DRIFT ≠ ERROR: RELIABILITY ANALYSIS OF AGRICULTURAL FOUNDATION MODELS UNDER DISTRIBUTION SHIFT

Shayan Nejadshamsi & Vahab Khoshdel
Department of Electrical and Computer Engineering
University of Manitoba, Winnipeg, Canada
vahab.khoshdel@umanitoba.ca

Yuanyuan Zhang, Brock Porth, & Shadi Zaki
AIRM Consulting Ltd., Winnipeg, Canada

Lysa Porth
AIRM Consulting Ltd., Winnipeg, Canada,
Economics and Finance, Lang School of Business and Economics, University of Guelph
Statistics and Actuarial Science, University of Waterloo

ABSTRACT

Recent geospatial foundation models (GFMs) enable accurate crop yield prediction across disparate regions and scales. Yet real deployments must remain reliable in the face of distribution drift, which can arise from temporal, label, concept, domain, or representation shifts. This paper focuses on prediction reliability: we build upon our previously published *Fine-Tuning Agricultural Regression Models* (FARM) framework—which adapts the Prithvi-EO-2.0-600M Vision Transformer for dense canola yield estimation over the Canadian Prairies—and develop a systematic drift monitoring pipeline. Our framework measures Mahalanobis, cosine, and Euclidean distances in both input and latent spaces to characterize distributional shifts and study their relationship with prediction error. Experiments on county-level and 30 m precision-agriculture datasets across normal (2022) and drift (2021) growing seasons show that both input and latent-space distances provide early warning of drift, but that large drift scores do not necessarily correspond to large prediction errors: some out-of-distribution samples are predicted accurately, while some in-distribution samples incur high error. These results reveal a critical gap between drift detection and uncertainty estimation, and highlight the need for reliability-aware monitoring strategies when deploying agricultural foundation models in non-stationary environments.

1 INTRODUCTION

Reliable crop yield prediction underpins decision making for farmers, governments, insurers, and agronomists Li et al. (2021); Basso & Liu (2019); Lobell et al. (2011). The Canadian Prairies—one of the world's major canola-growing regions—exhibit substantial climatic variability, making yield forecasts challenging Lesk et al. (2016). Conventional regression models trained on remotely sensed vegetation indices or shallow convolutional networks lack the capacity to generalize across geographies and to capture subtle intra-field patterns Khaki et al. (2020); Leng & Hall (2019). Recent geospatial foundation models (GFMs) pre-trained on petabyte-scale satellite imagery offer a powerful alternative: they can transfer rich representations to downstream tasks with limited labelled data Nejadshamsi et al. (2026). However, moving from research to real-world operations requires not only accurate predictions but also robustness to distribution drift, which can manifest as temporal shifts (e.g., extreme weather events), label drift (e.g., changes in reporting practices), concept drift (evolving crop phenology), domain shift (new sensors or regions) or representation shift (changes

in latent feature distributions). Detecting and responding to such drift is critical for reliable and trustworthy systems.

Building on the previously fine-tuned foundation model framework for canola yield prediction Nejadshamsi et al. (2026), we focus on monitoring and adaptation under distribution drift. The FARM model uses the Prithvi-EO-2.0-600M Vision Transformer (Vi-T) as its encoder Szwarcman et al. (2024) and was fine-tuned on high-resolution ground-truth yields in our prior study, achieving a Root Mean Square Error (RMSE) of 0.628 and an R^2 of 0.768. While this baseline achieves strong performance, existing work has not systematically explored how to detect and respond to distribution drift in operational settings. In particular, there is a gap in understanding whether distance metrics computed in input and latent spaces can reliably signal out-of-distribution inputs and whether such metrics align with prediction errors. We address this gap by conceptualizing a *Catch-Adapt-Operate* cycle for agricultural foundation models. In this work, we focus specifically on the critical *Catch* (drift detection) and *Operate* (prediction reliability assessment) phases. We investigate the following questions: (i) Can multi-space distance metrics reliably signal distribution drift across time and space? and (ii) Do elevated distances correspond to higher prediction error, thereby serving as valid proxies for model uncertainty?

