UNCERTAINTY-DRIVEN UNSUPERVISED DETECTION OF RELIABILITY-RELEVANT DISTRIBUTION DRIFT

Anonymous authors
Paper under double-blind review

ABSTRACT

Unsupervised concept drift detection commonly relies on identifying changes in the distribution of inputs to a trained machine learning model. Yet not every distribution shift should trigger an alarm: a deployment distribution may differ from recent observations, while remaining covered by prediction-relevant variations encountered during training. Drift thus should be triggered when incoming data introduce novel predictive behavior that was insufficiently supported by the training data. We show that this distinction is naturally captured by epistemic uncertainty (EU). Focusing on Bayesian neural networks (NNs) obtained through a Laplace approximation, we derive upper and lower bounds on EU in a local prediction-relevant geometry induced jointly by the trained model and its training data. The bounds reveal that EU along training-covered directions is attenuated according to the information accumulated during training, whereas components orthogonal to the training-supported subspace retain substantially larger EU. Guided by this characterization, we develop an unsupervised drift detector that monitors EU rather than input-distribution change directly, and can be applied post hoc to conventionally trained NNs. Experiments across non-stationary learning settings show that the proposed approach suppresses alarms on training-covered shifts while reliably detecting novel ones.

1 INTRODUCTION

Machine learning systems are increasingly deployed in non-stationary environments, where the data-generating conditions evolve over time (Giannini et al., 2024). Such changes can invalidate assumptions learned during training and lead to substantial degradation in predictive performance (Gama et al., 2004). This has motivated extensive research on *drift detection*, which aims to identify changes in a data stream that may require intervention, such as model adaptation (Gunasekara et al., 2023; Lu et al., 2018). Often in practice, labels are unavailable, and drifts must then be detected in an *unsupervised* manner, based on the observed inputs, typically by identifying changes in the input distribution.

Yet, a detectable change in the input distribution does not necessarily indicate that the deployed model has entered an unfamiliar regime (Webb et al., 2016). A model may be trained using data collected across multiple operating conditions, and deployment may transition between distributions that differ from one another while all being adequately represented during training. We refer to such changes as *training-covered* distribution shifts. In contrast, a *novel* distribution shift exposes the model to prediction-relevant variations that were insufficiently represented during training. Thus, for model monitoring, the relevant question is not merely whether the input distribution has changed, but whether it activates predictive behaviors and parameter directions that were insufficiently constrained during training.

Distinguishing between training-covered and novel shifts is challenging from unlabeled observations. A discrepancy between input distributions alone cannot resolve this distinction, as it quantifies statistical change without accounting for what the predictive model has learned. Predictive confidence by itself is also insufficient: a model may be uncertain on intrinsically difficult or ambiguous inputs even when such inputs are well represented during training (Kendall & Gal, 2017). What is needed instead is a model-aware notion of confidence that reflects specifically whether the current observations require predictive behavior that was insufficiently supported by the training data.

This perspective motivates using *epistemic uncertainty (EU)* as a measure of whether the observations are supported by the knowledge acquired in training (Gawlikowski et al., 2023; Ovadia et al., 2019). In this work, we make this intuition precise for Bayesian neural networks (NNs) obtained through a

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

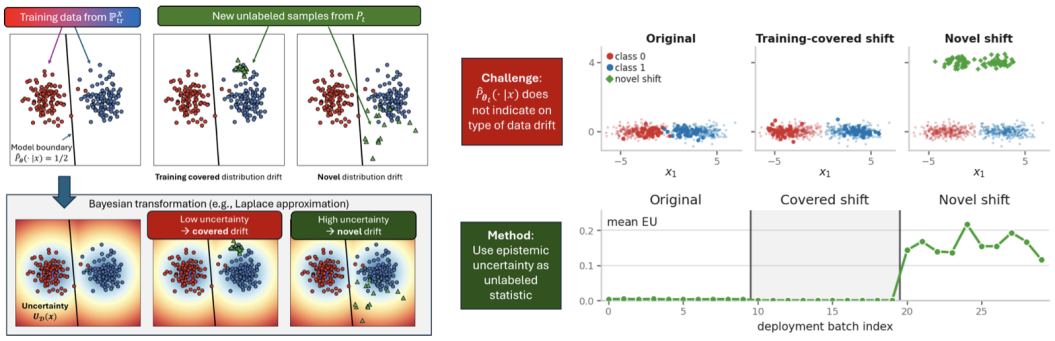


Figure 1: *Left:* Methodology illustration: inherent challenge in identifying drifts from unsupervised data (upper), and the proposed Bayesian transformation for uncertainty extraction as drift statistic (lower). *Right:* Drift statistic behavior for a 2D Bayesian logistic regression model with Laplace approximation.

Laplace approximation (Ritter et al., 2018). We characterize their EU in a local prediction-relevant geometry induced jointly by the model and its training data, and derive upper and lower bounds in terms of the components of its prediction-relevant representation that are aligned with, or orthogonal to, the directions represented during training. The characterization shows that uncertainty along training-covered directions is attenuated according to the information accumulated in training, whereas novel orthogonal directions retain substantially larger uncertainty. Thus, EU reflects novelty not merely in the raw input space, but in a model-induced geometry that captures which variations matter to the predictive task.

Guided by this characterization, we propose an unsupervised drift detection framework that monitors EU rather than changes in the input distribution, as illustrated in Fig. 1. For each incoming unlabeled batch, we extract EU from the predictor, aggregate it into a scalar monitoring statistic, and supply the resulting sequence to a sequential change detector. The detector is therefore designed to remain insensitive to transitions that are statistically different yet training-covered, while responding when the incoming data exhibit prediction-relevant variations that were not sufficiently supported during training. While the framework naturally applies to Bayesian predictors, we also instantiate it for conventionally trained NNs using a post-hoc Laplace approximation (Daxberger et al., 2021), allowing epistemic uncertainty to be extracted without modifying the original training procedure. Our main contributions are:

- We formulate unsupervised drift detection around the distinction between *training-covered* and *novel* distribution shifts, where the latter correspond to prediction-relevant variations that are insufficiently supported by the training data.
- We characterize EU under a Laplace approximation, deriving bounds that relate it to aspects of prediction-relevant variations lying within and outside the subspace supported by the training data. This rigorously explains why EU can distinguish training-covered from novel shifts.
- Building on this characterization, we develop an uncertainty-driven unsupervised drift detector that can be applied to conventionally trained NNs through post-hoc Laplace approximation, and show that it can suppress alarms on training-covered shifts while detecting novel ones.

2 Related Works

Unsupervised Concept Drift Detection. Concept drift detection accommodates multiple approaches differing in the signal used to identify changes and their dependence on labels (Gama et al., 2014; Khamassi et al., 2018). Classical performance-based detectors, such as DDM (Gama et al., 2004), monitor prediction errors and therefore require labels during deployment. *Unsupervised* drift detection refers to monitoring without online labels, and includes both fully unsupervised methods that operate solely on the observed stream and pre-trained methods that may exploit labeled data before deployment (Gemaque et al., 2020). Fully unsupervised detectors typically compare recent and historical observations, possibly after constructing statistical or learned representations, and detect changes through distributional discrepancies or hypothesis tests (Lukats et al., 2025); recent learning-based approaches such as MCD-DD (Wan et al., 2024) learn representations in which changes between portions of the stream can be identified without labels. Such methods are generally agnostic to whether the change introduces behav-

ior that is novel relative to the knowledge acquired by the predictor during training. A closer line of work therefore constructs model-dependent monitoring signals: STUDD (Cerqueira et al., 2023) monitors the discrepancy between a deployed teacher and an auxiliary student trained to mimic its predictions, UDD (Baier et al., 2023) applies a sequential detector to predictive uncertainty extracted through Monte Carlo dropout. Closely related to our motivation, UBDD (Doty et al., 2026) argues that distributional changes corresponding to operating conditions represented in training need not warrant intervention, and uses changes in predictive entropy to identify performance-degrading shifts. These uncertainty-based approaches, however, rely on predictive entropy or related confidence measures, which conflate uncertainty due to insufficient training support with uncertainty inherent in ambiguous observations. Moreover, they do not characterize when the monitored uncertainty reflects prediction-relevant coverage by the training data. Our work addresses this gap by theoretically relating EU to such training coverage and exploiting this relation to distinguish training-covered from novel distribution shifts.

Distribution-Shift and Out-of-Distribution Detection. Our setting is related to dataset-shift and out-of-distribution (OOD) detection (Yang et al., 2024). Dataset-shift detection seeks to determine whether deployment data differ statistically from a reference distribution, commonly through two-sample tests applied either to the inputs or to learned representations (Rabanser et al., 2019). OOD detection typically assigns individual observations a score indicating their compatibility with the training distribution. Such scores have been constructed from prediction confidence (Hendrycks & Gimpel, 2017), learned representations and their distances to the training data (Lee et al., 2018; Sun et al., 2022), and energy functions (Liu et al., 2020). More recent approaches increasingly exploit the geometry induced by the predictive model itself, for example through distances to its decision boundaries (Liu & Qin, 2024), feature geometry relative to classifier weights (Liu & Qin, 2025), or decompositions of learned activations according to their relevance to the classifier output (Zöngür et al., 2025). These methods provide model-aware notions of sample-wise novelty, but their objective remains to distinguish observations regarded as OOD. Our objective differs in that a deployment distribution may change substantially, and may even be statistically distinct from the training distribution, without constituting novel drift if its prediction-relevant variations remain supported by what the model learned during training. We therefore seek to detect persistent shifts that introduce prediction-relevant directions insufficiently constrained by the training data, rather than merely observations that receive a large OOD score.

Epistemic Uncertainty Estimation. Predictive uncertainty in NNs can arise from fundamentally different sources, commonly distinguished as *aleatoric uncertainty*, associated with ambiguity or noise inherent in the observations, and *EU*, associated with limited knowledge of the predictive model (Kendall & Gal, 2017; Depeweg et al., 2018). The latter is particularly relevant in our setting, as it is expected to decrease when prediction-relevant behavior is sufficiently supported by training data. Bayesian NNs provide a principled framework for representing such model uncertainty by placing a posterior distribution over the network parameters, although exact Bayesian inference is generally intractable for modern architectures. This has motivated approximate inference methods including variational Bayesian learning (Blundell et al., 2015), Monte Carlo dropout (Gal & Ghahramani, 2016), and Laplace approximations (Daxberger et al., 2021; Morad et al., 2026). The behavior of these uncertainty estimates under distribution shift has been studied extensively, with empirical evidence showing substantial differences in their ability to remain informative as test data move away from the training distribution (Ovadia et al., 2019). Our work differs from this literature in that we do not primarily seek a new uncertainty estimator. Rather, we theoretically characterize the EU induced by a Laplace posterior in terms of prediction-relevant coverage by the training data, and use this characterization to determine which distribution shifts should trigger an unsupervised drift alarm.

3 Problem Setup

3.1 Non-Stationary Data Model

We consider a C-class classification problem with input space $\mathcal{X}$ and finite label space $\mathcal{Y} = \{1, \dots, C\}$. The predictive model is trained offline using data collected from a family of K training environments. Environment $k \in \{1, \dots, K\}$ is characterized by a data-generating distribution $P_{\mathrm{tr}}^{(k)}(x, y) = P_{\mathrm{tr}}^{(k)}(y|x)P_{\mathrm{tr}}^{(k)}(x)$, and provides a training set of N_k labeled samples

$$\mathcal{D}_k = \left\{ \left(x_i^{(k)}, y_i^{(k)}\right) \right\}_{i=1}^{N_k}, \qquad \left(x_i^{(k)}, y_i^{(k)}\right) \overset{\text{i.i.d.}}{\sim} P_{\mathrm{tr}}^{(k)}. \tag{1}$$

The complete training set is $\mathcal{D}_{\mathrm{tr}} := \bigcup_{k=1}^{K} \mathcal{D}_k$. We use $\mathbb{P}_{\mathrm{tr}}^{X} := \left\{P_{\mathrm{tr}}^{(k)}(x)\right\}_{k=1}^{K}$ to denote the input distributions represented by the training environments. Whether a deployment distribution is training-covered will be defined relative to the prediction-relevant information learned from $\mathcal{D}_{\mathrm{tr}}$.

During deployment, data arrive sequentially in blocks indexed by $t \in \mathbb{N}$. Within block t, the data-generating distribution is $P_t(x, y) = P_t(y|x)P_t(x)$, which may vary across blocks. The monitoring mechanism has access only to an unlabeled batch of B samples, denoted

$$\mathcal{B}_t := \left\{x_i^{(t)}\right\}_{i=1}^{B}, \qquad x_i^{(t)} \overset{\text{i.i.d.}}{\sim} P_t(x). \tag{2}$$

3.2 Problem Formulation

Predictive Model. We consider a parameterized predictive model, e.g., a NN, with parameters $\boldsymbol{\theta} \in \mathbb{R}^p$ and predictive distribution $\hat{P}_{\boldsymbol{\theta}}(y|x)$. The model is trained using the aggregate dataset $\mathcal{D}_{\mathrm{tr}}$ collected from the K training environments defined in Subsection 3.1. We denote the resulting fixed parameter estimate by $\widehat{\boldsymbol{\theta}}$ and the deployed predictive distribution by $\hat{P}_{\widehat{\boldsymbol{\theta}}}(y|x)$.

Training-Covered and Novel Distribution Shifts. A change in the deployment distribution from $P_{t-1}(x)$ to $P_t(x)$ does not necessarily indicate that the model has encountered an unfamiliar operating regime. For example, $P_t(x)$ may correspond to a different training environment, or may constitute a statistically new distribution whose prediction-relevant variations were nevertheless sufficiently represented in $\mathcal{D}_{\mathrm{tr}}$. We refer to such a change as a *training-covered distribution shift*. Consequently, a conventional unsupervised detector that directly compares successive input distributions may signal drift even though the incoming data remain covered by what the predictor learned during training.

A *novel distribution shift* introduces prediction-relevant variations that were insufficiently represented by the training data in (1). Thus, the notion of coverage considered is model-dependent: it is determined not solely by statistical similarity between $P_t(x)$ and the preceding distributions, but by whether the incoming observations rely on prediction-relevant behavior that was sufficiently constrained during training. In Section 4 we formalize this notion through the local geometry of the trained predictor.

Unsupervised Monitoring Objective. Our objective is to construct a label-free monitoring statistic $T_t = \mathcal{T}\big(\mathcal{B}_t; \widehat{\boldsymbol{\theta}}\big)$, that is indicative of whether or not the current distribution represents a novel distribution shift. The statistic is then supplied to a sequential detector

$$d_t = \mathcal{A}(T_1, \ldots, T_t) \in \{0, 1\}, \tag{3}$$

where $d_t = 1$ indicates a detected novel distribution shift.

The desired detector should satisfy two complementary properties: (i) it should remain insensitive to transitions among training-covered environments, even when these transitions induce a substantial change in the marginal input distribution; and (ii) it should respond when the current observations activate prediction-relevant behavior that was insufficiently constrained by the training data.

4 Uncertainty-Based Detection of Novel Distribution Shifts

The monitoring objective in Subsection 3.2 calls for a model-aware unsupervised drift statistic. We use EU for this purpose. The underlying intuition is that a Bayesian posterior encodes which parameter directions were constrained by the training data and which remain uncertain. Consequently, when the prediction at a given input is primarily sensitive to well-constrained directions, uncertainty due to the model parameters should be limited; when it is sensitive to directions that remain poorly constrained after training, its EU should increase. Our analysis below makes this intuition precise and relates it to the prediction-relevant notion of training coverage introduced in Section 3.

We begin by equipping the predictor with a linearized Laplace approximation, which provides a tractable Bayesian representation of its parameter and predictive uncertainty. We then characterize EU geometrically, showing how it depends on the components of the local predictive behavior that lie within and outside the directions supported by the training data. This yields a model-induced notion of training coverage and provides the theoretical basis for distinguishing training-covered from novel distribution shifts. This insight is then translated into a monitoring statistic for the incoming batches and combined with a sequential change detector. Proofs of the theoretical results are provided in Appendix C.

