StreamLTT: Risk Certification for Policy Selection under Stream Dependence

Anonymous authors
Paper under double-blind review

Abstract

Learn-then-Test (LTT) provides a general framework for certifying deployment policies against a prescribed risk target, but its standard formulation targets a fixed population risk estimated from calibration losses. In sequential deployment, losses may instead be dependent and nonstationary, so the risk of a fixed policy can evolve with the observed history. We introduce *StreamLTT*, which replaces this fixed population risk with *average predictable risk*, defined by conditioning each decision's loss on the information available before its outcome is revealed. Conditional centering turns empirical deviations from this target into a martingale sum, enabling both a test that adapts to predictable variance and a finite-sample Hoeffding alternative. Combined with family-wise error rate control, StreamLTT yields a simultaneously certified policy region that supports subsequent deployment selection. Across seven forecasting datasets and seven deployment risks, StreamLTT places the certified frontier substantially closer to the prescribed risk target than IID, robust-variance, block-bootstrap, and concentration-based alternatives, while empirical family-wise error closely follows the prescribed level. A finite-sample ablation further shows that the predictable-risk formulation drives the tighter risk frontier, while adaptation to predictable variance makes more effective use of the available family-wise error budget. StreamLTT therefore extends risk-certified policy selection to dependent deployment streams by addressing both what risk should be certified and how it should be tested.

1 Introduction

Modern prediction systems are ultimately deployed through decisions rather than raw predictions (Elmachtoub & Grigas, 2022; Wilder et al., 2019; Donti et al., 2017). A practitioner may choose a confidence threshold, prediction set width, abstention rule, alarm threshold, intervention rule, or another operating policy subject to an explicit risk requirement (Geifman & El-Yaniv, 2019; Liu et al., 2019). A broad literature has developed statistical tools for uncertainty-aware and risk-controlled decision-making, including selective classification, conformal prediction, and conformal risk control (Geifman & El-Yaniv, 2017; Angelopoulos & Bates, 2023; Angelopoulos et al., 2024). Learn then Test (LTT) provides a natural framework for policy selection: given a finite candidate family, it tests each policy against a prescribed risk target and corrects for multiplicity so that the resulting policy region is certified simultaneously (Angelopoulos et al., 2025). For independent calibration losses, a common implementation constructs p values for individual policies using a central limit theorem (CLT) approximation to empirical risk; we refer to this instantiation as *IID-LTT*. This separation between prediction, certification, and subsequent selection makes LTT particularly attractive when deployment must satisfy an explicit risk budget.

A central difficulty arises when certification data arrive as a stream. Forecasts are temporally ordered, operating regimes evolve, and observations may be coupled through past outcomes, recurring entities, or shared structure. Even for a fixed policy, its conditional loss distribution can therefore depend on the accumulated history and evolve throughout deployment, so a single stationary marginal risk no longer describes the risks encountered over time. This challenge has motivated a broad literature on *online calibration*, which uses sequential feedback to maintain predictive calibration or coverage under distribution shift and nonexchangeability (Gibbs & Candes, 2021; Zaffran et al., 2022; Gibbs & Candès, 2024; Xu & Xie, 2023).

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.

Policy certification, however, poses a different statistical problem. In LTT, the candidate family is fixed before certification, and each policy must be tested against a risk constraint before multiplicity correction controls the probability of certifying any unsafe candidate (Angelopoulos et al., 2025). Under stream dependence, the challenge is therefore not only how to construct a valid test, but also what risk should be certified when the risk of a fixed policy can evolve along the stream. Existing tools address the inference side of this problem: long-run variance estimators adjust uncertainty for dependent averages (Newey & West, 1987), block resampling approximates sampling distributions under local dependence (Politis & Romano, 1992), and sequential concentration controls deviations for adapted observations (Howard et al., 2020). Yet these approaches still require a pre-specified risk target, leaving open how risk itself should be defined when it evolves with the observed history.

We address this gap by defining risk along the information flow of deployment. Immediately before each outcome is revealed, we consider the conditional risk of the corresponding decision given the available history and average these risks along the stream, yielding the *average predictable risk*. This target does not require a stationary marginal distribution and reduces to the standard population risk in the IID setting. Conditional centering further turns empirical deviations from this target into a martingale sum, providing the basis for ***StreamLTT***, our extension of LTT to policy certification over dependent deployment streams. We develop a test that adapts to predictable quadratic variation, together with a martingale Hoeffding alternative (Hoeffding, 1963) that requires only bounded losses and provides finite-sample validity. The resulting p-values plug directly into the multiple-testing layer of LTT (Angelopoulos et al., 2025), yielding a simultaneously certified policy region under family-wise error rate control (Holm, 1979). A deployment policy can then be selected from this region without repeating certification, retaining the modular selection principle of LTT while aligning certification with the information structure of an evolving stream.

Our experiments evaluate StreamLTT along two dimensions: the tightness of the certified risk frontier and the control of false certification. Across seven real-world forecasting datasets and seven deployment risks, StreamLTT produces substantially tighter risk certificates than IID-LTT and alternatives based on robust variance estimation, block resampling, and concentration bounds, while empirical family-wise error closely follows the prescribed level. A finite-sample ablation further separates the two sources of improvement: the predictable-risk formulation already yields a tight certified frontier, while adaptation to predictable variance makes more effective use of the available family-wise error budget. These results distinguish the benefit of changing the risk being certified from that of adapting the statistical test to the variability of the stream.

Our contributions are threefold:

- We formulate risk certification for policy selection under dependent and nonstationary deployment through *average predictable risk*, which follows the information available at each decision time and recovers the standard population risk in the IID setting.
- We develop StreamLTT by using conditional centering to obtain martingale fluctuations, leading to a test that adapts to predictable variance and a Hoeffding alternative with finite sample validity. Combined with control of the family wise error rate, StreamLTT certifies an entire policy region while preserving validity under subsequent policy selection.
- Across seven forecasting datasets and seven deployment risks, StreamLTT yields substantially tighter certification than IID, robust variance, bootstrap, and concentration based alternatives while empirical family wise error follows the prescribed level. A finite sample ablation further separates the benefit of the predictable risk formulation from that of predictable variance adaptation.

2 Problem Formulation and Preliminaries

Consider a prediction system deployed over an ordered stream $(X_1, Y_1), \ldots, (X_N, Y_N)$. At round n, the covariates X_n and the history from previous rounds are available before a deployment decision is made, whereas the outcome Y_n is revealed only afterward. Let $\{\mathcal{F}_n\}_{n\geq 0}$ denote the associated filtration, with the history, current covariates, model output, and deployment decision at round n measurable with respect to $\mathcal{F}_{n-1}$. Here, we make no assumptions about independence nor stationarity. The conditional law of the stream may vary over time and depend on previously observed outcomes, recurring entities, or other shared context.

A finite family of deployment policies $\Theta := \{\theta_1, \ldots, \theta_M\}$ is fixed before certification. Each policy may specify a confidence threshold, prediction-set width, alarm threshold, or another operating rule applied to a frozen predictive model. At round n, policy θ incurs an adapted bounded loss $L_n(\theta) \in [0, B(\theta)]$, where the deterministic bound $B(\theta) < \infty$ is fixed before certification. Given a prescribed risk target α and family-wise error level δ, the objective is to construct a certified policy region such that the probability of including any policy whose deployment risk exceeds α is controlled at level δ.

Learn-then-Test (LTT) casts risk-controlled policy selection as a multiple-testing problem (Angelopoulos et al., 2025). In the standard setting, each candidate θ has population risk $R(\theta) := \mathbb{E}[L(\theta)]$. A policy is considered safe when $R(\theta) \leq \alpha$, so LTT tests the complementary null

$$H_j : R(\theta_j) > \alpha, \qquad j = 1, \ldots, M. \tag{1}$$

A valid p-value is constructed for each candidate and passed to a multiple-testing procedure. If the rejection rule controls the family-wise error rate (FWER) at level δ, the rejected hypotheses define a certified region $\widehat{\Theta}_{\mathrm{safe}}$ satisfying

$$\mathbb{P}\Big(\widehat{\Theta}_{\mathrm{safe}} \subseteq \Theta_{\mathrm{safe}}\Big) \geq 1 - \delta, \qquad \Theta_{\mathrm{safe}} := \{\theta \in \Theta : R(\theta) \leq \alpha\}. \tag{2}$$

Thus, with probability at least $1-\delta$, every policy retained in the certified region satisfies the prescribed risk requirement. Since the guarantee holds simultaneously over the region, a final operating point may subsequently be selected from $\widehat{\Theta}_{\mathrm{safe}}$ without retesting the candidate family.

For independent calibration losses, LTT admits several policy level tests, including finite sample bounded loss tests and asymptotic approximations. In this work, we refer to the CLT based instantiation as *IID-LTT*. These standard constructions target a common population risk and rely on independent calibration observations. They therefore do not directly address the deployment setting considered here, where losses may be dependent and their conditional risks may evolve with the observed history.

Problem formulation. Let $\mathcal{R}_N(\theta)$ denote the stream-level deployment risk associated with policy θ, whose precise form will be specified in Section 3. For a prescribed risk level α, define the corresponding safe region as

$$\Theta_{\mathrm{safe},N} := \{\theta \in \Theta : \mathcal{R}_N(\theta) \leq \alpha\}. \tag{3}$$

Given the observed stream, the certification problem is to construct a data-dependent region $\widehat{\Theta}_{\mathrm{safe}} \subseteq \Theta$ satisfying

$$\mathbb{P}\Big(\widehat{\Theta}_{\mathrm{safe}} \not\subseteq \Theta_{\mathrm{safe},N}\Big) = \mathbb{P}\Big(\exists\, \theta \in \widehat{\Theta}_{\mathrm{safe}} : \mathcal{R}_N(\theta) > \alpha\Big) \leq \delta. \tag{4}$$

Thus, the statistical task is to specify a stream-level risk $\mathcal{R}_N(\theta)$ compatible with the deployment information structure and to construct valid policy-level tests for the unsafe hypotheses $H_j : \mathcal{R}_N(\theta_j) > \alpha$. Section 3 instantiates $\mathcal{R}_N$ as average predictable risk and develops the corresponding dependence-aware tests.

3 StreamLTT: Learn-then-Test under Stream Dependence

We now develop *StreamLTT* for the stream-level certification problem introduced in Section 2. Rather than imposing a common population risk, StreamLTT averages the conditional risks encountered along the deployment stream. Centering losses around this predictable target induces martingale fluctuations, enabling finite-sample and variance-adaptive policy tests under dependence. The resulting p-values are combined with family-wise error control to obtain a simultaneously certified policy region.

3.1 From Marginal to Predictable Risk

For a fixed candidate policy $\theta \in \Theta$, let $L_n(\theta)$ denote its bounded deployment loss at round n. Before Y_n is revealed, define the one-step predictable risk $p_n(\theta) := \mathbb{E}[L_n(\theta) \mid \mathcal{F}_{n-1}]$, the conditional risk of deploying θ given the information available when the decision is made. Unlike a fixed population risk, $p_n(\theta)$ may change with the observed history and operating environment.

