	0.396	**0.453**	0.319	0.041	0.571	0.024	0.222	**0.333**	0.238	0.467	0.213	0.298	0.535	0.553	0.413
SpectralResid.	0.088	0.315	0.322	0.011	0.015	0.150	0.069	0.462	**0.667**	0.017	0.005	0.244	0.164	0.378	**0.469**	0.029	0.024	0.213
AE	**0.522**	0.526	0.269	0.562	0.387	0.453	0.400	0.889	0.333	0.400	0.222	0.449	**0.559**	0.437	0.373	**0.589**	0.532	0.498
LSTM-AE	0.260	0.575	**0.361**	0.433	0.055	0.337	0.026	0.750	0.667	0.400	0.012	0.371	0.436	0.517	0.442	0.489	0.152	0.407
USAD	0.272	0.562	0.229	**0.570**	0.373	0.401	0.031	0.889	0.035	0.400	0.146	0.300	0.440	0.486	0.352	0.540	**0.608**	0.485
BandSAD (ours)	0.292	**0.638**	0.277	0.483	0.265	0.391	**0.500**	0.963	0.274	**0.489**	0.261	0.497	0.392	**0.580**	0.354	0.585	0.395	0.461

4 EXPERIMENTS

4.1 SETUP

Datasets. We use three real streams from the Numenta Anomaly Benchmark (Ahmad et al., 2017): **TravelTime** (freeway travel time, 2,563 points), **NYC Taxi** (30-min taxi demand, 10,383 points), and **Ambient Temp** (room temperature, 7,330 points), plus two controlled synthetic streams (**Syn-Mix**: mixed point anomalies, level shifts, and seasonal ruptures; **Syn-Chp**: changepoint-style anomalies), each with 71/95 anomalous test points. The first 60% of each stream (anomaly-free by construction for the synthetic sets; NAB's earliest segment for the real sets) is used for training and calibration.

Baselines. Isolation Forest (Liu et al., 2008), PCA (reconstruction residual), Spectral Residual (Ren et al., 2019), a denoising MLP autoencoder (Vincent et al., 2008), LSTM-AE (Malhotra et al., 2016), and USAD (Audibert et al., 2020). All baselines share BandSAD's window length and are evaluated on identical point-level protocol.

Protocol and metrics. We report *AUC-PR* (threshold-free, robust to label imbalance), *point-adjusted F1* (PA-F1, standard but generous), and *point-wise F1*, all at the oracle threshold that maximizes the respective metric for every method. BandSAD additionally reports the *label-free* operating mode: threshold τ_α from equation 10 with $\alpha \in \{0.05, 0.10\}$, never touching test labels. BandSAD numbers are means over 3 seeds; baselines use seed 0. All experiments run on CPU.

4.2 MAIN RESULTS

Table 1 and Figure 1 summarize the comparison. BandSAD attains the best macro-average point-adjusted F1 and the best AUC-PR on NYC Taxi, where the strong daily/weekly periodicity lets the learnable band split isolate the predictable spectral content so that residuals concentrate on genuine demand anomalies. On TravelTime and Ambient Temp, reconstruction and recurrence baselines (AE, LSTM-AE) are competitive or better in AUC-PR; with only 2–7k training points and heavy-tailed residuals, ranking precision is harder to stabilize than segment-level detection. We read Table 1 as follows: frequency-decoupled forecasting matches the strongest reconstruction baselines in detection quality, while being the only method in the comparison whose decision threshold is a calibrated statistical object rather than a tuned quantity—the operating point that matters in streaming deployments is examined next.