4.1 Bayesian Predictive Model and Epistemic Uncertainty

Bayesian Predictive Model. We first formulate EU for a general Bayesian predictive model. Let $q(\boldsymbol{\theta}|\mathcal{D}_{\text{tr}})$ denote a posterior distribution over the model parameters, with $\hat{P}_{\boldsymbol{\theta}}(y|x)$ being the predictive distribution associated with a parameter realization $\boldsymbol{\theta}$. The posterior predictive distribution is

$$\overline{P}_{\mathcal{D}_{\text{tr}}}(y|x) := \mathbb{E}_{q(\boldsymbol{\theta}|\mathcal{D}_{\text{tr}})}\left[\hat{P}_{\boldsymbol{\theta}}(y|x)\right]. \tag{4}$$

We quantify EU through the mutual information between the predicted label and the model parameters conditioned on the input and training data. For classification, it can be expressed as

$$U_{\mathcal{D}_{\text{tr}}}(x) := I\left(Y;\boldsymbol{\theta}|x,\mathcal{D}_{\text{tr}}\right) = H\left[\overline{P}_{\mathcal{D}_{\text{tr}}}(\cdot|x)\right] - \mathbb{E}_{q(\boldsymbol{\theta}|\mathcal{D}_{\text{tr}})}\left[H\left[\hat{P}_{\boldsymbol{\theta}}(\cdot|x)\right]\right]. \tag{5}$$

The first term captures the total predictive uncertainty after averaging over plausible model parameters, whereas the second captures the uncertainty within predictions with different parameter realizations. Their difference therefore reflects predictive variability induced by uncertainty about the model parameters. While EU can be estimated for any Bayesian predictive model from which parameter samples can be obtained, our theoretical analysis focuses on a linearized Laplace construction, detailed next.

Linearized Laplace Approximation. The Laplace approximation can equip a conventionally trained NN with a post-hoc parameter posterior, without requiring Bayesian training from scratch (Ritter et al., 2018; Daxberger et al., 2021). For the theoretical construction, we take $\widehat{\boldsymbol{\theta}}$ to be a local maximum a-posteriori probability (MAP) solution under a Gaussian parameter prior and approximate the posterior locally around this solution. Let $\boldsymbol{z}_{\boldsymbol{\theta}}(x) \in \mathbb{R}^C$ denote the model logits and define their Jacobian at $\widehat{\boldsymbol{\theta}}$ as

$$\boldsymbol{G}_x := \nabla_{\boldsymbol{\theta}} \boldsymbol{z}_{\boldsymbol{\theta}}(x)|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}. \tag{6}$$

The linearized model approximates the logits as $\boldsymbol{z}_{\boldsymbol{\theta}}(x) \approx \boldsymbol{z}_{\widehat{\boldsymbol{\theta}}}(x) + \boldsymbol{G}_x\big(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\big)$. We further denote the predictive probabilities at the expansion point by $\widehat{\boldsymbol{p}}(x) := \operatorname{softmax}\big(\boldsymbol{z}_{\widehat{\boldsymbol{\theta}}}(x)\big)$, and define the corresponding categorical Fisher matrix in logit space as $\boldsymbol{H}_x := \operatorname{diag}\left(\widehat{\boldsymbol{p}}(x)\right) - \widehat{\boldsymbol{p}}(x)\widehat{\boldsymbol{p}}(x)^\top$.

For a training sample $x_i^{(k)}$, let $\boldsymbol{G}_i^{(k)} := \boldsymbol{G}_{x_i^{(k)}}$ and $\boldsymbol{H}_i^{(k)} := \boldsymbol{H}_{x_i^{(k)}}$. Using the generalized Gauss–Newton approximation to the likelihood curvature, the resulting Laplace posterior is

$$q_{\text{LA}}\left(\boldsymbol{\theta} \mid \mathcal{D}_{\text{tr}}\right) = \mathcal{N}\left(\widehat{\boldsymbol{\theta}}, \boldsymbol{\Sigma}_{\mathcal{D}_{\text{tr}}}\right), \tag{7}$$

where the posterior precision matrix is given by

$$\boldsymbol{\Sigma}_{\mathcal{D}_{\text{tr}}}^{-1} = \boldsymbol{\Lambda}_0 + \sum_{k=1}^{K}\sum_{i=1}^{N_k}\left(\boldsymbol{G}_i^{(k)}\right)^\top \boldsymbol{H}_i^{(k)}\boldsymbol{G}_i^{(k)}, \tag{8}$$

with $\boldsymbol{\Lambda}_0 \succ \boldsymbol{0}$ denotes the Gaussian prior precision. The derivation and additional details are provided in Appendix B. For clarity, the theoretical analysis below considers the isotropic case $\boldsymbol{\Lambda}_0 = \lambda \boldsymbol{I}$; the general positive-definite case can be reduced to this form through an appropriate whitening transformation.

We focus on the Laplace approximation for two main reasons: (i) it enables the proposed monitoring framework to be applied post hoc to conventionally trained NNs, without requiring Bayesian training from scratch; and (ii) its Gaussian parameter posterior and curvature-based covariance admit an analytically tractable characterization of EU. This allows us to relate EU at a deployment input to the prediction-relevant information accumulated collectively from the training environments.

4.2 Geometric Characterization of Epistemic Uncertainty

Local EU. In the following, we define the prediction-relevant tangent feature matrix $\boldsymbol{B}_x := \boldsymbol{H}_x^{1/2}\boldsymbol{G}_x$. The matrix $\boldsymbol{B}_x$ captures the local sensitivity of the predictive distribution at x to perturbations in the model parameters, with logit variations weighted according to their effect on the categorical predictive distribution. The following result provides the local form of the EU induced by the posterior (7):

Proposition 4.1 (Local EU characterization)**.** *Consider the linearized Laplace predictive model in Subsection 4.1, and define $\boldsymbol{V}_x := \boldsymbol{G}_x\boldsymbol{\Sigma}_{\mathcal{D}_{\text{tr}}}\boldsymbol{G}_x^\top$ where $\rho_x := \sqrt{\operatorname{tr}(\boldsymbol{V}_x)}$. Then, along any sequence of Laplace posteriors for which $\rho_x \to 0$, the EU at every fixed $x \in \mathcal{X}$ satisfies*

$$U_{\mathcal{D}_{\text{tr}}}(x) = \frac{1}{2}\operatorname{tr}\left(\boldsymbol{H}_x\boldsymbol{V}_x\right) + \mathcal{O}(\rho_x^3) = \frac{1}{2}\operatorname{tr}\left(\boldsymbol{B}_x\boldsymbol{\Sigma}_{\mathcal{D}_{\text{tr}}}\boldsymbol{B}_x^\top\right) + \mathcal{O}(\rho_x^3). \tag{9}$$

Proposition 4.1 identifies the leading-order term of EU when the posterior-induced predictive perturbations are local. We therefore define $\widetilde{U}_{\mathcal{D}_{\mathrm{tr}}}(x) := \frac{1}{2}\,\mathrm{tr}\left(\boldsymbol{B}_x \boldsymbol{\Sigma}_{\mathcal{D}_{\mathrm{tr}}} \boldsymbol{B}_x^\top\right)$ as the *local EU*, which is determined by the interaction between the prediction-relevant sensitivity $\boldsymbol{B}_x$ at the input and the parameter directions that remain uncertain after training, as represented by $\boldsymbol{\Sigma}_{\mathcal{D}_{\mathrm{tr}}}$. In particular, uncertainty increases when the current input is sensitive to parameter directions that retain substantial posterior variance. Combined with the precision structure in (8), this observation enables us to characterize whether an input activates prediction-relevant directions that were constrained by the training environments.

Prediction-Relevant Training Coverage. To make this connection explicit, we characterize the parameter directions that capture prediction-relevant information from the training data. Recall from (8) that, under the isotropic prior $\boldsymbol{\Lambda}_0 = \lambda \boldsymbol{I}$, the posterior covariance can be written as

$$\boldsymbol{\Sigma}_{\mathcal{D}_{\mathrm{tr}}} = \left(\lambda \boldsymbol{I} + \boldsymbol{M}_{\mathcal{D}_{\mathrm{tr}}}\right)^{-1}, \qquad \boldsymbol{M}_{\mathcal{D}_{\mathrm{tr}}} := \sum_{k=1}^{K}\sum_{i=1}^{N_k} \left(\boldsymbol{B}_i^{(k)}\right)^\top \boldsymbol{B}_i^{(k)}, \qquad \boldsymbol{B}_i^{(k)} := \boldsymbol{B}_{x_i^{(k)}}. \tag{10}$$

The matrix $\boldsymbol{M}_{\mathcal{D}_{\mathrm{tr}}}$ therefore aggregates the local prediction-relevant curvature contributed by all training samples. We define the corresponding *training-supported tangent subspace* as

$$\mathcal{S}_{\mathcal{D}_{\mathrm{tr}}} := \mathrm{range}\left(\boldsymbol{M}_{\mathcal{D}_{\mathrm{tr}}}\right) = \mathrm{span}\left\{\mathrm{range}\left(\left(\boldsymbol{B}_i^{(k)}\right)^\top\right) : k = 1, \ldots, K,\ i = 1, \ldots, N_k\right\}. \tag{11}$$

Thus, $\mathcal{S}_{\mathcal{D}_{\mathrm{tr}}}$ contains the parameter directions along which the training data provide prediction-relevant information, while its orthogonal complement consists of directions to which the training likelihood contributes no local curvature. Let $\boldsymbol{\Pi}_{\mathcal{D}_{\mathrm{tr}}}$ denote the orthogonal projector onto $\mathcal{S}_{\mathcal{D}_{\mathrm{tr}}}$ and $\boldsymbol{\Pi}_{\mathcal{D}_{\mathrm{tr}}}^{\perp} := \boldsymbol{I} - \boldsymbol{\Pi}_{\mathcal{D}_{\mathrm{tr}}}$. For a deployment input x, we accordingly decompose its tangent features as

$$\boldsymbol{B}_x = \boldsymbol{B}_x \boldsymbol{\Pi}_{\mathcal{D}_{\mathrm{tr}}} + \boldsymbol{B}_x \boldsymbol{\Pi}_{\mathcal{D}_{\mathrm{tr}}}^{\perp}. \tag{12}$$

The first component captures prediction-relevant sensitivity along directions represented during training, whereas the second captures sensitivity along directions outside the support of the training information. We quantify their energies by $\kappa_{\mathcal{D}_{\mathrm{tr}}}(x) := \left\|\boldsymbol{B}_x \boldsymbol{\Pi}_{\mathcal{D}_{\mathrm{tr}}}\right\|_{\mathrm{F}}^2$ and $\nu_{\mathcal{D}_{\mathrm{tr}}}(x) := \left\|\boldsymbol{B}_x \boldsymbol{\Pi}_{\mathcal{D}_{\mathrm{tr}}}^{\perp}\right\|_{\mathrm{F}}^2$. We refer to $\kappa_{\mathcal{D}_{\mathrm{tr}}}(x)$ as the *training-covered tangent energy*, while $\nu_{\mathcal{D}_{\mathrm{tr}}}(x)$ measures the *prediction-relevant novelty* of x relative to the aggregate training data.

Let $\gamma_{\mathcal{D}_{\mathrm{tr}}}$ and $\Gamma_{\mathcal{D}_{\mathrm{tr}}}$ denote the smallest and largest strictly positive eigenvalues of $\boldsymbol{M}_{\mathcal{D}_{\mathrm{tr}}}$, respectively. The following result relates the local EU directly to prediction-relevant training coverage.

Theorem 4.2 (EU and prediction-relevant training coverage)**.** *Suppose that $\boldsymbol{\Lambda}_0 = \lambda \boldsymbol{I}$ and $\mathcal{S}_{\mathcal{D}_{\mathrm{tr}}} \neq \{\boldsymbol{0}\}$. Then, for every $x \in \mathcal{X}$,*

$$\widetilde{U}_{\mathcal{D}_{\mathrm{tr}}}(x) = \frac{1}{2}\,\mathrm{tr}\left(\boldsymbol{B}_x \boldsymbol{\Pi}_{\mathcal{D}_{\mathrm{tr}}} \boldsymbol{\Sigma}_{\mathcal{D}_{\mathrm{tr}}} \boldsymbol{\Pi}_{\mathcal{D}_{\mathrm{tr}}} \boldsymbol{B}_x^\top\right) + \frac{\nu_{\mathcal{D}_{\mathrm{tr}}}(x)}{2\lambda}, \tag{13}$$

and consequently

$$\frac{\kappa_{\mathcal{D}_{\mathrm{tr}}}(x)}{2(\lambda + \Gamma_{\mathcal{D}_{\mathrm{tr}}})} + \frac{\nu_{\mathcal{D}_{\mathrm{tr}}}(x)}{2\lambda} \leq \widetilde{U}_{\mathcal{D}_{\mathrm{tr}}}(x) \leq \frac{\kappa_{\mathcal{D}_{\mathrm{tr}}}(x)}{2(\lambda + \gamma_{\mathcal{D}_{\mathrm{tr}}})} + \frac{\nu_{\mathcal{D}_{\mathrm{tr}}}(x)}{2\lambda}. \tag{14}$$

Theorem 4.2 yields two direct consequences that clarify the role of training coverage.

Corollary 4.3 (Training-covered tangent features)**.** *If the tangent features of x are entirely contained in the training-supported subspace, i.e., $\nu_{\mathcal{D}_{\mathrm{tr}}}(x) = 0$, then $\widetilde{U}_{\mathcal{D}_{\mathrm{tr}}}(x) \leq \frac{\|\boldsymbol{B}_x\|_{\mathrm{F}}^2}{2\left(\lambda + \gamma_{\mathcal{D}_{\mathrm{tr}}}\right)}$.*

Thus, when the prediction-relevant sensitivity of x lies entirely along directions represented during training, its local EU is attenuated according to the information accumulated along the training-supported subspace. In particular, when these directions are strongly constrained, as reflected by a large $\gamma_{\mathcal{D}_{\mathrm{tr}}}$, the resulting EU is correspondingly small relative to the tangent sensitivity $\|\boldsymbol{B}_x\|_{\mathrm{F}}^2$.

Corollary 4.4 (Uncertainty induced by prediction-relevant novelty)**.** *For every $x \in \mathcal{X}$, it holds that $\widetilde{U}_{\mathcal{D}_{\mathrm{tr}}}(x) \geq \frac{\nu_{\mathcal{D}_{\mathrm{tr}}}(x)}{2\lambda}$. Moreover, if the tangent features of x lie entirely outside the training-supported subspace, i.e., if $\kappa_{\mathcal{D}_{\mathrm{tr}}}(x) = 0$, then $\widetilde{U}_{\mathcal{D}_{\mathrm{tr}}}(x) = \frac{\|\boldsymbol{B}_x\|_{\mathrm{F}}^2}{2\lambda}$.*

Algorithm 1 Uncertainty-Based Novel Drift Detection at Block t

Require: Fixed posterior $q(\boldsymbol{\theta} \mid \mathcal{D}_{\mathrm{tr}})$, sequential detector $\mathcal{A}$, posterior sample count S

for deployment blocks $t = 1, 2, \ldots$ **do**
  Receive unlabeled batch $\mathcal{B}_t = \{x_i^{(t)}\}_{i=1}^{B}$
  for $i = 1, \ldots, B$ **do**
    Estimate $\widehat{U}_{\mathcal{D}_{\mathrm{tr}}}(x_i^{(t)})$ using (17)
  end for
  $T_t \leftarrow \frac{1}{B}\sum_{i=1}^{B} \widehat{U}_{\mathcal{D}_{\mathrm{tr}}}(x_i^{(t)})$
  $d_t \leftarrow \mathcal{A}(T_1, \ldots, T_t)$
end for

Implications for Novel Distribution Shifts. The preceding results extend from individual inputs to a deployment distribution. In particular, for $X \sim P_t(x)$, Corollary 4.4 implies that

$$\mathbb{E}_{P_t}\left[\widetilde{U}_{\mathcal{D}_{\mathrm{tr}}}(X)\right] \geq \frac{1}{2\lambda}\mathbb{E}_{P_t}\left[\nu_{\mathcal{D}_{\mathrm{tr}}}(X)\right]. \tag{15}$$

Hence, a distribution shift that introduces prediction-relevant sensitivity outside the training-supported subspace necessarily increases the corresponding contribution to the average local EU. Conversely, if a deployment distribution is tangent-covered with $\nu_{\mathcal{D}_{\mathrm{tr}}}(X) = 0$ almost surely, then Corollary 4.3 yields

$$\mathbb{E}_{P_t}\left[\widetilde{U}_{\mathcal{D}_{\mathrm{tr}}}(X)\right] \leq \frac{\mathbb{E}_{P_t}\left[\|\boldsymbol{B}_X\|_{\mathrm{F}}^2\right]}{2\left(\lambda + \gamma_{\mathcal{D}_{\mathrm{tr}}}\right)}. \tag{16}$$

These statements formalize the distinction introduced in Subsection 3.2. A deployment distribution may be statistically different from both the preceding and training distributions while remaining training-covered if its prediction-relevant tangent features lie along directions already supported and sufficiently constrained by the training data. Conversely, a shift that persistently introduces tangent components outside this support yields a nonzero novelty energy and an unavoidable contribution to EU.

