RESTORETS: EXPLICIT INPUT RESTORATION FOR ROBUST TIME SERIES FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Time series forecasting supports scientific analysis and operational planning across systems that rely on sensors. Yet sensor errors can introduce spikes and sustained level shifts into observed histories, degrading downstream forecasts. Existing approaches reduce corruption effects by changing forecasting architectures or training procedures. However, forecasting robustness alone does not provide an explicitly restored history whose corrections can be inspected. To address this need, we introduce RestoreTS, an input restoration layer that corrects suspicious intervals before an unchanged forecasting backbone. It combines Bayesian online change point detection with temporal return criteria and exposes the affected intervals and applied corrections for inspection. Across eight datasets, five forecasting backbones, four horizons, and five levels of controlled combined corruption, restoration reduces MSE in 39 of the 40 dataset–backbone pairs, while MAE decreases in all 40. Mean reductions across the five backbones range from 6.8% to 19.3% for MSE and from 4.1% to 7.8% for MAE. These results suggest that input restoration can improve corruption robustness across the tested forecasting architectures. Code is available at: https://anonymous.4open.science/r/restorets-review-B9BB/.

1 INTRODUCTION

Time series forecasting uses sensor measurements to support scientific analysis and operational planning in energy infrastructure and environmental monitoring (Cheng et al., 2024; Xu et al., 2026). In practice, sensor measurements are susceptible to errors, which can appear as spikes or sustained level shifts and degrade downstream forecasts (Cheng et al., 2024; Kim et al., 2026). Existing approaches seek to mitigate these effects through changes to forecasting architectures or training procedures. LGA (Kim et al., 2026) modifies attention, RobustTSF (Cheng et al., 2024) addresses anomalous training data, and TameR (Xu et al., 2026) reduces excessive dependence on recent observations.

In scientific monitoring, the measurements themselves also inform scientific analysis and operational decisions. Preparing these measurements for use can involve inspecting suspected errors, correcting affected values, and documenting the changes. An example is SAPFLUXNET (Poyatos et al., 2021), an initiative to compile and curate a global database of sap flow measurements in plants. Its data quality control workflow combines automated checks with expert inspection, with data changes documented and quality flags retained. Improved forecasting accuracy from methods focused on the forecasting model does not by itself provide a corrected history or a record of input changes. We therefore seek to automate input restoration while retaining the corrected history and a record of its changes for inspection.

Such restoration must be selective because some abrupt changes reflect genuine events. Removing useful temporal variation can impair prediction (Liu et al., 2022). Earlier repair methods already recognize the importance of preserving valid observations. SCREEN (Song et al., 2015) and IMR (Zhang et al., 2017) follow a principle of minimizing changes when repairing errors. Our approach examines a particular temporal pattern: an abrupt departure from the surrounding level followed by a return towards it. We investigate whether these patterns can guide selective input corrections that improve forecasting across unchanged backbones. Evaluation therefore considers forecasting accuracy, which observations are changed, whether corrections reduce error relative to the uncorrupted

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.

benchmark histories retained as references, and how clean inputs are affected. Figure 1 connects this motivation to an illustrative SAPFLUXNET benchmark history.

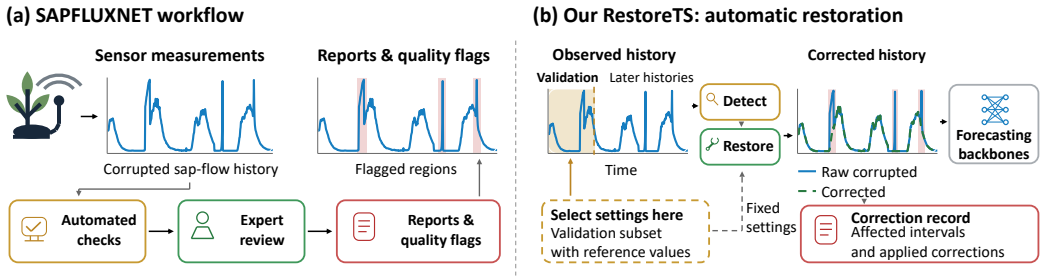


Figure 1: Conceptual workflow for inspectable restoration. (a) SAPFLUXNET addresses the quality of sensor data through automated checks, expert review, and documented flags (Poyatos et al., 2021). Shaded regions illustrate flagged observations. (b) RestoreTS applies settings selected on validation data to restore later histories automatically and records its changes for inspection.

To address this question, we propose **RestoreTS**. Bayesian online change point detection supplies candidate boundaries, and temporal return criteria identify intervals for median-return or refined-cycle correction. A subset of validation data with reference histories serves as development data for choosing restoration settings. Configurations are selected by restoration error on these data and then fixed for each dataset and supplied corruption severity. At inference, RestoreTS uses only the historical window, exposes its corrections, and passes the restored input to an unchanged forecasting backbone. To study this workflow under controlled errors resembling sensor faults, we use TSRBench (Kim et al., 2026) to inject spikes and level shifts into real benchmark histories while retaining the uncorrupted histories as references and leaving forecast targets unchanged.

We evaluate RestoreTS on established forecasting benchmarks using paired comparisons: the same trained model forecasts from either the original or the restored history, with identical forecast origins and targets. For the six datasets in Table 1, the mean relative reductions across datasets are 9.12% for MSE and 7.02% for MAE before PatchTST, averaging four horizons and five levels of combined corruption. Across all eight datasets, the corresponding reductions are 6.99% and 5.41%. Applying restoration before PatchLGA also improves its forecasts. Appendix B compares PatchTST and PatchLGA, each with and without restoration, to assess this complementarity.

Our contributions are twofold:

- We introduce **RestoreTS**, a **backbone-agnostic** input restoration layer that uses change point detection and temporal return criteria to correct suspicious historical intervals without modifying or retraining the forecasting backbone.
- We provide corrected histories and **explicit correction records**, exposing where and by how much observations were changed. These outputs support inspection alongside domain expertise and may help practitioners judge whether and how to use the resulting forecasts in scientific analysis and operational decision making.

2 Related Work

2.1 Robust Time Series Forecasting

Approaches to forecasting robustness act at different points in the pipeline. RobustTSF (Cheng et al., 2024) combines sample selection and robust losses for anomalous training data; LGA (Kim et al., 2026) adapts attention to local temporal geometry; and TameR (Xu et al., 2026) mitigates bias towards recent observations through sampling and modeling the global context. Related work on distribution shift includes RevIN (Kim et al., 2022), which normalizes input statistics and restores output statistics, and Non-stationary Transformers (Liu et al., 2022), which retain predictive nonstationary information. These methods address changing input distributions. RestoreTS instead uses

detected change points and temporal return patterns to select historical intervals for explicit value correction, retaining a record of each applied change. Its input restoration step can be combined with robustness provided by the forecasting model, as evaluated here with PatchLGA.

2.2 Time Series Cleaning and Restoration

Time series repair aims to correct erroneous observations while preserving valid temporal structure. SCREEN (Song et al., 2015) constrains the speed of temporal changes, while IMR (Zhang et al., 2017) combines temporal modeling, partially labeled references, and a principle of minimizing changes. Imputation of missing values offers a related task, where SAITS (Du et al., 2023) uses self-attention and CSDI (Tashiro et al., 2021) uses conditional diffusion to reconstruct missing entries. Imputation starts from identified missing entries, but restoration must also decide which observed values warrant modification. RestoreTS uses detected changes and subsequent returns towards the surrounding level to select intervals for correction.

2.3 Change Point Detection for Temporal Restoration

Bayesian online change point detection (BOCD) (Adams & MacKay, 2007) tracks the time since the most recent change through a posterior over run length. Robust segmentation reduces false change point detections caused by outliers (Fearnhead & Rigaill, 2019), and robust Bayesian extensions address outlier sensitivity and scalability (Altamirano et al., 2023). These methods locate changes in a series, which divide it into consecutive segments, or runs. Such a partition alone does not specify which observations to correct. RestoreTS uses BOCD to construct runs, then examines their levels and intervening jumps for temporal returns that guide restoration.

3 Method: RestoreTS

3.1 Formulation and Framework

We forecast from a possibly corrupted history $X \in \mathbb{R}^{L\times C}$, with L time steps and C measured channels. The target $Y \in \mathbb{R}^{H\times C}$ contains the next H steps. A forecaster f_ϕ with trained parameters ϕ produces the prediction from the raw input

$$\widehat{Y}^{\mathrm{raw}} = f_\phi(X). \tag{1}$$

The hat denotes a prediction, and the superscript raw indicates that it is computed from the uncorrected history. Forecasting error is measured against Y, which is unavailable during restoration.

RestoreTS first uses Bayesian online change point detection (BOCD) to partition the input into temporal runs. The selected restoration method, median-return or refined-cycle, then screens and translates candidate intervals. The unchanged forecaster receives the restored history, while correction locations and magnitudes remain available for inspection. Figure 2 separates selecting the method and parameters on a validation subset from applying the fixed choice to test histories. Sections 3.2–3.4 describe detection, correction, and selection.

For fixed configuration θ, RestoreTS estimates a correction matrix $\Delta_\theta(X) \in \mathbb{R}^{L\times C}$. The restored history and forecast are

$$\widetilde{X} = X + \Delta_\theta(X), \qquad \widehat{Y} = f_\phi(\widetilde{X}). \tag{2}$$

Here, the tilde denotes restoration, $\widetilde{X}$ retains the shape of X, and ϕ remains unchanged. Restoration acts independently on each channel after standardization using statistics computed from clean training data. Forecasters train on clean observations. Corrupted histories serve restoration selection and testing, with uncorrupted forecast targets. The configuration $\theta_{d,s}$ fixes the detector, restoration method, and acceptance thresholds for dataset d and supplied severity s. Both restoration methods preserve variation within corrected intervals and leave other values unchanged.

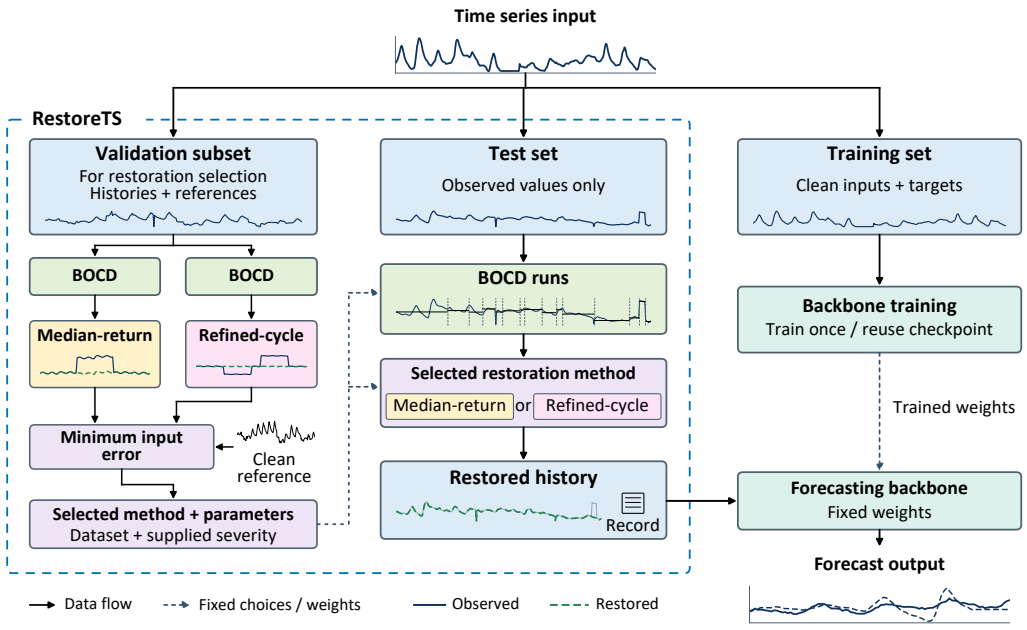


Figure 2: RestoreTS workflow. A validation subset selects the restoration method and parameters with the lowest mean squared error relative to reference histories. The selected configuration is fixed for each dataset and supplied severity, then applied to test histories. The forecasting backbone retains its trained weights.

3.2 Change Point Preliminaries and Candidate Runs

For one channel, let x_t be the value at position $t \in \{0, \ldots, L-1\}$ and $x_{0:t}$ the history through t. BOCD (Adams & MacKay, 2007) maintains a posterior over run length r_t, the number of observations since the most recent reset. We use a Gaussian predictive model with a local linear trend. Its reset probability $h = 1/\lambda$ is the prior probability of starting a new run at each step, with expected run duration λ. The observed warmup segment supplies prior and noise scales. Updates use centered joint log probabilities for numerical stability. The probability of a short run is

$$q_t = \sum_{r=0}^{R} p(r_t = r \mid x_{0:t}), \tag{3}$$

where p is the posterior probability mass function, r indexes run lengths, and R is the largest run length counted as recent. Thus, q_t measures the posterior probability of a change within the last R steps.

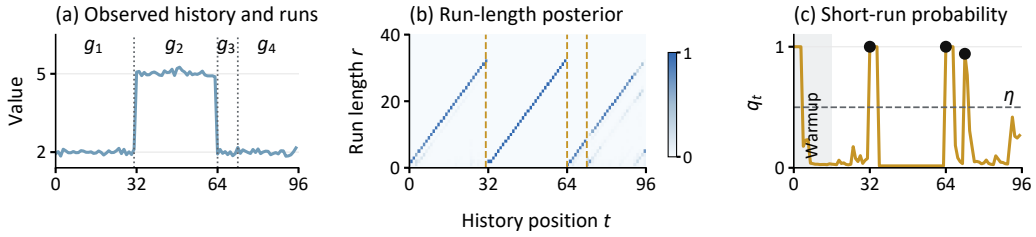


Figure 3: BOCD on a constructed history. (a) Consecutive runs g_i. (b) Posterior probability over run length (color). (c) Probability of a short run q_t, threshold η, and detected peaks (dots).

After warmup, local peaks above threshold η flag possible changes, separated by a minimum interval. Each flagged peak projects back to $b = t - \widehat{r} + 1$, where $\widehat{r}$ is the most probable run length

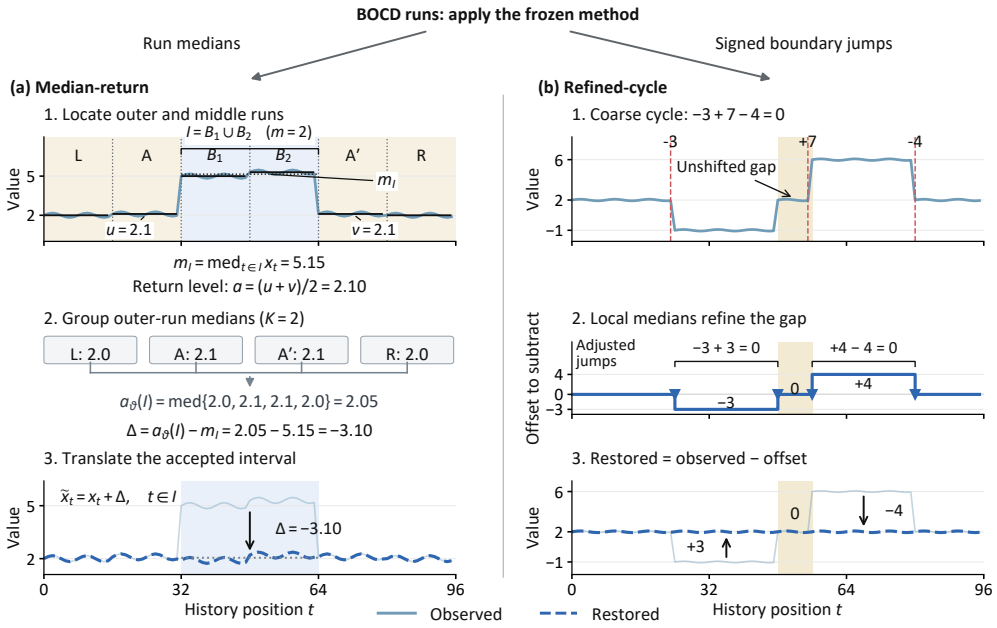


Figure 4: Restoration from BOCD run statistics. (a) Median-return identifies an excursion using agreement between the outer runs and translates it to a reference median. (b) Local refinement splits a coarse cycle into two corrections, preserving the unshifted gap and variation within each segment. These separate constructed examples use supplied partitions or coarse boundaries. The frozen configuration selects one method, subject to the acceptance conditions in the text.

among $1 \le r \le \min(R, t+1)$. When estimated change locations are close, we retain the one with the largest q_t.