Definition 3.1 (Average predictable risk)**.** The average predictable risk of policy θ over the first N rounds is

$$R_{\mathrm{pred},N}(\theta) := \frac{1}{N}\sum_{n=1}^{N} p_n(\theta). \tag{5}$$

The quantity $R_{\mathrm{pred},N}(\theta)$ therefore measures the average conditional risk encountered along the realized deployment history, rather than risk under a single stationary population. It allows the conditional risk to vary across rounds and requires neither independence nor stationarity. In the IID setting with common mean $R(\theta)$, $p_n(\theta) = R(\theta)$ for every n, so $R_{\mathrm{pred},N}(\theta) = R(\theta)$ and the standard LTT target is recovered.

This target also yields the statistical structure needed for certification. Let

$$\widehat{R}_N(\theta) := \frac{1}{N}\sum_{n=1}^{N} L_n(\theta), \qquad Z_n(\theta) := L_n(\theta) - p_n(\theta).$$

Since $\mathbb{E}[Z_n(\theta) \mid \mathcal{F}_{n-1}] = 0$, $\{Z_n(\theta)\}_{n\geq 1}$ is a martingale-difference sequence, and hence

$$\widehat{R}_N(\theta) - R_{\mathrm{pred},N}(\theta) = \frac{1}{N}\sum_{n=1}^{N} Z_n(\theta). \tag{6}$$

Thus, dependence in the losses is retained in the risk target, while fluctuations of empirical risk around that target admit martingale-based inference.

3.2 Stream-Aware Policy Testing

For each candidate $\theta_j \in \Theta$, StreamLTT tests $H_j : R_{\mathrm{pred},N}(\theta_j) > \alpha$, so rejecting H_j certifies $R_{\mathrm{pred},N}(\theta_j) \leq \alpha$. By Eq. (6), empirical risk differs from this target by a martingale average. This structure permits a finite-sample test under bounded losses without specifying the dependence among the losses.

Theorem 3.2 (Finite-sample StreamLTT test)**.** *Suppose $L_n(\theta_j)$ is adapted to $\mathcal{F}_n$ and $0 \leq L_n(\theta_j) \leq B_j$, where $B_j := B(\theta_j)$ is a deterministic bound fixed before certification. Then*

$$p_j^{\mathrm{H}} := \exp\left(-\frac{2N\big(\alpha - \widehat{R}_N(\theta_j)\big)_+^2}{B_j^2}\right) \tag{7}$$

is finite-sample super-uniform under $H_j : R_{\mathrm{pred},N}(\theta_j) > \alpha$.

Theorem 3.2 provides a finite-sample guarantee using only the deterministic loss bound B_j. The same robustness can make the test conservative: its normalization reflects the worst-case loss range rather than the conditional variability actually encountered along the stream. When the latter is substantially smaller, a variance-sensitive normalization can provide tighter certification. The proof is given in Appendix B.1.

Variance-adaptive testing. To capture this stream-specific variability, define the predictable quadratic variation

$$V_N^2(\theta) := \sum_{n=1}^{N} \mathbb{E}\big[Z_n^2(\theta) \mid \mathcal{F}_{n-1}\big]. \tag{8}$$

Whereas B_j gives a uniform worst-case bound, $V_N^2(\theta)$ accumulates the conditional variance of the martingale increments along the deployment stream. We require this variation to admit a stable asymptotic scale.

Assumption 3.3 (Predictable-variance stabilization)**.** For every fixed $\theta_j \in \Theta$, there exists a deterministic sequence $v_{N,j}^2 > 0$ such that $v_{N,j}^2 \to \infty$ and $V_N^2(\theta_j)/v_{N,j}^2 \xrightarrow{p} 1$.

Since bounded losses yield uniformly bounded martingale increments, $v_{N,j} \to \infty$ makes the conditional Lindeberg condition automatic. Together with Assumption 3.3, the martingale central limit theorem then yields the variance-adaptive test below.

Theorem 3.4 (Variance-adaptive StreamLTT test)**.** *Suppose $L_n(\theta_j)$ is adapted to $\mathcal{F}_n$ and $0 \le L_n(\theta_j) \le B_j$, where $B_j := B(\theta_j)$ is a deterministic bound, and Assumption 3.3 holds. Let $\widehat{V}_N(\theta_j)$ consistently estimate the predictable variation, such that $\widehat{V}_N(\theta_j)/V_N(\theta_j) \xrightarrow{p} 1$. Then*

$$p_j^{\mathrm{SL}} := \Phi\left(\frac{N\big(\widehat{R}_N(\theta_j) - \alpha\big)}{\widehat{V}_N(\theta_j)}\right) \tag{9}$$

is asymptotically super-uniform under $H_j : R_{\mathrm{pred},N}(\theta_j) > \alpha$; namely, for every $z \in [0,1]$,

$$\limsup_{N\to\infty} \mathbb{P}\big(p_j^{\mathrm{SL}} \le z\big) \le z.$$

Theorem 3.4 trades the finite-sample robustness of Theorem 3.2 for adaptation to the predictable variability of the stream. When this variability is well below the worst-case range, the adaptive normalization can reduce conservatism and yield tighter certification. The proof is given in Appendix B.2. Under IID losses, the construction recovers the standard population-risk test.

Corollary 3.5 (IID reduction)**.** *Suppose $\{L_n(\theta_j)\}_{n=1}^N$ are IID with $\mathbb{E}[L_n(\theta_j)] = R(\theta_j)$ and positive variance σ_j^2. Then $R_{\mathrm{pred},N}(\theta_j) = R(\theta_j)$ and $V_N^2(\theta_j) = N\sigma_j^2$. Consequently, the variance-adaptive StreamLTT statistic reduces asymptotically to the standard IID central-limit-theorem statistic for testing $H_j : R(\theta_j) > \alpha$.*

Estimating the predictable variance. The adaptive test depends on $V_N(\theta)$, which is generally unknown. For Bernoulli losses, $\mathbb{E}[Z_n^2(\theta) \mid \mathcal{F}_{n-1}] = p_n(\theta)(1 - p_n(\theta))$. Given an online estimate $\widehat{p}_n(\theta)$ of $p_n(\theta)$, we use the plug-in estimator

$$\widehat{V}_N^2(\theta) := \sum_{n=1}^{N} \widehat{p}_n(\theta)\big(1 - \widehat{p}_n(\theta)\big). \tag{10}$$

The next result shows that consistency of the one-step risk estimates is sufficient for the variance estimation required by Theorem 3.4.

Lemma 3.6 (Consistency of predictable variance estimation)**.** *Suppose $N^{-1}\sum_{n=1}^N (\widehat{p}_n(\theta) - p_n(\theta))^2 \xrightarrow{p} 0$ and $V_N^2(\theta)/N \xrightarrow{p} \sigma_\theta^2 > 0$. Then $\widehat{V}_N(\theta)/V_N(\theta) \xrightarrow{p} 1$. Moreover, Assumption 3.3 holds with $v_N^2 = N\sigma_\theta^2$.*

Lemma 3.6 therefore links estimation of the one-step predictable risks directly to the normalization used by the variance-adaptive test. The proof is given in Appendix B.3.

3.3 From Policy Tests to a Certified Region

The preceding results establish valid policy-level tests for predictable risk. We now lift these guarantees from individual candidates to the entire policy family. Let $\mathcal{R}_N \subseteq \{1, \dots, M\}$ denote the rejection set returned by the multiple-testing procedure and define $\widehat{\Theta}_{\mathrm{safe}} := \{\theta_j : j \in \mathcal{R}_N\}$. The corresponding safe region is $\Theta_{\mathrm{safe},N} := \{\theta \in \Theta : R_{\mathrm{pred},N}(\theta) \le \alpha\}$. Controlling false rejections across the candidate family then yields simultaneous certification of $\widehat{\Theta}_{\mathrm{safe}}$.

Theorem 3.7 (Simultaneous StreamLTT certification)**.** *Let $\widehat{\Theta}_{\mathrm{safe}}$ be obtained by applying FWER control at level δ to the finite-sample StreamLTT p-values in Theorem 3.2. Then, for any possibly data-dependent selection $\widehat{\theta} \in \widehat{\Theta}_{\mathrm{safe}}$ whenever $\widehat{\Theta}_{\mathrm{safe}} \neq \emptyset$,*

$$\mathbb{P}\Big(\widehat{\Theta}_{\mathrm{safe}} \subseteq \Theta_{\mathrm{safe},N}\Big) \ge 1 - \delta, \quad \mathbb{P}\Big(\widehat{\Theta}_{\mathrm{safe}} \neq \emptyset,\, R_{\mathrm{pred},N}(\widehat{\theta}) > \alpha\Big) \le \delta. \tag{11}$$

For a fixed finite candidate family, applying FWER control at level δ to the variance-adaptive p-values in Theorem 3.4 yields

$$\liminf_{N\to\infty} \mathbb{P}\Big(\widehat{\Theta}_{\mathrm{safe}} \subseteq \Theta_{\mathrm{safe},N}\Big) \ge 1 - \delta, \quad \limsup_{N\to\infty} \mathbb{P}\Big(\widehat{\Theta}_{\mathrm{safe}} \neq \emptyset,\, R_{\mathrm{pred},N}(\widehat{\theta}) > \alpha\Big) \le \delta. \tag{12}$$

Algorithm 1 StreamLTT for risk-certified policy selection

Require: Certification stream $\{(X_n, Y_n)\}_{n=1}^{N}$; candidate family $\Theta = \{\theta_1, \ldots, \theta_M\}$; loss $L_n(\theta)$; risk target α; bounds $\{B(\theta_j)\}_{j=1}^{M}$; FWER level δ; MODE $\in \{\text{ADAPTIVE}, \text{HOEFFDING}\}$.

Ensure: Certified region $\widehat{\Theta}_{\text{safe}}$; optional policy $\widehat{\theta}$.

```
Freeze the candidate family and all test-specific quantities.
Initialize empirical-risk summaries and, for ADAPTIVE mode, predictable-variance summaries.
for n = 1, ..., N do
    Observe F_{n-1} and form decisions for all θ_j ∈ Θ.
    Reveal Y_n and record L_n(θ_j) for all θ_j ∈ Θ̂.
    if MODE = ADAPTIVE then
        Estimate p̂_n(θ_j) and update the predictable-variance summaries via Eq. (10).
    end if
end for
for j = 1, ..., M do
    Compute R̂_N(θ_j) ← N^{-1} Σ_{n=1}^{N} L_n(θ_j).
    if MODE = ADAPTIVE then
        Compute p_j ← p_j^{SL} via Eq. (9).
    else
        Compute p_j ← p_j^{H} via Eq. (7).
    end if
end for
Apply FWER control at level δ to {p_j}_{j=1}^{M} and form Θ̂_safe from the rejected hypotheses.
if Θ̂_safe = ∅ then
    return "no certified policy".
end if
Optionally select θ̂ ∈ Θ̂_safe according to the deployment criterion.
return Θ̂_safe and optional θ̂.
```

Theorem 3.7 lifts the policy-level validity established above to simultaneous certification of the candidate family. The policy-level tests account for dependence along the deployment stream, whereas FWER control accounts for multiplicity across candidate policies. The first guarantee ensures that the entire certified region is safe with probability at least $1 - \delta$; the second shows that any subsequent data-dependent selection from this region inherits the same risk control without additional testing.