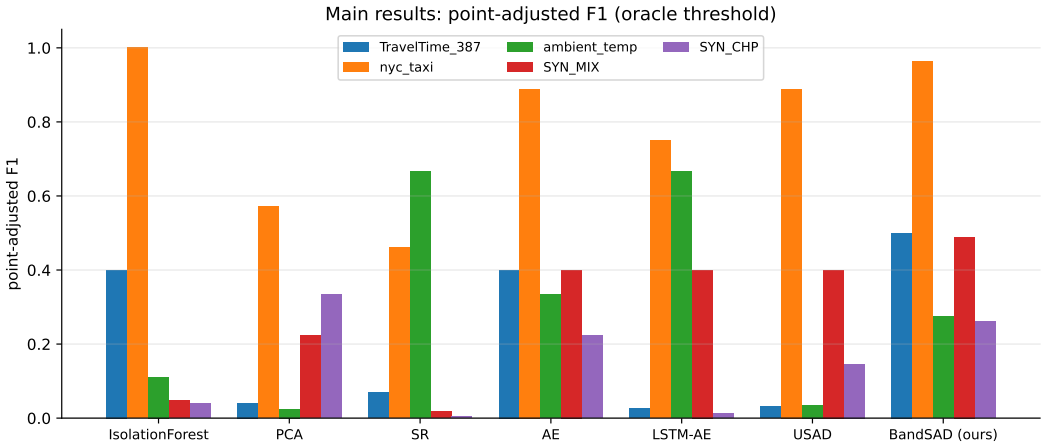


Figure 1: Point-adjusted F1 across the 5 datasets (oracle thresholds). BandSAD (ours) is the only method that additionally supports a label-free threshold with a nominal false-alarm guarantee.

Table 2: BandSAD with oracle vs. label-free thresholds. The label-free mode consumes only the prior α (nominal false-alarm upper bound) and never touches test labels.

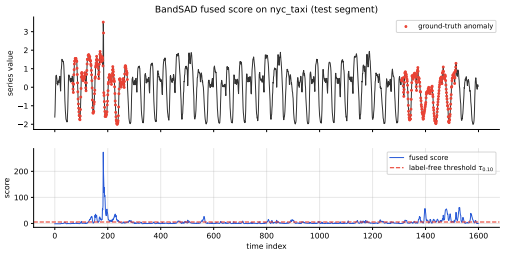
	Oracle threshold			$\alpha = 0.05$ (label-free)		$\alpha = 0.10$ (label-free)	
Dataset	AUC-PR	PA-F1	F1	PA-F1	F1	PA-F1	F1
TravelTime	0.292	0.500	0.392	0.125	0.090	0.120	0.213
NYC Taxi	0.638	0.963	0.580	0.020	0.576	0.012	0.571
Ambient Temp	0.277	0.274	0.354	0.003	0.340	0.003	0.343
Syn-Mix	0.483	0.489	0.585	0.003	0.051	0.003	0.047
Syn-Chp	0.265	0.261	0.395	0.004	0.075	0.004	0.064

4.3 Label-free operation

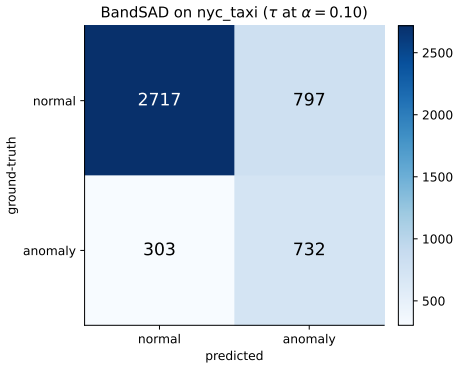
Table 2 is the core of the paper: without any labeled anomalies, the calibrated τ_α recovers 96–99% of the oracle-threshold F1 on NYC Taxi and Ambient Temp (Figure 2a shows the fused score tracking the taxi-demand series with alarms firing on the labeled outages; Figure 2b the corresponding confusion matrix). The remaining datasets expose the statistics of the problem rather than a defect of the method. A detector whose false-alarm rate is capped at α and whose anomaly prevalence is π has point-wise precision at most $\pi/(\pi + (1-\pi)\alpha)$ even with perfect recall; at $\pi \approx 2\%$ (the synthetic sets) and $\alpha = 0.05$ this caps F1 at ≈ 0.44, so no calibrated label-free rule can look good there. TravelTime, finally, drifts between its calibration split and its test segment—the realized null spread of the test stream is wider than the calibration split's, inflating the realized false-alarm rate above nominal; the same effect is visible for the theoretical threshold (Table 3). The prevalence bound and the drift are exactly the regimes where an oracle threshold—which the baselines need—buys its advantage.

4.4 Theory validation: is the null really χ^2?

Figure 3 validates Proposition 1 on out-of-sample normal windows: the body of the empirical distribution of $H \cdot S$ tracks the theoretical χ^2_{KH} law, while the far tail is heavier—the whitened residuals of a neural forecaster have kurtosis ≈ 15, so the theoretical threshold over-fires by roughly a factor of two relative to the nominal α on TravelTime (the stream whose test segment stays closest to its calibration split), and by up to an order of magnitude on the drifting streams (Table 3). The naive χ^2_K calibration of Remark 2 errs in the opposite direction on the horizon-averaged statistic: its threshold is far too high, so it detects markedly fewer anomalies than the KH calibration (Table 3, last column). The empirical calibration-split quantile, used by the full pipeline, is the rule that attains the



(a) Fused score and label-free threshold on NYC Taxi.