The boundaries define chronological runs $g_i = \{s_i, \ldots, e_i - 1\}$, with inclusive start s_i and exclusive end e_i (Figure 3). Each run supplies its median and, when sufficiently long, a robust Theil–Sen linear fit. The signed jump is the right fitted value minus the left fitted value at their boundary. Run medians and signed jumps support the two restoration methods in Section 3.3.

3.3 Two Temporal Return Restoration Methods

Both restoration methods seek an excursion followed by a return. Median-return uses run medians to move an interval towards its surrounding level. Refined-cycle uses signed jumps to estimate successive offsets. Figure 4 uses constructed numerical examples to illustrate the operations, and their values are not fixed method constants. One method is selected for each dataset and supplied severity.

Median-return restoration. We search chronologically for $A \to B_1 \to \cdots \to B_m \to A'$, where A, A' are the runs before and after the excursion, and the m intervening runs $B_1, \ldots, B_m$ form a candidate interval I. The number of middle runs satisfies $1 \le m \le M$, where M is the configured maximum, and the total span is also bounded. Let u, v be the medians of the outer runs, m_I the observation median within I, and $a = (u + v)/2$. A candidate requires

$$\frac{|v - u|}{\max(|m_I - a|, \epsilon)} \le \rho, \tag{4}$$

where $\epsilon > 0$ is a denominator floor and ρ the return tolerance. The numerator measures how closely the series returns to its earlier level, and the denominator measures the excursion from that level. A small ratio therefore favors a substantial departure followed by a close return. Medians of the middle runs must lie on the same side of a beyond a minimum contrast, with sufficiently large, oppositely signed entry and return changes. Selecting the earliest qualifying return for each left run, then advancing to the right outer run, produces nonoverlapping intervals.

The correction target $a_\theta(I)$ estimates the surrounding level to which the interval should return. We consider the immediate outer average, a median over K reference runs on each side for broader support, and a robust linear fit to those run medians evaluated at the interval midpoint to accommodate a local trend. Development search selects the estimator, K, and M, and these choices remain fixed for test intervals. This target can differ from the level a used in the return criterion. Insufficient reference context causes rejection. Acceptance bounds additionally check return agreement, surrounding support dispersion, and variation within the interval. The correction is

$$\widetilde{x}_t = x_t + a_\theta(I) - m_I, \qquad t \in I. \tag{5}$$

This constant shift moves the interval median to the target while preserving differences between its observations. Appendix A.3 gives the algebraic identities.

Refined-cycle restoration. Refined-cycle subtracts accumulated jumps from a excursion involving several changes. Starting with a jump that exceeds the configured magnitude threshold and has valid fits on both sides, we accumulate signed jumps j_i at ordered boundaries b_i. Here, J is the number of jumps in a candidate cycle, subject to limits on duration and boundary count. The earliest coarse cycle closes when $e = \sum_{i=1}^{J} j_i$ is small relative to the largest jump and the terminal jump opposes and reduces the preceding residual. To make the estimated offset return to zero, we subtract e/J from each jump, giving $\bar{j}_i = j_i - \frac{e}{J}$ and the correction

$$\widetilde{x}_t = x_t - \sum_{i=1}^{k} \bar{j}_i, \qquad b_k \le t < b_{k+1}, \quad 1 \le k < J. \tag{6}$$

Here, i indexes jumps and k indexes the segment after boundary b_k. The adjusted jumps sum to zero, so the inferred offset closes. Each segment receives a constant shift, preserving its internal differences.

Local median differences refine a coarse cycle only if all detected candidates form closed cycles whose outer endpoints align within the prescribed radius. Otherwise, the coarse cycle remains. A minimum segment length avoids corrections on very short segments. When active, a plateau bound limits the 10–90% residual width after robust detrending relative to the correction magnitude, screening out segments with large internal variation. Failure of any active bound rejects the whole cycle. The procedure accepts only nonoverlapping correction intervals. The illustrative jumps $-3, +7, -4$ in Figure 4(b) show how refinement can split a coarse cycle into two corrections while retaining an unshifted gap.

3.4 Frozen Selection and Inspectable Corrections

For each dataset and supplied severity, we choose the restoration method and settings that best reconstruct the uncorrupted reference histories, measured by mean squared error on a development subset of validation data. The selected choice is fixed before inference and shared across channels, horizons, backbones, and corruption types. Inference uses only the available input history and leaves it unchanged when no interval is accepted. Each correction records its channel, interval, and signed magnitude for inspection. Appendices A.4 and A.5 give selection details and pseudocode, and Sections 4.1 and 4.5 describe evaluation.

4 Experiments

4.1 Evaluation Setup

We evaluate electrical transformer measurements (Zhou et al., 2021), weather (Max Planck Institute for Biogeochemistry, 2020), air quality (Chen, 2017), and sap flow (Poyatos et al., 2021). Table 3 gives domains, sampling intervals, and sizes. Table 1 shows ETTm1, ETTm2, Weather, AQWan, CzeLan, and ZafNoo, with the complete comparison across eight datasets in Appendix B, Table 8. We compare PatchTST (Nie et al., 2023), PatchLGA (Kim et al., 2026), and RestoreTS + PatchTST, and assess transfer to DLinear (Zeng et al., 2023), TimesNet (Wu et al., 2023), TimeMixer (Wang et al., 2024), and iTransformer (Liu et al., 2024).

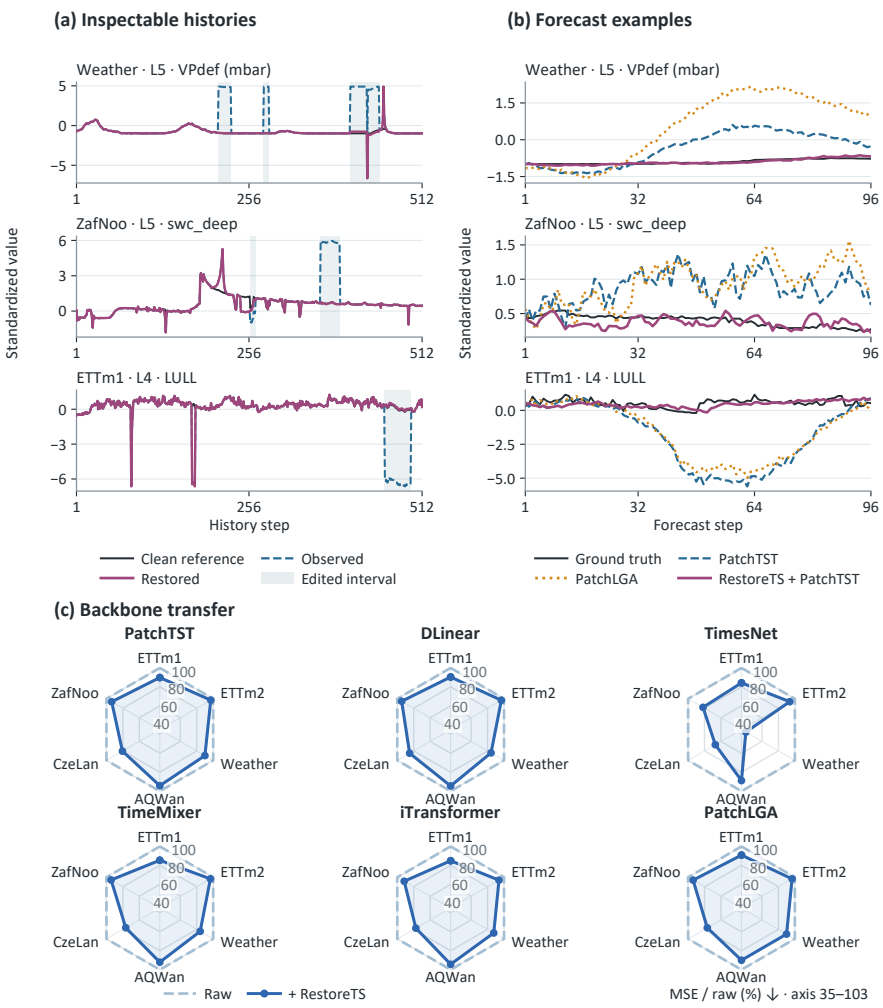


Figure 5: Matched histories, forecasts, and backbone transfer. Rows in (a,b) share channel and origin: Weather and ZafNoo at combined L5, ETTm1 at L4. (a) Full inputs of 512 steps with edited intervals shaded and clean references shown for evaluation. (b) Corresponding forecasts over 96 steps from selected successful cases. (c) MSE relative to raw inputs across six datasets, four horizons, and combined L1–L5.

Restoration parameters and the final method choice are selected by development input error. Restoration uses neither future forecast targets nor clean test references. Earlier searches required confirmation checks before accepting a configuration, and the final method comparison is evaluated on confirmation data without reranking or retuning. Test evaluation retains both improvements and regressions. The benchmark splits had already been evaluated in earlier runs. More details about the restoration parameters and the final restoration method selection can be seen in Appendix A.

The original TSRBench generator (Kim et al., 2026) injects spikes and sustained offsets while retaining reference histories and unmodified targets for evaluation. L0 denotes clean inputs and L1–L5 increasing corruption frequency, duration, and tail severity. Results in the main text use combined corruption over horizons $H \in \{96, 192, 336, 720\}$. Paired raw/restored forecasts share checkpoints, origins, and targets, scored by mean squared error (MSE) and mean absolute error (MAE). Restoration settings are fixed across horizons and backbones (Appendix A.1). L0 measures clean preservation and is excluded from corrupted aggregates. Percentage reductions first average errors over horizons and L1–L5 within each dataset, then average the relative reductions equally across datasets. Appendix A supplies corruption parameters, partitions, and seeds.

Table 1: Forecasting under combined corruption, averaged equally over $H \in \{96, 192, 336, 720\}$. RestoreTS (ours) denotes RestoreTS + PatchTST, and L0 is clean. Bold marks the best displayed value, and underlining the second distinct value.

	ETTm1						**ETTm2**						**Weather**					
	PatchTST		PatchLGA		RestoreTS (ours)		PatchTST		PatchLGA		RestoreTS (ours)		PatchTST		PatchLGA		RestoreTS (ours)	
Level	MSE↓	MAE↓	MSE↓	MAE↓	MSE↓	MAE↓	MSE↓	MAE↓	MSE↓	MAE↓	MSE↓	MAE↓	MSE↓	MAE↓	MSE↓	MAE↓	MSE↓	MAE↓
L0	0.352	0.382	**0.350**	**0.380**	0.352	0.382	**0.256**	**0.317**	0.259	**0.317**	0.257	**0.317**	0.226	0.264	0.226	0.264	**0.225**	**0.263**
L1	0.359	0.389	**0.358**	**0.387**	0.359	0.388	**0.264**	0.328	0.265	**0.326**	**0.264**	0.327	0.239	0.283	0.237	0.282	**0.236**	**0.277**
L2	0.375	0.401	**0.372**	**0.399**	0.374	**0.399**	0.277	0.338	0.278	0.336	**0.274**	**0.334**	0.264	0.311	0.264	0.312	**0.253**	**0.291**
L3	0.585	0.495	0.532	0.475	**0.509**	**0.455**	0.308	0.360	0.305	0.356	**0.300**	**0.352**	0.305	0.337	0.303	0.338	**0.278**	**0.306**
L4	0.667	0.546	0.640	0.536	**0.563**	**0.488**	0.361	0.400	0.358	0.396	**0.347**	**0.387**	0.369	0.382	0.362	0.389	**0.323**	**0.342**
L5	0.803	0.601	0.759	0.587	**0.702**	**0.543**	0.431	0.434	0.424	0.429	**0.408**	**0.416**	0.494	0.422	0.463	0.424	**0.413**	**0.371**

	AQWan						**CzeLan**						**ZafNoo**					
	PatchTST		PatchLGA		RestoreTS (ours)		PatchTST		PatchLGA		RestoreTS (ours)		PatchTST		PatchLGA		RestoreTS (ours)	
Level	MSE↓	MAE↓	MSE↓	MAE↓	MSE↓	MAE↓	MSE↓	MAE↓	MSE↓	MAE↓	MSE↓	MAE↓	MSE↓	MAE↓	MSE↓	MAE↓	MSE↓	MAE↓
L0	**0.837**	**0.506**	0.881	0.520	**0.837**	**0.506**	**0.227**	**0.272**	0.232	0.278	**0.227**	**0.272**	**0.513**	**0.456**	0.517	0.458	**0.513**	**0.456**
L1	0.851	0.525	0.886	0.533	**0.849**	**0.519**	0.251	0.301	0.253	0.302	**0.245**	**0.291**	0.530	0.474	0.532	0.474	**0.527**	**0.473**
L2	0.864	0.541	0.904	0.552	**0.858**	**0.530**	0.300	0.347	0.291	0.340	**0.263**	**0.309**	0.560	0.502	0.559	0.500	**0.551**	**0.493**
L3	0.974	0.607	1.044	0.621	**0.932**	**0.570**	0.374	0.393	0.358	0.379	**0.298**	**0.333**	0.616	0.529	0.615	0.527	**0.596**	**0.516**
L4	1.055	0.657	1.201	0.687	**0.991**	**0.608**	0.525	0.485	0.541	0.481	**0.392**	**0.385**	0.740	0.577	0.732	0.575	**0.669**	**0.554**
L5	1.344	0.757	1.655	0.811	**1.164**	**0.669**	1.265	0.629	1.325	0.618	**0.989**	**0.499**	0.976	0.652	0.974	0.648	**0.861**	**0.621**

4.2 Robustness and Complementarity

Across the six datasets in Table 1, RestoreTS reduces PatchTST's mean relative MSE by 9.12% and MAE by 7.02%, averaging four horizons and L1–L5 within each dataset. RestoreTS + PatchTST also has lower MSE than PatchLGA in 28 of 30 dataset–severity pairs and lower MAE in 27, using unrounded values. Across all eight datasets (Table 8), the reductions relative to PatchTST are 6.99% and 5.41%.

Before PatchLGA, RestoreTS reduces mean relative MSE by 7.77% and MAE by 5.41%, supporting its combination with a robust forecasting model. Appendix B, Table 9 reports PatchTST and PatchLGA, each with and without restoration, followed by results for separate corruption types.