Complete procedure. Algorithm 1 summarizes the resulting StreamLTT pipeline. Each candidate is evaluated using either the finite-sample Hoeffding test or the variance-adaptive test, and the resulting p-values are combined with FWER control to form $\widehat{\Theta}_{\text{safe}}$. A deployment policy may then be selected from the certified region according to the downstream objective. For M candidates, StreamLTT maintains $\mathcal{O}(M)$ running summaries and requires $\mathcal{O}(M)$ computation per stream round.

4 EXPERIMENTS

We evaluate StreamLTT along two complementary dimensions of risk certification: the tightness of the certified risk frontier and the control of false certification. We refer to the variance adaptive test in Theorem 3.4 as *StreamLTT-A* and the finite sample Hoeffding test in Theorem 3.2 as *StreamLTT-H*. Our primary comparison evaluates StreamLTT-A against IID, dependence aware, and sequential certification baselines. StreamLTT-H provides a finite sample counterpart that isolates the effect of predictable variance adaptation.

4.1 EXPERIMENTAL SETUP

Datasets and forecasting protocol. We consider seven real world forecasting datasets spanning three domains: Electricity (Trindade, 2015), Appliances Energy (Candanedo et al., 2017), and Household Power (Hebrail & Berard, 2006) for energy forecasting; PEMS04 and PEMS08 (Guo et al., 2019)

Table 1: Binary deployment risks used for certification. All auxiliary thresholds are determined from tuning data and fixed before certification.

Risk	Structure	Error event	α
Interval miss	Pointwise	$Y_i \notin C_m(X_i)$	.10
Lower-tail miss	Pointwise	$Y_i < \ell_i(m)$	.10
Severe miss	Pointwise	$\lvert Y_i - \widehat{\mu}_i \rvert > ms_i + e_0$	.10
Task-regime miss	Regime	$G_i = 1,\ Y_i \notin C_m(X_i)$	.05
Opposite direction	Temporal	$a_i(\theta) z_i^d = -1$	.10
Missed reversal	History	$r_i = 1,\ \widehat{r}_i(\theta) = 0$	.10
False reversal	History	$r_i = 0,\ \widehat{r}_i(\theta) = 1$	.10

for traffic forecasting; and Beijing Weather (Chen, 2017) and Jena Weather (Max Planck Institute for Biogeochemistry, 2016) for temperature forecasting. The datasets vary in sequence length and sampling frequency, and their temporal order is preserved throughout the experiment. Each dataset is split chronologically into a training set (20%), a validation set (2%), a calibration set (8%), and a certification stream (70%). A temporal convolutional network (TCN) (Bai et al., 2018)is trained for one step forecasting on the training set, selected on the validation set, and then frozen. The calibration set determines auxiliary quantities, such as severe miss tolerances and regime thresholds, which are fixed before certification. The remaining observations are processed sequentially as the certification stream, with the same frozen forecasts shared across all certification methods. At each round, only information available before the corresponding outcome is revealed is used; history dependent decisions may additionally use previously revealed outcomes. Model specifications, preprocessing, and further protocol details are provided in Appendix C.1.

Deployment risks. We consider seven binary deployment risks spanning pointwise, regime dependent, directional, and history dependent errors, as summarized in Table 1. These risks range from standard forecasting failures to decisions whose losses depend on the current operating regime or previously observed outcomes, providing diverse dependence structures for stream level certification. All auxiliary quantities are fixed before certification, and each risk defines a separate certification problem. Complete risk constructions are provided in Appendix C.2.

Methods. Our primary method is *StreamLTT-A*, the variance adaptive test in Theorem 3.4. We compare it with *IID-LTT* (Angelopoulos et al., 2025), *LTT-HB-Holm* (Hoeffding, 1963; Bentkus, 2004), *HAC-CLT-Holm* (Newey & West, 1987), *CircularBlockBootstrap-Holm* (Politis & Romano, 1992), and *EB-Howard-Bonferroni* (Howard et al., 2020), covering IID inference, bounded loss concentration, long run variance correction, block resampling, and sequential concentration, respectively. We also evaluate *StreamLTT-H*, the finite sample test in Theorem 3.2, which retains the same predictable risk target as StreamLTT-A while removing predictable variance adaptation. All methods use the same risk targets and family wise levels. Further details are provided in Appendix C.3.

Evaluation protocol and metrics. We evaluate all $7 \times 7 = 49$ dataset–risk combinations at $\delta \in \{0.10, 0.20\}$ using 2,000 independent repetitions per configuration. We assess certification through risk tightness and false certification control. Empirical FWER is the fraction of repetitions in which at least one certified policy has predictable risk above α. For each nonempty certified region, $R_{\max}$ denotes its largest predictable risk, and $\overline{R}_{\max}$ its mean across repetitions. We report the risk gap $\alpha - \overline{R}_{\max}$ and FWER gap $\delta - \widehat{\mathrm{FWER}}$, where values near zero indicate fuller use of the respective risk and family wise budgets. We additionally report the certified fraction, the proportion of candidate policies retained after certification. Further evaluation details, risk wise and dataset level results, and Monte Carlo uncertainty are provided in Appendix C.4.

4.2 Results

StreamLTT-A approaches both certification boundaries. Figure 1 shows that StreamLTT-A uses substantially more of both the risk and family wise error budgets. At $\delta = 0.10$, its mean risk gap is 0.0090, placing the certified frontier close to the prescribed risk boundary, compared with 0.0302 for IID-LTT and approximately 0.030–0.032 for the dependence and concentration based alternatives. This tighter frontier is not accompanied by excessive false certification: StreamLTT-

Table 2: Representative certification results at $\delta = 0.10$ across seven datasets. Each entry reports $\overline{R}_{\max}/\widehat{\mathrm{FWER}}$. The three risks represent two sided, one sided, and history dependent deployment errors.

Risk	Method	Electricity	Appliances	Household	PEMS04	PEMS08	Beijing	Jena
Interval ($\alpha = .10$)	IID-LTT	.051/.028	.046/.013	.048/.020	.051/.017	.044/.015	.047/.019	.045/.012
	LTT-HB-Holm	.050/.006	.045/.004	.047/.009	.051/.006	.043/.006	.046/.005	.045/.004
	HAC-CLT-Holm	.051/.035	.046/.014	.048/.025	.051/.021	.044/.019	.047/.024	.045/.017
	BlockBootstrap-Holm	.051/.034	.046/.015	.048/.025	.052/.021	.044/.019	.047/.023	.045/.018
	EB-Howard-Bonf.	.050/.000	.045/.000	.047/.000	.050/.000	.043/.000	.046/.000	.044/.000
	StreamLTT-A	**.096/.084**	**.096/.082**	**.097/.102**	**.096/.094**	**.095/.087**	**.096/.089**	**.096/.083**
Lower tail ($\alpha = .10$)	IID-LTT	.055/.045	.069/.031	.053/.032	.065/.025	.066/.026	.069/.020	.064/.025
	LTT-HB-Holm	.053/.014	.069/.010	.052/.009	.065/.009	.066/.009	.069/.004	.063/.007
	HAC-CLT-Holm	.055/.052	.069/.037	.053/.037	.066/.030	.066/.031	.070/.023	.064/.032
	BlockBootstrap-Holm	.055/.052	.069/.036	.053/.036	.066/.031	.066/.031	.070/.023	.064/.031
	EB-Howard-Bonf.	.052/.000	.068/.000	.052/.000	.065/.000	.065/.000	.069/.000	.063/.000
	StreamLTT-A	**.097/.095**	**.096/.097**	**.095/.093**	**.097/.080**	**.097/.088**	**.096/.081**	**.097/.085**
Missed reversal ($\alpha = .10$)	IID-LTT	.080/.087	.064/.026	.075/.040	.048/.016	.050/.018	.089/.091	.081/.047
	LTT-HB-Holm	.077/.030	.063/.008	.074/.015	.048/.005	.049/.006	.086/.031	.080/.018
	HAC-CLT-Holm	.081/.099	.064/.033	.075/.045	.049/.022	.050/.023	.090/.099	.081/.052
	BlockBootstrap-Holm	.081/.097	.064/.033	.075/.047	.049/.022	.050/.023	.090/.099	.081/.053
	EB-Howard-Bonf.	.073/.000	.063/.000	.074/.000	.048/.001	.049/.000	.079/.000	.080/.000
	StreamLTT-A	**.097/.088**	**.095/.092**	**.095/.085**	**.095/.090**	**.095/.089**	**.096/.093**	**.082/.092**

A attains empirical FWER 0.0902, close to the prescribed $\delta = 0.10$. In contrast, the competing procedures leave considerably more slack in the risk budget, the family wise budget, or both. StreamLTT-A therefore expands certification toward the admissible risk boundary while maintaining the intended family wise control.

Table 2 resolves this aggregate improvement at the dataset level for three representative risks; complete results for all seven risks are provided in the appendix. For interval miss, the baseline frontiers remain between 0.043 and 0.052, whereas StreamLTT-A reaches 0.095–0.097 across all seven datasets. The separation persists for lower tail miss, where the baseline values range from 0.052 to 0.070 while StreamLTT-A remains at 0.095–0.097. For the history dependent missed reversal risk, the gain is particularly pronounced on PEMS04 and PEMS08, where StreamLTT-A reaches 0.095 compared with approximately 0.048–0.050 for the competing methods. Across all 49 dataset–risk combinations, StreamLTT-A retains 76.7% of the candidate family while maintaining mean empirical FWER 0.0902 at the prescribed $\delta = 0.10$.

Dependence aware inference alone does not remove certification slack. The risk gaps help isolate the source of the improvement. At $\delta = 0.10$, IID-LTT has a mean gap of 0.0302, while HAC-CLT-Holm and CircularBlockBootstrap-Holm attain 0.0300 despite explicitly accounting for temporal dependence. StreamLTT-A reduces this gap to 0.0090. Thus, adjusting the sampling distribution for dependence has little effect on the certified frontier when the underlying risk target remains unchanged. StreamLTT instead certifies average predictable risk, conditioning risk on the information available at each decision time. The empirical separation therefore supports the distinction in Section 3.1: dependence aware inference changes how risk is tested, whereas StreamLTT also changes what risk is certified along the deployment stream.