The contributions of this work are summarized as follows:

- We design a multi-space drift detection framework that computes distance metrics (Mahalanobis, cosine and Euclidean) in both the input spectral space and the model's latent embedding space to identify out-of-distribution inputs.
- Through experiments on county-level and precision-agriculture datasets across 2021-2022, we evaluate the distribution differences and the ability of these distance metrics to detect drift and show that latent-space Mahalanobis distances can highlight drift years, although high distances do not consistently correspond to high prediction error.
- Our analysis reveals that some points flagged as out-of-distribution have prediction errors comparable to in-distribution points, indicating that distributional distance alone is insufficient for uncertainty quantification and motivating integrated adaptation strategies.

The paper is organized as follows: Section 2 provides details of the FARM baseline, the datasets, and the definition of our multi-space drift monitoring framework. Section 3 presents our experimental results, first validating the framework's ability to detect the 2021 drift event and then analyzing the disconnect between distributional distance and prediction error. Finally, Section 4 concludes with a discussion of implications for robust agricultural AI systems in dynamic, non-stationary environments.

2 METHODOLOGY

2.1 FARM BASELINE MODEL: ARCHITECTURE AND FINE-TUNING SUMMARY

Our base GFM is Prithvi-EO-2.0-600M, a ViT pre-trained using a masked autoencoder objective on Landsat-8 and Sentinel-2 imagery. To adapt Prithvi to crop yield regression, the FARM model takes as input a temporal stack of satellite images (chips) $X \in \mathbb{R}^{T\times C\times H\times W}$ representing the growing season (T) (May to September) and comprises six spectral bands (C) (Blue, Green, Red, NIR, SWIR-1, SWIR-2) at $\sim$ 30m resolution. The encoder tokenizes 3D spatiotemporal patches and processes them through 32 Transformer layers. The embeddings from the final Transformer block are globally averaged across temporal and spatial dimensions by the decoder, followed by three fully connected layers that map the representation to a single yield prediction.

Regarding the ground-truth crop yield values, our previous study proceeded in two phases. First, the architecture was trained on a large county-level yield dataset with lower spatial resolution (see Section 2.2 for more details). These county-level yield values were up-sampled to the pixel level to enable training, and the resulting model—denoted **FARM-C**—achieved an RMSE of 0.4368 and an R^2 of 0.8105. Because this coarse-resolution dataset provides broad coverage but lacks field-level detail, we next fine-tuned the FARM-C model on a smaller, high-resolution Precision-Agriculture (PA) yield dataset (Experiment 2 in Nejadshamsi et al. (2026)). The fine-tuned model denoted **FARM-PA**, achieved an RMSE of 0.628 and an R^2 of 0.768, representing a substantial improvement

over existing baseline models and demonstrating that the foundation model offers a strong starting point for fine-grained yield estimation. Throughout the remainder of this paper, we treat FARM-C and FARM-PA as fixed baselines and focus on assessing their prediction reliability under distribution drift.

2.2 DATASET

According to Section 2.1, our experiments are conducted using two ground-truth yield datasets with different spatial resolutions:

County-level yield dataset: This dataset comprises multi-year (2018–2023) yield values aggregated from official agricultural statistics for the Canadian Prairies. Each label represents the average yield for an entire county, providing broad geographic coverage but relatively coarse spatial detail. The FARM-C model is trained and evaluated using this dataset. While these administrative areas are officially designated as Rural Municipalities (RMs) in Manitoba and Saskatchewan, we refer to them as 'counties' throughout this work to maintain consistency with the terminology commonly used in agricultural remote sensing literature."

Precision-Agriculture (PA) yield dataset: This dataset contains high-resolution ($\sim$30 m) yield measurements collected by combine-mounted yield monitors within field boundaries. Labels capture intra-field yield variability rather than county-level aggregates and are therefore more suitable for fine-grained modeling. The FARM-PA model is obtained by fine-tuning the FARM-C model using this dataset.

Input representation: For both datasets, multi-temporal satellite imagery is processed into fixed-size image chips of 224×224 pixels, each covering the growing season from May through September. To capture localized spatial variation, each chip is further partitioned into non-overlapping 16×16 pixel patches. Chip-level representations are used for county-scale analyses, while patch-level representations enable fine-grained modeling and drift analysis for the PA dataset.