4.3 Restoration across Forecasting Backbones

Figure 5(c) shows paired raw/restored MSE ratios for six datasets, including PatchLGA. Across five backbones and all eight datasets (Appendix B, Table 7), restoration reduces MSE in 39 of 40 dataset–backbone pairs and MAE in all 40. Mean relative reductions range from 6.83% to 19.27% for MSE and 4.14% to 7.84% for MAE, with restoration settings shared across backbones.

4.4 Component Ablations

Figure 6 and Table 2 examine BOCD partitioning, local refinement, and selection of the restoration method. Uniform intervals replace BOCD, with interval width selected on development inputs and downstream settings fixed. Removing local refinement retains coarse cycles and the same checks on minimum segment length and plateaus. Variants with a fixed method use median-return or refined-cycle everywhere with their archived settings. All variants share PatchTST checkpoints and test inputs.

Replacing BOCD reduces the MSE gain from 6.99% to 1.88%. Removing refinement gives 6.85% MSE and 5.25% MAE reductions, indicating a small aggregate contribution. Refinement improves both metrics on six dataset averages, leaves ETTh2 unchanged, and worsens CzeLan. Method selection adds 0.33 MSE and 0.19 MAE percentage points over using refined-cycle alone. Figure 6

Table 2: Component ablations before PatchTST. Mean error reductions (%) relative to raw inputs across eight datasets, four horizons, and combined L1–L5.

Variant	ΔMSE↑	ΔMAE↑
Uniform intervals	1.88	2.10
No local refinement	6.85	5.25
All median-return	2.92	2.43
All refined-cycle	6.66	5.22
RestoreTS (ours)	**6.99**	**5.41**

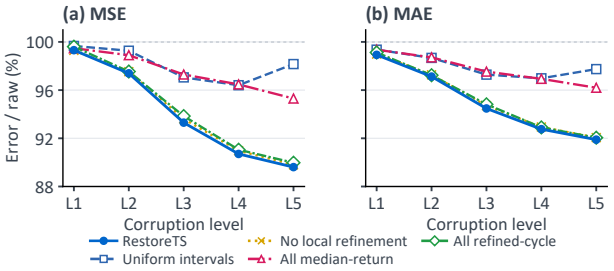


Figure 6: Component ablations across levels of combined corruption before PatchTST. At each level, errors average four horizons within each dataset, then variant/raw ratios average equally over eight datasets.

shows that the gains generally increase from L1 to L5: RestoreTS reaches 89.61% of raw MSE and 91.89% of raw MAE at L5. Uniform intervals remain closer to raw performance, while the curves for no local refinement and fixed refined-cycle stay close to RestoreTS. Appendix E.1 reports the protocols, paired results, and tradeoffs on clean inputs. Other restoration procedures appear in Appendix E.2. Separate ablations of the median-return acceptance bounds use seven datasets at $H = 336$ (Appendix C).

4.5 Inspecting Corrections and Preserving Clean Inputs

Figure 5(a,b) pairs histories of 512 steps with forecasts over the subsequent 96 steps for Weather/VPdef, ZafNoo/swc_deep, and ETTm1/LULL, using the same channel, window, and corruption level. The observed and restored curves expose the changes, where shading marks edited intervals and $\widetilde{X} - X$ gives their signed magnitudes. Clean references are shown only for evaluation. These selected successful cases illustrate how corrections can be inspected, and the input audit assesses restoration across the benchmark histories.

We use *restoration selectivity* to describe where corrections are applied and whether they reduce input error. Across the eight datasets, the input audit under combined corruption finds reductions in input MSE of 9.60–50.95%. Of the edited values, 39.32–84.51% move closer to the reference. On clean histories, 0–0.34332% of point exposures are edited, with nonzero introduced error on six datasets. The input audit uses the available cached forecast origins at $H = 96$. The separate study of acceptance bounds uses seven datasets at $H = 336$ and 128 fixed test origins. Appendix C reports both scopes, Appendix A.2 defines the tested bounds, and Appendix F discusses their implications. The restored histories and correction records make the interventions available for inspection, showing which observations changed and by how much. Together with the input audit, these outputs can help domain experts assess the corrections alongside forecasts when using the results to inform scientific analysis and decision making.

5 Conclusion

We presented RestoreTS, an input restoration layer that corrects suspicious historical intervals before an unchanged forecasting backbone. It combines Bayesian online change point detection with temporal return criteria and records the locations and magnitudes of applied corrections. Across eight datasets, five backbones, four horizons, and five levels of combined corruption, paired evaluations show lower MSE for 39 of the 40 combinations of dataset and forecasting backbone and lower MAE for all 40. Further experiments demonstrate complementarity with PatchLGA and consistent aggregate gains across separate studies varying training seeds and corruption seeds. Together, these findings support explicit input restoration as a complementary approach to robust forecasting. The corrected histories and correction records also provide a concrete basis for domain experts to inspect interventions alongside forecasts when considering their use in scientific analysis and operational planning.

AI USE STATEMENT

The human authors owned the original ideas, defined the research problem, and directed the scientific and editorial decisions. Under their supervision, generative AI tools assisted with parts of the method implementation and debugging, provided feedback on experimental procedures and help check mathematical explanations of the restoration operations. Additional assistance included improving organization and readability, and checking figure-generation code. We have reviewed all AI-assisted work and revised the material incorporated into the manuscript. The authors take responsibility for the final manuscript and accompanying artifacts, including all AI-assisted text, claims, code, and figures.

ETHICS STATEMENT

This work evaluates restoration and forecasting on existing energy and environmental time-series benchmarks. Dataset sources and processing are documented in Appendix A. Reuse and redistribution should respect the source licenses and attribution requirements. A detected change is not necessarily a sensor fault, and automatic correction can suppress genuine events or alter valid measurements. Corrected histories should therefore remain traceable to the original observations through correction records and receive domain review before use in consequential scientific or operational decisions. The controlled-corruption experiments do not establish safety in deployment or measure reductions in expert workload.

REPRODUCIBILITY STATEMENT

Code is available at: https://anonymous.4open.science/r/restorets-review-B9BB/. Section 3 specifies the restoration methods and frozen configuration selection, and Section 4.1 describes the paired evaluation. Appendix A documents datasets, chronological partitions, preprocessing, corruption settings, checkpoints, and seed choices. Sections 3.3–3.4 give the restoration and selection equations. To assess repeatability, we vary training seeds and corruption seeds in two separate three-seed studies covering all eight datasets and four horizons, with restoration settings fixed. Appendix D reports the study protocols, three-seed means, sample standard deviations, and individual results.

REFERENCES

Ryan Prescott Adams and David J. C. MacKay. Bayesian online changepoint detection, 2007. URL https://arxiv.org/abs/0710.3742.

Matias Altamirano, Francois-Xavier Briol, and Jeremias Knoblauch. Robust and scalable Bayesian online changepoint detection. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 642–663. PMLR, 2023. URL https://proceedings.mlr.press/v202/altamirano23a.html.

Song Chen. Beijing multi-site air quality. UCI Machine Learning Repository, 2017. URL https://archive.ics.uci.edu/dataset/501/beijing+multi+site+air+quality+data.

Hao Cheng, Qingsong Wen, Yang Liu, and Liang Sun. RobustTSF: Towards theory and design of robust time series forecasting with anomalies. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/179f5dcdeedc149443ebd3ba70811dbd-Abstract-Conference.html.

Wenjie Du, David Côté, and Yan Liu. SAITS: Self-attention-based imputation for time series. *Expert Systems with Applications*, 219:119619, 2023. doi: 10.1016/j.eswa.2023.119619. URL https://doi.org/10.1016/j.eswa.2023.119619.

Paul Fearnhead and Guillem Rigaill. Changepoint detection in the presence of outliers. *Journal of the American Statistical Association*, 114(525):169–183, 2019. doi: 10.1080/01621459.2017.1385466. URL https://doi.org/10.1080/01621459.2017.1385466.

Neal C. Gallagher and Gary L. Wise. A theoretical analysis of the properties of median filters. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29(6):1136–1141, 1981. doi: 10.1109/TASSP.1981.1163708.

Qihe Huang, Zhengyang Zhou, Kuo Yang, Zhongchao Yi, Xu Wang, and Yang Wang. TimeBase: The power of minimalism in efficient long-term time series forecasting. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 26227–26246, 2025. URL https://proceedings.mlr.press/v267/huang25az.html.

Dongbin Kim, Youngjoo Park, Woojin Jeong, and Jaewook Lee. Local geometry attention for time series forecasting under realistic corruptions. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/476ab8f369e489c04187ba84f68cfa68-Abstract-Conference.html.

Taesung Kim, Jinhee Kim, Yunwon Tae, Cheonbok Park, Jang-Ho Choi, and Jaegul Choo. Reversible instance normalization for accurate time-series forecasting against distribution shift. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=cGDAkQo1C0p.

Yong Liu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Non-stationary transformers: Exploring the stationarity in time series forecasting. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/4054556fcaa934b0bf76da52cf4f92cb-Abstract-Conference.html.

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. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/2ea18fdc667e0ef2ad82b2b4d65147ad-Abstract-Conference.html.

Max Planck Institute for Biogeochemistry. Weather station data, jena, 2020. URL https://www.bgc-jena.mpg.de/wetter/. Weather benchmark: 2020 observations.

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

Ronald K. Pearson. Outliers in process modeling and identification. *IEEE Transactions on Control Systems Technology*, 10(1):55–63, 2002. doi: 10.1109/87.974338.

Rafael Poyatos, Víctor Granda, Víctor Flo, et al. Global transpiration data from sap flow measurements: the SAPFLUXNET database. *Earth System Science Data*, 13:2607–2649, 2021. doi: 10.5194/essd-13-2607-2021. URL https://essd.copernicus.org/articles/13/2607/2021/.

Shaoxu Song, Aoqian Zhang, Jianmin Wang, and Philip S. Yu. SCREEN: Stream data cleaning under speed constraints. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, pp. 827–841, 2015. doi: 10.1145/2723372.2723730. URL https://doi.org/10.1145/2723372.2723730.

Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. CSDI: Conditional score-based diffusion models for probabilistic time series imputation. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/cfe8504bda37b575c70ee1a8276f3486-Abstract.html.

Shiyu Wang, Haixu Wu, Xiaoming Shi, Tengge Hu, Huakun Luo, Lintao Ma, James Y. Zhang, and Jun Zhou. TimeMixer: Decomposable multiscale mixing for time series forecasting. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/a7ac8a21e5a27e7ab31a5f42a0117bdb-Abstract-Conference.html.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. TimesNet: Temporal 2d-variation modeling for general time series analysis. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=ju_Uqw384Oq.

Longlong Xu, Zeyan Li, Xiao He, Zhaoyang Yu, Changhua Pei, Zhe Xie, Zijun Dou, Tieying Zhang, and Dan Pei. Taming the recent-data bias: Towards robust time series forecasting with global context. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://netman.aiops.org/wp-content/uploads/2026/05/ICML26_TameR_20260520.pdf.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 11121–11128, 2023. doi: 10.1609/aaai.v37i9.26317.

Aoqian Zhang, Shaoxu Song, Jianmin Wang, and Philip S. Yu. Time series data cleaning: From anomaly detection to anomaly repairing. *Proceedings of the VLDB Endowment*, 10(10):1046–1057, 2017. URL https://www.vldb.org/pvldb/vol10/p1046-song.pdf.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(12):11106–11115, 2021. doi: 10.1609/aaai.v35i12.17325. URL https://ojs.aaai.org/index.php/AAAI/article/view/17325.

A Evaluation Protocol and Scope

Table 3 summarizes the retained benchmark series. The ETT datasets contain transformer load and oil-temperature measurements, while Weather contains meteorological variables. AQWan is an air-quality benchmark series, while CzeLan and ZafNoo contain sap-flow and environmental measurements. The three latter files are the processed releases distributed with TimeBase (Huang et al., 2025). Their upstream sources are Beijing Multi-Site Air Quality and SAPFLUXNET (Chen, 2017; Poyatos et al., 2021).

Table 3: Dataset details. Counts refer to retained time steps. C is the channel count and L is the input length. Train/validation/test counts follow the executed chronological split.

Dataset	Sampling interval	C	L	Train / validation / test
ETTh1 / ETTh2	1 hour	7	336	8,640 / 2,880 / 2,880
ETTm1 / ETTm2	15 minutes	7	512	34,560 / 11,520 / 11,520
Weather	10 minutes	21	512	36,887 / 5,270 / 10,539
AQWan	1 hour	11	512	21,038 / 7,014 / 7,012
CzeLan	30 minutes	11	512	11,960 / 3,988 / 3,986
ZafNoo	30 minutes	11	512	11,535 / 3,845 / 3,845

Development is the validation portion used to search restoration settings and select the final family lookup by input error. *Confirmation* is a later portion used to evaluate fixed choices before test replay. Environmental datasets reserve an initial clean validation block for backbone training: forecasting MSE on this block selects the checkpoint and triggers early stopping after the configured patience period without improvement. The remaining validation data are divided chronologically into development and confirmation for restoration selection and evaluation. Legacy ETT/Weather experiments use archived partitions with the prior exposure described below. The final family lookup uses neither confirmation errors nor forecasting scores for ranking.

Two archived median-return searches used confirmation as an acceptance check. Each selected the best eligible development candidate and assessed that same candidate on confirmation. The reference search required introduced clean-input MSE divided by the mean squared clean-signal amplitude to be at most 0.01, and the mean combined-corruption restored/raw input-error ratio over L1–L5 to be at most 1.00. The later support/coherence-bound search retained the clean-input bound and required a mean combined ratio below 1.00, a mean spike ratio at most 1.05, each L1–L2 shift and combined ratio at most 1.10, and changes to at least one corrupted point. Failure prevented that search from writing a configuration lock and flagged further method development. It did not select a substitute using confirmation. These earlier configuration checks are separate from the final median-return/refined-cycle comparison and do not filter the final confirmation/test results.

TSRBench corruption settings. Controlled spikes and level shifts approximate short-lived excursions and sustained offsets, with combined corruption testing their interaction. Reference counterparts enable matched severity comparisons and development selection. This benchmark does not identify natural faults, reproduce every sensor failure, or measure annotation costs and manual correction time.

The original LGA generator defines levels L1–L5 using frequency parameters $(0.002, 0.004, 0.004, 0.008, 0.008)$, duration parameters $(6, 9, 12, 12, 15)$, and extreme-value tail-risk parameters $(0.0016, 0.0016, 0.0004, 0.0004, 0.0001)$ (Kim et al., 2026). Durations are sampled rather than fixed. Lower tail risk produces more extreme amplitudes. Combined corruption retains the larger absolute spike or shift perturbation at each location, rather than adding both. We reuse the saved perturbations from the original generator. Its whole-series amplitude calibration is part of benchmark construction. The forecasting standardizer is fitted only on training data. Reference histories and future targets are excluded from the restoration input.

The four-horizon tables use historical model seed 2021 and corruption seed 2025. The supplementary repetition study varies training seeds $\{2021, 2026, 2027\}$ at corruption seed 2025, and separately varies corruption seeds $\{2025, 2026, 2027\}$ with the original seed-2021 checkpoints. Both studies cover all eight datasets and four horizons. The two axes are reported separately in Appendix D.

Paired evaluation. The five-backbone transfer suite averages the twenty horizon–severity conditions per dataset/backbone. Complementarity (including PatchLGA), components, and both seed studies use four horizons and full retained test support. The acceptance-bound study instead uses $H = 336$ and 128 evenly spaced origins per phase. Each comparison pairs identical origins, checkpoints, and targets. Full support retains complete batches of 128, rather than 128 origins total. Historical architectures retain their training configurations. The historical and newly matched seed-2021 checkpoints are distinct and tracked separately.