4.3 Uncertainty-Based Novel Drift Detection

From Training Coverage to a Monitoring Statistic. The characterization in Subsection 4.2 provides a direct route from prediction-relevant training coverage to unsupervised drift detection. In particular, Theorem 4.2 and Corollary 4.4 show that tangent components outside the training-supported subspace contribute unavoidable uncertainty, while uncertainty along sufficiently constrained training-supported directions is attenuated. At the distribution level, this implies that persistent prediction-relevant novelty is reflected in the average EU of the incoming observations. We therefore use EU as the monitoring signal for the unlabeled stream. For deployment block t, we define the population-level monitoring quantity $T_t^\star := \mathbb{E}_{X\sim P_t}\left[U_{\mathcal{D}_{\mathrm{tr}}}(X)\right]$. Since P_t is unknown, we estimate $T_t^\star$ from the observed batch in (2).

Monte Carlo Estimation of EU. The Bayesian posterior used for monitoring is constructed offline from $\mathcal{D}_{\mathrm{tr}}$ and remains fixed during deployment. In our Laplace implementation, this is $q_{\mathrm{LA}}(\boldsymbol{\theta} \mid \mathcal{D}_{\mathrm{tr}})$ from (7). For each input $x_i^{(t)}$, we draw $\{\boldsymbol{\theta}^{(s)}\}_{s=1}^{S}$ from this posterior to evaluate $\boldsymbol{p}_i^{(s)} := \hat{P}_{\boldsymbol{\theta}^{(s)}}(\cdot \mid x_i^{(t)})$ and $\overline{\boldsymbol{p}}_i := \frac{1}{S}\sum_{s=1}^{S} \boldsymbol{p}_i^{(s)}$. Using the definition in (5), the sample-wise EU is estimated as

$$\widehat{U}_{\mathcal{D}_{\mathrm{tr}}}\left(x_i^{(t)}\right) = H\left[\overline{\boldsymbol{p}}_i\right] - \frac{1}{S}\sum_{s=1}^{S} H\left[\boldsymbol{p}_i^{(s)}\right]. \tag{17}$$

The corresponding batch-level monitoring statistic that empirically estimates $T_t^\star$ is

$$T_t := \frac{1}{B}\sum_{i=1}^{B} \widehat{U}_{\mathcal{D}_{\mathrm{tr}}}\left(x_i^{(t)}\right). \tag{18}$$

After the posterior has been constructed offline, evaluating T_t requires only the incoming unlabeled batch and the fixed Bayesian predictor; the training data are not needed during deployment.

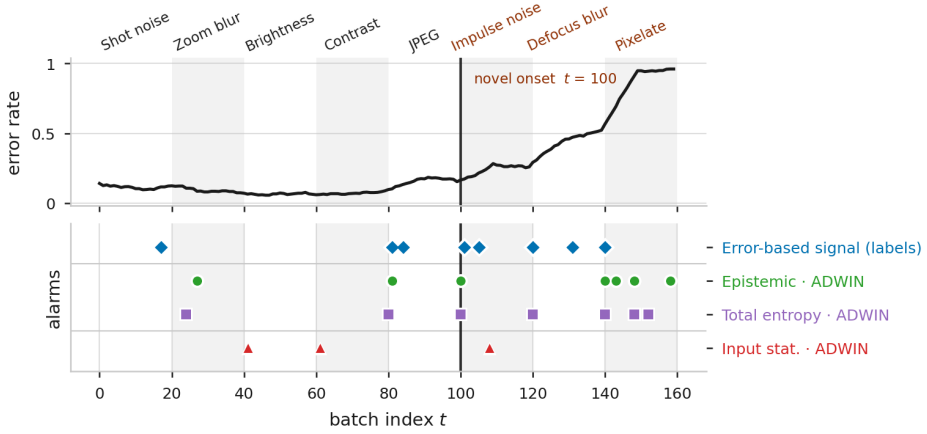


Figure 2: Unsupervised drift detection on CIFAR-100-C. The top panel shows the error rate across known and novel corruptions, with the vertical line indicating the onset of held-out corruptions. The bottom panel shows alarms produced by the supervised error-based detector using labels, uncertainty-based window detector using epistemic and total uncertainty, and model-free input-statistic window detector.

Sequential Novel Drift Detection. The scalar sequence $\{T_t\}_{t\geq 1}$ is supplied to a sequential change detector $\mathcal{A}$, yielding the decision rule (3) with the test statistic (18). The framework is agnostic to the particular sequential detector: the role of $\mathcal{A}$ is to identify persistent changes in the uncertainty sequence, while the model-aware distinction between training-covered and novel shifts is encoded in T_t. The drift detector is summarized as Algorithm 1. In our experiments, we instantiate $\mathcal{A}$ with three established sequential detectors: ADWIN (Bifet & Gavaldà, 2007), Page–Hinkley (Page, 1954), and KSWIN (Raab et al., 2020).

5 Experimental Study

Experimental Setup. We evaluate our framework with two objectives: (i) examine whether EU provides the behavior predicted by Subsection 4.2, i.e., remaining comparatively stable under training-covered distribution changes while increasing when deployment introduces prediction-relevant variations not represented during training; and (ii) test whether this distinction persists across datasets, architectures, and Bayesian constructions. We combine a controlled nonstationary corruption experiment, which exposes the temporal behavior of the test statistic, with a broader benchmark in which training-covered and novel operating conditions are varied across modalities and models. Full details of the setups considered here are provided in Appendix D, with additional supporting experimental results reported in Appendix E. We compare the EU statistic with predictive entropy, which measures total predictive uncertainty, and with a model-independent input-distribution statistic. For reference, we also report a supervised error-based rule (Gama et al., 2004).

Detecting Novel Drifts. We first illustrate the intended behavior in a controlled CIFAR-100-C stream. A ViT classifier is trained using several corruption types, while additional corruption types are held out and appear only during deployment. The stream first transitions among the training-covered corruptions and subsequently enters the held-out conditions. Figure 2 shows the resulting behavior. Input-based monitoring reacts to several transitions within the training-covered portion of the stream, illustrating its inability to identify novel drift. The uncertainty-based statistic remains substantially more stable throughout the covered conditions and changes prominently once held-out corruptions are introduced, coinciding with the deterioration in predictive performance. This behavior is consistent with the characterization in Subsection 4.2: novel conditions expose prediction-relevant behavior that is insufficiently constrained by the training data and therefore induce increased model uncertainty.

Generalization Across Tasks and Models. We next isolate the quality of the monitoring signal from the choice of detector. For each setting, we regard batches from training-covered conditions as negatives

Table 1: Drift detection performance across the evaluated settings. AUROC (↑) and FPR95 (↓), reported as mean ± standard deviation over 5 stream seeds. Best per row in **bold**, second best underlined. Error uses deployment labels and serves only as a supervised reference.

	AUROC ↑				FPR95 ↓			
	Error	EU	Entropy	Input	Error	EU	Entropy	Input
MNIST								
ResNet-20 Laplace	0.66±0.02	0.75±0.01	**0.79±0.01**	0.60±0.01	0.85±0.05	0.54±0.06	**0.46±0.06**	1.00±0.00
ViT Laplace	0.82±0.02	**0.88±0.01**	0.87±0.01	0.60±0.01	0.38±0.03	0.35±0.03	**0.30±0.01**	1.00±0.00
CIFAR-10								
ResNet-20 Laplace	**0.83±0.01**	0.80±0.01	0.81±0.02	0.28±0.03	**0.33±0.01**	**0.33±0.01**	**0.33±0.00**	1.00±0.00
ViT Laplace	0.80±0.01	**0.82±0.02**	**0.82±0.02**	0.28±0.03	0.51±0.03	**0.43±0.04**	**0.43±0.03**	1.00±0.00
ResNet-20 Laplace (sev.)	0.78±0.02	0.75±0.02	**0.81±0.02**	0.54±0.02	0.63±0.06	0.66±0.06	**0.53±0.03**	0.71±0.00
CIFAR-100								
ViT-B/16 Laplace	0.99±0.01	**1.00±0.00**	**1.00±0.00**	0.44±0.03	0.06±0.04	**0.00±0.00**	**0.00±0.00**	1.00±0.00
ESC-50								
CNN MC	**0.99±0.01**	0.98±0.01	0.59±0.02	0.30±0.01	**0.04±0.03**	0.11±0.03	0.73±0.08	1.00±0.00
ResNet-18 MC	**0.90±0.02**	0.72±0.03	0.60±0.01	0.30±0.01	**0.43±0.10**	0.65±0.08	0.75±0.07	1.00±0.00
CNN-wide MC	**1.00±0.00**	0.98±0.01	0.59±0.02	0.30±0.01	**0.03±0.03**	0.08±0.02	0.80±0.06	1.00±0.00
Speech Commands								
CNN MC	0.89±0.03	**1.00±0.00**	0.98±0.00	0.27±0.03	0.50±0.06	**0.01±0.01**	0.11±0.01	1.00±0.01
ResNet-18 MC	0.87±0.02	**0.96±0.01**	0.95±0.01	0.27±0.03	0.52±0.08	**0.18±0.04**	0.21±0.05	1.00±0.01
CNN-wide MC	0.90±0.02	**1.00±0.00**	0.98±0.00	0.27±0.03	0.35±0.04	**0.00±0.01**	0.09±0.04	1.00±0.01
UrbanSound8K								
CNN MC	**0.97±0.01**	**0.97±0.01**	0.83±0.01	0.32±0.02	**0.15±0.01**	0.16±0.05	0.62±0.07	0.98±0.01
ResNet-18 MC	**0.89±0.02**	0.86±0.01	0.72±0.03	0.32±0.02	**0.40±0.09**	0.49±0.04	0.82±0.04	0.98±0.01
CNN-wide MC	0.93±0.02	**0.98±0.00**	0.80±0.02	0.32±0.02	0.41±0.09	**0.10±0.04**	0.66±0.07	0.98±0.01
Wireless								
LOS-drift Laplace	0.79±0.05	**0.81±0.03**	0.78±0.04	0.44±0.05	**0.59±0.15**	0.60±0.10	0.74±0.16	0.97±0.02
LOS→NLOS Laplace	**0.78±0.03**	0.76±0.04	0.74±0.04	0.44±0.03	**0.47±0.02**	0.57±0.11	0.83±0.08	0.92±0.04

and batches from held-out conditions as positives, and evaluate how well each statistic separates the two using AUROC and FPR95. Table 1 reports the results across image, audio, speech, and wireless tasks, using multiple architectures and uncertainty approximations. The EU statistic consistently separates between covered and novel conditions and substantially outperforms the model-independent input statistic in most settings. This behavior is observed across both Laplace- and sampling-based uncertainty approximations, indicating that the effect is not tied to a single architecture or posterior construction.

We also observe in Table 1 that predictive entropy is competitive with EU in several image-based settings, where the contribution of intrinsic uncertainty is relatively limited and the two statistics behave similarly. The advantage of isolating EU is more pronounced for acoustic and wireless signals, which exhibit substantial intrinsic variability arising from factors such as noise, propagation, and source-dependent fluctuations. Since total predictive uncertainty combines this aleatoric variability with EU, it can provide a less specific indication of novel drift. In these cases, isolating EU yields a cleaner measure of insufficient predictive support. Thus, we do not suggest that total uncertainty is unusable for drift detection, but that EU is the more targeted statistic for identifying novelty attributable to limited model knowledge.

6 Conclusions

We proposed an uncertainty-driven framework for unsupervised detection of novel distribution shifts. Our theoretical analysis shows how EU reflects this distinction through the geometry of the Laplace posterior, while experiments across multiple modalities and model classes demonstrate that it provides a robust monitoring signal, particularly in settings with substantial intrinsic variability. These results motivate EU as a model-aware statistic for detecting reliability-relevant drift without requiring deployment labels, as experimentally demonstrated across different tasks and models.

AI DISCLOSURE

In this work, we used generative AI tools to assist with the implementation of experimental code, the plotting and presentation of experimental results, and the polishing of the manuscript for readability. Generative AI was not used to determine the datasets, research methodology, or theoretical contributions of the work. All AI-assisted outputs, including code, figures, and manuscript text, were reviewed and verified by the authors. The authors take full responsibility for the final content of this work, including all claims, results, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The code for reproducing the experiments is available in a public repository.[1] The experimental setups, model architectures, uncertainty-estimation procedures, corruption settings, detector configurations, and evaluation protocols are described in Appendices D and E. All datasets used in the experiments are identified and cited in the paper, while the wireless channel realizations are generated using the QuaDRiGa channel simulator.

REFERENCES

Lucas Baier, Tim Schlör, Jakob Schöffer, and Niklas Kühl. Detecting concept drift with neural network model uncertainty. In *Proceedings of the 56th Hawaii International Conference on System Sciences*, pp. 836–844, 2023. doi: 10.24251/HICSS.2023.104.

Albert Bifet and Ricard Gavaldà. Learning from time-changing data with adaptive windowing. In *Proceedings of the 2007 SIAM International Conference on Data Mining*, pp. 443–448. SIAM, 2007. doi: 10.1137/1.9781611972771.42.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *International Conference on Machine Learning*, pp. 1613–1622. PMLR, 2015.

Vitor Cerqueira, Heitor Murilo Gomes, Albert Bifet, and Luis Torgo. STUDD: A student–teacher method for unsupervised concept drift detection. *Machine Learning*, 112(11):4351–4378, 2023. doi: 10.1007/s10994-022-06188-7.

Gregory Cohen, Saeed Afshar, Jonathan Tapson, and André van Schaik. Emnist: Extending mnist to handwritten letters. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pp. 2921–2926. IEEE, 2017.

Erik Daxberger, Agustinus Kristiadi, Alexander Immer, Runa Eschenhagen, Matthias Bauer, and Philipp Hennig. Laplace redux-effortless Bayesian deep learning. *Advances in Neural Information Processing Systems*, 34, 2021.

Stefan Depeweg, Jose-Miguel Hernandez-Lobato, Finale Doshi-Velez, and Steffen Udluft. Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning. In *International Conference on Machine Learning*. PMLR, 2018.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Christina Doty, Bill Peria, Ryan Wood, and Jason E. Summers. From detection to correction: uncertainty-based drift detection for ml models for motion imagery. In *Proc. SPIE Assurance and Security for AI-enabled Systems*, volume 14046, pp. 140460M, 2026.

Tom Fawcett. An introduction to roc analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006. doi: 10.1016/j.patrec.2005.10.010.

[1] https://github.com/klpavrnqzmtx/uncertainty-driven-drift-detection

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pp. 1050–1059. PMLR, 2016.

Joao Gama, Pedro Medas, Gladys Castillo, and Pedro Rodrigues. Learning with drift detection. In *Brazilian Symposium on Artificial Intelligence*, pp. 286–295. Springer, 2004.

João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. A survey on concept drift adaptation. *ACM computing surveys (CSUR)*, 46(4):1–37, 2014.

Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, et al. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, 56(Suppl 1):1513–1589, 2023.

Rosana Noronha Gemaque, Albert França Josuá Costa, Rafael Giusti, and Eulanda Miranda Dos Santos. An overview of unsupervised drift detection methods. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(6):e1381, 2020.

Federico Giannini, Giacomo Ziffer, Andrea Cossu, and Vincenzo Lomonaco. Streaming continual learning for unified adaptive intelligence in dynamic environments. *IEEE Intell. Syst.*, 39(6):81–85, 2024.

Nuwan Gunasekara, Bernhard Pfahringer, Heitor Murilo Gomes, and Albert Bifet. Survey on online streaming continual learning. In *IJCAI*, pp. 6628–6637, 2023.

James A. Hanley and Barbara J. McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982. doi: 10.1148/radiology.143.1.7063747.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=HJz6tiCqYm`.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *International Conference on Learning Representations*, 2017.

Stephan Jaeckel, Leszek Raschkowski, Kai Börner, and Lars Thiele. QuaDRiGa: A 3-D multi-cell channel model with time evolution for enabling virtual field trials. *IEEE Trans. Antennas Propag.*, 62(6):3242–3256, 2014.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

Imen Khamassi, Moamar Sayed-Mouchaweh, Moez Hammami, and Khaled Ghédira. Discussion and review on evolving data streams and concept drift adapting. *Evolving systems*, 9(1):1–23, 2018.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in Neural Information Processing Systems*, 31, 2018.

Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *International Conference on Learning Representations*, 2018.

Litian Liu and Yao Qin. Fast decision boundary based out-of-distribution detector. In *International Conference on Machine Learning*. PMLR, 2024.

Litian Liu and Yao Qin. Detecting out-of-distribution through the lens of neural collapse. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15424–15433, 2025.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33:21464–21475, 2020.

Jie Lu et al. Learning under concept drift: A review. *IEEE Trans. Knowl. Data Eng.*, 31(12): 2346–2363, 2018.

Daniel Lukats, Oliver Zielinski, Axel Hahn, and Frederic Stahl. A benchmark and survey of fully unsupervised concept drift detectors on real-world data streams. *International Journal of Data Science and Analytics*, 19(1):1–31, 2025.