Additional details regarding spatial resolution, label construction, and train–test splits are provided in Nejadshamsi et al. (2026).

2.3 DRIFT MONITORING FRAMEWORK

Beyond assessing predictive accuracy, we evaluate prediction reliability in the inference phase by detecting deviations between new inputs and the training distribution. The first step is to compute a baseline distribution from the training set and define distance-based metrics relative to that reference. Throughout this study we monitor drift in two complementary feature spaces: the *input space*, consisting of multi-spectral satellite image features, and the *latent space*, consisting of learned representations extracted from the foundation model encoder.

2.3.1 BASELINE DISTRIBUTION AND METRIC SPECIFICATION

Baseline construction. We aggregate chip- and patch-level statistics across the training years to form a baseline distribution. Multi-temporal image chips (224×224) are partitioned into 16×16-pixel patches. For each chip or patch we compute the mean and standard deviation of reflectance across six spectral bands and the five monthly time steps of the growing season (May–September), producing a 12-dimensional feature vector that defines the baseline in the input space. For the latent space we pass the training chips through the Prithvi–EO–V2 backbone of the fine-tuned FARM model, extract features from the neck layer where transformer tokens are reshaped into a spatial grid and apply adaptive average pooling to obtain a 7×7 grid of latent vectors. These chip- and patch-level vectors constitute the baseline distribution for subsequent drift analysis.

Distance metrics. We quantify deviation from the baseline using Mahalanobis distance, cosine similarity and Euclidean distance Zamzmi et al. (2024). Denote by μ and Σ the mean vector and covariance matrix of the baseline distribution (in either input or latent space). For a sample v, we compute: (i) Mahalanobis distance $d_M(v) = \sqrt{(v-\mu)^\top \Sigma^{-1}(v-\mu)}$, which accounts for feature variances and covariances Mahalanobis (1936); (ii) cosine similarity $\cos(v, \mu) = \frac{v \cdot \mu}{\|v\| \, \|\mu\|}$, capturing directional (semantic) differences Singhal (2001); and (iii) Euclidean distance $d_E(v) = \|v - \mu\|_2$, capturing absolute magnitude differences Deza & Deza (2009). These metrics are applied to both

input and latent feature vectors at chip and patch resolutions. We will reference this baseline and metric specification in subsequent sections rather than repeating methodological details.

2.3.2 PERFORMANCE QUANTIFICATION

To assess operational prediction reliability, we quantify predictive performance (RMSE) at the same spatial granularities used for drift monitoring. The fine-tuned models (FARM-C and FARM-PA) are applied to multi-temporal imagery from a designated evaluation year for which drift status is assessed (the *drift year*). In this study, 2021 serves as the primary drift year, as it was identified through agronomic records as a non-typical growing season characterized by widespread drought conditions across the Canadian Prairies. Yield estimates generated for this year are compared against ground-truth measurements at both the county and PA resolutions. Calculating error metrics at chip and patch levels provides localized insight into model performance with the same spatial precision as the drift metrics.

2.4 CATCH: DETECTING DISTRIBUTION DRIFT

The Catch phase applies the baseline and metrics defined above to identify distributional shifts. For a given drift year y, we first construct a reference distribution by computing μ and Σ from all other years in the dataset, and then compute the distance metrics defined in Section 2.3.1. Comparing the distribution of year y to this baseline reveals whether it is out-of-distribution (OOD). Based on preliminary validation detailed in Section 3.1, we establish three criteria for the Catch mechanism. First, regarding metric sensitivity, we rely primarily on *latent-space Mahalanobis distance*, as it demonstrates higher sensitivity to true distribution drift than input-space metrics. Second, we normalize distances relative to the baseline to enable accurate cross-space and cross-year comparisons. Finally, a "Catch" event is triggered when there is a statistically significant shift in the normalized latent Mahalanobis distribution for a given year relative to the aggregate baseline, supported by increased alignment (correlation) between input and latent distances.

3 EXPERIMENTS AND RESULTS

3.1 VALIDATING DRIFT DETECTION

We first validate the sensitivity of our drift metrics by comparing a known anomalous year (2021) and a known normal year (2022) against the baseline constructed from all other years.