Scaling is fitted on training data. The centered-log-joint BOCD implementation, median-return/refined-cycle family settings, and dataset/severity lookup remain frozen. Clean development references support input-error selection, while clean evaluation references are used only for scoring. Restoration uses the available encoder window and supplied severity, including later observations inside that window when applying return criteria, but never the future forecast target. The same lookup is applied to spike-only, shift-only, and combined corruption without corruption-type-specific reselection.

Median and Hampel filtering operate along time with nearest-endpoint padding. Both search window widths $\{3, 5, 7, 11, 21\}$. Hampel additionally searches thresholds $\{2, 3, 4\}$ in MAD-scale units with factor 1.4826. Each dataset/severity choice minimizes mean restored-input MSE on 128 evenly spaced development windows between stopping and confirmation blocks, with full encoder support before confirmation. Identity is not a searched candidate. Choices are frozen before supplementary confirmation/test measurement, use no forecasting score, and are reused across all four horizons and corruption types.

For an odd width $w = 2r + 1$, let x^{pad} repeat the nearest endpoint outside the history. Define $m_t^{\mathrm{loc}} = \mathrm{median}_{|j|\le r}\, x_{t+j}^{\mathrm{pad}}$ and $d_t^{\mathrm{loc}} = \mathrm{median}_{|j|\le r}\, |x_{t+j}^{\mathrm{pad}} - m_t^{\mathrm{loc}}|$. Median filtering returns m_t^{loc} everywhere (Gallagher & Wise, 1981), whereas Hampel filtering returns

$$\widetilde{x}_t^{\mathrm{Hampel}} = \begin{cases} m_t^{\mathrm{loc}}, & |x_t - m_t^{\mathrm{loc}}| > \kappa\, 1.4826\, d_t^{\mathrm{loc}}, \\ x_t, & \text{otherwise.} \end{cases}$$

The factor 1.4826 gives Gaussian-consistent MAD scaling (Pearson, 2002). The strict test uses no MAD floor: when MAD is zero, any value differing from the local median is replaced. All neighborhood statistics use the original window in one pass, without recursive updates.

Let $\bar{e}_{d,k,v}$ denote the mean error for dataset d, backbone k, and input variant v over the conditions explicitly named in a table. The relative reduction is $100(1-\bar{e}_{d,k,v}/\bar{e}_{d,k,\mathrm{raw}})$. The reported aggregate averages these reductions equally over datasets. This differs from pooling raw errors across datasets or averaging condition-wise percentages. L0 is excluded from corrupted aggregates and reported as a separate condition in the same table. A tie after three-decimal rounding is a display tie, not a statistical equivalence result.

Data exposure. The reported results reuse benchmark splits evaluated in earlier runs, rather than a newly held-out temporal period. The final method choice is ranked by development input error, without reselection from confirmation or test scores. Additional corruption seeds change the perturbations while retaining the underlying observations.

A.1 Frozen Detector and Restoration Settings

The frozen dataset/severity lookup stores the family, detector and correction settings, and fixed implementation defaults. Boundaries, run statistics, reference levels, and corrections are computed from each observed history. Clean references enter development selection only, and neither clean references nor forecasting targets enter inference.

BOCD parameters and initialization. The warmup length W determines how many initial observations are used for empirical-Bayes initialization. The hazard parameter λ defines the constant reset probability $h = 1/\lambda$. The short-run cutoff is $R = 8$ in every selected configuration, and η is the alert threshold for q_t in Section 3.2. The refractory interval is 8 time steps, nearby boundary estimates are deduplicated within a radius of 4, and the minimum length for a run-level trend fit is 3. A numerical floor of 10^{-8} stabilizes scale calculations. Table 4 lists the varying values for every dataset/severity choice. The contrast threshold δ and maximum candidate span G belong to candidate correction, although they share the detector configuration dictionary in the implementation. The table reports the executed settings, not search ranges.

Each channel/window initializes its level from the median of the first W observations and its noise variance from the MAD of their consecutive differences. Prior level and slope scales use empirical variances and the numerical floor. Posterior probabilities, boundaries, run statistics, acceptance decisions, and translations therefore remain input-dependent despite frozen settings.

Table 4: Frozen family choices and varying detector/candidate settings. W is warmup length, λ the hazard scale, η the short-run alert threshold, δ the minimum contrast or major-jump threshold, and G the maximum candidate span. Time-related settings count input steps. Contrast thresholds are in standardized units and rounded to six decimals here. The saved configuration files retain full precision.

Dataset	Levels	Family	W	λ	η	δ	G
ETTh1	L0–L4	median-return	32	32	0.7	0.897174	72
ETTh1	L5	refined-cycle	32	32	0.6	1.196232	48
ETTh2	L0	refined-cycle	63	63	0.45	0.961768	64
ETTh2	L1–L5	median-return	63	63	0.45	0.961768	64
ETTm1	L0	median-return	96	96	0.45	0.953647	64
ETTm1	L1–L5	refined-cycle	96	96	0.45	0.953647	64
ETTm2	L0–L1	median-return	96	96	0.45	0.509495	64
ETTm2	L2–L5	refined-cycle	96	96	0.45	0.509495	64
Weather	L0	median-return	32	128	0.7	0.080446	48
Weather	L1–L5	refined-cycle	48	96	0.6	0.160892	48
AQWan	L0	median-return	48	96	0.5	0.902284	48
AQWan	L1, L3	refined-cycle	32	128	0.7	0.451142	48
AQWan	L2	refined-cycle	48	48	0.6	0.563927	48
AQWan	L4	refined-cycle	48	96	0.6	0.902284	48
AQWan	L5	refined-cycle	64	64	0.5	0.902284	80
CzeLan	L0	median-return	48	96	0.5	0.406151	48
CzeLan	L1, L4, L5	refined-cycle	48	96	0.6	0.406151	48
CzeLan	L2	refined-cycle	48	48	0.6	0.253844	48
CzeLan	L3	refined-cycle	96	96	0.45	0.406151	64
ZafNoo	L0, L2	median-return	48	96	0.7	0.374154	48
ZafNoo	L1	refined-cycle	32	128	0.7	0.149661	48
ZafNoo	L3	median-return	48	96	0.7	0.224492	48
ZafNoo	L4	refined-cycle	48	48	0.6	0.187077	48
ZafNoo	L5	refined-cycle	96	96	0.45	0.299323	64

Median-return parameters. Table 5 specifies the maximum number of middle runs M, return-ratio bound $\rho_{\max}$, reference estimator, and active support-dispersion and plateau-variation bounds. An immediate reference uses one outer run on each side. A grouped median pools K outer-run medians per side. A trend bridge fits a robust line through the outer-run medians. The statistics and their denominators are defined in Appendix A.2. "Off" means that the corresponding acceptance bound is not active. It does not remove candidate-return or contrast requirements. In particular, ETTh2 uses the saved current-gate choice without active support or plateau bounds.

Table 5: Frozen median-return settings at levels where that family is selected. Support and plateau columns give the maximum accepted ratios defined in Appendix A.2.

Dataset	Levels	M	$\rho_{\max}$	Reference	Support	Plateau
ETTh1	L0–L4	3	0.25	Grouped median, $K = 5$	0.25	0.5
ETTh2	L1–L5	3	0.25	Immediate	Off	Off
ETTm1	L0	3	0.25	Trend bridge, $K = 5$	0.25	0.25
ETTm2	L0–L1	1	0.25	Grouped median, $K = 2$	0.25	0.5
Weather	L0	3	0.25	Immediate	0.5	0.25
AQWan	L0	1	0.15	Immediate	0.1	0.25
CzeLan	L0	1	0.15	Immediate	0.1	0.25
ZafNoo	L0	1	0.15	Immediate	0.1	0.25
ZafNoo	L2	3	0.25	Immediate	0.25	0.25
ZafNoo	L3	3	0.35	Immediate	0.25	0.25

Refined-cycle parameters. The coarse-cycle closure tolerance varies as shown in Table 6. All selected refined-cycle configurations allow at most 6 coarse-cycle boundaries and 8 local-cycle

boundaries. Local refinement computes differences between medians using a window of 2 steps, a refinement radius of 5, and a nonmaximum-suppression radius of 2. The local closure tolerance is 0.15 and the local stability ratio is 0.5. Minimum segment length and the maximum detrended plateau-residual ratio vary by configuration. An inactive plateau bound does not disable the other cycle requirements. The coarse boundaries and jumps are retained when local refinement fails its matching conditions. A failed active segment-length or plateau bound rejects the candidate cycle, as described in Section 3.3.

Table 6: Frozen refined-cycle settings at levels where that family is selected. The plateau bound applies to robustly detrended segment residuals and is distinct from the median-return plateau statistic. All other local-refinement settings are given in the preceding paragraph.

Dataset	Levels	Coarse closure	Minimum length	Plateau bound
ETTh1	L5	0.08	4	0.35
ETTh2	L0	0.15	4	0.6
ETTm1	L1–L5	0.15	4	0.35
ETTm2	L2–L5	0.15	4	0.6
Weather	L1–L5	0.08	1	Off
AQWan	L1	0.06	4	0.35
AQWan	L2	0.08	4	0.35
AQWan	L3	0.06	1	Off
AQWan	L4	0.08	1	Off
AQWan	L5	0.1	1	Off
CzeLan	L1–L2	0.08	4	0.35
CzeLan	L3	0.15	4	0.35
CzeLan	L4–L5	0.08	1	Off
ZafNoo	L1	0.06	4	0.35
ZafNoo	L4	0.08	4	0.35
ZafNoo	L5	0.15	4	0.6

The tables specify detector and correction settings, while BOCD supplies run statistics computed from the current history. The chosen restoration method derives accepted intervals and translations from those statistics. The same frozen choice applies across channels, horizons, backbones, and corruption types.

A.2 Median-Return Acceptance Statistics

The acceptance-bound ablation in Appendix C removes one observed-history rejection test at a time. With $a = (u + v)/2$, pooled median m_I, and $c_I = \max(|m_I - a|, \epsilon)$, define

$$V_I = \frac{Q_{0.9}(X_I) - Q_{0.1}(X_I)}{c_I}, \qquad D_I = \frac{\max_{j \in \mathcal{S}} |z_j - \mathrm{median}_{\ell \in \mathcal{S}}\, z_\ell|}{c_I}. \tag{7}$$

Here, Q_p is an empirical quantile, $\mathcal{S}$ contains two surrounding runs per side, and z_j is a run median. Active thresholds require $V_I \le \tau_V$ and $D_I \le \tau_D$. Missing support rejects the candidate when the support bound is active. The first ratio measures the pooled interval's raw spread, while the second measures surrounding-level disagreement, both relative to the immediate outer contrast. The final translation can use a different reference estimator and reference-run count. Refined-cycle uses robustly detrended residuals for its plateau check. Removing a bound retains candidate construction, reference estimation, and all other active tests, without retuning on test results.

A.3 Translation Properties

For median-return, write $\delta_I = a_\theta(I) - m_I$. Translating every observation in I by this constant gives

$$\mathrm{median}_{t \in I}\, \widetilde{x}_t = m_I + \delta_I = a_\theta(I), \qquad \widetilde{x}_t - \widetilde{x}_{t'} = x_t - x_{t'} \quad (t, t' \in I).$$

The first identity verifies that the interval reaches the chosen target median; the second shows that internal differences are preserved, including when a local trend supplies the target. For refined-cycle, $\sum_{i=1}^{J} \bar{j}_i = \sum_{i=1}^{J} j_i - J(e/J) = 0$, so the accumulated offset returns to zero at the cycle endpoint. These identities describe the implemented translations and do not establish that the estimated corrections equal the unknown sensor errors.

A.4 Frozen Selection and Correction Records

The restoration family F belongs to $\mathcal{F} = \{\text{median-return}, \text{refined-cycle}\}$. Development parameter searches supply a configuration $\theta_{F,d,s}$ for each family, dataset d, and supplied severity s. For development windows $\mathcal{D}_{d,s}$ with observed histories X_n and clean counterparts $X_n^\star$, selection minimizes pooled input error:

$$E_{d,s}(F) = \sum_{n \in \mathcal{D}_{d,s}} \|g_{F,\theta_{F,d,s}}(X_n) - X_n^\star\|_{\mathrm{F}}^2, \qquad F_{d,s}^\star = \arg\min_{F \in \mathcal{F}} E_{d,s}(F). \tag{8}$$

Here, g is the restoration operator and the squared Frobenius norm sums errors over time and channels. Dividing by pooled raw-input error gives the same ranking whenever that denominator is positive. At clean L0, the numerator directly measures introduced error. The frozen lookup stores $F_{d,s}^\star$ and its configuration before inference, shared across channels, horizons, backbones, and corruption types. It uses neither forecasting scores nor per-window family comparisons. If the selected method finds no accepted interval, the history is unchanged. Section 4.1 explains confirmation and test replay.

Each applied translation is specified by its channel, interval, and signed magnitude. The difference $\widetilde{X} - X$ exposes the net intervention. A reader can align these records with the observed and restored histories to inspect exactly which values changed. Restoration operates on the complete available historical window, so the return criteria can use observations after a candidate boundary within that window. Future forecasting targets and clean evaluation references are used only for evaluation. Section 4.5 explains how we assess the resulting corrections.

A.5 RESTORETS PSEUDOCODE

Algorithms 1–3 summarize the implemented procedure using the notation of Section 3.3. All histories are standardized using training statistics. Algorithm 1 takes the configurations supplied by the earlier development searches (Appendix A) and selects between the two restoration methods by Eq. (8). Its restoration routine also supplies the operator g used when scoring development inputs. Clean references enter this selection step only; restoration takes an observed history and a fixed configuration.

Algorithm 1 RestoreTS: development selection and restoration with fixed settings

procedure SELECTMETHOD($\mathcal{D}_{d,s}, \{\theta_{F,d,s}\}_{F\in\mathcal{F}}$)
  for each $F \in \mathcal{F} = \{$median-return, refined-cycle$\}$ **do**
    $E(F) \leftarrow 0$
    for each $(X_n, X_n^\star) \in \mathcal{D}_{d,s}$ **do**
      $(\widetilde{X}_n, _) \leftarrow$ RESTOREHISTORY($X_n, F, \theta_{F,d,s}$)
      $E(F) \leftarrow E(F) + \|\widetilde{X}_n - X_n^\star\|_{\mathrm{F}}^2$
    end for
  end for
  $F^\star \leftarrow \arg\min_{F\in\mathcal{F}} E(F)$
  return fixed lookup entry $(F^\star, \theta_{F^\star,d,s})$
end procedure
procedure RESTOREHISTORY(X, F, θ)
  $\widetilde{X} \leftarrow X$; correction records $\mathcal{R} \leftarrow \emptyset$
  for each channel $c = 1, \ldots, C$ **do**
    $x \leftarrow X_{:,c}$
    Initialize BOCD from the observed warmup; update centered log-joint probabilities
    Compute q_t by Eq. (3); retain separated peaks above η after warmup
    Project each peak to $b = t - \widehat{r} + 1$; merge nearby estimates using the largest q_t
    Form runs $\mathcal{G}$; compute run medians and fitted signed jumps $\mathcal{B}$
    if $F =$ median-return **then**
      $\mathcal{T} \leftarrow$ MEDIANRETURN($x, \mathcal{G}, \theta$) ▷ Algorithm 2
    else
      $\mathcal{T} \leftarrow$ REFINEDCYCLE($x, \mathcal{B}, \theta$) ▷ Algorithm 3
    end if
    for each interval and signed translation $(I, \delta) \in \mathcal{T}$ **do**
      $\widetilde{X}_{I,c} \leftarrow X_{I,c} + \delta$; append (c, I, δ) to $\mathcal{R}$
    end for
  end for
  return $(\widetilde{X}, \mathcal{R})$
end procedure

For test-time forecasting, retrieve $(F^\star, \theta_{F^\star,d,s})$ once for the supplied dataset and severity, call RESTOREHISTORY, and compute $\widehat{Y} = f_\phi(\widetilde{X})$ with unchanged weights. No family comparison or parameter search occurs for a test window. Both restoration routines return disjoint intervals with signed translations; an empty list leaves the input unchanged. Run statistics and acceptance tests are computed from the original observed history, rather than from values already corrected. The record $\mathcal{R}$ and the net change $\widetilde{X} - X$ support inspection as described in Appendix A.4.