Itai Morad, Nir Shlezinger, and Yonina C Eldar. SGD-based knowledge distillation with Bayesian teachers: Theory and guidelines. In *International Conference on Learning Representations (ICLR)*, 2026.

Norman Mu and Justin Gilmer. Mnist-c: A robustness benchmark for computer vision. *arXiv preprint arXiv:1906.02337*, 2019. URL https://arxiv.org/abs/1906.02337.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in Neural Information Processing Systems*, 32, 2019.

Ewan S Page. Continuous inspection schemes. *Biometrika*, 41(1/2):100–115, 1954.

Karol J. Piczak. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM International Conference on Multimedia*, pp. 1015–1018, 2015. doi: 10.1145/2733373.2806390.

Christoph Raab, Moritz Heusinger, and Frank-Michael Schleif. Reactive soft prototype computing for concept drift streams. *Neurocomputing*, 416:340–351, 2020. doi: 10.1016/j.neucom.2019.11.111.

Stephan Rabanser, Stephan Günnemann, and Zachary Lipton. Failing loudly: An empirical study of methods for detecting dataset shift. *Advances in Neural Information Processing Systems*, 32, 2019.

Hippolyt Ritter, Aleksandar Botev, and David Barber. A scalable Laplace approximation for neural networks. In *International Conference on Learning Representations (ICLR)*, volume 6, 2018.

Justin Salamon, Christopher Jacoby, and Juan Pablo Bello. A dataset and taxonomy for urban sound research. In *Proceedings of the 22nd ACM International Conference on Multimedia*, pp. 1041–1044, 2014. doi: 10.1145/2647868.2655045.

Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your vit? data, augmentation, and regularization in vision transformers. *Transactions on Machine Learning Research*, 2022. URL https://openreview.net/forum?id=4nPswr1KcP.

Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *International Conference on Machine Learning*. PMLR, 2022.

Ke Wan, Yi Liang, and Susik Yoon. Online drift detection with maximum concept discrepancy. In *ACM SIGKDD conference on knowledge discovery and data mining*, pp. 2924–2935, 2024.

Pete Warden. Speech commands: A dataset for limited-vocabulary speech recognition. *arXiv preprint arXiv:1804.03209*, 2018. doi: 10.48550/arXiv.1804.03209.

Geoffrey I Webb, Roy Hyde, Hong Cao, Hai Long Nguyen, and Francois Petitjean. Characterizing concept drift. *Data Mining and Knowledge Discovery*, 30(4):964–994, 2016.

Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *International Journal of Computer Vision*, 132(12):5635–5662, 2024.

Barış Zöngür, Robin Hesse, and Stefan Roth. Activation subspaces for out-of-distribution detection. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3509–3519, 2025.

A Additional Related Work

Concept-Drift Taxonomies. Concept drift can be characterized along several complementary dimensions (Gama et al., 2014; Webb et al., 2016; Gemaque et al., 2020; Lukats et al., 2025). According to its temporal evolution, drift may occur abruptly, incrementally, gradually, or recur when a previously observed concept reappears. A separate distinction concerns the component of the data-generating process that changes. *Virtual drift* refers to changes in the input distribution $P(X)$, whereas *real drift* involves changes in the predictive relationship $P(Y \mid X)$ and can therefore directly affect model performance. These taxonomies describe how non-stationarity manifests in the data stream, but do not explicitly characterize whether the resulting observations remain supported by a particular model's training experience. Our distinction between *training-covered* and *novel* shifts is therefore complementary: different forms of real or virtual drift may either remain within the model's learned support or expose previously unsupported behavior that is relevant to its reliability.

Sequential Drift Detection. Drift detection methods generally monitor a stream of observations and compare recent behavior with a reference representing the current or previously stable concept (Gemaque et al., 2020; Lukats et al., 2025). Common approaches maintain fixed, sliding, or adaptive reference windows and signal drift when a discrepancy between the reference and recent observations becomes sufficiently large. The discrepancy may be evaluated using distance or density measures, likelihood-based statistics, or statistical tests such as Page–Hinkley, while adaptive-window approaches such as ADWIN dynamically modify the reference window as the stream evolves. These mechanisms are largely agnostic to the quantity being monitored: supervised methods commonly rely on classifier performance or prediction errors, whereas unsupervised methods instead monitor statistics that can be obtained without online labels. Our approach follows the latter setting, using epistemic uncertainty as the monitored statistic and standard sequential detectors to convert changes in this signal into drift alarms.

Unsupervised Distribution-Based Drift Detection. A major class of unsupervised drift detectors identifies changes by directly comparing the distributions of reference and recent data (Gemaque et al., 2020; Lukats et al., 2025). Existing methods differ primarily in how these windows are constructed, how the observations are represented, and how their dissimilarity is quantified. The literature includes approaches based on distances, density and likelihood estimates, clustering, dimensionality reduction, and hypothesis tests. While these methods enable drift detection without online labels, they primarily characterize changes in the observed data distribution and are therefore not necessarily tied to the predictive behavior of a particular model. In contrast, our framework retains unsupervised operation while monitoring a model-dependent quantity: epistemic uncertainty. This allows the detection process to focus specifically on changes indicating insufficient model knowledge, rather than on distributional changes alone.

B Derivation of the Laplace Posterior

This appendix derives the Gaussian parameter posterior used in (7)–(8). We first derive the standard Laplace approximation using the exact curvature of the negative log-posterior, and then specialize it to the generalized Gauss–Newton curvature employed in the main body.

B.1 Posterior Distribution and MAP Estimate

Recall that the aggregate training set consists of samples collected from K training environments:

$$\mathcal{D}_{\mathrm{tr}} = \bigcup_{k=1}^{K} \mathcal{D}_k, \qquad \mathcal{D}_k = \left\{ \left(x_i^{(k)}, y_i^{(k)} \right) \right\}_{i=1}^{N_k}. \tag{19}$$

Conditioned on the model parameters $\boldsymbol{\theta}$, we assume that the training labels are independent, such that

$$p\left(\mathcal{D}_{\mathrm{tr}} \mid \boldsymbol{\theta}\right) = \prod_{k=1}^{K} \prod_{i=1}^{N_k} \hat{P}_{\boldsymbol{\theta}} \left(y_i^{(k)} \mid x_i^{(k)} \right). \tag{20}$$

We place a Gaussian prior over the model parameters:

$$p(\boldsymbol{\theta}) = \mathcal{N}\left(\boldsymbol{\theta}; \boldsymbol{m}_0, \boldsymbol{\Lambda}_0^{-1}\right), \qquad \boldsymbol{\Lambda}_0 \succ \mathbf{0}, \tag{21}$$

where $\boldsymbol{m}_0$ and $\boldsymbol{\Lambda}_0$ denote the prior mean and precision, respectively. The commonly used isotropic zero-mean prior is obtained by setting $\boldsymbol{m}_0 = \mathbf{0}$ and $\boldsymbol{\Lambda}_0 = \lambda \boldsymbol{I}$.

By Bayes' rule, the parameter posterior satisfies

$$p\left(\boldsymbol{\theta} \mid \mathcal{D}_{\mathrm{tr}}\right) = \frac{p(\mathcal{D}_{\mathrm{tr}} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})}{p(\mathcal{D}_{\mathrm{tr}})}. \tag{22}$$

Up to an additive constant independent of $\boldsymbol{\theta}$, its negative logarithm is

$$\begin{aligned} \mathcal{L}_{\mathrm{post}}(\boldsymbol{\theta}) &:= -\log p\left(\boldsymbol{\theta} \mid \mathcal{D}_{\mathrm{tr}}\right) \\ &= \sum_{k=1}^{K} \sum_{i=1}^{N_k} \ell_i^{(k)}(\boldsymbol{\theta}) + \frac{1}{2} \left(\boldsymbol{\theta} - \boldsymbol{m}_0\right)^{\top} \boldsymbol{\Lambda}_0 \left(\boldsymbol{\theta} - \boldsymbol{m}_0\right) + \text{const.}, \end{aligned} \tag{23}$$

where

$$\ell_i^{(k)}(\boldsymbol{\theta}) := -\log \hat{P}_{\boldsymbol{\theta}}\left(y_i^{(k)} \mid x_i^{(k)}\right) \tag{24}$$

is the sample-wise negative log-likelihood. For softmax classification, $\ell_i^{(k)}$ is the cross-entropy loss.

The MAP estimate is consequently

$$\begin{aligned} \widehat{\boldsymbol{\theta}} &:= \arg\max_{\boldsymbol{\theta}} p\left(\boldsymbol{\theta} \mid \mathcal{D}_{\mathrm{tr}}\right) \\ &= \arg\min_{\boldsymbol{\theta}} \left\{ \sum_{k=1}^{K} \sum_{i=1}^{N_k} \ell_i^{(k)}(\boldsymbol{\theta}) + \frac{1}{2} \left(\boldsymbol{\theta} - \boldsymbol{m}_0\right)^{\top} \boldsymbol{\Lambda}_0 \left(\boldsymbol{\theta} - \boldsymbol{m}_0\right) \right\}. \end{aligned} \tag{25}$$

Thus, training by cross-entropy combined with the quadratic regularizer induced by (21) corresponds to MAP estimation. Cross-entropy training without the regularization term instead corresponds to maximum-likelihood estimation.

B.2 Second-Order Approximation Around the Posterior Mode

Suppose that $\widehat{\boldsymbol{\theta}}$ is a local posterior mode and that $\mathcal{L}_{\mathrm{post}}$ is twice continuously differentiable in its neighborhood. Since the MAP estimate is a stationary point,

$$\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathrm{post}}(\boldsymbol{\theta})\big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} = \mathbf{0}. \tag{26}$$

Define the posterior curvature at the mode as

$$\boldsymbol{H}_{\mathrm{post}} := \nabla_{\boldsymbol{\theta}}^2 \mathcal{L}_{\mathrm{post}}(\boldsymbol{\theta})\big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}. \tag{27}$$

For $\Delta\boldsymbol{\theta} := \boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}$, a second-order Taylor expansion gives

$$\mathcal{L}_{\mathrm{post}}\left(\widehat{\boldsymbol{\theta}} + \Delta\boldsymbol{\theta}\right) = \mathcal{L}_{\mathrm{post}}\left(\widehat{\boldsymbol{\theta}}\right) + \frac{1}{2} \Delta\boldsymbol{\theta}^{\top} \boldsymbol{H}_{\mathrm{post}} \Delta\boldsymbol{\theta} + o\left(\|\Delta\boldsymbol{\theta}\|^2\right). \tag{28}$$

The linear term is absent precisely because of the stationarity condition in (26).

Neglecting the higher-order remainder and exponentiating the negative quadratic approximation yields

$$\begin{aligned} p\left(\boldsymbol{\theta} \mid \mathcal{D}_{\mathrm{tr}}\right) &\propto \exp\left\{-\mathcal{L}_{\mathrm{post}}(\boldsymbol{\theta})\right\} \\ &\approx \exp\left\{-\frac{1}{2} \left(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\right)^{\top} \boldsymbol{H}_{\mathrm{post}} \left(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\right)\right\}. \end{aligned} \tag{29}$$

Provided that $\boldsymbol{H}_{\mathrm{post}} \succ \mathbf{0}$, normalizing (29) gives the Laplace approximation

$$q_{\mathrm{LA}}\left(\boldsymbol{\theta} \mid \mathcal{D}_{\mathrm{tr}}\right) = \mathcal{N}\left(\widehat{\boldsymbol{\theta}}, \boldsymbol{H}_{\mathrm{post}}^{-1}\right). \tag{30}$$

Consequently, the exact Laplace covariance is

$$\boldsymbol{\Sigma}_{\mathcal{D}_{\mathrm{tr}}} = \boldsymbol{H}_{\mathrm{post}}^{-1}. \tag{31}$$

The MAP property is required to center the quadratic approximation at $\widehat{\boldsymbol{\theta}}$. Had the expansion been performed around an arbitrary parameter vector $\boldsymbol{\theta}_0$ for which

$$\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathrm{post}}(\boldsymbol{\theta}_0) \neq \mathbf{0}, \tag{32}$$

the Taylor expansion would contain the linear term

$$\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\text{post}}(\boldsymbol{\theta}_0)^\top (\boldsymbol{\theta} - \boldsymbol{\theta}_0) . \tag{33}$$

Completing the square would then produce a Gaussian centered at

$$\boldsymbol{\theta}_0 - \boldsymbol{H}_0^{-1} \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\text{post}}(\boldsymbol{\theta}_0), \tag{34}$$

rather than at $\boldsymbol{\theta}_0$, where $\boldsymbol{H}_0$ is the local Hessian. Expanding at a posterior mode therefore ensures that the approximating Gaussian is centered at the trained parameters themselves.

B.3 Generalized Gauss–Newton Curvature

The exact posterior curvature in (27) can be written as

$$\boldsymbol{H}_{\text{post}} = \boldsymbol{\Lambda}_0 + \sum_{k=1}^{K} \sum_{i=1}^{N_k} \nabla_{\boldsymbol{\theta}}^2 \ell_i^{(k)}(\boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} . \tag{35}$$

Computing and storing the exact neural-network Hessian is generally impractical. Moreover, its likelihood component includes curvature arising from the nonlinear parameterization of the network. We therefore employ the generalized Gauss–Newton approximation.

For a training sample $\left(x_i^{(k)}, y_i^{(k)}\right)$, write

$$\ell_i^{(k)}(\boldsymbol{\theta}) = \widetilde{\ell}\left(\boldsymbol{z}_{\boldsymbol{\theta}}\left(x_i^{(k)}\right), y_i^{(k)}\right), \tag{36}$$

where $\widetilde{\ell}(\boldsymbol{z}, y)$ denotes the loss as a function of the logits. Recall that

$$\boldsymbol{G}_i^{(k)} := \nabla_{\boldsymbol{\theta}} \boldsymbol{z}_{\boldsymbol{\theta}}\left(x_i^{(k)}\right)\Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}, \tag{37}$$

as well as the logit-space gradient and curvature

$$\boldsymbol{g}_i^{(k)} := \nabla_{\boldsymbol{z}} \widetilde{\ell}\left(\boldsymbol{z}, y_i^{(k)}\right)\Big|_{\boldsymbol{z}=\boldsymbol{z}_{\widehat{\boldsymbol{\theta}}}(x_i^{(k)})}, \tag{38}$$

$$\boldsymbol{H}_i^{(k)} := \nabla_{\boldsymbol{z}}^2 \widetilde{\ell}\left(\boldsymbol{z}, y_i^{(k)}\right)\Big|_{\boldsymbol{z}=\boldsymbol{z}_{\widehat{\boldsymbol{\theta}}}(x_i^{(k)})} . \tag{39}$$

Applying the chain rule twice gives the exact sample-wise Hessian decomposition

$$\begin{aligned} \nabla_{\boldsymbol{\theta}}^2 \ell_i^{(k)}(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} = &\left(\boldsymbol{G}_i^{(k)}\right)^\top \boldsymbol{H}_i^{(k)} \boldsymbol{G}_i^{(k)} \\ &+ \sum_{c=1}^{C} \left[\boldsymbol{g}_i^{(k)}\right]_c \nabla_{\boldsymbol{\theta}}^2 z_{\boldsymbol{\theta},c}\left(x_i^{(k)}\right)\Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} . \end{aligned} \tag{40}$$

The first term propagates the curvature of the output-space loss through the logit Jacobian. The second term depends on the second derivatives of the network logits with respect to the parameters.