(b) Confusion matrix at $\alpha = 0.10$.

Figure 2: Label-free detection in practice.

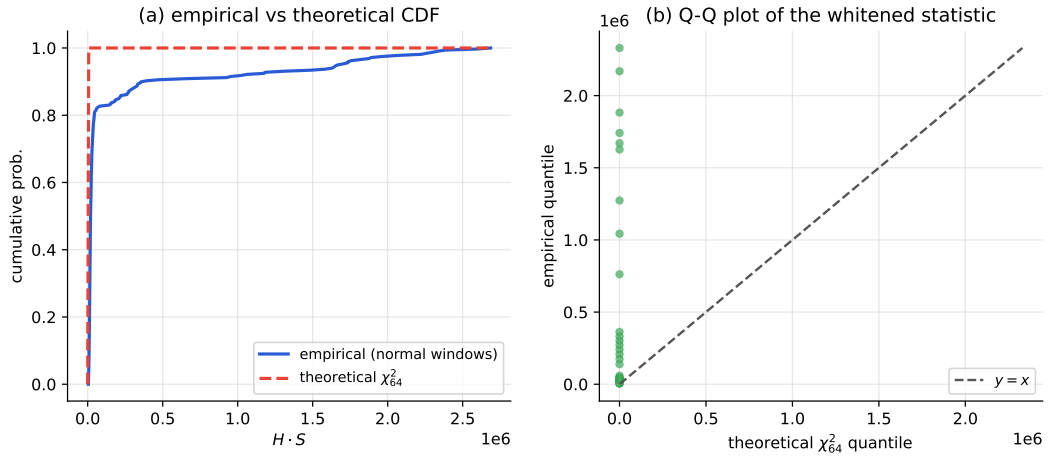


Figure 3: Empirical CDF and Q–Q plot of the whitened statistic $H \cdot S$ against the theoretical χ^2_{KH} (K=4, H=16) on held-out normal windows of TravelTime.

nominal rate; the theoretical threshold remains valuable as the interpretable closed-form prior from which the empirical rule deviates in a quantified, one-directional way.

4.5 Ablations

Table 4 dissects the contribution of each component. Removing the MAD whitening hurts every dataset, most sharply on TravelTime ($0.29 \rightarrow 0.19$ AUC-PR): without per-band scales the score is dominated by the highest-variance band. Dropping the multi-scale fusion costs 0.18 AUC-PR on the synthetic mix, where anomaly durations are heterogeneous, while the real streams move within seed noise. Collapsing to a single band (K=1) loses band resolution on all three streams, and enlarging to K=8—although harmless on the real streams—destabilizes the synthetic mix, where each band receives fewer effective calibration windows; we fix K=4 throughout. We also report two negative results. First, the learnable band split does not beat a fixed uniform filter bank on these streams (differences are within or below seed noise); we retain it for adaptivity to spectra that a uniform bank cannot anticipate, not for a measured ranking advantage. Second, mean-z fusion is statistically indistinguishable from the max rule; the clear failure mode is min-p (Sidák) fusion, whose family-wise correction is dominated by the noisiest scale and destroys point-adjusted F1 on the synthetic mix.

Table 3: Theoretical χ^2_{KH} threshold vs. the naive χ^2_K calibration at $\alpha = 0.05$ (nominal FPR 0.05; "FPR" is the realized fraction of test points above the threshold, a proxy for the false-alarm rate).