Algorithm 2 Median-return translations for one channel

Require: Observed series x, chronological runs $\mathcal{G} = (g_1, \ldots, g_N)$, fixed settings θ
Ensure: List $\mathcal{T}$ of disjoint intervals and signed translations
$\mathcal{T} \leftarrow \emptyset$; $i \leftarrow 1$
while $i \leq N - 2$ **do**
  Find the earliest $r \in \{i + 2, \ldots, \min(N, i + M + 1)\}$ whose middle interval I meets the span limit, Eq. (4), and the same-side and entry/return tests
  if no such r exists **then**
    $i \leftarrow i + 1$; **continue**
  end if
  Retain the candidate's outer runs (g_i, g_r); set $i \leftarrow r$
  Compute $m_I = \mathrm{median}_{t \in I}\, x_t$ and the fixed reference target $a_\theta(I)$
  if the required reference context is unavailable **then**
    continue
  end if
  if every active acceptance bound in θ passes **then**
    Append $(I, a_\theta(I) - m_I)$ to $\mathcal{T}$
  end if
end while
return $\mathcal{T}$

The candidate tests and reference estimators are defined in Section 3.3; Appendix A.2 defines the support and variation statistics. The scan advances to the right outer run before reference and acceptance checks, matching candidate construction in the implementation. Rejecting that candidate does not restart the search from its left run.

Algorithm 3 Refined-cycle translations for one channel

Require: Observed series x, fitted signed jumps $\mathcal{B}$, fixed settings θ
Ensure: List $\mathcal{T}$ of disjoint intervals and signed translations
$\mathcal{T} \leftarrow \emptyset$; discard jumps with invalid fits or magnitude below the threshold
$\mathcal{C} \leftarrow$ ClosedCycles$(\mathcal{B}, \theta)$ ▷ Earliest disjoint coarse cycles
for each coarse cycle $C_0 \in \mathcal{C}$ **do**
  Detect stable local median steps around C_0, within its assigned search region
  Keep the largest-magnitude steps within the suppression radius; sort by time
  Group the retained steps into local closed cycles using the local settings in θ
  if all retained steps belong to local cycles and the outer endpoints align with C_0 **then**
    $\mathcal{C}_{\mathrm{use}} \leftarrow$ local cycles
  else
    $\mathcal{C}_{\mathrm{use}} \leftarrow \{C_0\}$
  end if
  for each cycle with jumps $(j_1, \ldots, j_J)$ at $(b_1, \ldots, b_J)$ in $\mathcal{C}_{\mathrm{use}}$ **do**
    $e \leftarrow \sum_{i=1}^{J} j_i$; $\bar{j}_i \leftarrow j_i - e/J$ for $i = 1, \ldots, J$
    Form translations $([b_k, b_{k+1}), -\sum_{i=1}^{k} \bar{j}_i)$ for $k = 1, \ldots, J - 1$
    if every segment meets the minimum length and the active plateau bound **then**
      Append all translations of this cycle to $\mathcal{T}$
    end if
  end for
end for
return $\mathcal{T}$

ClosedCycles scans chronologically within the duration and jump-count limits, taking the earliest cycle whose residual is small relative to its largest jump and whose terminal jump opposes and reduces the preceding residual (Section 3.3). The next scan starts after that cycle's last jump. Local search regions are clipped at the midpoints between neighboring coarse cycles to prevent overlap. Local steps compare medians on either side and must meet the configured magnitude and within-window stability tests. Endpoint alignment uses the configured refinement radius. The plateau test uses the detrended 10–90% residual width divided by the correction magnitude. Failure on one segment rejects its whole chosen cycle; the selected settings appear in Appendix A.1.

B BACKBONE TRANSFER AND COMPLEMENTARITY

Table 7 reports the complete paired backbone comparison summarized in Figure 5(c), including ETTh1 and ETTh2. The PatchLGA comparison in that panel comes from Table 9. Table 8 preserves the complete three-method comparison, including ETTh1 and ETTh2. Table 9 additionally reports both raw/restored pairs at each severity and supports the complementarity claim in the main text.

Table 7: Transfer across five backbones. Each cell averages four horizons and combined-corruption levels L1–L5 on the original test support. + R denotes RestoreTS (ours). Bold marks the better displayed value in each adjacent pair. Δ averages the eight dataset-relative error reductions (positive is better).

Dataset	PatchTST		DLinear		TimesNet		TimeMixer		iTransformer	
	Raw	+ R	Raw	+ R	Raw	+ R	Raw	+ R	Raw	+ R
MSE										
ETTh1	0.517	**0.516**	0.522	**0.520**	0.849	**0.820**	0.558	**0.554**	0.601	**0.595**
ETTh2	0.373	**0.369**	**0.478**	0.479	0.578	**0.539**	0.422	**0.412**	0.528	**0.496**
ETTm1	0.558	**0.502**	0.536	**0.485**	0.884	**0.745**	0.656	**0.561**	0.674	**0.573**
ETTm2	0.328	**0.319**	0.346	**0.335**	0.501	**0.470**	0.362	**0.349**	0.398	**0.374**
Weather	0.334	**0.301**	0.445	**0.375**	1.671	**0.673**	0.379	**0.319**	0.389	**0.340**
AQWan	1.018	**0.959**	0.992	**0.937**	1.454	**1.291**	1.089	**1.003**	0.969	**0.914**
CzeLan	0.543	**0.437**	0.617	**0.524**	0.707	**0.473**	0.536	**0.410**	0.486	**0.376**
ZafNoo	0.684	**0.641**	0.664	**0.628**	1.337	**1.095**	0.688	**0.649**	0.793	**0.726**
Δ (%)	–	6.99	–	6.83	–	19.27	–	9.24	–	9.70
MAE										
ETTh1	0.487	**0.486**	0.496	**0.495**	0.613	**0.604**	0.502	**0.500**	0.528	**0.525**
ETTh2	0.412	**0.409**	0.466	**0.464**	0.525	**0.508**	0.442	**0.435**	0.489	**0.473**
ETTm1	0.486	**0.455**	0.480	**0.452**	0.618	**0.566**	0.525	**0.482**	0.536	**0.493**
ETTm2	0.372	**0.363**	0.384	**0.377**	0.454	**0.438**	0.388	**0.378**	0.411	**0.396**
Weather	0.347	**0.317**	0.399	**0.363**	0.491	**0.420**	0.379	**0.334**	0.365	**0.331**
AQWan	0.617	**0.579**	0.603	**0.571**	0.754	**0.696**	0.638	**0.595**	0.592	**0.561**
CzeLan	0.431	**0.363**	0.494	**0.454**	0.509	**0.419**	0.417	**0.351**	0.406	**0.357**
ZafNoo	0.547	**0.531**	0.537	**0.524**	0.692	**0.649**	0.548	**0.534**	0.590	**0.572**
Δ (%)	–	5.41	–	4.14	–	7.84	–	6.27	–	5.66

Table 8: Forecasting under combined corruption, averaged equally over $H \in \{96, 192, 336, 720\}$. RestoreTS (ours) denotes RestoreTS + PatchTST. L0 is clean. Lower is better. Bold: best displayed value. Rounded ties share rank.

		PatchTST		PatchLGA		RestoreTS (ours)	
Dataset	Level	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	L0	0.416	**0.429**	0.416	**0.429**	**0.414**	**0.429**
	L1	0.416	0.431	0.417	0.431	**0.415**	**0.430**
	L2	**0.421**	**0.436**	0.422	**0.436**	**0.421**	**0.436**
	L3	0.474	0.468	**0.471**	**0.465**	**0.471**	0.466
	L4	0.549	0.513	0.549	**0.511**	**0.547**	**0.511**
	L5	0.722	0.587	**0.717**	**0.582**	0.724	0.584
ETTh2	L0	0.341	**0.386**	**0.339**	**0.386**	0.341	**0.386**
	L1	0.341	0.388	**0.340**	0.388	**0.340**	**0.386**
	L2	0.347	0.396	**0.346**	0.397	**0.346**	**0.394**
	L3	0.361	0.407	**0.359**	0.407	**0.359**	**0.403**
	L4	0.390	0.426	**0.381**	0.423	0.385	**0.422**
	L5	0.426	0.445	**0.404**	**0.435**	0.416	0.438
ETTm1	L0	0.352	0.382	**0.350**	**0.380**	0.352	0.382
	L1	0.359	0.389	**0.358**	**0.387**	0.359	0.388
	L2	0.375	0.401	**0.372**	**0.399**	0.374	**0.399**
	L3	0.585	0.495	0.532	0.475	**0.509**	**0.455**
	L4	0.667	0.546	0.640	0.536	**0.563**	**0.488**
	L5	0.803	0.601	0.759	0.587	**0.702**	**0.543**

Continued on next page

Table 8 (continued)

Dataset	Level	PatchTST MSE	PatchTST MAE	PatchLGA MSE	PatchLGA MAE	RestoreTS (ours) MSE	RestoreTS (ours) MAE
ETTm2	L0	**0.256**	**0.317**	0.259	**0.317**	0.257	**0.317**
	L1	**0.264**	0.328	0.265	**0.326**	**0.264**	0.327
	L2	0.277	0.338	0.278	0.336	**0.274**	**0.334**
	L3	0.308	0.360	0.305	0.356	**0.300**	**0.352**
	L4	0.361	0.400	0.358	0.396	**0.347**	**0.387**
	L5	0.431	0.434	0.424	0.429	**0.408**	**0.416**
Weather	L0	0.226	0.264	0.226	0.264	**0.225**	**0.263**
	L1	0.239	0.283	0.237	0.282	**0.236**	**0.277**
	L2	0.264	0.311	0.264	0.312	**0.253**	**0.291**
	L3	0.305	0.337	0.303	0.338	**0.278**	**0.306**
	L4	0.369	0.382	0.362	0.389	**0.323**	**0.342**
	L5	0.494	0.422	0.463	0.424	**0.413**	**0.371**
AQWan	L0	**0.837**	**0.506**	0.881	0.520	**0.837**	**0.506**
	L1	0.851	0.525	0.886	0.533	**0.849**	**0.519**
	L2	0.864	0.541	0.904	0.552	**0.858**	**0.530**
	L3	0.974	0.607	1.044	0.621	**0.932**	**0.570**
	L4	1.055	0.657	1.201	0.687	**0.991**	**0.608**
	L5	1.344	0.757	1.655	0.811	**1.164**	**0.669**
CzeLan	L0	**0.227**	**0.272**	0.232	0.278	**0.227**	**0.272**
	L1	0.251	0.301	0.253	0.302	**0.245**	**0.291**
	L2	0.300	0.347	0.291	0.340	**0.263**	**0.309**
	L3	0.374	0.393	0.358	0.379	**0.298**	**0.333**
	L4	0.525	0.485	0.541	0.481	**0.392**	**0.385**
	L5	1.265	0.629	1.325	0.618	**0.989**	**0.499**
ZafNoo	L0	**0.513**	**0.456**	0.517	0.458	**0.513**	**0.456**
	L1	0.530	0.474	0.532	0.474	**0.527**	**0.473**
	L2	0.560	0.502	0.559	0.500	**0.551**	**0.493**
	L3	0.616	0.529	0.615	0.527	**0.596**	**0.516**
	L4	0.740	0.577	0.732	0.575	**0.669**	**0.554**
	L5	0.976	0.652	0.974	0.648	**0.861**	**0.621**

Table 9: Forecasting under combined corruption, averaged equally over $H \in \{96, 192, 336, 720\}$ on the full retained test support. L0 is clean. Lower is better. Bold marks the better displayed value within each raw/restored pair. Rounded ties are both bold.

Dataset	Level	PatchTST Raw MSE	PatchTST Raw MAE	PatchTST + RestoreTS (ours) MSE	PatchTST + RestoreTS (ours) MAE	PatchLGA Raw MSE	PatchLGA Raw MAE	PatchLGA + RestoreTS (ours) MSE	PatchLGA + RestoreTS (ours) MAE
ETTh1	L0	0.416	**0.429**	**0.414**	**0.429**	0.416	**0.429**	**0.415**	**0.429**
	L1	0.416	0.431	**0.415**	**0.430**	0.417	0.431	**0.416**	**0.430**
	L2	**0.421**	**0.436**	**0.421**	**0.436**	0.422	0.436	**0.421**	**0.435**
	L3	0.474	0.468	**0.471**	**0.466**	0.471	0.465	**0.468**	**0.464**
	L4	0.549	0.513	**0.547**	**0.511**	0.549	0.511	**0.547**	**0.509**
	L5	**0.722**	0.587	0.724	**0.584**	**0.717**	0.582	0.719	**0.580**
ETTh2	L0	**0.341**	**0.386**	**0.341**	**0.386**	**0.339**	**0.386**	0.340	**0.386**
	L1	0.341	0.388	**0.340**	**0.386**	**0.340**	**0.388**	0.341	**0.388**
	L2	0.347	0.396	**0.346**	**0.394**	**0.346**	0.397	0.347	**0.396**
	L3	0.361	0.407	**0.359**	**0.403**	0.359	0.407	**0.358**	**0.404**
	L4	0.390	0.426	**0.385**	**0.422**	**0.381**	0.423	**0.381**	**0.420**
	L5	0.426	0.445	**0.416**	**0.438**	**0.404**	0.435	**0.404**	**0.432**

Continued on next page

Table 9 (continued)