Detecting the 2021 Drift. Figure 1 presents a comparison between the baseline distribution and the 2021 growing season at the PA patch level. In the input space, the overall mean distribution for 2021 is right-shifted relative to the baseline, with the Q-Q plot exhibiting systematic deviations and heavier upper tails. This indicates a pronounced covariate shift. Band-wise analyses reveal coherent mean shifts across all spectral channels (particularly NIR and SWIR), though substantial overlap remains. In the latent space, the distributional shift becomes more structured. Principal Component Analysis (PCA) projections show that the separation between baseline and 2021 samples is driven primarily by the second principal component (PC2), while PC1 distributions largely overlap. The PC2 distribution for 2021 exhibits a clear shift in both location and shape. This indicates that the foundation model encodes the 2021 drift as a directional change within a specific latent subspace rather than as a uniform shift.

Verifying Stability in 2022. In contrast to the pronounced shifts observed for 2021, the 2022 growing season closely aligns with the baseline distribution in both input and latent spaces (Figure 2). The normalized distance distributions for 2022 largely overlap with those of the reference years. This is reflected in substantially smaller dispersion of latent-space distances and a lower correlation between input- and latent-space drift metrics ($r \approx 0.46$), compared to the stronger coupling observed for 2021 ($r \approx 0.75$). These results confirm that our framework correctly identifies 2022 as an in-distribution reference year and 2021 as an OOD year.

These observations lead to three properties of the Catch mechanism. First, latent-space Mahalanobis distance is more sensitive to true distribution drift than input-space metrics. Second, normalization and baseline construction are essential for cross-space and cross-year comparisons. Third, comparing a single year against the aggregate of other years provides a label-free signal for detecting OOD years. We define the Catch trigger as a statistically significant shift in the normalized latent Mahalanobis distribution for a given year relative to its baseline, supported by increased alignment between input and latent distances. Nevertheless, detection of distributional drift alone does not

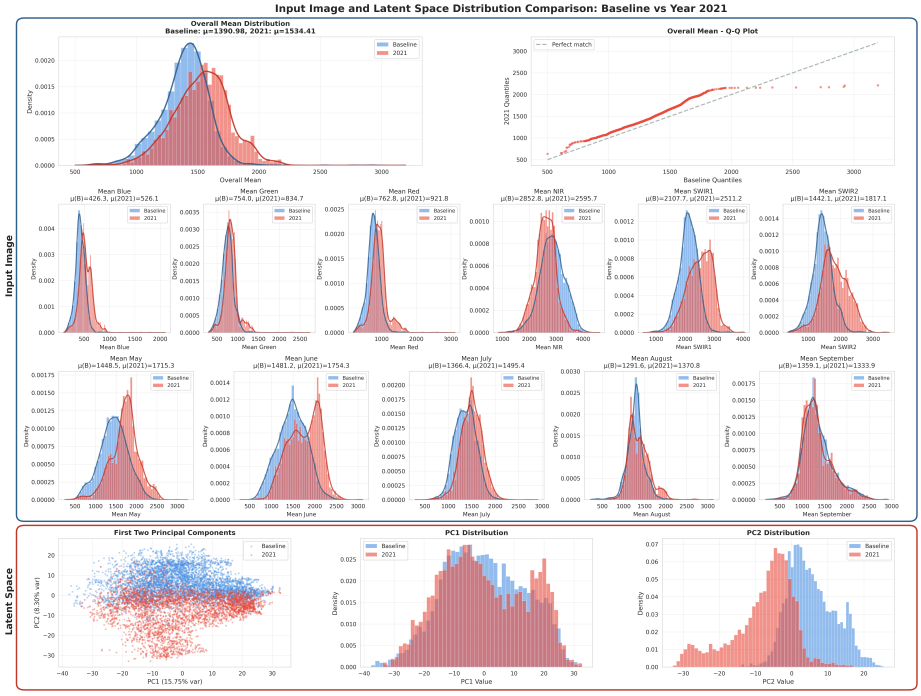

Figure 1: Comprehensive distribution comparison between the baseline years and the **2021** growing season at the PA patch level. **Top row:** Overall input-space mean distribution and Q-Q plot. **Second row:** Band-wise input-space distributions. **Third row:** Month-wise input-space distributions. **Bottom row:** Latent-space analysis (PCA scatter and marginal distributions).