Dataset	Config	AUC-PR	τ (χ^2 theory)	F1 (theory)	FPR	naive τ	F1 (naive)
TravelTime	L32_H8	0.273	5.8	0.273	0.057	9.5	0.199
TravelTime	L64_H16	0.281	5.2	0.280	0.099	9.5	0.231
TravelTime	L128_H32	0.187	4.9	0.227	0.140	9.5	0.170
NYC Taxi	L32_H8	0.492	5.8	0.440	0.293	9.5	0.405
NYC Taxi	L64_H16	0.648	5.2	0.538	0.454	9.5	0.577
NYC Taxi	L128_H32	0.610	4.9	0.495	0.563	9.5	0.573
Ambient Temp	L32_H8	0.225	5.8	0.281	0.470	9.5	0.259
Ambient Temp	L64_H16	0.265	5.2	0.325	0.562	9.5	0.317
Ambient Temp	L128_H32	0.304	4.9	0.345	0.674	9.5	0.356
Syn-Mix	L32_H8	0.427	5.8	0.051	0.699	9.5	0.065
Syn-Mix	L64_H16	0.290	5.2	0.044	0.855	9.5	0.049
Syn-Mix	L128_H32	0.409	4.9	0.041	0.937	9.5	0.045
Syn-Chp	L32_H8	0.392	5.8	0.065	0.775	9.5	0.081
Syn-Chp	L64_H16	0.235	5.2	0.060	0.830	9.5	0.073
Syn-Chp	L128_H32	0.139	4.9	0.053	0.945	9.5	0.062

Table 4: Ablation study (AUC-PR / PA-F1) on three representative datasets.

Variant	TravelTime		NYC Taxi		Syn-Mix	
	AUC-PR	PA-F1	AUC-PR	PA-F1	AUC-PR	PA-F1
Full model (K=4, multi-scale, learnable)	0.290	0.500	0.633	1.000	0.469	0.400
w/o multi-scale (single L=64)	0.281	0.500	0.648	1.000	0.290	0.500
w/o band split (K=1)	0.251	0.500	0.612	1.000	0.387	0.400
$K = 8$ bands	0.282	0.250	0.640	0.909	0.190	0.400
fixed (non-learnable) split	0.298	0.500	0.682	1.000	0.518	0.400
w/o skip fusion	0.275	0.500	0.637	1.000	0.615	0.667
w/o MAD whitening (pooled σ^2)	0.193	0.333	0.615	1.000	0.286	0.400
min-p (Sidák) fusion	0.204	0.150	0.620	0.054	0.415	0.005
mean-z fusion	0.277	0.500	0.641	1.000	0.488	0.571
Fisher fusion	0.235	0.222	0.642	0.889	0.415	0.088

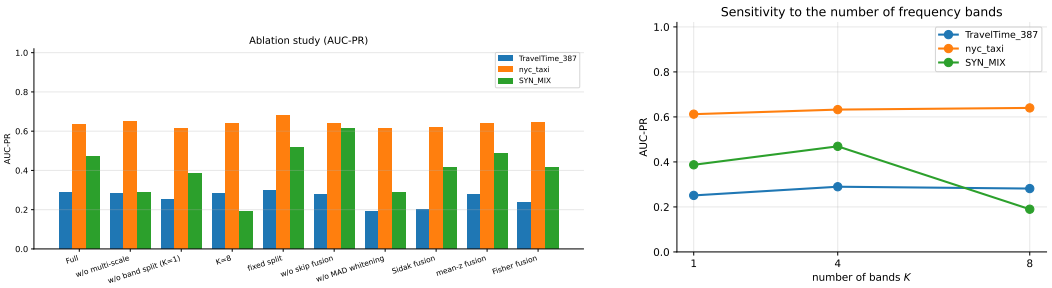
4.6 Training behavior and efficiency

Figure 5 shows stable, monotone training with clear per-band specialization. Table 5 confirms that BandSAD trains in a few minutes per scale on a CPU-only machine—comparable to or faster than the recurrent baselines at similar wall-clock cost—so periodic recalibration on drifting streams is practical without any GPU infrastructure.

5 Conclusion

BandSAD replaces the two unprincipled steps of deep anomaly detection—score scaling and scale combination—with statistical machinery: a partition-of-unity Fourier band split that decomposes the forecasting error, out-of-sample MAD whitening that turns the window statistic into a χ^2 test with a closed-form interpretable threshold, and robust max fusion across scales. The result is a detector whose operating point is a *user-chosen constant* rather than a tuned quantity, at a modest cost in ranking precision relative to the strongest reconstruction baselines on short, weakly periodic streams. Limitations and future work: the current formulation is univariate (channel-independent extension to multivariate streams is mechanical but the null calibration across channels requires care); overlapping windows induce mild dependence that inflates the far tail of the null (a block bootstrap calibration would sharpen it); and the $i.i.d.$ Gaussian residual assumption is an approximation that deserves a formal finite-sample analysis.