		PatchTST				PatchLGA			
		Raw		+ RestoreTS (ours)		Raw		+ RestoreTS (ours)	
Dataset	Level	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTm1	L0	**0.352**	**0.382**	**0.352**	**0.382**	**0.350**	**0.380**	**0.350**	**0.380**
	L1	**0.359**	0.389	**0.359**	**0.388**	**0.358**	0.387	**0.358**	**0.386**
	L2	0.375	0.401	**0.374**	**0.399**	0.372	0.399	**0.371**	**0.397**
	L3	0.585	0.495	**0.509**	**0.455**	0.532	0.475	**0.476**	**0.443**
	L4	0.667	0.546	**0.563**	**0.488**	0.640	0.536	**0.547**	**0.483**
	L5	0.803	0.601	**0.702**	**0.543**	0.759	0.587	**0.666**	**0.532**
ETTm2	L0	**0.256**	**0.317**	0.257	**0.317**	**0.259**	**0.317**	**0.259**	**0.317**
	L1	**0.264**	0.328	**0.264**	**0.327**	**0.265**	0.326	**0.265**	**0.325**
	L2	0.277	0.338	**0.274**	**0.334**	0.278	0.336	**0.275**	**0.333**
	L3	0.308	0.360	**0.300**	**0.352**	0.305	0.356	**0.297**	**0.349**
	L4	0.361	0.400	**0.347**	**0.387**	0.358	0.396	**0.341**	**0.383**
	L5	0.431	0.434	**0.408**	**0.416**	0.424	0.429	**0.397**	**0.410**
Weather	L0	0.226	0.264	**0.225**	**0.263**	**0.226**	**0.264**	**0.226**	**0.264**
	L1	0.239	0.283	**0.236**	**0.277**	0.237	0.282	**0.234**	**0.275**
	L2	0.264	0.311	**0.253**	**0.291**	0.264	0.312	**0.251**	**0.290**
	L3	0.305	0.337	**0.278**	**0.306**	0.303	0.338	**0.277**	**0.305**
	L4	0.369	0.382	**0.323**	**0.342**	0.362	0.389	**0.317**	**0.342**
	L5	0.494	0.422	**0.413**	**0.371**	0.463	0.424	**0.385**	**0.369**
AQWan	L0	**0.837**	**0.506**	**0.837**	**0.506**	**0.881**	**0.520**	**0.881**	**0.520**
	L1	0.851	0.525	**0.849**	**0.519**	0.886	0.533	**0.885**	**0.528**
	L2	0.864	0.541	**0.858**	**0.530**	0.904	0.552	**0.895**	**0.539**
	L3	0.974	0.607	**0.932**	**0.570**	1.044	0.621	**0.971**	**0.578**
	L4	1.055	0.657	**0.991**	**0.608**	1.201	0.687	**1.059**	**0.622**
	L5	1.344	0.757	**1.164**	**0.669**	1.655	0.811	**1.314**	**0.702**
CzeLan	L0	**0.227**	**0.272**	**0.227**	**0.272**	**0.232**	**0.278**	**0.232**	**0.278**
	L1	0.251	0.301	**0.245**	**0.291**	0.253	0.302	**0.247**	**0.293**
	L2	0.300	0.347	**0.263**	**0.309**	0.291	0.340	**0.263**	**0.310**
	L3	0.374	0.393	**0.298**	**0.333**	0.358	0.379	**0.294**	**0.331**
	L4	0.525	0.485	**0.392**	**0.385**	0.541	0.481	**0.396**	**0.388**
	L5	1.265	0.629	**0.989**	**0.499**	1.325	0.618	**0.921**	**0.490**
ZafNoo	L0	**0.513**	**0.456**	**0.513**	**0.456**	**0.517**	**0.458**	**0.517**	**0.458**
	L1	0.530	0.474	**0.527**	**0.473**	0.532	0.474	**0.530**	**0.473**
	L2	0.560	0.502	**0.551**	**0.493**	0.559	0.500	**0.551**	**0.492**
	L3	0.616	0.529	**0.596**	**0.516**	0.615	0.527	**0.596**	**0.514**
	L4	0.740	0.577	**0.669**	**0.554**	0.732	0.575	**0.665**	**0.551**
	L5	0.976	0.652	**0.861**	**0.621**	0.974	0.648	**0.862**	**0.614**

Table 10 retains both improvements and regressions. RestoreTS + PatchLGA tests whether restoration helps a model that already incorporates a robustness mechanism. Comparing RestoreTS + PatchTST against raw PatchLGA instead compares complete pipelines. Neither comparison is an ablation of the restoration-family selector.

Table 10: Dataset-relative reductions (%) from RestoreTS. Errors are averaged equally over $H \in \{96, 192, 336, 720\}$ and L1–L5 before computing each reduction. Positive is better.

Dataset	Backbone	Combined		Shift		Spike	
		MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	PatchTST	0.16	0.27	0.46	0.41	-0.16	-0.19
	PatchLGA	0.14	0.27	0.44	0.38	-0.21	-0.20
ETTh2	PatchTST	0.99	0.90	0.95	0.99	0.36	0.38
	PatchLGA	-0.01	0.51	0.05	0.64	-0.23	0.01
ETTm1	PatchTST	10.13	6.53	10.38	6.97	-0.25	-0.15
	PatchLGA	9.13	6.00	9.61	6.58	-0.25	-0.14

Continued on next page

Table 10 (continued)

Dataset	Backbone	Combined		Shift		Spike	
		MSE	MAE	MSE	MAE	MSE	MAE
ETTm2	PatchTST	2.92	2.28	2.94	2.52	-0.33	-0.15
	PatchLGA	3.38	2.35	3.50	2.61	-0.23	-0.11
Weather	PatchTST	10.07	8.58	9.93	9.21	-0.45	0.12
	PatchLGA	10.10	9.34	10.17	10.12	-0.43	0.21
AQWan	PatchTST	5.78	6.21	5.75	6.62	-0.60	-0.49
	PatchLGA	9.96	7.35	9.64	7.66	-0.73	-0.62
CzeLan	PatchTST	19.48	15.66	22.85	18.02	-0.14	-0.15
	PatchLGA	23.36	14.47	26.06	16.21	-0.09	-0.12
ZafNoo	PatchTST	6.38	2.85	6.33	2.99	-0.43	-0.64
	PatchLGA	6.10	2.97	5.37	2.88	-0.42	-0.60

C Clean Preservation and Input-Restoration Selectivity

Table 11 separates input changes from forecast changes. Clean-input MSE and MAE measure introduced distortion, whereas the last two columns report relative changes in forecasting error. For CzeLan and ZafNoo, the evaluated stored float32 histories contain exactly zero edited values, so their input edit rates, MSE and MAE are exactly zero rather than rounded small errors or missing measurements. Restoration was applied to L0 histories without an automatic clean-input bypass. These observations establish preservation only on the evaluated support, not on all possible clean inputs. The forecast-change entries of 0.000% instead represent tiny nonzero recorded differences rounded to three decimal places, as detailed below the table. They do not mean that forecasting errors themselves are zero. Across all five backbones and eight datasets, clean forecast MSE changes range from a 0.62% reduction to a 1.28% increase.

Table 11: Clean-input preservation. Input metrics use the complete existing $H = 96$ cache-origin support. Forecast changes use PatchTST over four horizons on the broad support. Positive forecast change means worse. Input MSE/MAE measure edits relative to the clean reference. Forecast changes are rounded to three decimal places, with no sign on values that round to zero.

Dataset	Edited (%)	Input MSE	Input MAE	Δ MSE (%)	Δ MAE (%)
ETTh1	0.07032	0.0061122	0.0018776	-0.270	-0.080
ETTh2	0.34332	0.0061451	0.0044949	+0.070	+0.005
ETTm1	0.00118	0.0000774	0.0000293	0.000	+0.002
ETTm2	0.19065	0.0025979	0.0021508	+0.197	+0.091
Weather	0.04375	0.0023841	0.0006945	-0.076	-0.067
AQWan	0.00235	0.0002209	0.0000566	-0.002	-0.002
CzeLan	0.00000	0.0000000	0.0000000	0.000	0.000
ZafNoo	0.00000	0.0000000	0.0000000	0.000	0.000

Additional precision for CzeLan and ZafNoo. The logged relative forecast changes are shown below in percent. Their magnitudes are below $5\times10^{-7}\%$ and do not indicate a meaningful improvement or degradation. Raw and restored forecasts were evaluated in separate forward passes. The specific numerical source of these residual differences has not been isolated.

Dataset	Δ MSE (%)	Δ MAE (%)
CzeLan	$+2.979604 \times 10^{-7}$	$+4.194667 \times 10^{-7}$
ZafNoo	-2.598108×10^{-7}	-6.378104×10^{-8}

Table 12 uses the complete existing cache-origin support for $H = 96$ inputs, pooling encoder-point exposures across L1–L5. Windows overlap, so exposures are not independent observations. Perturbed support is the observable difference between normalized float32 raw and clean values. Original injection-region annotations were unavailable. The support must not be interpreted as verified sensor-failure intervals. The improved-edit fraction checks whether the absolute error to the reference decreases. This is distinct from detecting a change point, recording an edit, or proving a physical measurement fault.

Table 12: Input restoration on combined L1–L5. Counts and squared errors are summed over encoder-point exposures, including overlapping windows. P is observed perturbation support ($x^{\mathrm{raw}} \neq x^{\mathrm{clean}}$), U its complement. Improved/worsened edits compare absolute error to the reference. All entries are percentages. These are not injection-mask precision/recall.

Dataset	Edited	Edited/P	Edited/U	Improved	Worsened	MSE reduction
ETTh1	0.45	2.75	0.13	71.90	28.10	9.60
ETTh2	2.35	10.32	1.46	39.32	60.68	14.92
ETTm1	1.56	12.22	0.23	84.51	15.49	36.43
ETTm2	1.90	12.93	0.67	66.59	33.41	27.04
Weather	2.85	21.48	0.89	69.07	30.93	50.95
AQWan	2.86	14.30	1.50	46.73	53.27	44.70
CzeLan	2.01	17.52	0.34	82.86	17.14	48.48
ZafNoo	1.42	8.15	0.76	46.78	53.22	31.31

The acceptance-bound comparison in Table 13 tests median-return alone, rather than the selected two-family pipeline. It uses seven datasets throughout at $H = 336$ on 128 fixed test origins per dataset. Appendix A.2 defines the plateau-variation and support-dispersion ratios and their rejection thresholds. ETTh2 is omitted because its frozen median-return configuration has no active acceptance thresholds. Each ablation removes one finite acceptance bound, retaining candidate generation, references, and other bounds. Removing the plateau bound or support-dispersion bound increases clean distortion even though the corrupted-input forecasting aggregate improves. No threshold was changed after observing these results.

Table 13: Acceptance-gate trade-off on the same seven datasets. ETTh2 has no active median-return acceptance thresholds and is excluded. Forecast columns are combined L1–L5 reductions (%). Clean damage is mean dataset input MSE at L0. All use the supplementary support.

Procedure	MSE reduction	MAE reduction	Clean input MSE
Median-return	3.184	2.442	0.001395
Without plateau bound	3.914	2.947	0.021989
Without support bound	6.729	4.737	0.007257

D Repetition over Training and Corruption Seeds

Both seed studies cover all eight datasets and horizons $H \in \{96, 192, 336, 720\}$, using the full retained stride-one test origins in complete batches of 128. Raw and restored inputs share the checkpoint, origins, and targets. Earlier scores on 128 sampled origins are replaced by these full-support evaluations. Restoration configurations remain fixed throughout.

The training-seed study uses seeds 2021, 2026, and 2027 with corruption seed 2025. PatchTST and PatchLGA use identical clean train/stop supports, Adam with learning rate 10^{-4}, OneCycleLR with `pct_start`=0.3, batch size 128, a maximum of 100 epochs, and patience 10. Each dataset retains its documented architecture. Of the 192 required checkpoints, 64 matched seed-2021 checkpoints were reused and 128 checkpoints were trained with seeds 2026 and 2027. These matched training checkpoints, including seed 2021, are distinct from the historical checkpoints used in the main comparisons.

The corruption-seed study uses seeds 2025, 2026, and 2027 with 64 fixed historical model-seed-2021 checkpoints. New perturbations use the original corruption generator. These are two separate three-seed studies, not a 3×3 factorial experiment. All 2,048 historical-checkpoint rows at corruption seed 2025 reproduce the corresponding main-suite scores exactly. The 2,048 reused matched-training-seed-2021 rows reproduce the previous full-support evaluation. Across both axes, the collection contains 12,288 test rows spanning clean L0 and combined, shift-only, and spike-only corruption at L1–L5.

For each dataset, backbone, corruption type, and seed, we first average raw and restored errors separately over the four horizons and L1–L5, then compute the percentage reduction relative to the raw error. Tables 14 and 15 report the mean and sample standard deviation of these reductions across three seeds. Tables 16–19 give the individual values. Standard deviations describe seed variation. Neither horizons nor overlapping windows are treated as replicates. Aggregate seed-study percentages average the dataset-level reductions equally across datasets and seeds.

Under combined corruption, mean MSE reductions before PatchTST and PatchLGA are 7.65% and 8.58% across training seeds, and 7.69% and 8.38% across corruption seeds. For combined corrup-

tion, training-seed variation gives positive MSE and MAE reductions in all 48 dataset–backbone–seed aggregates. Corruption-seed variation gives 46 positive MSE reductions and 48 positive MAE reductions. The MSE exceptions are ETTh2 with PatchLGA at corruption seeds 2025 and 2026 (increases of 0.0086% and 0.1847%, respectively). Across the four axis/backbone combinations, spike-only mean MSE increases by 0.24–0.35%. The largest four-horizon clean-input MSE deterioration is 0.34% for ETTm2/PatchTST with matched training seed 2027. The largest MAE deterioration is 0.15% for ETTh2/PatchLGA with the historical checkpoint in the corruption-seed study. L0 is the same clean input across corruption seeds. These clean-input figures are maxima over dataset–backbone–seed aggregates, not individual horizons. Appendix F discusses the corresponding future directions.

Table 14: Mean ± sample standard deviation of RestoreTS relative error reductions (%) across three training seeds, on all eight datasets and four horizons. Errors are averaged over horizons and L1–L5 within each seed before computing the paired reduction. Positive is better. Negative values indicate deterioration.

Dataset	Backbone	Combined		Shift		Spike	
		MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	PatchTST	0.39 ± 0.15	0.34 ± 0.04	0.57 ± 0.24	0.45 ± 0.07	-0.08 ± 0.03	-0.16 ± 0.02
ETTh1	PatchLGA	0.94 ± 0.13	0.59 ± 0.04	1.26 ± 0.21	0.76 ± 0.04	-0.05 ± 0.02	-0.11 ± 0.01
ETTh2	PatchTST	1.30 ± 0.28	1.14 ± 0.12	1.43 ± 0.29	1.31 ± 0.12	0.28 ± 0.10	0.37 ± 0.06
ETTh2	PatchLGA	0.74 ± 0.07	0.83 ± 0.04	0.87 ± 0.06	0.99 ± 0.03	0.09 ± 0.07	0.20 ± 0.03
ETTm1	PatchTST	11.00 ± 0.58	6.89 ± 0.21	11.02 ± 0.56	7.25 ± 0.23	-0.25 ± 0.01	-0.15 ± 0.01
ETTm1	PatchLGA	11.03 ± 0.60	6.68 ± 0.16	11.50 ± 0.70	7.26 ± 0.18	-0.27 ± 0.04	-0.15 ± 0.01
ETTm2	PatchTST	3.51 ± 0.52	2.73 ± 0.27	3.31 ± 0.18	2.93 ± 0.21	-0.36 ± 0.03	-0.15 ± 0.02
ETTm2	PatchLGA	3.22 ± 0.17	2.40 ± 0.06	3.44 ± 0.13	2.67 ± 0.07	-0.08 ± 0.06	-0.03 ± 0.02
Weather	PatchTST	13.76 ± 2.15	11.25 ± 0.94	13.96 ± 2.69	12.14 ± 1.16	-0.44 ± 0.08	0.13 ± 0.11
Weather	PatchLGA	10.98 ± 1.01	10.15 ± 1.37	10.66 ± 0.77	10.85 ± 1.37	-0.52 ± 0.03	0.14 ± 0.04
AQWan	PatchTST	5.84 ± 0.42	6.29 ± 0.23	5.93 ± 0.50	6.80 ± 0.27	-0.55 ± 0.10	-0.44 ± 0.06
AQWan	PatchLGA	10.05 ± 2.64	7.74 ± 0.74	10.01 ± 2.96	8.23 ± 0.90	-0.62 ± 0.22	-0.51 ± 0.17
CzeLan	PatchTST	19.39 ± 1.47	14.73 ± 1.01	21.88 ± 2.04	16.78 ± 1.11	-0.14 ± 0.00	-0.18 ± 0.01
CzeLan	PatchLGA	25.53 ± 0.69	15.67 ± 0.19	26.99 ± 0.57	17.13 ± 0.13	-0.09 ± 0.01	-0.12 ± 0.01
ZafNoo	PatchTST	6.03 ± 0.72	2.71 ± 0.26	5.74 ± 0.90	2.75 ± 0.35	-0.45 ± 0.07	-0.68 ± 0.08
ZafNoo	PatchLGA	6.13 ± 0.90	2.91 ± 0.17	5.72 ± 0.76	2.87 ± 0.12	-0.37 ± 0.02	-0.55 ± 0.02

Table 15: Mean ± sample standard deviation of RestoreTS relative error reductions (%) across three corruption seeds, on all eight datasets and four horizons. Errors are averaged over horizons and L1–L5 within each seed before computing the paired reduction. Positive is better. Negative values indicate deterioration.