The generalized Gauss–Newton approximation discards the second term in (40), yielding

$$\nabla_{\boldsymbol{\theta}}^2 \ell_i^{(k)}(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} \approx \left(\boldsymbol{G}_i^{(k)}\right)^\top \boldsymbol{H}_i^{(k)} \boldsymbol{G}_i^{(k)}. \tag{41}$$

For softmax classification with cross-entropy loss, let

$$\widehat{\boldsymbol{p}}_i^{(k)} := \text{softmax}\left(\boldsymbol{z}_{\widehat{\boldsymbol{\theta}}}\left(x_i^{(k)}\right)\right). \tag{42}$$

The corresponding logit-space gradient and Hessian are

$$\boldsymbol{g}_i^{(k)} = \widehat{\boldsymbol{p}}_i^{(k)} - \boldsymbol{e}_{y_i^{(k)}}, \tag{43}$$

$$\boldsymbol{H}_i^{(k)} = \text{diag}\left(\widehat{\boldsymbol{p}}_i^{(k)}\right) - \widehat{\boldsymbol{p}}_i^{(k)} \left(\widehat{\boldsymbol{p}}_i^{(k)}\right)^\top, \tag{44}$$

where $\boldsymbol{e}_{y_i^{(k)}}$ is the one-hot vector associated with the label $y_i^{(k)}$. Since

$$\boldsymbol{H}_i^{(k)} \succeq \mathbf{0}, \tag{45}$$

every generalized Gauss–Newton contribution satisfies

$$\left(\boldsymbol{G}_i^{(k)}\right)^\top \boldsymbol{H}_i^{(k)} \boldsymbol{G}_i^{(k)} \succeq \mathbf{0}. \tag{46}$$

Substituting (41) into (35) gives the approximate posterior precision

$$\boldsymbol{\Sigma}_{\mathcal{D}_{\mathrm{tr}}}^{-1} = \boldsymbol{\Lambda}_0 + \sum_{k=1}^{K} \sum_{i=1}^{N_k} \left(\boldsymbol{G}_i^{(k)}\right)^\top \boldsymbol{H}_i^{(k)} \boldsymbol{G}_i^{(k)}. \tag{47}$$

Consequently, the generalized Gauss–Newton Laplace posterior is

$$q_{\mathrm{LA}}\left(\boldsymbol{\theta} \mid \mathcal{D}_{\mathrm{tr}}\right) = \mathcal{N}\left(\widehat{\boldsymbol{\theta}}, \boldsymbol{\Sigma}_{\mathcal{D}_{\mathrm{tr}}}\right), \qquad \boldsymbol{\Sigma}_{\mathcal{D}_{\mathrm{tr}}} = \left[\boldsymbol{\Lambda}_0 + \sum_{k=1}^{K} \sum_{i=1}^{N_k} \left(\boldsymbol{G}_i^{(k)}\right)^\top \boldsymbol{H}_i^{(k)} \boldsymbol{G}_i^{(k)}\right]^{-1}. \tag{48}$$

For the isotropic prior $\boldsymbol{\Lambda}_0 = \lambda \boldsymbol{I}$ with $\lambda > 0$, the approximate precision in (47) is positive definite, and the covariance in (48) is therefore well-defined.

C Detailed Proofs

For brevity, throughout this appendix we use $\mathcal{D}$ to denote the collective training set $\mathcal{D}_{\mathrm{tr}}$.

C.1 Proof of Proposition 4.1

Fix an input $x \in \mathcal{X}$, and let

$$\boldsymbol{z}_0(x) := \boldsymbol{z}_{\widehat{\boldsymbol{\theta}}}(x), \qquad \boldsymbol{p}_0(x) := \mathrm{softmax}\left(\boldsymbol{z}_0(x)\right). \tag{49}$$

Under the Laplace posterior, write

$$\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}} + \Delta\boldsymbol{\theta}, \qquad \Delta\boldsymbol{\theta} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}_{\mathcal{D}}\right). \tag{50}$$

For the linearized predictive model of Subsection 4.1, i.e., $\boldsymbol{z}_{\boldsymbol{\theta}}(x) \approx \boldsymbol{z}_0(x) + \boldsymbol{G}_x\left(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\right)$, the corresponding perturbation of the logits is

$$\Delta\boldsymbol{z}_x := \boldsymbol{G}_x \Delta\boldsymbol{\theta} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{V}_x\right), \qquad \boldsymbol{V}_x = \boldsymbol{G}_x \boldsymbol{\Sigma}_{\mathcal{D}} \boldsymbol{G}_x^\top. \tag{51}$$

Thus, within the linearized model,

$$\boldsymbol{p}_{\boldsymbol{\theta}}(x) = \mathrm{softmax}\left(\boldsymbol{z}_0(x) + \Delta\boldsymbol{z}_x\right). \tag{52}$$

Recall that $\rho_x^2 = \mathrm{tr}(\boldsymbol{V}_x)$.

The posterior predictive distribution is

$$\overline{\boldsymbol{p}}_{\mathcal{D}}(x) := \mathbb{E}_{\boldsymbol{\theta}}\left[\boldsymbol{p}_{\boldsymbol{\theta}}(x)\right]. \tag{53}$$

By the mutual-information representation of EU,

$$U_{\mathcal{D}}(x) = \mathbb{E}_{\boldsymbol{\theta}}\left[D_{\mathrm{KL}}\left(\boldsymbol{p}_{\boldsymbol{\theta}}(x) \,\|\, \overline{\boldsymbol{p}}_{\mathcal{D}}(x)\right)\right]. \tag{54}$$

For any fixed reference distribution $\boldsymbol{p}_0(x)$, we have the exact decomposition

$$\mathbb{E}_{\boldsymbol{\theta}}\left[D_{\mathrm{KL}}\left(\boldsymbol{p}_{\boldsymbol{\theta}}(x) \,\|\, \boldsymbol{p}_0(x)\right)\right] = U_{\mathcal{D}}(x) + D_{\mathrm{KL}}\left(\overline{\boldsymbol{p}}_{\mathcal{D}}(x) \,\|\, \boldsymbol{p}_0(x)\right), \tag{55}$$

which follows by expanding the KL divergences and using $\mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{p}_{\boldsymbol{\theta}}(x)] = \overline{\boldsymbol{p}}_{\mathcal{D}}(x)$.

We first characterize the term on the left-hand side of (55). Define

$$\Psi_x(\boldsymbol{z}) := D_{\mathrm{KL}}\left(\mathrm{softmax}(\boldsymbol{z}) \,\|\, \mathrm{softmax}(\boldsymbol{z}_0(x))\right). \tag{56}$$

By construction,

$$\Psi_x(\boldsymbol{z}_0(x)) = 0, \qquad \nabla_{\boldsymbol{z}} \Psi_x(\boldsymbol{z}_0(x)) = \mathbf{0}, \tag{57}$$

and direct differentiation gives

$$\nabla^2_{\boldsymbol{z}} \Psi_x(\boldsymbol{z}_0(x)) = \boldsymbol{H}_x. \tag{58}$$

Hence, a third-order Taylor expansion around $\boldsymbol{z}_0(x)$ yields

$$\Psi_x\left(\boldsymbol{z}_0(x) + \Delta \boldsymbol{z}_x\right) = \frac{1}{2} \Delta \boldsymbol{z}_x^\top \boldsymbol{H}_x \Delta \boldsymbol{z}_x + \mathcal{O}\left(\|\Delta \boldsymbol{z}_x\|^3\right) \tag{59}$$

in a neighborhood of $\Delta \boldsymbol{z}_x = \mathbf{0}$.

Since $\Delta \boldsymbol{z}_x$ is a C-dimensional zero-mean Gaussian vector with covariance $\boldsymbol{V}_x$, its moments satisfy

$$\mathbb{E}\left[\|\Delta \boldsymbol{z}_x\|^m\right] = \mathcal{O}(\rho_x^m), \qquad m \geq 1, \tag{60}$$

for fixed C. In particular, $\mathbb{E}[\|\Delta \boldsymbol{z}_x\|^3] = \mathcal{O}(\rho_x^3)$. Therefore, taking expectations in (59) gives

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\theta}}\left[D_{\mathrm{KL}}\left(\boldsymbol{p}_{\boldsymbol{\theta}}(x) \,\|\, \boldsymbol{p}_0(x)\right)\right] &= \frac{1}{2} \mathbb{E}\left[\Delta \boldsymbol{z}_x^\top \boldsymbol{H}_x \Delta \boldsymbol{z}_x\right] + \mathcal{O}(\rho_x^3) \\ &= \frac{1}{2} \operatorname{tr}\left(\boldsymbol{H}_x \boldsymbol{V}_x\right) + \mathcal{O}(\rho_x^3). \end{aligned} \tag{61}$$

It remains to account for the fact that EU in (54) uses the posterior predictive distribution $\overline{\boldsymbol{p}}_{\mathcal{D}}(x)$ rather than $\boldsymbol{p}_0(x)$ as its reference. A second-order expansion of the softmax map around $\boldsymbol{z}_0(x)$ gives

$$\operatorname{softmax}\left(\boldsymbol{z}_0(x) + \Delta \boldsymbol{z}_x\right) = \boldsymbol{p}_0(x) + \boldsymbol{H}_x \Delta \boldsymbol{z}_x + \mathcal{O}\left(\|\Delta \boldsymbol{z}_x\|^2\right). \tag{62}$$

Taking expectations and using $\mathbb{E}[\Delta \boldsymbol{z}_x] = \mathbf{0}$ and $\mathbb{E}[\|\Delta \boldsymbol{z}_x\|^2] = \operatorname{tr}(\boldsymbol{V}_x) = \rho_x^2$ yields

$$\overline{\boldsymbol{p}}_{\mathcal{D}}(x) - \boldsymbol{p}_0(x) = \mathcal{O}(\rho_x^2). \tag{63}$$

Since every component of $\boldsymbol{p}_0(x)$ is strictly positive, the KL divergence is locally quadratic around $\overline{\boldsymbol{p}}_{\mathcal{D}}(x) = \boldsymbol{p}_0(x)$, and thus

$$D_{\mathrm{KL}}\left(\overline{\boldsymbol{p}}_{\mathcal{D}}(x) \,\|\, \boldsymbol{p}_0(x)\right) = \mathcal{O}\left(\|\overline{\boldsymbol{p}}_{\mathcal{D}}(x) - \boldsymbol{p}_0(x)\|^2\right) = \mathcal{O}(\rho_x^4). \tag{64}$$

Combining (55), (61), and (64), and noting that $\mathcal{O}(\rho_x^4) \subseteq \mathcal{O}(\rho_x^3)$ as $\rho_x \to 0$, gives

$$U_{\mathcal{D}}(x) = \frac{1}{2} \operatorname{tr}\left(\boldsymbol{H}_x \boldsymbol{V}_x\right) + \mathcal{O}(\rho_x^3). \tag{65}$$

Using $\boldsymbol{V}_x = \boldsymbol{G}_x \boldsymbol{\Sigma}_{\mathcal{D}} \boldsymbol{G}_x^\top$ and the cyclic property of the trace,

$$\begin{aligned} \operatorname{tr}\left(\boldsymbol{H}_x \boldsymbol{V}_x\right) &= \operatorname{tr}\left(\boldsymbol{H}_x \boldsymbol{G}_x \boldsymbol{\Sigma}_{\mathcal{D}} \boldsymbol{G}_x^\top\right) \\ &= \operatorname{tr}\left(\boldsymbol{B}_x \boldsymbol{\Sigma}_{\mathcal{D}} \boldsymbol{B}_x^\top\right), \end{aligned} \tag{66}$$

where $\boldsymbol{B}_x = \boldsymbol{H}_x^{1/2} \boldsymbol{G}_x$. Consequently,

$$U_{\mathcal{D}}(x) = \frac{1}{2} \operatorname{tr}\left(\boldsymbol{B}_x \boldsymbol{\Sigma}_{\mathcal{D}} \boldsymbol{B}_x^\top\right) + \mathcal{O}(\rho_x^3), \tag{67}$$

which proves the result. ■

C.2 Proof of Theorem 4.2

Recall that $\boldsymbol{M}_{\mathcal{D}} = \sum_{k=1}^{K} \sum_{i=1}^{N_k} \left(\boldsymbol{B}_i^{(k)}\right)^\top \boldsymbol{B}_i^{(k)}$ is symmetric and positive semidefinite, and that $\mathcal{S}_{\mathcal{D}} = \operatorname{range}(\boldsymbol{M}_{\mathcal{D}})$. For a symmetric matrix, the orthogonal complement of its range equals its null space. Hence,

$$\boldsymbol{M}_{\mathcal{D}} \boldsymbol{\Pi}_{\mathcal{D}}^{\perp} = \mathbf{0}, \qquad \boldsymbol{\Pi}_{\mathcal{D}}^{\perp} \boldsymbol{M}_{\mathcal{D}} = \mathbf{0}. \tag{68}$$

Let $r = \operatorname{rank}(\boldsymbol{M}_{\mathcal{D}})$. Since $\boldsymbol{M}_{\mathcal{D}}$ is positive semidefinite, it admits an eigendecomposition

$$\boldsymbol{M}_{\mathcal{D}} = \boldsymbol{V} \operatorname{diag}\left(\mu_1, \ldots, \mu_r, 0, \ldots, 0\right) \boldsymbol{V}^\top, \tag{69}$$

where

$$\Gamma_{\mathcal{D}} = \mu_1 \geq \cdots \geq \mu_r = \gamma_{\mathcal{D}} > 0. \tag{70}$$

The first r columns of $\boldsymbol{V}$ span $\mathcal{S}_{\mathcal{D}}$, while the remaining columns span $\mathcal{S}_{\mathcal{D}}^{\perp}$. It follows that

$$\boldsymbol{\Pi}_{\mathcal{D}} = \boldsymbol{V} \operatorname{diag}(\boldsymbol{I}_r, \boldsymbol{0})\, \boldsymbol{V}^{\top}. \tag{71}$$

Using $\boldsymbol{\Sigma}_{\mathcal{D}} = (\lambda \boldsymbol{I} + \boldsymbol{M}_{\mathcal{D}})^{-1}$, the same eigenbasis gives

$$\boldsymbol{\Sigma}_{\mathcal{D}} = \boldsymbol{V} \operatorname{diag}\left(\frac{1}{\lambda+\mu_1}, \ldots, \frac{1}{\lambda+\mu_r}, \frac{1}{\lambda}, \ldots, \frac{1}{\lambda}\right) \boldsymbol{V}^{\top}. \tag{72}$$

Therefore, the covariance restricted to the training tangent subspace satisfies

$$\frac{1}{\lambda+\Gamma_{\mathcal{D}}} \boldsymbol{\Pi}_{\mathcal{D}} \preceq \boldsymbol{\Pi}_{\mathcal{D}} \boldsymbol{\Sigma}_{\mathcal{D}} \boldsymbol{\Pi}_{\mathcal{D}} \preceq \frac{1}{\lambda+\gamma_{\mathcal{D}}} \boldsymbol{\Pi}_{\mathcal{D}}. \tag{73}$$

On the orthogonal complement, (72) yields

$$\boldsymbol{\Pi}_{\mathcal{D}}^{\perp} \boldsymbol{\Sigma}_{\mathcal{D}} \boldsymbol{\Pi}_{\mathcal{D}}^{\perp} = \frac{1}{\lambda} \boldsymbol{\Pi}_{\mathcal{D}}^{\perp}. \tag{74}$$

Moreover, because $\boldsymbol{\Sigma}_{\mathcal{D}}$ shares eigenvectors with $\boldsymbol{M}_{\mathcal{D}}$,

$$\boldsymbol{\Pi}_{\mathcal{D}} \boldsymbol{\Sigma}_{\mathcal{D}} \boldsymbol{\Pi}_{\mathcal{D}}^{\perp} = \boldsymbol{0}, \qquad \boldsymbol{\Pi}_{\mathcal{D}}^{\perp} \boldsymbol{\Sigma}_{\mathcal{D}} \boldsymbol{\Pi}_{\mathcal{D}} = \boldsymbol{0}. \tag{75}$$

We now decompose the leading-order epistemic uncertainty. Inserting $\boldsymbol{I} = \boldsymbol{\Pi}_{\mathcal{D}} + \boldsymbol{\Pi}_{\mathcal{D}}^{\perp}$ on both sides of $\boldsymbol{\Sigma}_{\mathcal{D}}$ and using (75) gives

$$\begin{aligned} 2\widetilde{U}_{\mathcal{D}}(x) &= \operatorname{tr}\left(\boldsymbol{B}_x \boldsymbol{\Sigma}_{\mathcal{D}} \boldsymbol{B}_x^{\top}\right) \\ &= \operatorname{tr}\left(\boldsymbol{B}_x \boldsymbol{\Pi}_{\mathcal{D}} \boldsymbol{\Sigma}_{\mathcal{D}} \boldsymbol{\Pi}_{\mathcal{D}} \boldsymbol{B}_x^{\top}\right) + \operatorname{tr}\left(\boldsymbol{B}_x \boldsymbol{\Pi}_{\mathcal{D}}^{\perp} \boldsymbol{\Sigma}_{\mathcal{D}} \boldsymbol{\Pi}_{\mathcal{D}}^{\perp} \boldsymbol{B}_x^{\top}\right). \end{aligned} \tag{76}$$

Applying (74) to the second term yields

$$2\widetilde{U}_{\mathcal{D}}(x) = \operatorname{tr}\left(\boldsymbol{B}_x \boldsymbol{\Pi}_{\mathcal{D}} \boldsymbol{\Sigma}_{\mathcal{D}} \boldsymbol{\Pi}_{\mathcal{D}} \boldsymbol{B}_x^{\top}\right) + \frac{1}{\lambda} \left\| \boldsymbol{B}_x \boldsymbol{\Pi}_{\mathcal{D}}^{\perp} \right\|_{\mathrm{F}}^2. \tag{77}$$

For the lower bound, the left inequality in (73) implies

$$\operatorname{tr}\left(\boldsymbol{B}_x \boldsymbol{\Pi}_{\mathcal{D}} \boldsymbol{\Sigma}_{\mathcal{D}} \boldsymbol{\Pi}_{\mathcal{D}} \boldsymbol{B}_x^{\top}\right) \geq \frac{1}{\lambda+\Gamma_{\mathcal{D}}} \operatorname{tr}\left(\boldsymbol{B}_x \boldsymbol{\Pi}_{\mathcal{D}} \boldsymbol{B}_x^{\top}\right) = \frac{\kappa_{\mathcal{D}}(x)}{\lambda+\Gamma_{\mathcal{D}}}. \tag{78}$$

Similarly, the right inequality in (73) gives

$$\operatorname{tr}\left(\boldsymbol{B}_x \boldsymbol{\Pi}_{\mathcal{D}} \boldsymbol{\Sigma}_{\mathcal{D}} \boldsymbol{\Pi}_{\mathcal{D}} \boldsymbol{B}_x^{\top}\right) \leq \frac{\kappa_{\mathcal{D}}(x)}{\lambda+\gamma_{\mathcal{D}}}. \tag{79}$$

Combining (77), (78), and (79), and dividing by two, yields

$$\frac{\kappa_{\mathcal{D}}(x)}{2(\lambda+\Gamma_{\mathcal{D}})} + \frac{\nu_{\mathcal{D}}(x)}{2\lambda} \leq \widetilde{U}_{\mathcal{D}}(x) \leq \frac{\kappa_{\mathcal{D}}(x)}{2(\lambda+\gamma_{\mathcal{D}})} + \frac{\nu_{\mathcal{D}}(x)}{2\lambda}, \tag{80}$$

which proves (14). ■

D Implementation Details of the Experiments in Section 5

CIFAR-100-C. We construct the deployment stream from the CIFAR-100 test set. Each batch contains 64 images sampled with replacement and corrupted on the fly using a single CIFAR-C-style corruption (Hendrycks & Dietterich, 2019) at severity 3. The stream comprises eight consecutive phases of 20 batches each (160 batches in total). The training-covered portion contains shot noise, zoom blur, brightness, contrast, and JPEG compression (batches 0–99), followed by the held-out impulse noise, defocus blur, and pixelation corruptions (batches 100–159). Hence, the novel phase begins at $t = 100$.