StreamLTT-A adapts to the available family-wise budget. Increasing δ from 0.10 to 0.20 provides a direct test of whether the certification procedure responds to a larger tolerance for false certification. StreamLTT-A's empirical FWER increases from 0.0902 to 0.1888, closely tracking the corresponding prescribed levels, while its mean risk gap remains small and decreases from 0.0090 to 0.0078. The baselines respond much less strongly: at $\delta = 0.20$, IID-LTT, HAC-CLT-Holm, and CircularBlockBootstrap-Holm reach FWER only 0.1125, 0.1217, and 0.1208, respectively, while LTT-HB-Holm remains at 0.0376 and EB-Howard-Bonferroni near zero. Their risk gaps also remain between 0.0283 and 0.0321. Thus, increasing the family wise budget primarily expands certification under StreamLTT-A, rather than leaving the additional error budget unused.

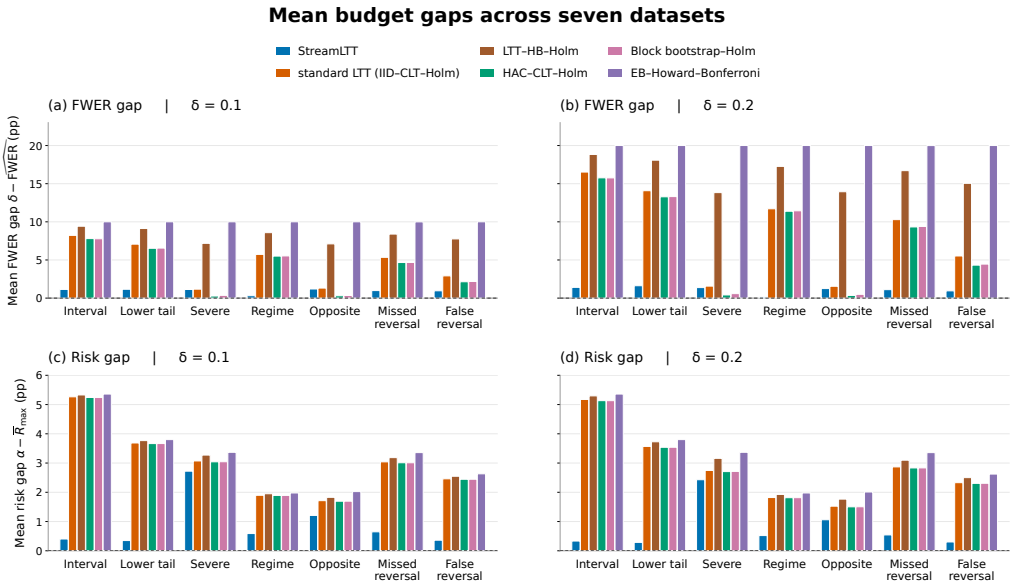

Figure 1: Mean family wise error rate (FWER) and risk gaps across seven deployment risks. The top row reports $\delta - \widehat{\mathrm{FWER}}$ at $\delta = 0.10$ and 0.20; the bottom row reports $\alpha - \overline{R}_{\max}$ at the corresponding family wise levels. Each bar averages over the seven datasets. Values closer to zero indicate tighter utilization of the corresponding certification budget.

Ablation: predictable variance adaptation primarily improves family wise budget utilization. StreamLTT-H and StreamLTT-A certify the same average predictable risk and differ only in the policy level test, allowing us to isolate the effect of predictable variance adaptation. Their certified frontiers are already similar: at $\delta = 0.10$ and 0.20, StreamLTT-H attains mean risk gaps 0.0118 and 0.0112, compared with 0.0090 and 0.0078 for StreamLTT-A. The certified fractions are likewise close, increasing from 75.3% to 76.7% at $\delta = 0.10$ and from 75.7% to 77.1% at $\delta = 0.20$.

The main difference appears in false certification. StreamLTT-H has empirical FWER only 0.0001 and 0.0006 at the two family wise levels, leaving nearly the entire error budget unused, whereas StreamLTT-A reaches 0.0902 and 0.1888. Thus, the tight frontier is already largely present when predictable risk is paired with the finite sample test; predictable variance adaptation provides a modest additional improvement in frontier tightness and certified fraction, but its dominant empirical effect is to reduce the conservatism of policy level testing and make substantially fuller use of the available family wise budget.

5 Conclusion

We introduced StreamLTT for risk certified policy selection over dependent and nonstationary deployment streams. StreamLTT replaces a fixed population risk with average predictable risk, aligning certification with the information available at each decision time and yielding martingale fluctuations for policy level testing. This structure supports both a finite sample bounded loss test and a variance adaptive test, which combine with family wise error control to certify the candidate policy region. Across seven forecasting datasets and seven deployment risks, StreamLTT-A produced substantially tighter certified frontiers than IID, dependence aware, bootstrap, and sequential concentration baselines, while empirical false certification closely tracked the prescribed family wise level. The finite sample ablation further separated the two sources of improvement: the predictable risk formulation accounted for most of the frontier tightening, while predictable variance adaptation primarily reduced testing conservatism and made fuller use of the available family wise budget. Together, these results extend LTT to evolving deployment streams by addressing both what risk should be certified and how it should be tested under dependence.

Reproducibility Statement

Complete proofs and experimental details are provided in the appendix to support reproduction of the theoretical and empirical results.

Ethics Statement

This study used public datasets and needs no ethical approval.

AI Use Statement

Generative AI was used only to assist with writing and language polishing.

References

Donald W. K. Andrews. Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica*, 59(3):817–858, 1991. doi: 10.2307/2938229.

Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Found. Trends Mach. Learn.*, 16(4):494–591, March 2023.

Anastasios N Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *Proc. Int. Conf. Learn. Represent.*, 2024.

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2):1641 – 1662, 2025.

Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018.

Vidmantas Bentkus. On Hoeffding's inequalities. *The Annals of Probability*, 32(2):1650–1673, 2004. doi: 10.1214/009117904000000360.

Luis M. Candanedo, Véronique Feldheim, and Dominique Deramaix. Data driven prediction models of energy use of appliances in a low-energy house. *Energy and Buildings*, 140:81–97, 2017.

Song Chen. Beijing Multi-Site Air Quality. UCI Machine Learning Repository, 2017.

C. K. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970. doi: 10.1109/TIT.1970.1054406.

Victor H. de la Peña, Tze Leung Lai, and Qi-Man Shao. *Self-Normalized Processes: Limit Theory and Statistical Applications*. Springer, Berlin, Heidelberg, 2009. doi: 10.1007/978-3-540-85636-8.

Priya L. Donti, Brandon Amos, and J. Zico Kolter. Task-based end-to-end model learning in stochastic optimization. In *Proc. Adv. Neural Inf. Process. Syst.*, volume 30, 2017.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *J. Mach. Learn. Res.*, 11(53):1605–1641, 2010.

Adam N. Elmachtoub and Paul Grigas. Smart "predict, then optimize". *Management Science*, 68(1): 9–26, 2022.

David A. Freedman. On tail probabilities for martingales. *The Annals of Probability*, 3(1):100–118, 1975. doi: 10.1214/aop/1176996452.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Proc. Adv. Neural Inf. Process. Syst.*, volume 30, 2017.

Yonatan Geifman and Ran El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. In *Proc. Int. Conf. Mach. Learn.*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2151–2159. PMLR, 2019.

Isaac Gibbs and Emmanuel Candes. Adaptive conformal inference under distribution shift. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Proc. Adv. Neural Inf. Process. Syst.*, 2021.

Isaac Gibbs and Emmanuel Candès. Conformal inference for online prediction with arbitrary distribution shifts. *J. Mach. Learn. Res.*, 25(162):1–36, 2024.

Shengnan Guo, Youfang Lin, Ning Feng, Chao Song, and Huaiyu Wan. Attention based spatial-temporal graph convolutional networks for traffic flow forecasting. In *Proc. AAAI Conf. Artif. Intel.*, volume 33, pp. 922–929, 2019. doi: 10.1609/aaai.v33i01.3301922.

Georges Hebrail and Alice Berard. Individual household electric power consumption. UCI Machine Learning Repository, 2006.

Yosef Hochberg. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, 75 (4):800–802, 1988. doi: 10.1093/biomet/75.4.800.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform Chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 17:257–317, 2020. doi: 10.1214/18-PS321.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, non-parametric, nonasymptotic confidence sequences. *Ann. Stat.*, 49(2):1055–1080, 2021. doi: 10.1214/20-AOS1991.

Hans R. Künsch. The jackknife and the bootstrap for general stationary observations. *Ann. Stat.*, 17 (3):1217–1241, 1989. doi: 10.1214/aos/1176347265.

S. N. Lahiri. *Resampling Methods for Dependent Data*. Springer, New York, 2003. doi: 10.1007/978-1-4757-3803-2.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523): 1094–1111, 2018. doi: 10.1080/01621459.2017.1307116.

Ziyin Liu, Zhikang Wang, Paul Pu Liang, Russ R Salakhutdinov, Louis-Philippe Morency, and Masahito Ueda. Deep gamblers: Learning to abstain with portfolio theory. In *Proc. Adv. Neural Inf. Process. Syst.*, volume 32, 2019.

Ruth Marcus, Eric Peritz, and K. R. Gabriel. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 63(3):655–660, 1976. doi: 10.1093/biomet/63.3.655.

Max Planck Institute for Biogeochemistry. Jena climate and weather data. `https://www.bgc-jena.mpg.de/wetter/`, 2016. 2009–2016 subset. Accessed: 2026-09-24.

Whitney K. Newey and Kenneth D. West. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3):703–708, 1987. doi: 10.2307/1913610.

Dimitris N. Politis and Joseph P. Romano. A circular block-resampling procedure for stationary data. In Raoul LePage and Lynne Billard (eds.), *Exploring the Limits of Bootstrap*, pp. 263–270. John Wiley & Sons, 1992.

Dimitris N. Politis and Joseph P. Romano. The stationary bootstrap. *J. Amer. Stat. Assoc.*, 89(428): 1303–1313, 1994. doi: 10.1080/01621459.1994.10476870.

Yaniv Romano, Matteo Sesia, and Emmanuel J. Candès. Classification with valid and adaptive coverage. In *Proc. Adv. Neural Inf. Process. Syst.*, 2020.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9:371–421, 2008.

Artur Trindade. ElectricityLoadDiagrams20112014. UCI Machine Learning Repository, 2015.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer Science & Business Media, 2005.

Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(1):1–27, 2024. doi: 10.1093/jrsssb/qkad009.

Bryan Wilder, Bistra Dilkina, and Milind Tambe. Melding the data-decisions pipeline: Decision-focused learning for combinatorial optimization. In *Proc. AAAI Conf. Artif. Intel.*, volume 33, pp. 1658–1665, 2019.

Chen Xu and Yao Xie. Sequential predictive conformal inference for time series. In *Proc. Int. Conf. Mach. Learn.*, volume 202 of *Proc. Mach. Learn. Res.*, pp. 38707–38727. PMLR, 2023.

Margaux Zaffran, Olivier Feron, Yannig Goude, Julie Josse, and Aymeric Dieuleveut. Adaptive conformal predictions for time series. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proc. Int. Conf. Mach. Learn.*, volume 162, pp. 25834–25866. PMLR, 17–23 Jul 2022.