Dataset	Backbone	Combined		Shift		Spike	
		MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	PatchTST	0.30 ± 0.15	0.40 ± 0.14	0.48 ± 0.48	0.46 ± 0.08	-0.17 ± 0.01	-0.20 ± 0.01
ETTh1	PatchLGA	0.38 ± 0.23	0.43 ± 0.15	0.55 ± 0.56	0.47 ± 0.13	-0.23 ± 0.02	-0.22 ± 0.01
ETTh2	PatchTST	0.88 ± 0.32	1.16 ± 0.28	0.75 ± 0.20	1.19 ± 0.18	0.39 ± 0.05	0.45 ± 0.07
ETTh2	PatchLGA	-0.01 ± 0.17	0.60 ± 0.15	0.05 ± 0.27	0.74 ± 0.18	-0.27 ± 0.12	0.06 ± 0.05
ETTm1	PatchTST	9.57 ± 1.74	6.23 ± 0.63	9.81 ± 1.40	6.62 ± 0.54	-0.30 ± 0.05	-0.18 ± 0.03
ETTm1	PatchLGA	8.94 ± 1.56	5.82 ± 0.57	9.43 ± 1.14	6.31 ± 0.47	-0.30 ± 0.05	-0.16 ± 0.03
ETTm2	PatchTST	2.52 ± 0.54	2.36 ± 0.42	2.52 ± 0.53	2.64 ± 0.50	-0.29 ± 0.04	-0.14 ± 0.01
ETTm2	PatchLGA	2.98 ± 0.49	2.48 ± 0.38	3.06 ± 0.56	2.78 ± 0.48	-0.21 ± 0.02	-0.11 ± 0.00
Weather	PatchTST	10.33 ± 0.60	8.74 ± 0.17	9.78 ± 0.65	9.33 ± 0.19	-0.40 ± 0.05	0.13 ± 0.03
Weather	PatchLGA	9.88 ± 0.26	9.32 ± 0.08	9.77 ± 0.41	10.13 ± 0.13	-0.35 ± 0.08	0.25 ± 0.03
AQWan	PatchTST	5.03 ± 1.32	5.50 ± 0.83	5.60 ± 0.65	6.04 ± 0.69	-0.74 ± 0.14	-0.59 ± 0.10
AQWan	PatchLGA	8.33 ± 2.53	6.54 ± 0.97	9.34 ± 0.66	7.04 ± 0.65	-0.87 ± 0.20	-0.69 ± 0.11
CzeLan	PatchTST	20.29 ± 2.70	14.91 ± 1.09	21.62 ± 3.67	16.44 ± 1.70	-0.18 ± 0.09	-0.18 ± 0.06
CzeLan	PatchLGA	22.77 ± 0.76	13.24 ± 1.06	23.98 ± 1.99	14.38 ± 1.59	-0.14 ± 0.08	-0.14 ± 0.04
ZafNoo	PatchTST	12.57 ± 5.50	4.79 ± 1.69	12.58 ± 5.60	5.12 ± 1.84	-0.42 ± 0.03	-0.64 ± 0.01
ZafNoo	PatchLGA	13.80 ± 7.33	5.12 ± 1.87	12.98 ± 7.22	5.16 ± 1.98	-0.40 ± 0.04	-0.60 ± 0.01

Table 16: Individual training-seed relative error reductions (%) across four horizons and L1–L5 on full retained test support (ETT datasets). Positive is better. Negative values indicate deterioration.

Dataset	Backbone	Seed	Combined		Shift		Spike	
			MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	PatchTST	2021	0.55	0.39	0.80	0.52	-0.05	-0.14
ETTh1	PatchTST	2026	0.37	0.33	0.58	0.44	-0.09	-0.16
ETTh1	PatchTST	2027	0.24	0.30	0.33	0.37	-0.11	-0.17
ETTh1	PatchLGA	2021	1.00	0.57	1.43	0.78	-0.04	-0.12
ETTh1	PatchLGA	2026	1.04	0.64	1.34	0.79	-0.07	-0.12
ETTh1	PatchLGA	2027	0.79	0.57	1.02	0.72	-0.04	-0.10
ETTh2	PatchTST	2021	1.03	1.05	1.15	1.22	0.16	0.31
ETTh2	PatchTST	2026	1.59	1.28	1.72	1.45	0.36	0.44
ETTh2	PatchTST	2027	1.28	1.09	1.43	1.25	0.32	0.35
ETTh2	PatchLGA	2021	0.78	0.85	0.93	1.01	0.10	0.21
ETTh2	PatchLGA	2026	0.66	0.79	0.81	0.95	0.01	0.17
ETTh2	PatchLGA	2027	0.78	0.86	0.87	1.00	0.15	0.23
ETTm1	PatchTST	2021	11.66	7.09	11.66	7.46	-0.24	-0.15
ETTm1	PatchTST	2026	10.61	6.67	10.61	7.00	-0.24	-0.15
ETTm1	PatchTST	2027	10.72	6.91	10.79	7.31	-0.26	-0.16
ETTm1	PatchLGA	2021	10.36	6.51	10.72	7.06	-0.31	-0.16
ETTm1	PatchLGA	2026	11.49	6.70	12.06	7.31	-0.27	-0.15
ETTm1	PatchLGA	2027	11.25	6.82	11.72	7.41	-0.24	-0.13
ETTm2	PatchTST	2021	3.04	2.49	3.12	2.75	-0.37	-0.15
ETTm2	PatchTST	2026	4.06	3.01	3.48	3.16	-0.32	-0.13
ETTm2	PatchTST	2027	3.42	2.68	3.34	2.88	-0.39	-0.16
ETTm2	PatchLGA	2021	3.26	2.46	3.48	2.74	-0.02	-0.01
ETTm2	PatchLGA	2026	3.03	2.33	3.29	2.60	-0.13	-0.05
ETTm2	PatchLGA	2027	3.38	2.41	3.55	2.66	-0.10	-0.04

Table 17: Individual training-seed relative error reductions (%) across four horizons and L1–L5 on full retained test support (Weather and environmental datasets). Positive is better. Negative values indicate deterioration.

Dataset	Backbone	Seed	Combined		Shift		Spike	
			MSE	MAE	MSE	MAE	MSE	MAE
Weather	PatchTST	2021	12.11	10.34	11.91	11.02	-0.39	0.21
Weather	PatchTST	2026	12.99	11.19	12.97	12.07	-0.40	0.18
Weather	PatchTST	2027	16.19	12.22	17.01	13.33	-0.54	0.00
Weather	PatchLGA	2021	10.71	9.68	10.50	10.42	-0.50	0.15
Weather	PatchLGA	2026	10.13	9.07	9.98	9.74	-0.49	0.10
Weather	PatchLGA	2027	12.11	11.69	11.50	12.37	-0.55	0.17
AQWan	PatchTST	2021	5.64	6.17	5.70	6.68	-0.52	-0.44
AQWan	PatchTST	2026	6.33	6.55	6.50	7.12	-0.66	-0.50
AQWan	PatchTST	2027	5.57	6.14	5.58	6.61	-0.47	-0.37
AQWan	PatchLGA	2021	9.21	7.17	8.78	7.46	-0.74	-0.63
AQWan	PatchLGA	2026	13.00	8.58	13.39	9.21	-0.75	-0.59
AQWan	PatchLGA	2027	7.93	7.47	7.88	8.01	-0.37	-0.32
CzeLan	PatchTST	2021	21.09	15.59	24.01	17.73	-0.14	-0.17
CzeLan	PatchTST	2026	18.58	13.62	21.69	15.56	-0.15	-0.19
CzeLan	PatchTST	2027	18.49	14.98	19.93	17.05	-0.14	-0.17
CzeLan	PatchLGA	2021	25.43	15.61	27.28	17.19	-0.07	-0.11
CzeLan	PatchLGA	2026	24.89	15.52	26.33	16.99	-0.09	-0.13
CzeLan	PatchLGA	2027	26.26	15.88	27.36	17.22	-0.10	-0.12
ZafNoo	PatchTST	2021	6.74	2.99	6.70	3.14	-0.41	-0.61
ZafNoo	PatchTST	2026	5.30	2.47	4.91	2.47	-0.41	-0.67
ZafNoo	PatchTST	2027	6.06	2.68	5.62	2.64	-0.53	-0.76
ZafNoo	PatchLGA	2021	5.67	2.82	5.29	2.80	-0.38	-0.55
ZafNoo	PatchLGA	2026	5.55	2.80	5.29	2.80	-0.38	-0.57
ZafNoo	PatchLGA	2027	7.17	3.11	6.60	3.00	-0.34	-0.53

Table 18: Individual corruption-seed relative error reductions (%) across four horizons and L1–L5 on full retained test support (ETT datasets). Positive is better. Negative values indicate deterioration.

Dataset	Backbone	Seed	Combined		Shift		Spike	
			MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	PatchTST	2025	0.16	0.27	0.46	0.41	-0.16	-0.19
ETTh1	PatchTST	2026	0.45	0.38	0.97	0.55	-0.19	-0.20
ETTh1	PatchTST	2027	0.30	0.56	0.01	0.42	-0.17	-0.21
ETTh1	PatchLGA	2025	0.14	0.27	0.44	0.38	-0.21	-0.20
ETTh1	PatchLGA	2026	0.60	0.45	1.15	0.62	-0.26	-0.22
ETTh1	PatchLGA	2027	0.38	0.57	0.04	0.42	-0.23	-0.23
ETTh2	PatchTST	2025	0.99	0.90	0.95	0.99	0.36	0.38
ETTh2	PatchTST	2026	1.13	1.45	0.57	1.28	0.44	0.51
ETTh2	PatchTST	2027	0.52	1.14	0.72	1.31	0.37	0.47
ETTh2	PatchLGA	2025	-0.01	0.51	0.05	0.64	-0.23	0.01
ETTh2	PatchLGA	2026	-0.18	0.51	-0.23	0.63	-0.17	0.12
ETTh2	PatchLGA	2027	0.16	0.77	0.32	0.96	-0.41	0.04
ETTm1	PatchTST	2025	10.13	6.53	10.38	6.97	-0.25	-0.15
ETTm1	PatchTST	2026	10.96	6.65	10.83	6.90	-0.35	-0.18
ETTm1	PatchTST	2027	7.62	5.51	8.21	6.00	-0.30	-0.22
ETTm1	PatchLGA	2025	9.13	6.00	9.61	6.58	-0.25	-0.14
ETTm1	PatchLGA	2026	10.40	6.27	10.46	6.58	-0.33	-0.16
ETTm1	PatchLGA	2027	7.30	5.18	8.21	5.76	-0.32	-0.20
ETTm2	PatchTST	2025	2.92	2.28	2.94	2.52	-0.33	-0.15
ETTm2	PatchTST	2026	2.74	2.81	2.69	3.19	-0.30	-0.14
ETTm2	PatchTST	2027	1.91	1.98	1.92	2.20	-0.25	-0.14
ETTm2	PatchLGA	2025	3.38	2.35	3.50	2.61	-0.23	-0.11
ETTm2	PatchLGA	2026	3.12	2.90	3.25	3.32	-0.21	-0.10
ETTm2	PatchLGA	2027	2.44	2.18	2.43	2.41	-0.19	-0.11

Table 19: Individual corruption-seed relative error reductions (%) across four horizons and L1–L5 on full retained test support (Weather and environmental datasets). Positive is better. Negative values indicate deterioration.

Dataset	Backbone	Seed	Combined		Shift		Spike	
			MSE	MAE	MSE	MAE	MSE	MAE
Weather	PatchTST	2025	10.07	8.58	9.93	9.21	-0.45	0.12
Weather	PatchTST	2026	11.02	8.92	10.34	9.55	-0.42	0.12
Weather	PatchTST	2027	9.90	8.73	9.06	9.23	-0.34	0.16
Weather	PatchLGA	2025	10.10	9.34	10.17	10.12	-0.43	0.21
Weather	PatchLGA	2026	9.94	9.37	9.79	10.26	-0.32	0.26
Weather	PatchLGA	2027	9.59	9.23	9.35	10.00	-0.28	0.27
AQWan	PatchTST	2025	5.78	6.21	5.75	6.62	-0.60	-0.49
AQWan	PatchTST	2026	3.51	4.58	4.89	5.28	-0.72	-0.57
AQWan	PatchTST	2027	5.81	5.71	6.17	6.21	-0.89	-0.69
AQWan	PatchLGA	2025	9.96	7.35	9.64	7.66	-0.73	-0.62
AQWan	PatchLGA	2026	5.41	5.47	8.58	6.36	-0.76	-0.64
AQWan	PatchLGA	2027	9.61	6.79	9.80	7.10	-1.10	-0.81
CzeLan	PatchTST	2025	19.48	15.66	22.85	18.02	-0.14	-0.15
CzeLan	PatchTST	2026	18.09	13.66	17.50	14.63	-0.13	-0.14
CzeLan	PatchTST	2027	23.30	15.40	24.52	16.67	-0.29	-0.24
CzeLan	PatchLGA	2025	23.36	14.47	26.06	16.21	-0.09	-0.12
CzeLan	PatchLGA	2026	21.92	12.63	22.09	13.42	-0.10	-0.12
CzeLan	PatchLGA	2027	23.02	12.63	23.78	13.50	-0.22	-0.19
ZafNoo	PatchTST	2025	6.38	2.85	6.33	2.99	-0.43	-0.64
ZafNoo	PatchTST	2026	14.42	5.94	14.28	6.30	-0.38	-0.63
ZafNoo	PatchTST	2027	16.90	5.57	17.14	6.05	-0.44	-0.65
ZafNoo	PatchLGA	2025	6.10	2.97	5.37	2.88	-0.42	-0.60
ZafNoo	PatchLGA	2026	14.61	6.10	13.84	6.17	-0.36	-0.59
ZafNoo	PatchLGA	2027	20.68	6.30	19.73	6.44	-0.43	-0.60

E Component Ablations and Restoration Comparisons

E.1 BOCD and Local-Refinement Ablations

These experiments isolate the partitioning and local-refinement steps while retaining the selected restoration family and its downstream settings. They use all eight datasets, the original combined-corruption seed 2025, and the saved seed-2021 PatchTST checkpoints at $H \in \{96, 192, 336, 720\}$. The corrupted-input forecast comparison covers L1–L5 and all retained test origins. L0 is evaluated separately for clean-input preservation. No forecasting model is retrained. The benchmark histories had previously been inspected, and these single-seed comparisons do not provide additional estimates of variation across seeds.