The classifier is a ViT-B/16 (Dosovitskiy et al., 2021) initialized from the ImageNet-21k AugReg checkpoint of Steiner et al. (2022) and fully fine-tuned on the CIFAR-100 for 10 epochs using SGD with momentum 0.9, an initial learning rate of 0.01 with cosine decay, batch size 64, and gradient clipping at 1.0. Inputs are bilinearly resized to 224×224. During fine-tuning, each mini-batch is corrupted with probability 0.8 using a corruption sampled uniformly from seven training corruptions. Thus, every corruption appearing in the training-covered stream is represented during training, whereas none of the novel-phase corruptions is observed.

Predictive uncertainty is obtained from a last-layer Laplace approximation (Daxberger et al., 2021): the fine-tuned backbone is held fixed, and the weights of the final linear layer are given a Gaussian posterior centered at their fine-tuned values, with a diagonal generalized Gauss–Newton precision estimated on 2,048 corrupted CIFAR-100 training images plus a prior precision λ. We set λ to the smallest value in $\{10^0, \ldots, 10^7\}$ that keeps the posterior-averaged accuracy within 2 percentage points of the MAP accuracy. For each batch we draw $S = 20$ posterior samples and compute per-sample mutual information (epistemic) and predictive entropy (total), averaged over the batch.

For sequential detection, we compare a supervised DDM reference (Gama et al., 2004), which receives per-sample binary prediction errors and therefore requires deployment labels, with ADWIN (Bifet & Gavaldà, 2007) applied separately to epistemic uncertainty, total uncertainty, and a model-free input statistic, $\sqrt{\mu_x^2 + \sigma_x^2}$, where μ_x and σ_x denote the mean and standard deviation of an image's pixel values. All detectors are updated once per sample, and a batch is marked as containing an alarm if a detector signals at any update within that batch. Hyperparameters are selected over five stream seeds to minimize detection delay subject to a known-phase false-alarm rate of at most 5%, yielding a DDM threshold of 2.0 with a 200-sample warm start, ADWIN $\delta = 0.1$ for the epistemic and total-uncertainty signals, and $\delta = 10^{-4}$ for the input statistic. The figure shows one representative stream realization; for visualization, the error-rate curve in the top panel is smoothed using a trailing mean over the preceding 10 batches.

AUROC / FPR95 Protocol. For every arm, we report AUROC and FPR95 as complementary measures of how well each monitored signal separates the known and novel phases of the stream. AUROC provides a threshold-independent measure of separability across all possible operating points (Fawcett, 2006), and can equivalently be interpreted through the Mann–Whitney ranking statistic (Hanley & McNeil, 1982). Within each stream, batches before the annotated drift boundary form the known phase (label 0) and batches from the boundary onward form the novel phase (label 1). For each of the four monitored signals we use the per-batch statistic logged during evaluation, which is the batch average of a per-sample quantity: (i) the misclassification rate, i.e., the mean of the per-sample 0/1 errors which requires labels and serves as our supervised reference; (ii) epistemic uncertainty, measured as the mutual information over S posterior (Laplace) or MC-dropout samples; (iii) total predictive entropy; and (iv) a model-free input statistic, $\sqrt{\mu_x^2 + \sigma_x^2}$, where μ_x and σ_x are the mean and standard deviation of each input's features. AUROC is computed directly on this statistic, with larger values scored as more novel. Scores are not re-oriented, so an AUROC below 0.5 indicates a signal that decreases under the novel shift.

We additionally report FPR95, an operating-point measure commonly used for evaluating distribution-shift and OOD detection methods (Liang et al., 2018). Following the OOD convention, in which the in-distribution class is treated as positive, we set the threshold at the 95th percentile of the known-phase statistic, so that 95% of known-phase batches are correctly accepted, and report the fraction of novel-phase batches whose statistic does not exceed this threshold. These batches are incorrectly classified as known and therefore correspond to missed novel drifts.

To reduce sensitivity to a particular stream realization, both metrics are computed independently over five random seeds per arm. Each seed re-draws the stream (the samples assigned to each batch and the random corruption realizations) together with the posterior or MC-dropout sampling, while the trained model is held fixed. We report the mean and standard deviation across the five seeds.

E Additional Experiments

We further evaluate whether the proposed uncertainty-based monitoring principle generalizes beyond the experiments in Section 5, considering both different data modalities and distribution shifts generated by physical dynamical systems. In particular, we extend the evaluation beyond image classification to wireless communication and audio settings, allowing us to assess whether epistemic uncertainty remains informative across substantially different signal structures and sources of non-stationarity.

Across these additional experiments, we follow the same evaluation protocol as in the main study: epistemic uncertainty is aggregated at the batch level and used as the monitoring signal for ADWIN, KSWIN, and Page–Hinkley. As a supervised reference, DDM operates on the true prediction-error stream, while input-based statistics serve as model-independent unsupervised baselines.

E.1 Uncertainty Decomposition Ablation

Throughout our experiments, we compare unsupervised drift detection based on epistemic uncertainty with detection based on total predictive uncertainty, which combines both epistemic and aleatoric components and has previously been considered in uncertainty-based drift detection methods such as UDD (Baier et al., 2023). Both provide informative label-free monitoring signals, but our results generally favor epistemic uncertainty for reliability-drift detection, as it achieves stronger or more stable detection performance in most settings. In addition to this empirical advantage, we favor explicitly distinguishing the source of the detected drift. An increase in epistemic uncertainty points specifically to insufficient model knowledge, whereas an increase in total predictive uncertainty may also be driven by aleatoric effects such as noise or intrinsic ambiguity. Total uncertainty can therefore still detect reliability-relevant shifts, but its response is less specific about their origin. For this reason, we view epistemic uncertainty as a particularly suitable signal for reliability-drift monitoring, combining strong detection performance with a clearer interpretation of why model reliability has changed.

E.2 CIFAR-10-C

We further evaluate uncertainty-based drift detection on CIFAR-10-C (Hendrycks & Dietterich, 2019) as an additional vision benchmark. The stream consists of severity-1 corruptions, with seven corruption types represented during training followed by three previously unseen corruptions. The transition to the novel phase occurs at batch 140. We consider both ResNet-20 (He et al., 2016) and a 1.8M-parameter ViT (Dosovitskiy et al., 2021), both trained from scratch, and estimate epistemic uncertainty using a last-layer Laplace approximation (Daxberger et al., 2021). The trained model remains fixed throughout the stream.

Figures 3 and 4 show a consistent pattern across the two architectures. Epistemic uncertainty remains comparatively stable while the stream contains training-covered corruptions and increases after the transition to previously unseen corruption types. This increase coincides with a degradation in predictive performance, indicating that the epistemic signal responds specifically when the stream moves toward regions that are less supported by the model's training experience. The consistent behavior across ResNet-20 and ViT further indicates that last-layer Laplace can provide useful epistemic uncertainty estimates across substantially different model architectures.

We additionally consider a severity-shift setting in which the corruption mechanisms remain familiar while their intensity increases. Specifically, the stream first contains the same seven corruption types at severity 1 and transitions at batch 140 to these corruptions at severity 3. Unlike the previous setting, the post-drift phase therefore does not introduce previously unseen corruption types, but instead strengthens perturbations that are already represented during training. As shown in Fig. 5, the increased severity is accompanied by higher prediction error, while the epistemic signal is less sharply separated across the drift boundary. This is consistent with our reliability-drift interpretation: because the corruption mechanisms themselves remain familiar, the shift is less directly associated with insufficient model knowledge and more strongly reflects increased observation degradation. Consequently, total predictive uncertainty can become comparatively informative in this setting because it also captures the aleatoric component. The severity experiment therefore complements the known-to-novel setting by illustrating that epistemic uncertainty is most selective when reduced reliability is driven by conditions that are insufficiently represented during training, rather than by stronger instances of already familiar perturbations.

E.3 MNIST-C

We next evaluate the same monitoring principle on MNIST-C (Mu & Gilmer, 2019), providing a second image domain with substantially different visual statistics. The stream is composed of severity-3 corruptions, with four training-covered corruption types followed by three previously unseen corruptions and a known-to-novel transition at batch 80. We again consider ResNet-20 and ViT with

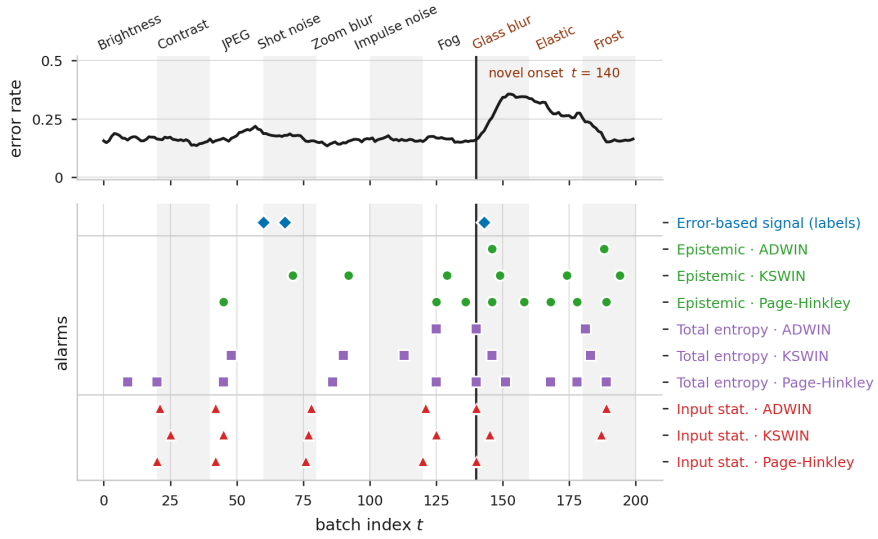


Figure 3: **CIFAR-10-C with ViT and last-layer Laplace.** The stream transitions from training-covered to previously unseen corruptions at batch 140. Epistemic uncertainty increases after the transition together with the prediction error, while the corresponding detectors operate without access to labels.

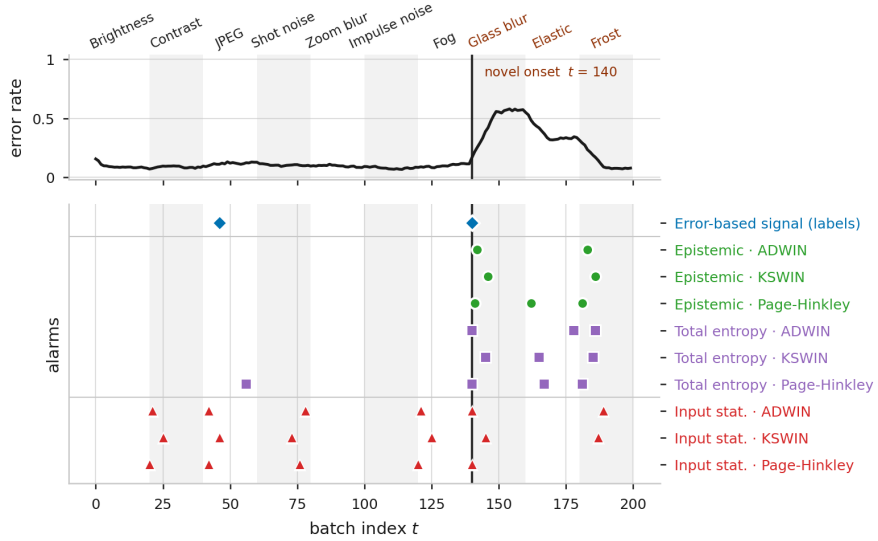


Figure 4: **CIFAR-10-C with ResNet-20 and last-layer Laplace.** The transition to unseen corruption types occurs at batch 140, producing an increase in epistemic uncertainty that accompanies the deterioration in predictive performance.

a last-layer Laplace approximation. ResNet-20 is modified only to accommodate the single-channel MNIST input; the remaining evaluation protocol is unchanged.

As shown in Figures 6 and 7, the behavior observed on CIFAR-10-C largely transfers to MNIST-C. Epistemic uncertainty remains low or comparatively stable during the training-covered portion of the stream and rises when previously unseen corruptions are introduced. The increase is again associated with a deterioration in prediction accuracy, supporting the use of epistemic uncertainty as a label-free indicator of reliability-relevant drift rather than merely a detector of arbitrary changes in the input distribution.

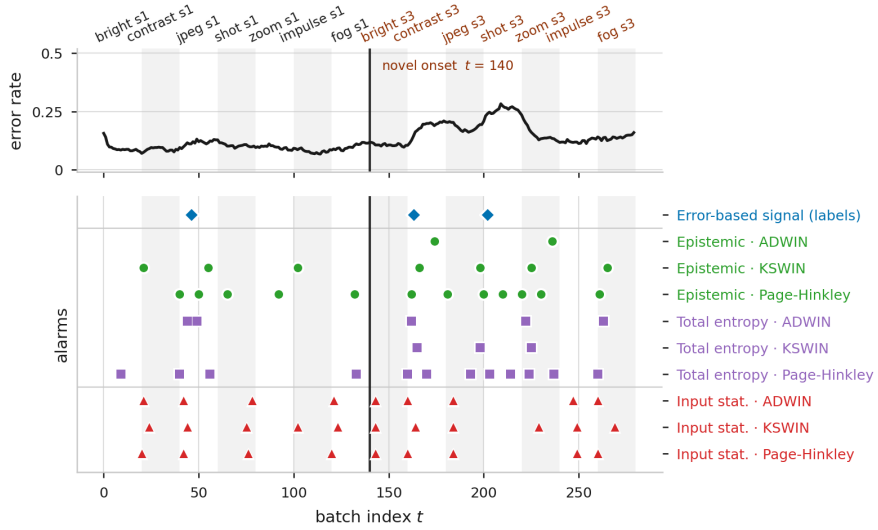


Figure 5: **CIFAR-10-C severity drift with ResNet-20 and last-layer Laplace.** The stream transitions at batch 140 from familiar corruption types at severity 1 to the same corruption types at severity 3. The stronger corruptions increase prediction error, while the epistemic signal is less sharply separated because the corruption mechanisms remain training-covered.