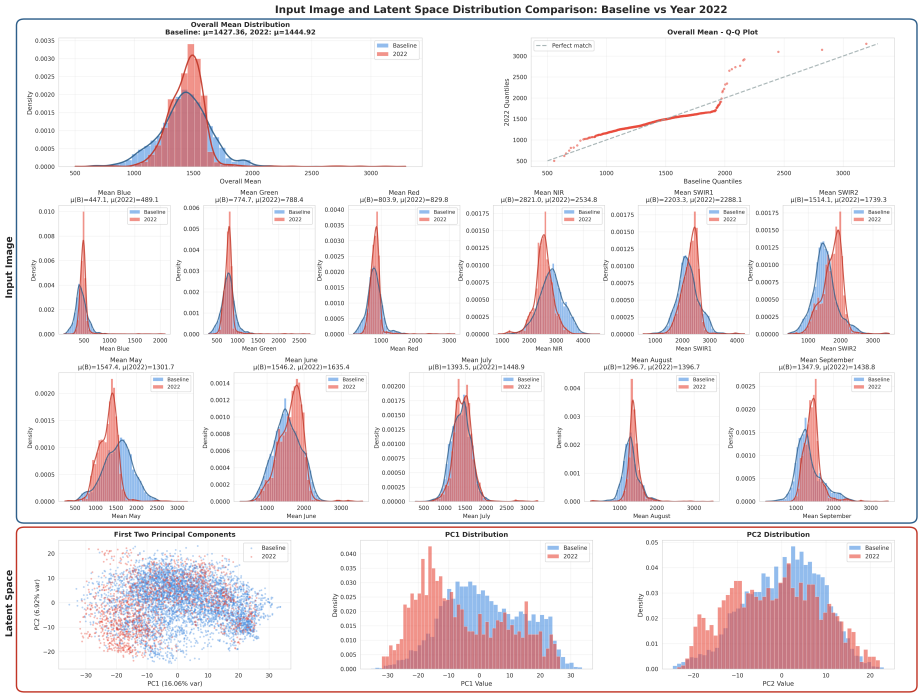

Figure 2: Comprehensive distribution comparison between the baseline years and the **2022** growing season at the PA patch level.

necessarily imply that a sample will incur a high prediction error. In our next experiment we explore the relationship between drift metrics and prediction error across both county and patch scales.

3.2 The Disconnect: Drift vs. Prediction Error

Having established that our metrics can detect drift, we next evaluate whether distributional distance metrics can serve as reliable indicators of prediction error. A common assumption in drift-aware monitoring is that model reliability degrades as inputs deviate from the training manifold; here, we explicitly examine whether samples exhibiting larger statistical distances from the baseline distribution also incur higher yield estimation error. To this end, chip- and patch-level RMSE is compared against Mahalanobis, Euclidean, and cosine distance metrics computed in both the input spectral space and the learned latent representation space. These analyses are conducted across spatial scales and for both county-level and PA-level datasets.

County level. Figure 3 summarizes the results for the 2021 growing season—identified as a drift year—using the FARM-C model and county-level ground-truth yields at the chip level. Contrary to the expectation that larger distributional distance corresponds to increased error, we observe no strong positive relationship between drift magnitude and RMSE. Scatter plots of RMSE versus Mahalanobis distance in both the input space and latent space (Figure 3(a)) show that some counties with the largest drift scores incur low prediction error, while several samples that are close to the training distribution exhibit large RMSE.

Spearman rank correlations between distance metrics and RMSE are weak or negative, indicating that distance alone explains little of the observed error variability. While latent-space Mahalanobis distance exhibits greater sensitivity to distributional shift than its input-space counterpart, this increased sensitivity does not translate into reliable error prediction. Similar behavior is observed for cosine similarity and input-space Euclidean distance (Figure 3(b)). Together, these results indicate that, at the county scale, the FARM-C model is able to generalize to distributionally distant inputs, and that drift detection metrics alone are insufficient proxies for prediction uncertainty.