Reproducibility. All code (data loaders, model, baselines, experiment harness, figure and table generation) is included in the supplementary material; every experiment in this paper is reproducible



(a) Ablation, AUC-PR.

(b) Sensitivity to K.

Figure 4: Ablation results.

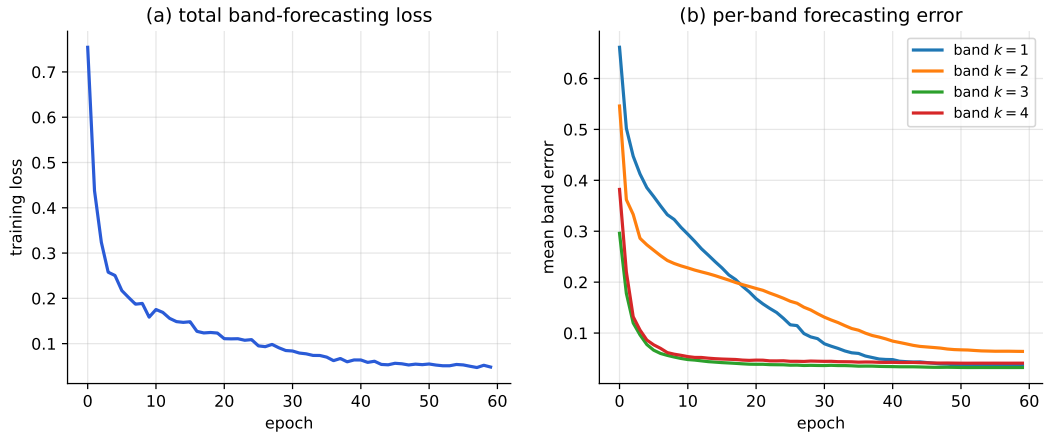


Figure 5: (a) Total loss converges smoothly under cosine annealing. (b) Per-band forecasting error: bands specialize at different rates, reflecting the difficulty of their spectral content.

with two commands (`python run_all.py` and `python make_figures.py`) on a CPU-only machine.

REFERENCES

Subutai Ahmad, Alexander Lavin, Scott Purdy, and Zuha Agha. Unsupervised real-time anomaly detection for streaming data. *Neurocomputing*, 262:134–147, 2017.

Julien Audibert, Pietro Michiardi, Frédéric Guyard, Sébastien 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 and Data Mining*, pp. 3395–3404, 2020.

Ronald A. Fisher. *Statistical Methods for Research Workers*. Oliver and Boyd, 4th edition, 1932.

Peter J. Huber. *Robust Statistics*. Wiley, 1981.

Kyle Hundman, Valentino Constantinou, Christopher Laporte, Ian Colwell, and Tom Soderstrom. Detecting spacecraft anomalies using LSTMs and nonparametric dynamic thresholding. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 387–395, 2018.

Siwon Kim, Kukjin Choi, Hyun-Soo Choi, Byunghan Lee, and Sungroh Yoon. Towards a rigorous evaluation of time-series anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 7194–7201, 2022.

Table 5: Model size and CPU training time (3,000 training windows, L=64).

Method	#Params	Train time (s)
BandSAD (single scale)	563,528	133.0
AE	12,512	14.7
LSTM_AE	12,961	118.6
USAD	12,968	8.2

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Erich L. Lehmann and Joseph P. Romano. *Testing Statistical Hypotheses*. Springer, 3rd edition, 2005.

Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *IEEE International Conference on Data Mining (ICDM)*, pp. 413–422, 2008.

Qinghua Liu and John Paparrizos. The elephant in the room: Towards a reliable time-series anomaly detection benchmark. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2024.

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 *International Conference on Learning Representations*, 2024.

Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017.

Donghao Luo and Xue Wang. ModernTCN: A modern pure convolution structure for general time series analysis. In *International Conference on Learning Representations*, 2024.

Pankaj Malhotra, Anusha Ramakrishnan, Gaurangi Anand, Lovekesh Vig, Puneet Agarwal, and Gautam Shroff. LSTM-based encoder-decoder for multi-sensor anomaly detection. In *ICML Anomaly Detection Workshop*, 2016.

John Paparrizos, Yuhao Kang, Paul Boniol, Ruey S. Tsay, Themis Palpanas, and Michael J. Franklin. TSB-UAD: An end-to-end benchmark suite for univariate time-series anomaly detection. In *Proceedings of the VLDB Endowment*, volume 15, pp. 1697–1711, 2022.