Supplementary File

A Related Works

StreamLTT is related to the following works on risk controlled prediction, sequential calibration, inference under dependence, and multiple testing.

A.1 Risk Control and Sequential Calibration

Conformal prediction provides distribution free predictive uncertainty under exchangeability (Vovk et al., 2005; Shafer & Vovk, 2008; Lei et al., 2018; Angelopoulos & Bates, 2023), with extensions to regression, classification, and more general predictive settings (Romano et al., 2020). Related work on selective prediction studies decisions with abstention and the tradeoff between prediction frequency and reliability (Chow, 1970; El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2017; 2019). More recently, conformal risk control extends calibration from coverage to general monotone losses (Angelopoulos et al., 2024), while Learn then Test (LTT) casts risk controlled policy selection as a multiple testing problem (Angelopoulos et al., 2025). StreamLTT retains the LTT architecture of testing a fixed candidate family and certifying the rejected policies simultaneously, but replaces the stationary population risk with average predictable risk.

A complementary literature studies conformal calibration when data arrive sequentially or exchangeability fails. Adaptive Conformal Inference and its extensions use sequential feedback to maintain coverage under distribution shift (Gibbs & Candes, 2021; Zaffran et al., 2022; Gibbs & Candès, 2024), while SPCI exploits temporal structure in sequential conformity scores (Xu & Xie, 2023). These methods primarily adapt predictive calibration as new observations arrive. StreamLTT instead keeps the candidate policy family fixed and asks which policies satisfy a deployment risk constraint. Online calibration and StreamLTT therefore operate at complementary stages: prediction can adapt over time, while certification evaluates a fixed family of deployment policies.

A.2 Inference under Dependence

Classical inference under temporal dependence often retains the target estimand while modifying its sampling approximation. HAC methods estimate long run variance under serial dependence (Newey & West, 1987; Andrews, 1991), while block bootstrap methods preserve dependence by resampling consecutive observations (Künsch, 1989; Politis & Romano, 1992; 1994; Lahiri, 2003). These approaches motivate our HAC and block bootstrap baselines. They address how an empirical statistic should be normalized or resampled under dependence, but do not by themselves specify which deployment risk should replace a stationary marginal expectation when conditional risk changes along the stream.

Martingale concentration provides a complementary route to sequential inference. Classical bounded increment and variance sensitive inequalities include Hoeffding, Bentkus, and Freedman type bounds (Hoeffding, 1963; Bentkus, 2004; Freedman, 1975), with self normalized extensions developed for martingale processes (de la Peña et al., 2009). More recent work uses nonnegative supermartingales to obtain time uniform concentration and confidence sequences (Howard et al., 2020; 2021), including betting based inference for bounded means (Waudby-Smith & Ramdas, 2024). StreamLTT uses this martingale perspective after defining the risk target: centering each loss by its one step predictable risk yields a martingale difference sequence. Boundedness then gives the finite sample test, while predictable quadratic variation supports the variance adaptive construction.

A.3 Simultaneous Policy Certification

Certification across a finite policy family additionally requires multiplicity control. Classical procedures such as Bonferroni and Holm provide control of the family wise error rate, with related stepwise and closed testing procedures providing further refinements (Holm, 1979; Marcus et al., 1976; Hochberg, 1988). LTT connects these tools to policy selection by interpreting rejection of an unsafe risk null as certification of the corresponding policy (Angelopoulos et al., 2025). StreamLTT preserves this multiple testing layer: once valid p values for predictable risk are available, standard

family wise correction yields a simultaneously certified region and permits subsequent selection within that region.

B PROOFS FOR STREAMLTT

This appendix provides the proofs and supporting arguments for Section 3. We follow the order of the main text: the finite-sample StreamLTT test, the variance-adaptive StreamLTT test, consistency of the predictable-variance estimator, and simultaneous certification with post-certification selection.

B.1 PROOF OF THE FINITE-SAMPLE STREAMLTT TEST

Proof of Theorem 3.2. Fix $\theta_j \in \Theta$ and suppress the argument θ_j. Recall $p_n = \mathbb{E}[L_n \mid \mathcal{F}_{n-1}]$ and $Z_n = L_n - p_n$. By construction, $\mathbb{E}[Z_n \mid \mathcal{F}_{n-1}] = 0$, so $\{Z_n\}_{n\geq 1}$ is a martingale-difference sequence with respect to $\{\mathcal{F}_n\}_{n\geq 0}$.

Since $0 \leq L_n \leq B_j$ and conditional expectation preserves these bounds, $0 \leq p_n \leq B_j$. Conditional on $\mathcal{F}_{n-1}$,

$$-p_n \leq Z_n \leq B_j - p_n,$$

so Z_n lies in an interval of length B_j. Hoeffding's lemma therefore gives, for every $s > 0$,

$$\mathbb{E}\big[e^{-sZ_n} \mid \mathcal{F}_{n-1}\big] \leq \exp\left(\frac{s^2 B_j^2}{8}\right).$$

Define

$$M_k := \exp\left(-s\sum_{n=1}^{k} Z_n - \frac{ks^2 B_j^2}{8}\right), \qquad M_0 := 1.$$

Then $\{M_k, \mathcal{F}_k\}_{k=0}^N$ is a nonnegative supermartingale, since

$$\mathbb{E}[M_k \mid \mathcal{F}_{k-1}] = M_{k-1} e^{-s^2 B_j^2/8} \mathbb{E}\big[e^{-sZ_k} \mid \mathcal{F}_{k-1}\big] \leq M_{k-1}.$$

Hence $\mathbb{E}[M_N] \leq 1$.

For any $t > 0$, on the event $\{\sum_{n=1}^N Z_n \leq -Nt\}$,

$$M_N \geq \exp\left(sNt - \frac{Ns^2 B_j^2}{8}\right).$$

Markov's inequality therefore yields

$$\mathbb{P}\left(\sum_{n=1}^{N} Z_n \leq -Nt\right) \leq \exp\left(-sNt + \frac{Ns^2 B_j^2}{8}\right).$$

Optimizing over $s > 0$ at $s = 4t/B_j^2$ gives

$$\mathbb{P}\Big(\widehat{R}_N - R_{\mathrm{pred},N} \leq -t\Big) \leq \exp\left(-\frac{2Nt^2}{B_j^2}\right), \tag{13}$$

where Eq. (6) was used.

Fix $z \in (0,1)$ and define $t_z := B_j\sqrt{\log(1/z)/(2N)}$. By Eq. (7), $p_j^{\mathrm{H}} \leq z$ implies $\widehat{R}_N \leq \alpha - t_z$. Under $H_j : R_{\mathrm{pred},N} > \alpha$,

$$\widehat{R}_N - R_{\mathrm{pred},N} < -t_z.$$

Thus

$$\{p_j^{\mathrm{H}} \leq z\} \subseteq \{\widehat{R}_N - R_{\mathrm{pred},N} \leq -t_z\}.$$

Applying Eq. (13),

$$\mathbb{P}_{H_j}\big(p_j^{\mathrm{H}} \leq z\big) \leq \exp\left(-\frac{2Nt_z^2}{B_j^2}\right) = z.$$

The cases $z = 0$ and $z = 1$ are immediate. Hence p_j^{H} is finite-sample super-uniform under $H_j : R_{\mathrm{pred},N}(\theta_j) > \alpha$. □

B.2 Proof of the Variance-Adaptive StreamLTT Test

Proof of Theorem 3.4. Fix $\theta_j \in \Theta$ and suppress the argument θ_j. Recall $p_n = \mathbb{E}[L_n \mid \mathcal{F}_{n-1}]$ and $Z_n = L_n - p_n$. By construction, $\{Z_n\}_{n\geq 1}$ is a martingale-difference sequence with respect to $\{\mathcal{F}_n\}_{n\geq 0}$, and Eq. (6) gives

$$N\big(\widehat{R}_N - R_{\mathrm{pred},N}\big) = \sum_{n=1}^{N} Z_n.$$

Since $0 \leq L_n \leq B_j$, we have $0 \leq p_n \leq B_j$ and therefore $|Z_n| \leq B_j$. By Assumption 3.3, there exists a deterministic sequence $v_{N,j}^2 > 0$ such that $v_{N,j}^2 \to \infty$ and

$$\frac{V_N^2}{v_{N,j}^2} \xrightarrow{p} 1,$$

where V_N^2 is defined in Eq. 8.

For any $\varepsilon > 0$, $v_{N,j} \to \infty$ implies $\varepsilon v_{N,j} > B_j$ for all sufficiently large N. Hence

$$\frac{1}{v_{N,j}^2} \sum_{n=1}^{N} \mathbb{E}\big[Z_n^2 \mathbb{I}\{|Z_n| > \varepsilon v_{N,j}\} \mid \mathcal{F}_{n-1}\big] = 0$$

eventually. Thus the conditional Lindeberg condition holds.

Combining the conditional Lindeberg condition with predictable-variance stabilization, the martingale central limit theorem gives

$$\frac{\sum_{n=1}^{N} Z_n}{v_{N,j}} \Rightarrow \mathcal{N}(0,1).$$

Moreover, $V_N / v_{N,j} \xrightarrow{p} 1$, so Slutsky's theorem yields

$$\frac{\sum_{n=1}^{N} Z_n}{V_N} \Rightarrow \mathcal{N}(0,1).$$

By the variance-consistency condition in Theorem 3.4, $\widehat{V}_N / V_N \xrightarrow{p} 1$. Applying Slutsky's theorem again,

$$T_N := \frac{N\big(\widehat{R}_N - R_{\mathrm{pred},N}\big)}{\widehat{V}_N} \Rightarrow \mathcal{N}(0,1). \tag{14}$$

The statistic underlying Eq. (9) can be written as

$$\frac{N(\widehat{R}_N - \alpha)}{\widehat{V}_N} = T_N + \frac{N(R_{\mathrm{pred},N} - \alpha)}{\widehat{V}_N}.$$

Under $H_j : R_{\mathrm{pred},N} > \alpha$, the second term is positive. Since Φ is increasing,

$$p_j^{\mathrm{SL}} \geq \Phi(T_N).$$

Consequently, for every $z \in (0,1)$,

$$\{p_j^{\mathrm{SL}} \leq z\} \subseteq \{T_N \leq \Phi^{-1}(z)\}.$$

Using Eq. (14),

$$\limsup_{N\to\infty} \mathbb{P}_{H_j}\big(p_j^{\mathrm{SL}} \leq z\big) \leq \lim_{N\to\infty} \mathbb{P}\big(T_N \leq \Phi^{-1}(z)\big) = \Phi\big(\Phi^{-1}(z)\big) = z.$$

The endpoints $z = 0$ and $z = 1$ are immediate. Hence p_j^{SL} is asymptotically super-uniform under $H_j : R_{\mathrm{pred},N}(\theta_j) > \alpha$. □

B.3 Proof of Predictable-Variance Consistency

Proof of Lemma 3.6. Fix $\theta_j \in \Theta$ and suppress the argument θ_j. Since $L_n \in \{0, 1\}$,

$$\mathbb{E}[Z_n^2 \mid \mathcal{F}_{n-1}] = p_n(1 - p_n).$$

Therefore,

$$V_N^2 = \sum_{n=1}^{N} p_n(1 - p_n),$$

while Eq. (10) gives

$$\widehat{V}_N^2 = \sum_{n=1}^{N} \widehat{p}_n(1 - \widehat{p}_n).$$

Let $g(x) := x(1 - x)$ for $x \in [0, 1]$. Since $|g(x) - g(y)| \le |x - y|$ for all $x, y \in [0, 1]$,

$$\left| \frac{\widehat{V}_N^2 - V_N^2}{N} \right| \le \frac{1}{N} \sum_{n=1}^{N} |\widehat{p}_n - p_n| \le \left[\frac{1}{N} \sum_{n=1}^{N} (\widehat{p}_n - p_n)^2 \right]^{1/2} = o_p(1),$$

where the second inequality follows from Cauchy–Schwarz.