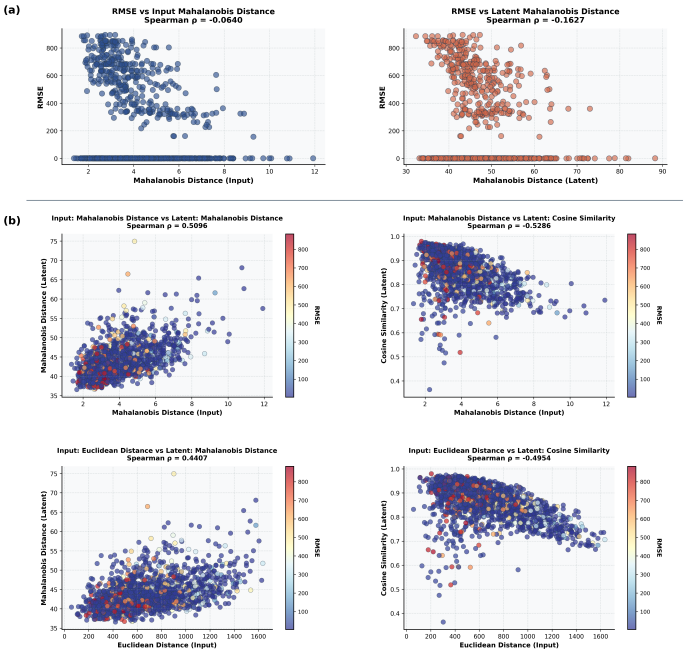

Figure 3: Reliability analysis of the FARM-C model under 2021 distribution drift at the county chip level. (a) Error-Distance Relationship: Scatter plots of chip-level RMSE against Mahalanobis distance in the input spectral space (left) and the latent embedding space (right). (b) Multi-Space Metric Comparison: Correlation between input-space metrics and latent-space metrics. Points are color-coded by RMSE to visualize the distribution of error across the drift spectrum.

PA level. At the PA scale, we observe stronger monotonic relationships between drift metrics and prediction error compared to the county-level analysis, particularly in the latent space. Figure 4(a)

plots patch-level RMSE against Mahalanobis distance in both the input and latent spaces for the 2021 growing season. While the input-space Mahalanobis distance exhibits a moderate negative Spearman correlation with RMSE ($\rho \approx -0.36$), the latent-space Mahalanobis distance shows a substantially stronger negative association ($\rho \approx -0.50$). This indicates that patches that are more distant in latent space are, on average, predicted with lower error.

Despite this stronger correlation, the scatter remains broad, with considerable overlap across the drift spectrum: patches with large latent distances can still incur low or moderate error, and vice versa. Similar behavior is observed when comparing input-space distances to latent-space cosine similarity (Figure 4(b)), where moderate correlations reflect structured drift without yielding a sharp separation between low- and high-error samples. Overall, although latent-space metrics are more sensitive to distributional shift at the PA level, distributional distance alone remains insufficient as a reliable proxy for prediction error. These results further suggest that the FARM-PA model can extrapolate into drifted regions of the input space without consistently degrading performance.

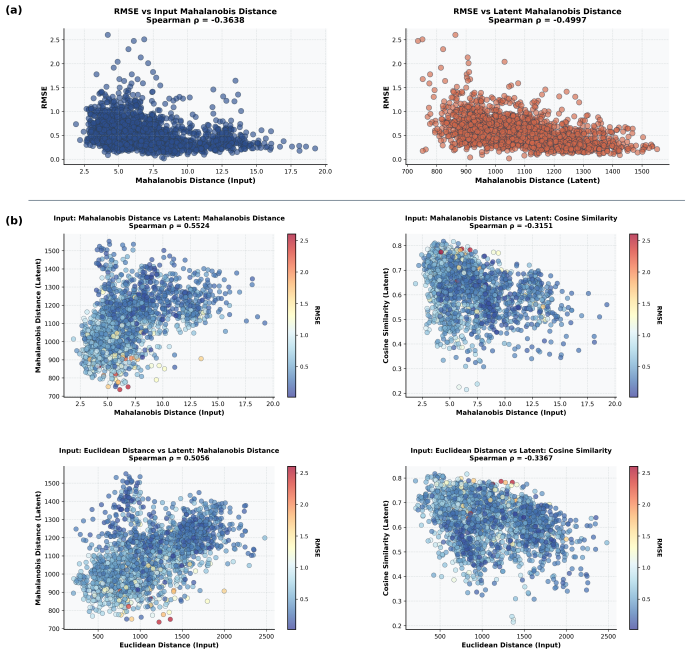

Figure 4: Reliability analysis of the FARM-PA model under 2021 distribution drift at the PA patch level. (a) Error-Distance Relationship: Scatter plots of patch-level RMSE against Mahalanobis distance in the input spectral space (left) and the latent embedding space (right). (b) Multi-Space Metric Comparison: Correlation between input-space metrics and latent-space metrics. Points are color-coded by RMSE to visualize the distribution of error across the drift spectrum.