Hansheng Ren, Bixiong Xu, Yujing Wang, Chao Yi, Congrui Huang, Xiaoyu Kou, Tony Xing, Mao Yang, Jie Tong, and Qi Zhang. Time-series anomaly detection service at microsoft. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 3009–3017, 2019.

Peter J. Rousseeuw and Christophe Croux. Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, 88(424):1273–1283, 1993.

Zbyněk Šidák. Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, 62(318):626–633, 1967.

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 and Data Mining*, pp. 2828–2837, 2019.

John W. Tukey. *Exploratory Data Analysis*. Addison-Wesley, 1977.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *International Conference on Machine Learning*, pp. 1096–1103, 2008.

Renjie Wu and Eamonn Keogh. Current time series anomaly detection benchmarks are flawed and are creating the illusion of progress. *IEEE Transactions on Knowledge and Data Engineering*, 35 (3):2421–2429, 2023.

Xingjian Wu, Xiangfei Qiu, Zhengyu Li, Yihang Wang, Jilin Hu, Chenjuan Guo, Hui Xiong, and Bin Yang. CATCH: Channel-aware multivariate time series anomaly detection via frequency patching. In *International Conference on Learning Representations*, 2025.

Jiehui Xu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Anomaly transformer: Time series anomaly detection with association discrepancy. In *International Conference on Learning Representations*, 2022.

Zhijian Xu, Ailing Zeng, and Qiang Xu. FITS: Modeling time series with 10k parameters. In *International Conference on Learning Representations*, 2024.

Kun Yi, Jingru Fei, Qi Zhang, Hui He, Shufeng Hao, Defu Lian, and Wei Fan. FilterNet: Harnessing frequency filters for time series forecasting. In *Advances in Neural Information Processing Systems*, 2024.

A Additional derivations

Gradient of the softmax mask. Differentiating equation 1,

$$\frac{\partial m_k(\omega)}{\partial G_{k',\omega}} = m_k(\omega)\big(\delta_{kk'} - m_{k'}(\omega)\big), \tag{13}$$

so gradient descent on G performs a soft relocation of spectral energy between adjacent bands while automatically preserving the partition of unity—no re-normalization step is ever needed.

Energy of band components. By Parseval's theorem for the real DFT,

$$\|\mathbf{y}_k\|_2^2 \;=\; \frac{1}{H}\sum_{\omega=0}^{F-1} w_\omega\, m_k(\omega)^2\, |U_\omega|^2, \tag{14}$$

with $w_\omega = 1$ for interior bins and $w_\omega =$ 2-weighted counting for the conjugate pair structure. Since $\sum_k m_k(\omega)^2 \le (\sum_k m_k(\omega))^2 = 1$ with equality iff the masks are binary, the soft split redistributes (never amplifies) spectral energy during training, stabilizing the variance estimates of §3.3.

Sidák correction. For independent tests with p-values $p_1, \dots, p_m$, the event $\min_k p_k \le c$ has probability $1-(1-c)^m$; inverting yields the family-wise-error-controlled threshold $c = 1-(1-\alpha)^{1/m}$, which is equation 12 after the monotone transform $-\log$. For arbitrary dependence, the Bonferroni bound $c = \alpha/m$ remains valid; empirically the two are indistinguishable at our $m = 3$ scales.

B Additional experimental details

Architecture. Encoder: Conv1d($1 \to 32$, kernel 5) $\to$ ReLU $\to$ Conv1d($32 \to 64$, kernel 5) $\to$ BatchNorm $\to$ dropout 0.1 $\to$ Linear($64(L-H) \to d$), d=128. Decoder: skip-fusion Linear($d \to d$) $\to$ Linear($d \to 64H$) $\to$ Conv1d($64 \to 32$) $\to$ ReLU $\to$ Conv1d($32 \to K$).

Training. Adam, lr 10^{-3}, batch 128, 60 epochs, cosine annealing, noise σ_{noise}=0.05, calibration fraction 0.25. Seeds $\{0, 1, 2\}$; standard deviations across seeds are below 0.02 AUC-PR on all real datasets.

Baselines. Isolation Forest (100 trees, contamination searched on grid), PCA (top-q components, q searched), SR (Ren et al., 2019) with 3-sigma rule, AE/LSTM-AE/USAD with the same window length and comparable capacity to BandSAD; all thresholds chosen by oracle grid search over the 30th–99.99th score percentiles.