By the assumed nondegenerate variance limit $V_N^2/N \xrightarrow{p} \sigma_j^2 > 0$,

$$\frac{\widehat{V}_N^2}{V_N^2} - 1 = \frac{(\widehat{V}_N^2 - V_N^2)/N}{V_N^2/N} \xrightarrow{p} 0.$$

Hence $\widehat{V}_N^2/V_N^2 \xrightarrow{p} 1$, and the continuous mapping theorem gives

$$\frac{\widehat{V}_N}{V_N} \xrightarrow{p} 1.$$

Finally, define $v_{N,j}^2 := N\sigma_j^2$. Then $v_{N,j}^2 \to \infty$ and

$$\frac{V_N^2}{v_{N,j}^2} = \frac{V_N^2/N}{\sigma_j^2} \xrightarrow{p} 1.$$

Thus Assumption 3.3 also holds. □

B.4 Proof of Simultaneous StreamLTT Certification

Proof of Theorem 3.7. At horizon N, define the unsafe-policy index set

$$\mathcal{H}_{0,N} := \{j \in \{1, \dots, M\} : R_{\mathrm{pred},N}(\theta_j) > \alpha\},$$

and let $\mathcal{R}_N$ denote the rejection set returned by the multiple-testing procedure. By definition,

$$\widehat{\Theta}_{\mathrm{safe}} = \{\theta_j : j \in \mathcal{R}_N\}, \qquad \Theta_{\mathrm{safe},N} = \{\theta_j : R_{\mathrm{pred},N}(\theta_j) \le \alpha\}.$$

Under this convention, false certification occurs exactly when at least one unsafe policy is rejected:

$$\left\{\widehat{\Theta}_{\mathrm{safe}} \not\subseteq \Theta_{\mathrm{safe},N}\right\} = \{\mathcal{R}_N \cap \mathcal{H}_{0,N} \neq \emptyset\}. \tag{15}$$

Suppose first that the policy-level p-values are finite-sample super-uniform under their corresponding null hypotheses and that the multiple-testing procedure controls the FWER at level δ. Then

$$\mathbb{P}(\mathcal{R}_N \cap \mathcal{H}_{0,N} \neq \emptyset) \le \delta.$$

Together with Eq. (15), this gives

$$\mathbb{P}\left(\widehat{\Theta}_{\mathrm{safe}} \not\subseteq \Theta_{\mathrm{safe},N}\right) \le \delta,$$

or equivalently,

$$\mathbb{P}\Big(\widehat{\Theta}_{\mathrm{safe}} \subseteq \Theta_{\mathrm{safe},N}\Big) \geq 1-\delta.$$

We next establish the post-certification statement. Let $\widehat{\theta}$ be any possibly data-dependent selection satisfying $\widehat{\theta} \in \widehat{\Theta}_{\mathrm{safe}}$ whenever $\widehat{\Theta}_{\mathrm{safe}} \neq \emptyset$. On the event $\widehat{\Theta}_{\mathrm{safe}} \subseteq \Theta_{\mathrm{safe},N}$, every policy in the certified region satisfies the predictable-risk constraint. Therefore,

$$\Big\{\widehat{\Theta}_{\mathrm{safe}} \neq \emptyset,\ R_{\mathrm{pred},N}(\widehat{\theta}) > \alpha\Big\} \subseteq \Big\{\widehat{\Theta}_{\mathrm{safe}} \not\subseteq \Theta_{\mathrm{safe},N}\Big\}. \tag{16}$$

It follows that

$$\mathbb{P}\Big(\widehat{\Theta}_{\mathrm{safe}} \neq \emptyset,\ R_{\mathrm{pred},N}(\widehat{\theta}) > \alpha\Big) \leq \delta.$$

This proves both finite-sample statements.

For the variance-adaptive StreamLTT test, consider Bonferroni correction, which rejects candidate j whenever $p_j^{\mathrm{SL}} \leq \delta/M$. By the union bound,

$$\mathbb{P}(\mathcal{R}_N \cap \mathcal{H}_{0,N} \neq \emptyset) \leq \sum_{j=1}^{M} \mathbb{P}\left(j \in \mathcal{H}_{0,N},\ p_j^{\mathrm{SL}} \leq \frac{\delta}{M}\right). \tag{17}$$

For each fixed candidate θ_j, Theorem 3.4 gives

$$\limsup_{N\to\infty} \mathbb{P}\left(j \in \mathcal{H}_{0,N},\ p_j^{\mathrm{SL}} \leq \frac{\delta}{M}\right) \leq \frac{\delta}{M}.$$

Since M is fixed and finite, taking the limit superior in Eq. (17) yields

$$\limsup_{N\to\infty} \mathbb{P}(\mathcal{R}_N \cap \mathcal{H}_{0,N} \neq \emptyset) \leq \sum_{j=1}^{M} \limsup_{N\to\infty} \mathbb{P}\left(j \in \mathcal{H}_{0,N},\ p_j^{\mathrm{SL}} \leq \frac{\delta}{M}\right) \leq \sum_{j=1}^{M} \frac{\delta}{M} = \delta.$$

Using Eq. (15),

$$\limsup_{N\to\infty} \mathbb{P}\Big(\widehat{\Theta}_{\mathrm{safe}} \not\subseteq \Theta_{\mathrm{safe},N}\Big) \leq \delta,$$

which is equivalent to

$$\liminf_{N\to\infty} \mathbb{P}\Big(\widehat{\Theta}_{\mathrm{safe}} \subseteq \Theta_{\mathrm{safe},N}\Big) \geq 1-\delta.$$

Finally, Eq. (16) holds for every N. Hence

$$\limsup_{N\to\infty} \mathbb{P}\Big(\widehat{\Theta}_{\mathrm{safe}} \neq \emptyset,\ R_{\mathrm{pred},N}(\widehat{\theta}) > \alpha\Big) \leq \limsup_{N\to\infty} \mathbb{P}\Big(\widehat{\Theta}_{\mathrm{safe}} \not\subseteq \Theta_{\mathrm{safe},N}\Big) \leq \delta.$$

This proves both asymptotic statements. □

C Experimental Details

This section provides additional details for the experiments in Section 4, including the forecasting protocol, construction of the deployment risks, certification methods, and evaluation metrics. All model configurations, auxiliary quantities, and test specific choices are fixed before certification.

C.1 Datasets and Forecasting Protocol

We consider seven real world forecasting datasets spanning energy, traffic, and temperature forecasting. The energy datasets are Electricity (Trindade, 2015), Appliances Energy (Candanedo et al., 2017), and Household Power (Hebrail & Berard, 2006); the traffic datasets are PEMS04 and PEMS08 (Guo et al., 2019); and the temperature datasets are Beijing Weather (Chen, 2017) and Jena Weather (Max Planck Institute for Biogeochemistry, 2016). The datasets differ in sequence length, sampling frequency, and

Table 3: Summary of the forecasting datasets used in the experiments.

Dataset	Domain	Forecasting task
Electricity	Energy	Load forecasting
Appliances	Energy	Energy use forecasting
Household Power	Energy	Power forecasting
PEMS04	Traffic	Traffic forecasting
PEMS08	Traffic	Traffic forecasting
Beijing	Temperature	Weather forecasting
Jena	Temperature	Weather forecasting

temporal structure, while their chronological order is preserved throughout preprocessing, forecasting, and certification.

Each dataset is split chronologically into a training set (20%), a validation set (2%), a calibration set (8%), and a certification stream (70%). A temporal convolutional network (TCN) is trained for one step forecasting on the training set, selected using the validation set, and then frozen. The calibration set determines auxiliary quantities required by the deployment risks, such as severe miss tolerances and regime thresholds, which remain fixed throughout certification. The final 70% of observations form the certification stream and are processed sequentially in their original temporal order.

Forecasting model. We use the same TCN forecasting architecture across all seven datasets. The model takes the previous 24 observations of the target variable as input and predicts the next outcome; Electricity additionally includes temporal and customer information. Inputs and targets are standardized using statistics from the training set. Models are trained by minimizing mean squared error with validation based early stopping, and the selected model is frozen before calibration and certification. The resulting forecasts are shared across all certification methods.

At each certification round, predictions and deployment decisions use only information available before the corresponding outcome is revealed. History dependent decisions may additionally use previously revealed outcomes, whereas unrevealed outcomes remain unavailable.

Table 3 summarizes the application domains and forecasting tasks.

C.2 Deployment Risk Construction

We consider the seven binary deployment risks summarized in Table 1. They span pointwise forecasting failures, regime dependent events, directional decisions, and history dependent reversal errors. Each loss takes values in $\{0, 1\}$, matching the Bernoulli setting used for predictable variance estimation in Section 3.2.

For the interval based risks, let $\widehat{\mu}_i$ denote the frozen point forecast and define

$$C_m(X_i) = [\widehat{\mu}_i - ms_i,\ \widehat{\mu}_i + ms_i],$$

where s_i is a calibration based residual scale fixed before certification, and m indexes the candidate interval widths.

Interval miss. The interval miss loss is $L_i^{\mathrm{int}}(m) = \mathbb{I}\{Y_i \notin C_m(X_i)\}$, with target $\alpha = 0.10$.

Lower tail miss. The lower tail loss isolates outcomes below the lower prediction limit, $L_i^{\mathrm{low}}(m) = \mathbb{I}\{Y_i < \ell_i(m)\}$, with $\alpha = 0.10$.

Severe interval miss. Let e_0 denote an additional tolerance determined from the calibration set. We define $L_i^{\mathrm{sev}}(m) = \mathbb{I}\{|Y_i - \widehat{\mu}_i| > ms_i + e_0\}$, with $\alpha = 0.10$. This risk captures errors exceeding the ordinary uncertainty scale by the prescribed tolerance.