3.3 Fine-Grained Analysis

This experiment was motivated by our prior exploratory interpretability analysis of the FARM model, which showed that certain months and spectral bands contribute more strongly to yield prediction Nejadshamsi et al. (2026). We therefore examined whether the degree of drift in each individual spectral band or monthly composite would exhibit a different relationship with RMSE when analyzed in isolation. To this end, we computed Mahalanobis, Euclidean, and cosine distances separately for each of the six spectral bands and for each monthly composite from May through September, and correlated these per-band and per-month drift measures with prediction error.

Although space constraints prevent us from presenting all results, Figures 5 and 6 illustrate representative cases. In particular, Figure 5 shows RMSE plotted against Mahalanobis distance in both the input (image) space and the latent space for the July composite, while Figure 6 presents the same analysis for the NIR band. In both cases, Mahalanobis distance in the latent space exhibits a substantially stronger negative Spearman correlation with RMSE than the corresponding image-space Mahalanobis distance, while input-space drift shows only weak or negligible association with error

magnitude despite statistical significance driven by sample size. Color-coding points by normalized true yield in Figure 5(a,b) reveals no clear relationship between yield magnitude and RMSE: both low- and high-yield samples achieve low prediction error across the latent drift spectrum. This indicates that the observed latent-space correlation with RMSE is not driven by extreme yield values, but instead reflects properties of the learned representation.

Importantly, despite the stronger latent-space correlation, the overall trend remains inconsistent and non-predictive: large drift values—whether computed per band or per month—do not reliably correspond to high RMSE. This observation holds across all spectral bands and temporal composites. These results reinforce our earlier findings that distributional drift, even when isolated to specific bands or phenological stages, is not a reliable proxy for performance degradation. The FARM model is therefore able to extrapolate into drifted spectral–temporal regimes without systematically incurring large prediction errors.

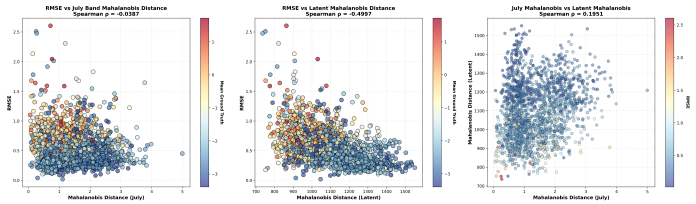

Figure 5: PA-level patch-wise analysis for the 2021 drift year (**July composite**). (a) RMSE vs. Input-space Mahalanobis distance. (b) RMSE vs. Latent-space Mahalanobis distance. (c) Input-space vs. Latent-space Mahalanobis distance. Points in (a) and (b) are color-coded by normalized true yield; points in (c) are color-coded by RMSE.

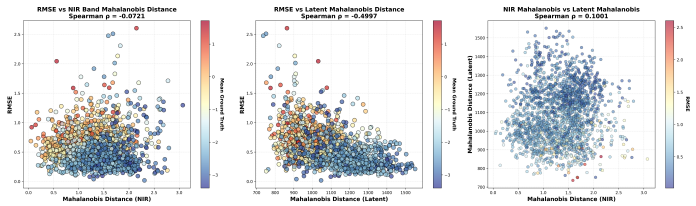

Figure 6: PA-level patch-wise analysis for the 2021 drift year (**NIR**). (a) RMSE vs. Input-space Mahalanobis distance. (b) RMSE vs. Latent-space Mahalanobis distance. (c) Input-space vs. Latent-space Mahalanobis distance. Points in (a) and (b) are color-coded by normalized true yield; points in (c) are color-coded by RMSE.