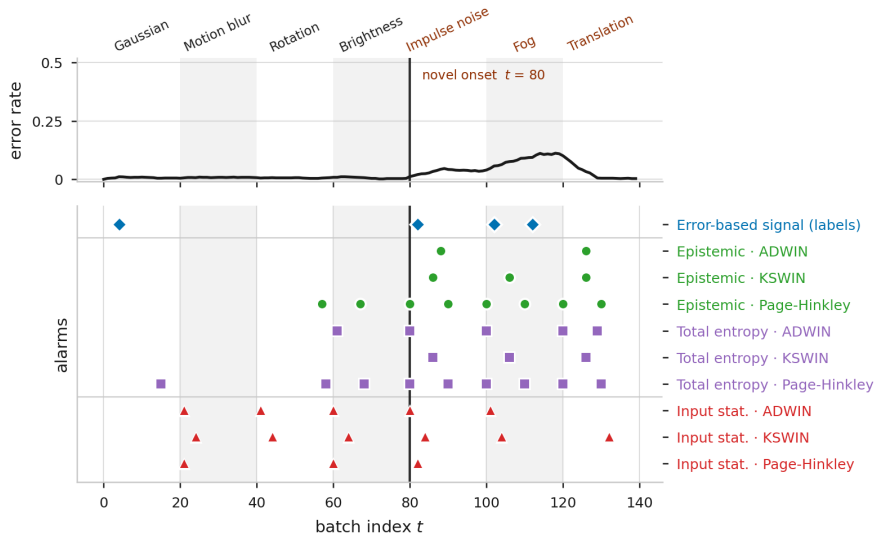


Figure 6: **MNIST-C with ViT and last-layer Laplace.** The stream transitions to previously unseen corruptions at batch 80. The novel phase is accompanied by increased epistemic uncertainty and increased prediction error.

E.4 Wireless Channel Drift

We further evaluate the proposed monitoring principle in a wireless communication setting, where non-stationarity arises from changes in the physical propagation environment. Channel impulse responses are generated using QuaDRiGa (Jaeckel et al., 2014). The receiver performs differential-QPSK symbol detection over multipath fading channels, where delayed signal components introduce inter-symbol interference (ISI). We use a single-layer LSTM receiver with a last-layer Laplace approximation and keep the receiver fixed throughout deployment. We consider two

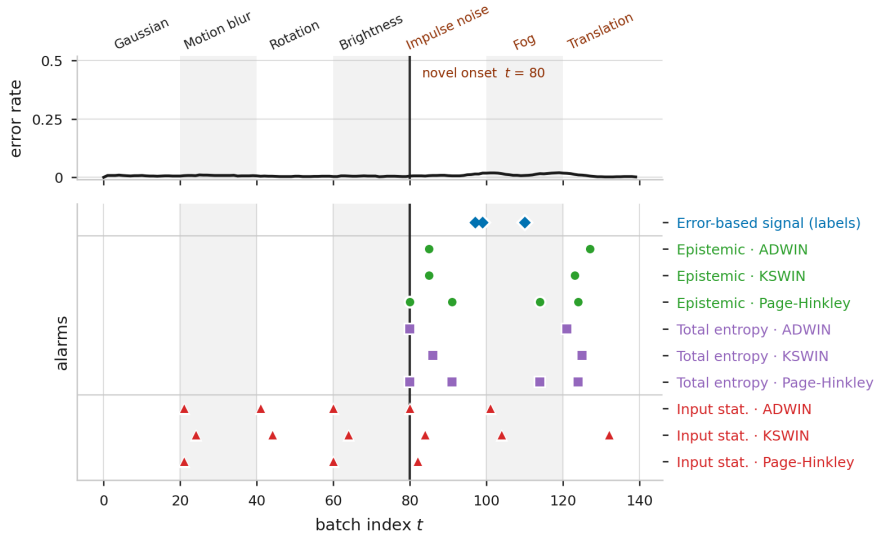


Figure 7: **MNIST-C with ResNet-20 and last-layer Laplace.** Epistemic uncertainty responds to the transition from training-covered to unseen corruption types at batch 80 and follows the corresponding loss in predictive reliability.

complementary experiments designed to separate changes in propagation environment from changes in line-of-sight (LOS) and non-line-of-sight (NLOS) conditions.

Propagation-Environment Drift. In the first experiment, the receiver is trained exclusively on Highway LOS channels and evaluated on a stream that transitions from Highway LOS to Urban LOS and subsequently to Indoor LOS. The SNR is fixed at 15 dB throughout the stream. All three scenarios retain a line-of-sight component, while the propagation geometry changes relative to the Highway environment used for training. Highway LOS therefore represents the training-covered regime, whereas Urban LOS and Indoor LOS introduce previously unseen propagation environments.

As shown in Fig. 8, the transition from the training-covered Highway LOS regime to the unseen Urban LOS environment at batch 30 is accompanied by a clear increase in BER. The epistemic-uncertainty detectors respond more selectively to this transition, with alarms becoming more concentrated in the novel Urban and Indoor regimes. In contrast, detectors based on total predictive uncertainty trigger frequently throughout the stream, including within the training-covered Highway regime. This behavior is consistent with total uncertainty also capturing aleatoric variability induced by noisy and fluctuating channel conditions, making it less specific to changes in model knowledge. The experiment therefore highlights the advantage of isolating epistemic uncertainty when the objective is to detect propagation changes that correspond to reduced model familiarity rather than general variability in the received signal.

LOS-to-NLOS Drift. In the second experiment, we isolate a previously unseen propagation condition while keeping the physical environments themselves training-covered. The receiver is trained on Highway LOS and Urban LOS channels and is subsequently evaluated on Highway NLOS and Urban NLOS channels. Thus, the Highway and Urban environments are both represented during training, while the absence of a direct propagation path constitutes the novel condition. The SNR and modulation remain fixed throughout the experiment.

Figure 9 shows that epistemic uncertainty identifies the transition to the previously unseen NLOS regime. In contrast, the input-based statistic does not consistently reflect this reliability-relevant change, including the absence of a corresponding Page–Hinkley alarm in this experiment. This setting is particularly useful because the environmental context remains familiar while only the propagation condition changes, allowing us to test whether epistemic uncertainty captures model-relevant novelty beyond generic changes in the received-signal distribution.

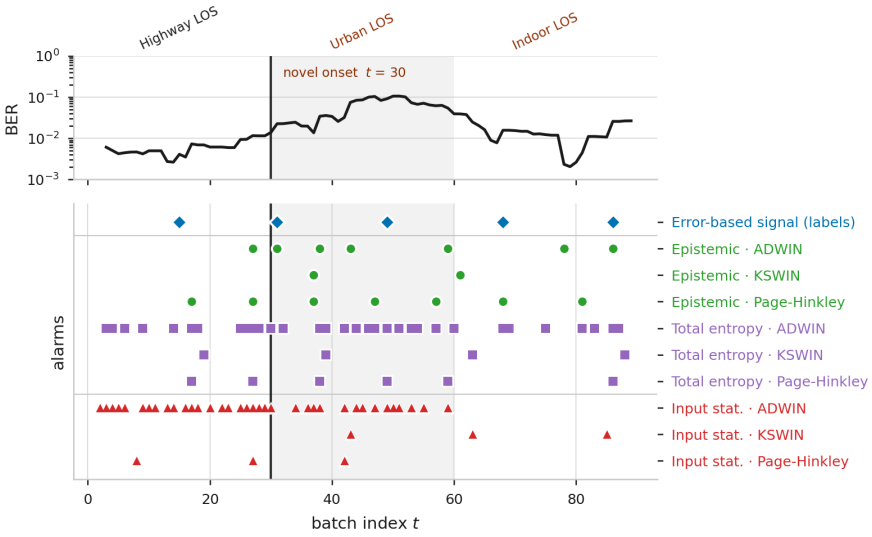


Figure 8: **Wireless propagation-environment drift with last-layer Laplace.** The receiver is trained on Highway LOS channels and evaluated on a stream transitioning from Highway LOS to previously unseen Urban LOS and Indoor LOS environments, with the SNR fixed at 15 dB.

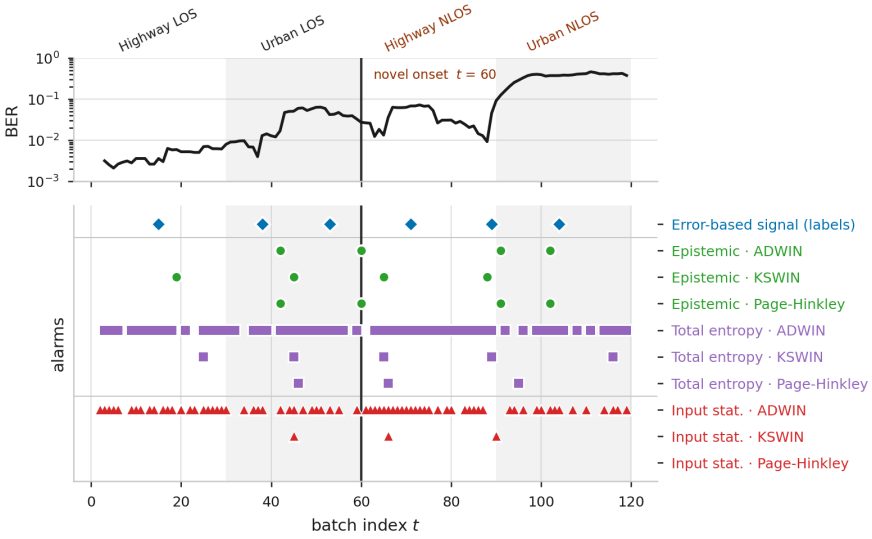


Figure 9: **Wireless LOS-to-NLOS drift with last-layer Laplace.** The receiver is trained on Highway and Urban LOS channels and evaluated on Highway and Urban NLOS channels, such that the propagation environments remain familiar while the NLOS condition is previously unseen.

Together, the two experiments examine complementary sources of wireless non-stationarity: changes to previously unseen propagation environments and transitions to an unseen NLOS condition within familiar environments. The results support the use of epistemic uncertainty as a model-aware signal for identifying channel conditions that are insufficiently represented during training, rather than reacting indiscriminately to every change in the received-signal distribution.

E.5 Uncertainty Estimator Ablation: MC-Dropout

The main experiments use last-layer Laplace to estimate epistemic uncertainty. To assess whether the observed reliability-drift behavior depends on this particular approximation, we repeat all experimental settings—CIFAR-100-C, CIFAR-10-C, MNIST-C, and the wireless channel-drift experiments—using MC-dropout (Gal & Ghahramani, 2016). At inference time, multiple stochastic forward passes are used to approximate the predictive distribution and derive the epistemic mutual-information signal. Across datasets, model architectures, and modalities, MC-dropout reproduces the same overall qualitative behavior observed with last-layer Laplace: epistemic uncertainty remains comparatively lower or more stable under training-covered conditions and increases when the stream enters previously unseen regimes associated with reduced predictive reliability. This consistency across experiments indicates that the proposed monitoring principle is not specific to a particular Bayesian approximation, model architecture, or data modality, despite being evaluated with substantially different predictors and input domains. Rather, the results support epistemic uncertainty itself as the relevant monitoring quantity, with both last-layer Laplace and MC-dropout providing viable mechanisms for estimating it.

E.6 Audio Modality Ablation

To further assess cross-modal generalization, we apply the proposed monitoring principle to audio classification to examine whether the observed behavior extends to a substantially different signal modality. We use ESC-50 (Piczak, 2015), Speech Commands (Warden, 2018), and UrbanSound8K (Salamon et al., 2014), and for each construct a corruption stream containing six corruption types represented during training, including additive-noise, filtering, and level/quality degradations applied to the raw waveform, followed by three previously unseen corruptions. The transition to the novel phase occurs at batch 120. We consider three classifiers operating on log-mel spectrograms: a compact CNN and a wider variant of it (CNN-wide), with 247K and 1.58M parameters on ESC-50, respectively, and a ResNet-18 (He et al., 2016). For this modality, epistemic uncertainty is estimated using MC-dropout (Gal & Ghahramani, 2016). The trained classifiers remain fixed throughout the stream.

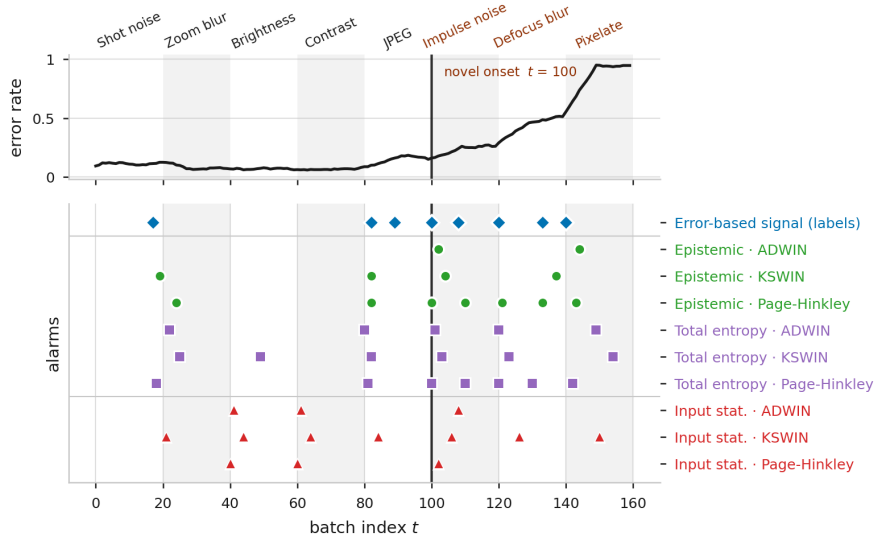
Figures 11 and 12, together with Tables 1 and 2, show that the behavior observed in the other modalities also extends to audio. Epistemic uncertainty remains comparatively stable over training-covered corruptions and increases after the transition to previously unseen corruption types, together with a deterioration in predictive performance. The distinction between epistemic and total predictive uncertainty is particularly pronounced on ESC-50 and UrbanSound8K. Audio corruptions can introduce substantial observation-level variability even within training-covered conditions, causing total uncertainty to respond strongly to noise and ambiguity that do not necessarily indicate insufficient model knowledge. As a result, total uncertainty provides a noisier monitoring signal, whereas the epistemic component more selectively reflects the transition to conditions that are insufficiently represented during training. This experiment therefore further supports epistemic uncertainty as a more informative signal for reliability-drift detection, while the consistent behavior across the three architectures, including the wider CNN, indicates that the effect is not tied to a particular audio architecture or to limited model capacity.

E.7 Ambiguous Label-Prior Shift

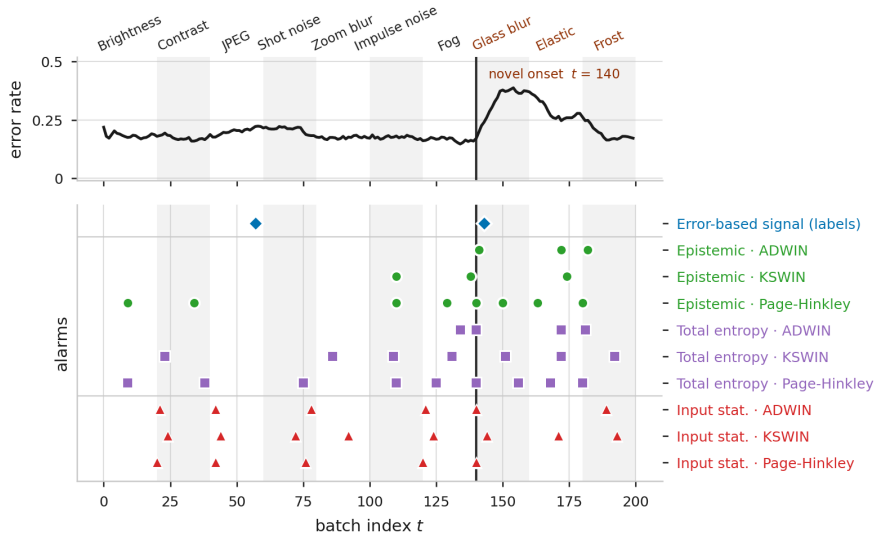
The preceding experiments examine shifts that move the model beyond its training support. We now consider the complementary case: a shift that remains entirely within the training-covered region and leaves predictive performance unchanged, for which the desired detector behavior is to remain silent, while more and more ambiguous data appear.

We construct such a stream on EMNIST-Balanced (Cohen et al., 2017) using ResNet-20 with MC-dropout, retaining the evaluation protocol described above. EMNIST is suited to this purpose because several of its classes are structurally ambiguous rather than merely difficult: the glyphs for `1`/`I`/`L` and `0`/`O` coincide under handwriting, so the label is underdetermined by the input while the input itself remains typical and densely represented in the training distribution. Such classes exhibit high aleatoric but low epistemic uncertainty.

The stream presents 40 batches drawn under a uniform class prior, followed by 60 batches under a skew towards ambiguous classes, beginning at batch 40. The skewed prior upweights ambiguous-but-accurate classes (`1`, `9`) alongside high-accuracy classes (`H`, `R`, `N`, `G`), such that mean predictive entropy increases while accuracy remains unchanged.