Task regime miss. Let $G_i \in \{0, 1\}$ indicate an extreme operating regime determined from the calibration set. The corresponding loss is $L_i^{\mathrm{reg}}(m) = \mathbb{I}\{G_i = 1,\ Y_i \notin C_m(X_i)\}$, with $\alpha = 0.05$. This is a joint event risk rather than miscoverage conditional on $G_i = 1$.

Opposite direction. Let $a_i(\theta) \in \{-1, 0, 1\}$ denote the action issued by policy θ and $z_i^d \in \{-1, 0, 1\}$ the realized direction. We define $L_i^{\mathrm{dir}}(\theta) = \mathbb{I}\{a_i(\theta) z_i^d = -1\}$, with $\alpha = 0.10$.

Missed reversal. Let $r_i \in \{0, 1\}$ denote the realized trend reversal indicator and $\widehat{r}_i(\theta) \in \{0, 1\}$ the corresponding policy prediction. The loss is $L_i^{\text{miss}}(\theta) = \mathbb{I}\{r_i = 1, \widehat{r}_i(\theta) = 0\}$, with $\alpha = 0.10$.

False reversal. The complementary false alert loss is $L_i^{\text{false}}(\theta) = \mathbb{I}\{r_i = 0, \widehat{r}_i(\theta) = 1\}$, with $\alpha = 0.10$. Together, the missed and false reversal risks distinguish the two asymmetric failure modes of reversal detection.

All quantities defining these risks are either fixed using precertification data or constructed from information available before the corresponding outcome is revealed.

C.3 Certification Methods

Our primary implementation, ***StreamLTT-A***, uses the variance adaptive test in Theorem 3.4. We compare it with five baselines: ***IID-LTT*** (Angelopoulos et al., 2025), ***LTT-HB-Holm*** (Hoeffding, 1963; Bentkus, 2004), ***HAC-CLT-Holm*** (Newey & West, 1987), ***CircularBlockBootstrap-Holm*** (Politis & Romano, 1992), and ***EB-Howard-Bonferroni*** (Howard et al., 2020). We additionally evaluate ***StreamLTT-H***, which uses the same average predictable risk target as StreamLTT-A but replaces the adaptive test with the finite sample test in Theorem 3.2.

IID-LTT and LTT-HB-Holm both retain the standard LTT population risk under independent calibration losses. IID-LTT uses a CLT approximation to empirical risk, whereas LTT-HB-Holm uses the finite sample Hoeffding–Bentkus bounded loss test. HAC-CLT-Holm instead adjusts the long run variance for temporal dependence, while CircularBlockBootstrap-Holm approximates the dependent sampling distribution through circular block resampling. EB-Howard-Bonferroni uses an empirical Bernstein concentration construction.

For StreamLTT-A, the one step predictable risk $p_n(\theta) = \mathbb{E}[L_n(\theta) \mid \mathcal{F}_{n-1}]$ is estimated using only information available before Y_n is revealed. Given $\widehat{p}_n(\theta)$, the predictable quadratic variation is estimated as

$$\widehat{V}_N^2(\theta) = \sum_{n=1}^{N} \widehat{p}_n(\theta)\big(1 - \widehat{p}_n(\theta)\big),$$

as in Eq. equation 10. Pointwise risks use the current forecasting information and policy quantities, while regime and history dependent risks add the corresponding information available in $\mathcal{F}_{n-1}$. All estimator configurations are fixed before certification.

StreamLTT-H retains the same predictable risk formulation but uses the finite sample martingale Hoeffding p value in Eq. equation 7. Comparing StreamLTT-H with StreamLTT-A therefore isolates predictable variance adaptation while holding the certified risk target fixed.

C.4 Evaluation Protocol and Metrics

We evaluate all $7 \times 7 = 49$ dataset–risk combinations at $\delta \in \{0.10, 0.20\}$ using 2,000 independent repetitions per configuration.

A family wise error occurs when at least one certified policy exceeds the prescribed predictable risk target,

$$\exists\, \theta \in \widehat{\Theta}_{\text{safe}} \quad \text{such that} \quad R_{\text{pred},N}(\theta) > \alpha.$$

Empirical FWER is the frequency of this event across repetitions.

For each nonempty certified region, define

$$R_{\max} := \max_{\theta \in \widehat{\Theta}_{\text{safe}}} R_{\text{pred},N}(\theta),$$

and let $\overline{R}_{\max}$ denote its mean across nonempty repetitions. We measure certification tightness by the risk gap $\alpha - \overline{R}_{\max}$ and family wise budget utilization by the FWER gap $\delta - \widehat{\text{FWER}}$. Values near zero indicate that the corresponding certification boundary is approached closely.

As a complementary measure, the certified fraction $|\widehat{\Theta}_{\text{safe}}|/|\Theta|$ records the proportion of candidate policies retained after certification. Monte Carlo uncertainty is computed from the 2,000 repetitions. For empirical FWER $\widehat{q}$, its Monte Carlo standard error is $\sqrt{\widehat{q}(1-\widehat{q})/2000}$; we additionally compute

Table 4: Dataset level certification results at $\delta = 0.10$ for all seven deployment risks. Each entry reports $\overline{R}_{\max}/\widehat{\mathrm{FWER}}$.

Risk	Method	Electricity	Appliances	Household	PEMS04	PEMS08	Beijing	Jena
Interval miss ($\alpha = .10$)	IID-LTT	.051/.028	.046/.013	.048/.020	.051/.017	.044/.015	.047/.019	.045/.012
	LTT-HB-Holm	.050/.006	.045/.004	.047/.009	.051/.006	.043/.006	.046/.005	.045/.004
	HAC-CLT-Holm	.051/.035	.046/.014	.048/.025	.051/.021	.044/.019	.047/.024	.045/.017
	BlockBootstrap-Holm	.051/.034	.046/.015	.048/.025	.052/.021	.044/.019	.047/.023	.045/.018
	EB-Howard-Bonf.	.050/.000	.045/.000	.047/.000	.050/.000	.043/.000	.046/.000	.044/.000
	StreamLTT-A	**.096/.084**	**.096/.082**	**.097/.102**	**.096/.094**	**.095/.087**	**.096/.089**	**.096/.083**
Lower tail miss ($\alpha = .10$)	IID-LTT	.055/.045	.069/.031	.053/.032	.065/.025	.066/.026	.069/.020	.064/.025
	LTT-HB-Holm	.053/.014	.069/.010	.052/.009	.065/.009	.066/.009	.069/.004	.063/.007
	HAC-CLT-Holm	.055/.052	.069/.037	.053/.037	.066/.030	.066/.031	.070/.023	.064/.032
	BlockBootstrap-Holm	.055/.052	.069/.036	.053/.036	.066/.031	.066/.031	.070/.023	.064/.031
	EB-Howard-Bonf.	.052/.000	.068/.000	.052/.000	.065/.000	.065/.000	.069/.000	.063/.000
	StreamLTT-A	**.097/.095**	**.096/.097**	**.095/.093**	**.097/.080**	**.097/.088**	**.096/.081**	**.097/.085**
Severe miss ($\alpha = .10$)	IID-LTT	.054/.083	.080/.082	.063/.079	.075/.090	.069/.091	.072/.097	.073/.096
	LTT-HB-Holm	.051/.029	.078/.026	.060/.024	.073/.029	.066/.028	.070/.030	.071/.031
	HAC-CLT-Holm	.054/.092	.080/.094	.063/.087	.075/.097	.069/.099	.073/.105	.073/.107
	BlockBootstrap-Holm	.054/.091	.080/.093	.063/.086	.075/.096	.069/.098	.073/.105	.073/.105
	EB-Howard-Bonf.	.050/.000	.078/.000	.059/.000	.072/.000	.065/.000	.069/.000	.070/.000
	StreamLTT-A	**.072/.085**	**.086/.082**	**.063/.080**	**.075/.090**	**.069/.091**	**.072/.097**	**.073/.096**
Task regime miss ($\alpha = .05$)	IID-LTT	.025/.097	.016/.016	.032/.017	.040/.048	.040/.059	.025/.015	.038/.050
	LTT-HB-Holm	.024/.033	.016/.005	.032/.005	.040/.017	.040/.018	.025/.004	.038/.017
	HAC-CLT-Holm	.026/.099	.016/.018	.032/.021	.040/.049	.040/.060	.025/.018	.038/.051
	BlockBootstrap-Holm	.026/.100	.016/.017	.032/.021	.040/.048	.040/.060	.025/.017	.038/.050
	EB-Howard-Bonf.	.023/.000	.016/.000	.032/.000	.039/.000	.040/.000	.025/.000	.038/.000
	StreamLTT-A	**.034/.097**	**.044/.117**	**.047/.086**	**.046/.093**	**.045/.105**	**.048/.084**	**.045/.096**
Opposite direction ($\alpha = .10$)	IID-LTT	.095/.085	.094/.086	.083/.085	.088/.087	.091/.088	.074/.088	.054/.089
	LTT-HB-Holm	.095/.032	.094/.029	.082/.027	.088/.026	.090/.029	.072/.026	.052/.034
	HAC-CLT-Holm	.095/.095	.094/.097	.083/.091	.089/.099	.091/.098	.074/.101	.055/.099
	BlockBootstrap-Holm	.095/.097	.094/.096	.084/.092	.089/.098	.091/.098	.074/.098	.055/.099
	EB-Howard-Bonf.	.087/.000	.091/.000	.082/.000	.087/.000	.090/.000	.071/.000	.050/.000
	StreamLTT-A	**.097/.087**	**.094/.087**	**.083/.086**	**.097/.087**	**.097/.088**	**.092/.090**	**.055/.091**
Missed reversal ($\alpha = .10$)	IID-LTT	.080/.087	.064/.026	.075/.040	.048/.016	.050/.018	.089/.091	.081/.047
	LTT-HB-Holm	.077/.030	.063/.008	.074/.015	.048/.005	.049/.006	.086/.031	.080/.018
	HAC-CLT-Holm	.081/.099	.064/.033	.075/.045	.049/.022	.050/.023	.090/.099	.081/.052
	BlockBootstrap-Holm	.081/.097	.064/.033	.075/.047	.049/.022	.050/.023	.090/.099	.081/.053
	EB-Howard-Bonf.	.073/.000	.063/.000	.074/.000	.048/.001	.049/.000	.079/.000	.080/.000
	StreamLTT-A	**.097/.088**	**.095/.092**	**.095/.085**	**.095/.090**	**.095/.089**	**.096/.093**	**.082/.092**
False reversal ($\alpha = .10$)	IID-LTT	.089/.088	.094/.086	.084/.081	.049/.018	.045/.020	.091/.105	.076/.098
	LTT-HB-Holm	.089/.029	.094/.033	.083/.028	.048/.004	.044/.006	.090/.028	.074/.030
	HAC-CLT-Holm	.090/.098	.094/.096	.084/.091	.049/.022	.045/.023	.091/.110	.076/.111
	BlockBootstrap-Holm	.090/.100	.094/.096	.084/.089	.049/.022	.045/.024	.091/.110	.076/.107
	EB-Howard-Bonf.	.088/.000	.090/.000	.082/.000	.048/.000	.044/.000	.090/.000	.073/.000
	StreamLTT-A	**.097/.088**	**.097/.086**	**.096/.081**	**.096/.086**	**.096/.088**	**.098/.105**	**.094/.100**

binomial confidence intervals for FWER and Monte Carlo uncertainty for $\overline{R}_{\max}$ and the certified fraction.