Uniform intervals replace only BOCD, starting at encoder index zero. The final shorter interval remains subject to the minimum-fit rule. Observations, run medians, Theil–Sen fits, signed jumps, correction methods, and acceptance criteria are retained. For each dataset/severity, $w \in \{8, 16, 32, 64\}$ minimizes reconstruction MSE on 192 equally spaced encoders wholly between stopping and confirmation, with ties favouring smaller widths. Widths are frozen before confirmation/test (Table 21). The unchanged family lookup and thresholds make this a partition replacement, not a fully retuned system or detector comparison.

The refinement ablation passes the original coarse cycles through the same minimum-length and plateau checks, bypassing local median-step refinement. Checks act on each arm's actual segments, so acceptance can differ. Median-return cells are unchanged. The earlier coarse closed-cycle procedure also removes acceptance checks (Appendix E.2) and does not isolate refinement.

For each dataset d, let E_d denote MSE or MAE averaged equally over the four horizons and five corrupted severities. Table 2 reports the mean of $100(1 - E_d^{\text{variant}}/E_d^{\text{raw}})$. Table 20 instead reports $100(1 - E_d^{\text{original}}/E_d^{\text{ablated}})$ for each dataset, followed by the equal mean over datasets. The denominators differ, so paired advantages should not be read as differences between the reduction percentages in Table 2.

BOCD improves both metrics on all eight dataset averages, by a mean 5.27% MSE and 3.44% MAE relative to uniform intervals. Refinement contributes 0.12% and 0.16%: six datasets improve, ETTh2 is unchanged (all L1–L5 choices are median-return), and CzeLan worsens by 1.51% MSE and 0.04% MAE relative to removal. Separate all-refined-cycle diagnostics give 0.17% and 0.21%. Restricting to the existing lookup's refined-cycle cells gives 0.20% and 0.21% across seven datasets. That subset is fixed by the lookup, not test gains. These comparisons are descriptive, without significance tests.

Table 20: Paired component effects in the selected pipeline. Positive values favour the original RestoreTS pipeline. The denominator is the corresponding ablated alternative, unlike Table 2, which uses raw inputs. Errors first average four horizons and combined L1–L5 within each dataset.

	BOCD vs. uniform intervals		Refinement on vs. off	
Dataset	MSE (%)	MAE (%)	MSE (%)	MAE (%)
ETTh1	0.25	0.22	0.03	0.02
ETTh2	0.80	0.55	0.00	0.00
ETTm1	6.66	4.16	0.30	0.31
ETTm2	2.06	1.45	0.15	0.16
Weather	6.24	4.62	1.13	0.45
AQWan	5.09	4.30	0.25	0.25
CzeLan	16.39	9.54	-1.51	-0.04
ZafNoo	4.66	2.70	0.58	0.16
Mean	5.27	3.44	0.12	0.16

Table 21: Uniform interval widths selected using development reconstruction MSE. Widths are measured in time steps. L0 is clean input. No confirmation or test score selects the width.

Dataset	L0	L1	L2	L3	L4	L5
ETTh1	16	16	16	16	8	8
ETTh2	8	8	8	8	8	8
ETTm1	8	8	8	8	8	16
ETTm2	16	8	16	8	8	16
Weather	64	8	8	16	8	16
AQWan	8	8	8	8	16	16
CzeLan	8	8	8	8	16	8
ZafNoo	16	8	8	8	8	16

On L0, removing refinement changes only ETTh2 inputs, reducing MSE from 0.006145 to 0.005844. Other selected cells use median-return. Uniform intervals preserve Weather clean inputs exactly, versus input MSE 0.002384 for BOCD (Table 22), showing that stronger corrupted-input forecasts do not imply uniformly better clean preservation. Diagnostics use complete retained $H = 96$ encoder support, with overlapping windows not treated as independent replicates. Summed worker times for test L1–L5 are approximately 103,789 seconds (uniform) and 208,588 seconds (BOCD), measured under concurrent execution rather than as wall-clock latency or dedicated throughput.

Table 22: Clean-input distortion and edited-value fractions for the selected pipeline and its component ablations. Input MSE uses training-standardized values and the complete retained $H = 96$ encoder support. Edited fractions are percentages of input values. Rounded zeros can include smaller nonzero values. These are input metrics, not forecast errors.

	RestoreTS		Uniform intervals		No local refinement	
Dataset	Input MSE	Edited (%)	Input MSE	Edited (%)	Input MSE	Edited (%)
ETTh1	0.006112	0.070	0.000054	0.004	0.006112	0.070
ETTh2	0.006145	0.343	0.005050	0.295	0.005844	0.312
ETTm1	0.000077	0.001	0.000059	0.002	0.000077	0.001
ETTm2	0.002598	0.191	0.000567	0.043	0.002598	0.191
Weather	0.002384	0.044	0.000000	0.000	0.002384	0.044
AQWan	0.000221	0.002	0.000298	0.006	0.000221	0.002
CzeLan	0.000000	0.000	0.000000	0.000	0.000000	0.000
ZafNoo	0.000000	0.000	0.000100	0.030	0.000000	0.000

E.2 Comparisons with Other Restoration Procedures

Median filtering (Gallagher & Wise, 1981) replaces each value by its local window median. Hampel filtering (Pearson, 2002) uses the local median and median absolute deviation to replace only values exceeding its threshold. The searches for both filters are specified in Appendix A. Paired-jump correction restricts candidate returns to an oppositely signed boundary pair. Coarse closed-cycle correction groups accumulated signed jumps into return cycles using the refined family's frozen detector and coarse-cycle settings. It omits local refinement and the refined branch's minimum-length and plateau checks, so this comparison does not isolate refinement alone. All-median-return and all-refined-cycle apply their named family at every dataset/severity, using that family's archived settings. RestoreTS selects between those two families using development input error. The fixed-family variants ablate family selection. The other comparisons replace complete correction procedures and retain their frozen settings. The acceptance-bound removals in Appendix C isolate individual gates.

Table 23 reports the full procedure comparison for combined, shift-only, and spike-only corruption. Its combined-corruption family-selection rows are retained in main-text Table 2. Negative reductions are retained. The subsequent tables report the combined-corruption comparison at each severity.

Table 23: Restoration procedures before PatchTST on full retained test support. Errors are averaged over four horizons $(96, 192, 336, 720)$ and L1–L5 within each dataset. Entries average the eight dataset-relative reductions (%). Negative values indicate deterioration. Family-specific settings remain frozen across horizons and corruption types.

	Combined		Shift only		Spike only	
Correction	MSE	MAE	MSE	MAE	MSE	MAE
Median filter	-2.88	0.71	-2.88	0.16	-1.38	0.97
Hampel filter	-2.20	1.43	-2.43	0.82	-1.76	1.17
Paired jumps	1.42	2.01	1.45	2.34	-8.03	-5.06
Coarse closed-cycle	5.58	4.08	5.81	4.48	-2.25	-1.95
All median-return	2.92	2.43	2.42	2.21	0.06	0.05
All refined-cycle	6.66	5.22	7.12	5.80	-0.51	-0.39
RestoreTS (ours)	6.99	5.41	7.45	5.97	-0.25	-0.16

Tables 24 and 25 give the corresponding reductions at each severity, using the same fixed settings as the aggregate comparison.

Table 24: Combined-corruption MSE reductions (%) before PatchTST. Each row averages errors over $H \in \{96, 192, 336, 720\}$ at the specified severity before computing relative reductions. Positive is better.

Dataset	Level	Selected	Refined-cycle	Median-return	Paired	Closed	Median	Hampel
ETTh1	1	0.27	-1.81	0.27	-22.27	-5.04	-0.22	-0.07
	2	0.18	-1.73	0.18	-24.06	-5.17	-0.12	0.09
	3	0.59	-1.78	0.59	-19.31	-5.33	0.16	-3.79
	4	0.29	-1.12	0.29	-18.75	-3.36	1.22	0.42
	5	-0.29	-0.29	0.61	-12.99	-3.08	1.08	-1.08
ETTh2	1	0.36	0.01	0.36	-1.11	-0.61	-0.11	-1.02
	2	0.27	0.13	0.27	-0.29	-0.58	-0.43	-0.99
	3	0.52	-0.28	0.52	-2.88	-0.65	-8.08	-0.99
	4	1.27	-0.15	1.27	-2.45	-0.76	-0.19	-0.02
	5	2.23	-0.69	2.23	-1.57	-1.45	-4.08	0.52
ETTm1	1	0.13	0.13	0.08	-6.15	-2.64	-0.06	0.24
	2	0.24	0.24	0.02	-5.75	-2.26	0.42	1.11
	3	13.04	13.04	3.61	12.62	12.72	1.52	2.57
	4	15.56	15.56	2.43	14.11	15.34	0.39	2.45
	5	12.60	12.60	1.88	13.82	13.17	-3.01	-0.90
ETTm2	1	0.08	0.18	0.08	-1.46	-0.48	0.29	-0.43
	2	1.01	1.01	0.38	-0.63	0.31	0.43	0.04
	3	2.62	2.62	0.61	1.13	2.21	-0.19	0.69
	4	3.99	3.99	1.39	3.13	3.41	-0.47	-0.12
	5	5.19	5.19	2.09	5.34	5.54	-1.02	-0.52
Weather	1	1.18	1.18	1.03	0.58	1.14	0.02	0.73
	2	4.48	4.48	2.07	3.70	3.71	0.07	0.98
	3	8.76	8.76	3.48	7.85	7.08	0.26	2.13
	4	12.44	12.44	5.98	11.48	11.66	2.67	4.02
	5	16.40	16.40	5.78	15.46	15.13	2.83	4.07
AQWan	1	0.30	0.30	0.52	-3.33	-0.49	-0.42	-0.30
	2	0.66	0.66	0.46	-2.59	-0.04	-1.06	-0.26
	3	4.34	4.34	2.98	3.88	4.07	-5.00	0.15
	4	6.07	6.07	2.91	5.29	5.73	-3.86	-0.03
	5	13.36	13.36	9.98	5.38	13.02	-1.13	2.28
CzeLan	1	2.58	2.58	1.58	1.73	2.43	0.25	0.81
	2	12.42	12.42	3.84	11.98	11.94	-1.00	-1.14
	3	20.41	20.41	6.54	20.57	20.34	-0.76	-0.40
	4	25.22	25.22	9.24	28.75	25.73	-14.66	-13.19
	5	21.85	21.85	8.21	30.94	24.39	-33.37	-32.63
ZafNoo	1	0.59	0.59	0.27	-16.37	-1.92	0.64	1.01
	2	1.59	2.45	1.59	-42.99	-6.77	1.64	-18.44
	3	3.31	2.12	3.31	-24.60	-3.94	2.30	1.89
	4	9.55	9.55	4.71	-2.98	7.79	4.87	-3.96
	5	11.80	11.80	6.83	0.66	10.33	-9.46	2.65

Table 25: Combined-corruption MAE reductions (%) before PatchTST. Each row averages errors over $H \in \{96, 192, 336, 720\}$ at the specified severity before computing relative reductions. Positive is better.

Dataset	Level	Selected	Refined-cycle	Median-return	Paired	Closed	Median	Hampel
ETTh1	1	0.07	-1.44	0.07	-10.08	-3.49	-0.14	-0.02
	2	0.03	-1.27	0.03	-10.70	-3.33	-0.00	0.14
	3	0.36	-1.25	0.36	-8.15	-3.26	0.25	1.17
	4	0.28	-0.35	0.28	-8.52	-1.74	0.70	1.87
	5	0.53	0.53	0.56	-4.76	-0.95	0.88	1.89
ETTh2	1	0.55	0.12	0.55	-0.42	-0.21	0.53	0.49
	2	0.48	0.16	0.48	-0.10	-0.11	0.23	0.77
	3	0.78	0.06	0.78	-0.51	-0.16	-3.01	1.05
	4	1.02	0.48	1.02	-0.77	0.12	0.62	1.11
	5	1.57	0.29	1.57	-0.09	-0.05	-1.41	1.67
ETTm1	1	0.18	0.18	0.09	-1.58	-0.76	0.06	0.47
	2	0.47	0.47	0.07	-1.09	-0.39	0.34	0.88
	3	8.02	8.02	2.53	8.40	7.92	2.37	2.91
	4	10.64	10.64	2.25	10.56	10.74	3.09	4.20
	5	9.70	9.70	1.76	11.26	10.47	1.19	2.35
ETTm2	1	0.29	0.67	0.29	-0.04	0.35	0.43	1.04
	2	0.93	0.93	0.44	0.26	0.62	0.63	1.92
	3	2.08	2.08	0.74	1.65	1.95	2.43	2.71
	4	3.21	3.21	1.25	3.24	3.10	1.83	1.71
	5	4.16	4.16	1.67	4.60	4.38	1.43	1.44
Weather	1	2.40	2.40	1.37	2.64	2.34	0.63	2.66
	2	6.50	6.50	2.35	6.87	6.07	0.66	2.52
	3	9.18	9.18	3.50	9.48	8.52	2.37	4.93
	4	10.57	10.57	5.06	10.92	10.18	6.72	6.89
	5	11.97	11.97	5.45	12.46	11.52	5.98	6.20
AQWan	1	1.29	1.29	0.88	-1.09	0.77	0.38	1.28
	2	2.04	2.04	1.12	0.21	1.53	-0.02	1.62
	3	6.12	6.12	3.60	6.25	5.68	-1.13	1.88
	4	7.52	7.52	3.96	7.94	7.27	-0.41	2.36
	5	11.52	11.52	7.92	8.44	11.27	-0.26	2.38
CzeLan	1	3.50	3.50	1.63	2.83	3.29	1.29	2.44
	2	10.87	10.87	4.09	11.12	10.81	0.35	1.10
	3	15.20	15.20	5.77	15.42	15.29	5.10	4.64
	4	20.63	20.63	7.85	22.73	20.70	0.55	0.96
	5	20.59	20.59	7.39	24.60	21.23	-3.29	-3.15
ZafNoo	1	0.29	0.29	0.30	-15.53	-3.62	1.27	2.43
	2	1.67	2.41	1.67	-28.91	-8.35	1.81	-8.99
	3	2.46	2.04	2.46	-20.14	-7.65	2.17	1.71
	4	4.06	4.06	2.87	-7.95	0.66	3.58	-4.64
	5	4.85	4.85	4.12	-5.87	0.81	-9.98	-1.61

F Limitations and Future Directions

RestoreTS is evaluated under controlled corruption with supplied severity. Its temporal-return restoration methods are less effective for isolated spikes: mean MSE increases by 0.25% before PatchTST and 0.32% before PatchLGA (Appendix B). Corrections can also alter clean observations. In the median-return ablation, removing either the plateau-variation bound (which limits variation inside a proposed interval) or the support-dispersion bound (which requires agreement among surrounding run medians) improves aggregate corrupted-input forecasting but increases clean-input damage (Appendices C and A.2). Future work could combine interval restoration with spike-specific handling, strengthen criteria for preserving genuine temporal changes, and select configurations without supplied severity.

Earlier method development included inspection of the validation and test histories, as detailed in Appendix A. Evaluation on an untouched temporal period would test whether the observed benefits extend beyond these previously inspected histories. Finally, the correction records expose the locations and magnitudes of interventions, while their value for downstream decisions remains to be evaluated. Future studies should examine how domain experts use these records to assess restored histories and decide whether and how to use the resulting forecasts.