(a) CIFAR-100-C.

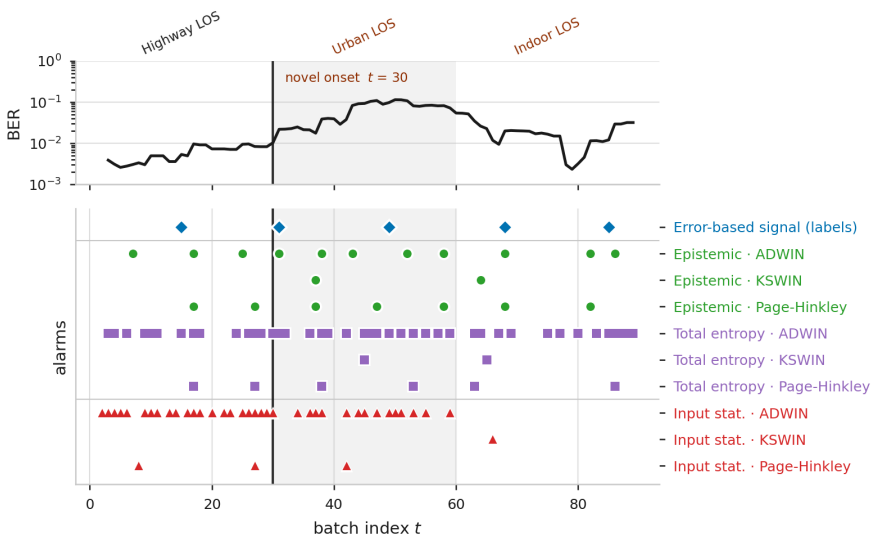


(b) CIFAR-10-C.

Figure 10: **MC-dropout ablation across experimental settings.** (a) CIFAR-100-C using a ViT-B/16 classifier and (b) CIFAR-10-C using a ViT classifier. Figure continued on the next page.

As shown in Fig. 13, prediction error remains flat across the transition and epistemic uncertainty is unchanged, whereas total predictive uncertainty increases. Consequently, ADWIN and KSWIN applied to epistemic uncertainty raise no alarms over the entire stream, while the same detectors applied to total uncertainty raise alarms.

This provides a controlled instance where the model remains well within its training support and its predictive performance is unaffected, whereas isolating the epistemic component yields a statistic that abstains from signaling shifts for which no corrective action, such as retraining, would be warranted. The total entropy detector fails at this.



(c) Wireless propagation-environment drift.

Figure 10: **MC-dropout ablation across experimental settings (continued).** (c) Wireless propagation-environment drift using a single-layer LSTM receiver. Across these different datasets, architectures, and modalities, epistemic uncertainty estimated with MC-dropout exhibits behavior consistent with the corresponding last-layer Laplace results.

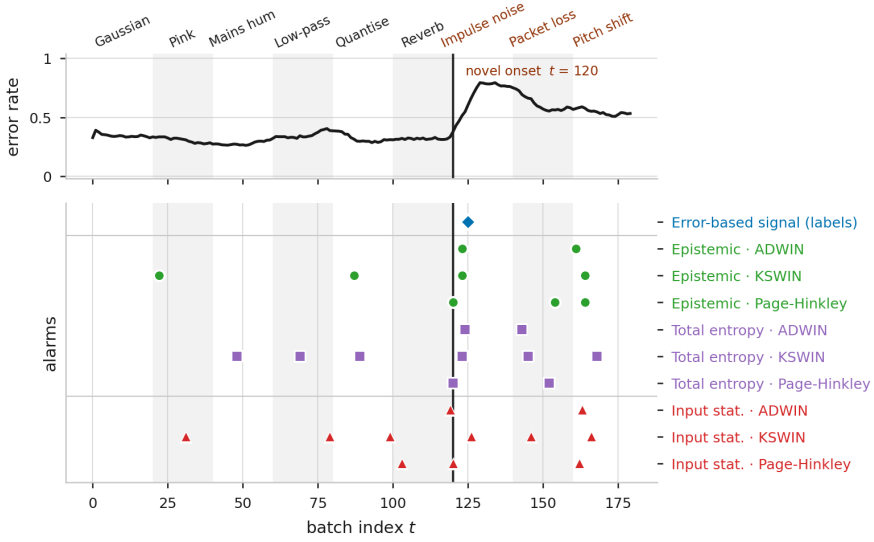


Figure 11: **ESC-50 with a 247K-parameter CNN and MC-dropout.** The stream transitions from training-covered to previously unseen audio corruptions at batch 120. The novel phase is accompanied by increased epistemic uncertainty and a degradation in predictive performance.

E.8 FPR Operating-Point Ablation

To assess whether the conclusions depend on the particular operating point used for FPR95, we additionally report FPR99 and FPR100 in Table 2, obtained by setting the threshold such that 99% and 100% of known-phase batches, respectively, are accepted as known, and measuring the fraction of novel-phase batches that are then also accepted as known. As expected, requiring increasingly complete acceptance of the known phase increases the false-positive rate for all signals. Nevertheless, the relative behavior

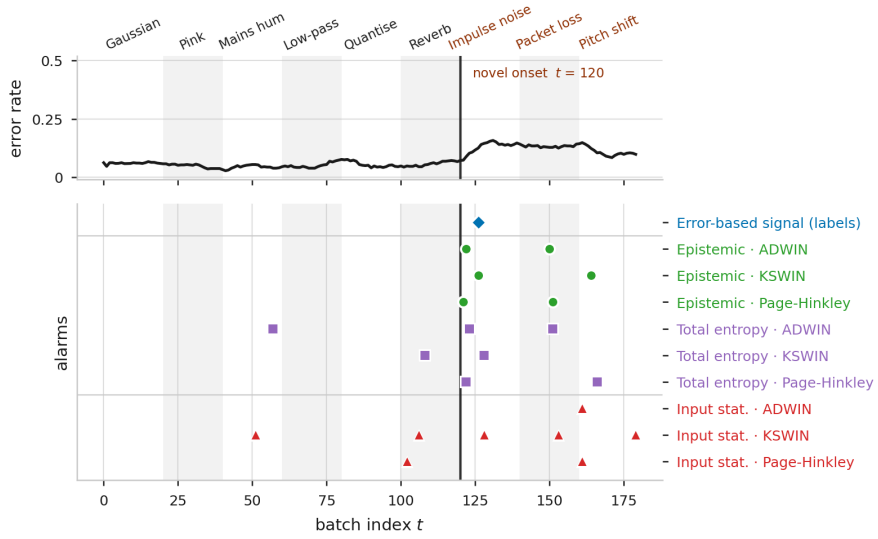


Figure 12: **Speech Commands with a 245K-parameter CNN and MC-dropout.** The stream transitions from training-covered to previously unseen audio corruptions at batch 120. Epistemic detectors raise no alarms during the training-covered phase and respond within six batches of the transition, whereas total-uncertainty and input-statistic detectors also fire on training-covered corruptions.

Table 2: FPR at three known-phase operating points (95%, 99%, and 100%), reported as mean ± standard deviation over 5 stream seeds. Lower is better; best per row and operating point in **bold**, second best underlined. Error uses deployment labels and serves only as a supervised reference.

	FPR95 ↓				FPR99 ↓				FPR100 ↓			
	Error	EU	Entropy	Input	Error	EU	Entropy	Input	Error	EU	Entropy	Input
MNIST												
ResNet-20 Laplace	0.85±0.05	0.54±0.06	**0.46±0.06**	1.00±0.00	0.87±0.07	0.62±0.07	**0.56±0.03**	1.00±0.00	0.89±0.07	0.68±0.09	**0.59±0.04**	1.00±0.00
ViT Laplace	0.38±0.03	0.35±0.03	**0.30±0.01**	1.00±0.00	0.45±0.06	0.40±0.04	**0.31±0.01**	1.00±0.00	0.47±0.07	0.46±0.04	**0.33±0.02**	1.00±0.00
CIFAR-10												
ResNet-20 Laplace	**0.33±0.01**	**0.33±0.01**	**0.33±0.00**	1.00±0.00	0.34±0.01	**0.33±0.00**	**0.33±0.00**	1.00±0.00	0.35±0.02	0.34±0.01	**0.33±0.00**	1.00±0.00
ViT Laplace	0.51±0.03	**0.43±0.04**	**0.43±0.03**	1.00±0.00	0.64±0.05	**0.57±0.07**	**0.57±0.08**	1.00±0.00	0.69±0.05	**0.65±0.04**	0.69±0.08	1.00±0.00
ResNet-20 Laplace (sev.)	0.63±0.06	0.66±0.06	**0.53±0.03**	0.71±0.00	0.71±0.06	0.78±0.08	**0.64±0.04**	0.73±0.01	0.82±0.05	0.86±0.03	**0.73±0.05**	0.76±0.03
CIFAR-100												
ViT-B/16 Laplace	0.06±0.04	**0.00±0.00**	**0.00±0.00**	1.00±0.00	0.15±0.06	**0.00±0.00**	**0.00±0.00**	1.00±0.00	0.18±0.09	**0.00±0.00**	**0.00±0.01**	1.00±0.00
ESC-50												
CNN MC	**0.04±0.03**	0.11±0.03	0.73±0.08	1.00±0.00	**0.12±0.11**	0.19±0.03	0.87±0.07	1.00±0.00	**0.22±0.12**	0.24±0.04	0.93±0.05	1.00±0.00
ResNet-18 MC	**0.43±0.10**	0.65±0.08	0.75±0.07	1.00±0.00	**0.55±0.07**	0.82±0.07	0.88±0.04	1.00±0.00	**0.60±0.03**	0.90±0.08	0.92±0.05	1.00±0.00
CNN-wide MC	**0.03±0.03**	0.08±0.02	0.80±0.06	1.00±0.00	**0.08±0.09**	0.15±0.05	0.91±0.05	1.00±0.00	**0.12±0.10**	0.17±0.06	0.94±0.04	1.00±0.00
Speech Commands												
CNN MC	0.50±0.06	**0.01±0.01**	0.11±0.01	1.00±0.01	0.63±0.08	**0.06±0.05**	0.16±0.04	1.00±0.00	0.79±0.05	**0.14±0.15**	0.39±0.23	1.00±0.00
ResNet-18 MC	0.52±0.08	**0.18±0.04**	0.21±0.05	1.00±0.01	0.67±0.10	**0.35±0.09**	0.42±0.08	1.00±0.00	0.86±0.07	**0.54±0.18**	0.59±0.17	1.00±0.00
CNN-wide MC	0.35±0.04	**0.00±0.01**	0.09±0.04	1.00±0.01	0.46±0.04	**0.06±0.06**	0.15±0.06	1.00±0.00	0.51±0.07	**0.11±0.11**	0.30±0.01	1.00±0.00
UrbanSound8K												
CNN MC	**0.15±0.01**	0.16±0.05	0.62±0.07	0.98±0.01	0.30±0.13	**0.28±0.09**	0.86±0.05	1.00±0.01	**0.44±0.19**	0.45±0.11	0.93±0.08	1.00±0.00
ResNet-18 MC	**0.40±0.09**	0.49±0.04	0.82±0.04	0.98±0.01	**0.63±0.05**	0.64±0.05	0.92±0.02	1.00±0.01	0.76±0.15	**0.68±0.05**	0.98±0.01	1.00±0.00
CNN-wide MC	0.41±0.09	**0.10±0.04**	0.66±0.07	0.98±0.01	0.51±0.14	**0.21±0.02**	0.92±0.04	1.00±0.01	0.63±0.12	**0.25±0.06**	0.95±0.04	1.00±0.00
Wireless												
LOS-drift Laplace	**0.59±0.15**	0.60±0.10	0.74±0.16	0.97±0.02	0.76±0.18	**0.75±0.14**	0.86±0.14	0.99±0.01	**0.78±0.17**	**0.78±0.17**	0.87±0.14	0.99±0.01
LOS→NLOS Laplace	**0.47±0.02**	0.57±0.11	0.83±0.08	0.92±0.04	**0.55±0.05**	0.82±0.08	0.94±0.03	0.96±0.03	**0.60±0.12**	0.86±0.09	0.95±0.02	0.96±0.03

remains consistent across the three operating points: epistemic uncertainty is generally the strongest or second-strongest unsupervised signal, while the input-based statistic remains largely uninformative. The distinction is particularly clear in the audio experiments, where total predictive uncertainty deteriorates substantially as more known-phase batches must be accepted, whereas epistemic uncertainty remains markedly more selective. The CIFAR-10 severity experiment remains an expected exception, since the post-drift regime strengthens already familiar corruptions rather than introducing a new source of model uncertainty. Overall, the FPR99 and FPR100 results support the same conclusions as FPR95 and indicate that the observed advantage of epistemic uncertainty is not specific to a single choice of operating point. As with FPR95, all values are averaged independently over five random seeds.

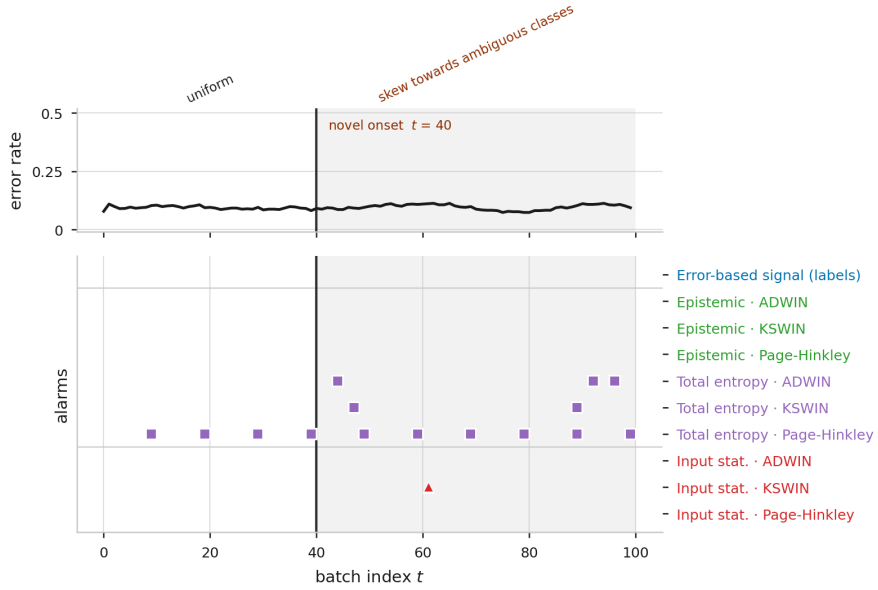


Figure 13: **Ambiguous label-prior shift on EMNIST-Balanced with ResNet-20 and MC-dropout.** The class prior begins to shift at batch 40 while prediction error remains unchanged. Total predictive uncertainty increases and triggers alarms, whereas epistemic uncertainty remains flat and raises none, as does the error-based reference.

E.9 Threshold Detection Ablation

Detector Calibration. To ensure a fair comparison across detectors, we select all alarm thresholds using the same matched false-alarm budget. Since detector hyperparameters are not directly comparable across algorithms, we avoid manual tuning and instead sweep the primary sensitivity parameter of each detector over a predefined grid. For every grid point, we evaluate the false-alarm rate (FAR), defined as the fraction of known-phase batches that trigger an alarm, together with the detection delay, measured as the number of batches between the true drift boundary and the first subsequent alarm. Results are averaged across multiple random seeds. We then select, for each detector, the operating point with the minimum detection delay subject to a common FAR constraint of 5%, or the lowest-FAR operating point when no grid point satisfies this constraint.

F Limitations

Our approach relies on epistemic uncertainty being meaningfully estimated by the underlying predictive model. In practice, approximate Bayesian constructions such as Laplace approximations or MC-dropout may require calibration, and their quality can depend on the architecture and data modality. In addition, equipping a pre-trained model with a post-hoc last-layer Laplace approximation requires access to training data again in order to fit the posterior approximation, which may be inconvenient or infeasible in some deployment settings. Poorly calibrated uncertainty estimates may therefore reduce the separability between training-covered and novel conditions. While our experiments consider multiple uncertainty estimators, a more general characterization of how posterior approximation, calibration, and data availability affect drift detection remains an important direction for future work.

The proposed notion of novelty is defined relative to the support of a fixed predictive task and model. Accordingly, the method is intended to identify previously unsupported variations within the task for which the model was trained, rather than the appearance of entirely new tasks, label spaces, or semantic classes.

Finally, the proposed framework addresses *detection* rather than adaptation. Once a novel shift is detected, deciding whether and how to update the predictor is left to the downstream system. If the model is retrained or adapted, its notion of training support changes, and the Bayesian approximation

and associated monitoring statistics should consequently be re-estimated for the updated model. Developing joint detection-and-adaptation mechanisms that update both the predictor and its uncertainty representation online is a natural extension of this work.