D Additional Experimental Results

This section complements Section 4.2 with complete dataset level results for all seven deployment risks at both family wise levels, the finite sample StreamLTT ablation, and Monte Carlo diagnostics.

D.1 Dataset Level Results at $\delta = 0.10$

Table 4 reports the complete dataset level comparison at $\delta = 0.10$. Each entry gives $\overline{R}_{\max}/\widehat{\mathrm{FWER}}$, allowing frontier tightness and false certification to be compared jointly across datasets and risks.

The complete results show that the aggregate behavior in Section 4.2 is not driven by a single dataset or risk. The largest separation appears for interval and lower tail miss, where StreamLTT-A consistently moves the frontier close to $\alpha = 0.10$ while the competing methods remain substantially below it. Similar gains appear for missed and false reversal, particularly on datasets where the baseline frontier is far from the target.

Table 5: Dataset level certification results at $\delta = 0.20$. Each entry reports $\overline{R}_{\max}/\widehat{\mathrm{FWER}}$.

Risk	Method	Electricity	Appliances	Household	PEMS04	PEMS08	Beijing	Jena
Interval miss ($\alpha = .10$)	IID-LTT	.052/.054	.046/.022	.049/.038	.052/.035	.044/.032	.048/.033	.046/.029
	LTT-HB-Holm	.050/.016	.046/.009	.048/.016	.051/.011	.043/.011	.046/.011	.045/.008
	HAC-CLT-Holm	.053/.060	.047/.029	.049/.046	.053/.044	.045/.041	.048/.038	.046/.038
	BlockBootstrap-Holm	.053/.061	.047/.028	.049/.046	.053/.044	.045/.042	.048/.038	.046/.038
	EB-Howard-Bonf.	.050/.000	.045/.000	.047/.000	.050/.000	.043/.000	.046/.000	.044/.000
	StreamLTT-A	**.097/.185**	**.096/.178**	**.097/.211**	**.097/.186**	**.095/.181**	**.097/.177**	**.097/.182**
Lower tail miss ($\alpha = .10$)	IID-LTT	.057/.095	.070/.063	.055/.062	.066/.050	.067/.051	.070/.035	.065/.059
	LTT-HB-Holm	.054/.030	.069/.021	.052/.017	.065/.018	.066/.018	.069/.014	.063/.018
	HAC-CLT-Holm	.057/.102	.070/.070	.055/.068	.067/.058	.068/.065	.070/.044	.065/.064
	BlockBootstrap-Holm	.057/.101	.070/.070	.055/.068	.067/.057	.068/.065	.070/.043	.065/.065
	EB-Howard-Bonf.	.052/.000	.068/.000	.052/.000	.065/.000	.065/.000	.069/.000	.063/.000
	StreamLTT-A	**.098/.184**	**.096/.183**	**.096/.199**	**.098/.175**	**.098/.185**	**.097/.175**	**.097/.184**
Severe miss ($\alpha = .10$)	IID-LTT	.059/.180	.082/.173	.067/.179	.078/.188	.072/.189	.075/.195	.076/.185
	LTT-HB-Holm	.053/.056	.079/.057	.062/.058	.074/.067	.068/.066	.071/.067	.072/.063
	HAC-CLT-Holm	.060/.194	.082/.186	.067/.191	.078/.202	.072/.198	.076/.202	.076/.199
	BlockBootstrap-Holm	.059/.192	.082/.185	.067/.188	.078/.199	.072/.197	.076/.203	.076/.198
	EB-Howard-Bonf.	.050/.000	.078/.000	.059/.000	.072/.000	.065/.000	.069/.000	.070/.000
	StreamLTT-A	**.075/.185**	**.088/.175**	**.067/.180**	**.078/.188**	**.072/.190**	**.075/.196**	**.076/.189**
Task regime ($\alpha = .05$)	IID-LTT	.028/.189	.017/.033	.032/.034	.040/.093	.041/.105	.026/.032	.039/.096
	LTT-HB-Holm	.025/.064	.016/.007	.032/.010	.040/.029	.040/.038	.025/.009	.038/.034
	HAC-CLT-Holm	.028/.190	.017/.035	.033/.038	.040/.098	.041/.111	.026/.034	.039/.099
	BlockBootstrap-Holm	.028/.188	.017/.035	.033/.040	.040/.096	.041/.108	.026/.034	.039/.100
	EB-Howard-Bonf.	.023/.000	.016/.000	.032/.000	.039/.000	.040/.000	.025/.000	.038/.000
	StreamLTT-A	**.036/.189**	**.045/.217**	**.048/.193**	**.046/.191**	**.045/.212**	**.048/.192**	**.046/.197**
Opposite direction ($\alpha = .10$)	IID-LTT	.096/.187	.095/.178	.085/.172	.090/.180	.092/.190	.077/.189	.060/.196
	LTT-HB-Holm	.095/.061	.094/.064	.083/.060	.088/.057	.090/.063	.073/.056	.053/.063
	HAC-CLT-Holm	.096/.198	.095/.191	.085/.186	.090/.191	.092/.203	.077/.199	.060/.207
	BlockBootstrap-Holm	.096/.196	.095/.189	.085/.183	.090/.190	.092/.203	.077/.199	.060/.206
	EB-Howard-Bonf.	.087/.000	.092/.000	.082/.000	.087/.000	.090/.000	.071/.000	.050/.000
	StreamLTT-A	**.097/.188**	**.095/.180**	**.085/.179**	**.098/.182**	**.098/.192**	**.093/.192**	**.060/.200**
Missed reversal ($\alpha = .10$)	IID-LTT	.084/.173	.065/.058	.076/.085	.050/.042	.051/.041	.092/.191	.082/.090
	LTT-HB-Holm	.079/.061	.064/.018	.074/.030	.048/.011	.049/.011	.088/.064	.081/.035
	HAC-CLT-Holm	.085/.182	.065/.066	.076/.096	.050/.051	.051/.047	.093/.206	.082/.100
	BlockBootstrap-Holm	.084/.180	.065/.066	.076/.095	.050/.051	.051/.046	.093/.204	.082/.100
	EB-Howard-Bonf.	.073/.000	.063/.000	.074/.000	.048/.001	.049/.000	.079/.000	.080/.000
	StreamLTT-A	**.098/.173**	**.095/.191**	**.095/.192**	**.096/.186**	**.096/.185**	**.097/.195**	**.084/.201**
False reversal ($\alpha = .10$)	IID-LTT	.091/.183	.095/.178	.085/.163	.050/.043	.046/.039	.092/.209	.079/.199
	LTT-HB-Holm	.089/.066	.094/.058	.083/.057	.048/.010	.045/.012	.091/.075	.075/.070
	HAC-CLT-Holm	.091/.197	.095/.193	.085/.178	.050/.050	.047/.043	.092/.226	.079/.210
	BlockBootstrap-Holm	.091/.195	.095/.192	.085/.177	.050/.049	.047/.043	.092/.224	.079/.208
	EB-Howard-Bonf.	.088/.000	.091/.000	.082/.000	.048/.000	.044/.000	.090/.000	.073/.000
	StreamLTT-A	**.098/.186**	**.097/.179**	**.097/.163**	**.097/.192**	**.097/.198**	**.098/.214**	**.095/.202**

Other risks illustrate why frontier tightness and FWER must be interpreted jointly. For severe miss and opposite direction, HAC-CLT-Holm and the block bootstrap often attain empirical FWER near 0.10, but their frontiers remain close to IID-LTT on several datasets. Conversely, StreamLTT-A maintains FWER near the prescribed level while improving the frontier where additional admissible policies are available. The attainable frontier can itself vary by dataset, as seen for opposite direction on Jena and for severe miss on several streams.

D.2 Dataset Level Results at $\delta = 0.20$

Table 5 reports the corresponding dataset level results after increasing the family wise level to $\delta = 0.20$. This comparison shows how each procedure responds when a larger false certification budget is available.

The larger family wise budget produces two distinct behaviors. StreamLTT-A responds across nearly all dataset–risk pairs: empirical FWER generally moves toward 0.20, while the certified frontier remains close to the corresponding risk target whenever the candidate family contains policies near that boundary. For example, the interval and lower tail frontiers remain around 0.095–0.098 across all seven datasets, while their empirical FWER values move from the neighborhood of 0.10 at $\delta = 0.10$ toward 0.20.

Table 6: Risk wise results for StreamLTT-H, averaged across the seven datasets. $\overline{R}_{\max}$ denotes the mean maximum predictable risk, FWER the empirical family wise error rate, and Frac. the mean certified fraction.

	$\delta = 0.10$			$\delta = 0.20$		
Risk	$\overline{R}_{\max}$	FWER	Frac.	$\overline{R}_{\max}$	FWER	Frac.
Interval miss	.092	.000	.322	.093	.001	.325
Lower tail miss	.094	.000	.699	.094	.001	.704
Severe miss	.070	.000	.981	.070	.001	.982
Task regime miss	.042	.000	.647	.043	.000	.652
Opposite direction	.086	.000	.947	.086	.001	.953
Missed reversal	.091	.000	.864	.091	.001	.866
False reversal	.093	.000	.809	.094	.001	.814

The baseline response is less uniform. HAC-CLT-Holm and BlockBootstrap-Holm approach the larger FWER budget for some risks, such as severe miss and opposite direction, but remain substantially below it for interval, lower tail, task regime, and several reversal settings. LTT-HB-Holm remains more conservative throughout, while EB-Howard-Bonferroni leaves essentially the entire family wise budget unused. Moreover, increasing δ changes many baseline frontiers only modestly. For interval miss, for example, their $\overline{R}_{\max}$ values remain around 0.04–0.05, whereas StreamLTT-A remains near 0.10.

D.3 Finite Sample StreamLTT Ablation

To complement the aggregate ablation in Section 4.2, Table 6 reports the risk wise results for StreamLTT-H after averaging across the seven datasets. StreamLTT-H retains the same average predictable risk target as StreamLTT-A but uses the finite sample martingale Hoeffding test, isolating the effect of removing predictable variance adaptation.

Across risks, StreamLTT-H already produces relatively tight certified frontiers, while its empirical FWER remains essentially zero at both family wise levels. Increasing δ from 0.10 to 0.20 changes the frontier and certified fraction only modestly, consistent with the conservatism of the finite sample Hoeffding test. Together with the StreamLTT-A results reported above, this risk wise breakdown supports the aggregate ablation: the predictable risk formulation accounts for most of the frontier tightening, whereas predictable variance adaptation primarily enables the procedure to use more of the available family wise budget.