4 Conclusion

We have presented a systematic reliability analysis framework for the FARM model, evaluating its robustness under distribution drift. By computing Mahalanobis, cosine, and Euclidean distances in both input and latent spaces, we demonstrate that representation-based metrics can effectively *sense* distributional drift—such as the anomalous 2021 growing season—across.

Our experiments further show that large drift scores do not necessarily imply large prediction errors: samples that are far from the training manifold can still be predicted accurately, while some in-distribution samples incur high RMSE. This reveals a critical disconnect between drift detection and error estimation, highlighting the limitations of relying on distributional distance alone for operational decision-making. In this sense, our methodology aligns with the workshop's themes by (i) *sensing drift* through multi-space, representation-aware metrics, (ii) *characterizing the implications of drift* to lay the groundwork for a Catch–Adapt–Operate conceptual framework, and (iii) *evaluating robustness at scale* using large, multi-year datasets spanning multiple spatial resolutions.

Future work will focus on the *Adapt* phase by developing error-aware drift indicators, integrating uncertainty estimation with representation-space metrics, and designing adaptation strategies that distinguish harmful shifts from benign distributional changes. We will also investigate causal drivers underlying the observed mismatch between drift magnitude and prediction error.

REFERENCES

Bruno Basso and Lin Liu. Chapter four - seasonal crop yield forecast: Methods, applications, and accuracies. volume 154 of *Advances in Agronomy*, pp. 201–255. Academic Press, 2019. doi: https://doi.org/10.1016/bs.agron.2018.11.002. URL https://www.sciencedirect.com/science/article/pii/S0065211318300944.

Elena Deza and Michel Marie Deza. *Encyclopedia of Distances*. Springer, 2009. doi: 10.1007/978-3-642-00234-2.

Saeed Khaki, Lizhi Wang, and Sotirios V Archontoulis. A cnn-rnn framework for crop yield prediction. *Frontiers in Plant Science*, 10:1750, 2020.

Guoyong Leng and Jim Hall. Crop yield sensitivity of global major agricultural countries to droughts and the projected changes in the future. *Science of the Total Environment*, 654:811–821, 2019.

Corey Lesk, Pedram Rowhani, and Navin Ramankutty. Influence of extreme weather disasters on global crop production. *Nature*, 529(7584):84–87, 2016.

Hong Li, Lysa Porth, Ken Seng Tan, and Wenjun Zhu. Improved index insurance design and yield estimation using a dynamic factor forecasting approach. *Insurance: Mathematics and Economics*, 96:208–221, 2021.

David B. Lobell, Wolfram Schlenker, and Justin Costa-Roberts. Climate trends and global crop production since 1980. *Science*, 333(6042):616–620, 2011. doi: 10.1126/science.1204531. URL https://www.science.org/doi/abs/10.1126/science.1204531.

Prasanta Chandra Mahalanobis. On the generalized distance in statistics. *Proceedings of the National Institute of Sciences of India*, 2:49–55, 1936.

Shayan Nejadshamsi, Yuanyuan Zhang, Brock Porth, Shadi Zaki, Lysa Porth, and Vahab Khoshdel. Farm: Crop yield prediction via regression on prithvi's encoder for satellite sensing. *AgriEngineering*, 8(1), 2026. ISSN 2624-7402. doi: 10.3390/agriengineering8010002. URL https://www.mdpi.com/2624-7402/8/1/2.

Amit Singhal. Modern information retrieval: A brief overview. *IEEE Data Engineering Bulletin*, 24(4):35–43, 2001.

Daniela Szwarcman, Sujit Roy, Paolo Fraccaro, orsteinn Elí Gíslason, Benedikt Blumenstiel, Rinki Ghosal, Pedro Henrique de Oliveira, Joao Lucas de Sousa Almeida, Rocco Sedona, Yanghui Kang, et al. Prithvi-eo-2.0: A versatile multi-temporal foundation model for earth observation applications. *arXiv preprint arXiv:2412.02732*, 2024.

Ghada Zamzmi, Krithika Venkatesh, Benjamin Nelson, et al. Out-of-distribution detection and radiological data monitoring using statistical process control. *Journal of Imaging Informatics in Medicine*, 2024. doi: 10.1007/s10278-024-01212-9.