Advanced Forecast Evaluation Metrics: A Comprehensive Guide

Introduction

Overview

This report details a collection of advanced forecast evaluation metrics, designed to provide a comprehensive assessment of forecasting model performance. While standard metrics like Mean Absolute Percentage Error (MAPE) or Root Mean Square Error (RMSE) offer valuable summaries of point forecast accuracy, they often fail to capture the full picture of a model's behavior and its implications for business decisions. Robust evaluation extends beyond simple accuracy, encompassing aspects like systematic bias, handling of outliers, prediction of dynamics, calibration of uncertainty, and stability over time. Such comprehensive evaluation is critical not only for selecting the best forecasting model but also for ongoing monitoring, diagnosing performance issues, and understanding the practical consequences of forecast quality.[1]

The Need for Multifaceted Evaluation

Relying on a single error metric can be misleading, as different metrics illuminate distinct facets of forecast performance. For example, a forecast might exhibit low average error (e.g., low MAE) but possess a significant systematic bias, or it might accurately predict stable periods but fail catastrophically during volatile times or at turning points. A multifaceted approach, employing a suite of metrics, is necessary to gain a holistic understanding of a forecast's strengths and weaknesses.[1] This deeper understanding is crucial because forecast quality directly impacts key business outcomes. Accurate and reliable forecasts are foundational for effective inventory management, resource allocation, waste reduction, cost optimization, financial planning, and ultimately, profitability and customer satisfaction.[1]

Structure of the Report

This report is organized into five main sections, mirroring the logical grouping of the evaluation metrics:

1. **Systematic Bias Assessment:** Metrics designed to detect and quantify consistent over- or under-forecasting tendencies.
2. **Outlier & Anomaly Behavior Analysis:** Metrics focused on identifying and characterizing unusual behavior in the actual data or forecast errors.
3. **Direction & Velocity Dynamics Evaluation:** Metrics assessing the forecast's

ability to capture the time series' movement, speed, acceleration, and turning points.

4. **Probabilistic Calibration & Sharpness Assessment:** Metrics for evaluating forecasts that provide uncertainty estimates, focusing on their reliability and precision.
5. **Distributional Drift & Stability Monitoring:** Metrics aimed at detecting changes in the statistical properties of data or forecast errors over time.

Within each section, every metric is detailed, covering its technical definition, mathematical calculation, business interpretation supported by research findings, deeper implications and connections to other concepts, and inherent limitations.

Target Audience

The intended audience for this report includes Data Scientists, Quantitative Analysts, Machine Learning Engineers, and technical managers who are involved in developing, deploying, evaluating, or utilizing forecasting systems. A solid technical background is assumed, but the report aims to provide clear explanations of the underlying concepts, interpretations, and business relevance of each metric.

Summary Table of Metrics

The following table provides a high-level overview of the 25 forecast evaluation metrics discussed in this report, categorizing them by the primary aspect of performance they measure and their typical use case. This serves as a quick reference guide to the capabilities covered.

Group	Metric Name	Brief Description	Key Aspect Measured	Primary Use Case
1. Systematic Bias	Mean Bias (MFE)	Average difference between predictions and actuals.	Average Error Direction & Magnitude	Overall Bias Check
1. Systematic Bias	Tracking Signal	Cumulative error scaled by mean absolute deviation of errors.	Error Persistence vs. Magnitude	Monitoring Bias Stability Over Time

1. Systematic Bias	Residual Sign Count Difference	Difference between counts of positive and negative errors.	Frequency of Over/Under Prediction	Non-parametric Directional Bias Check
1. Systematic Bias	Area Under Sparsification Curve	Area under the curve of error magnitude threshold vs. coverage.	Cumulative Distribution of Errors	Comparing Overall Error Magnitudes
2. Outlier & Anomaly Behavior	Data Anomaly Rate (MAD-based)	Fraction of actuals deemed anomalous based on MAD.	Inherent Data Volatility	Assessing Baseline Data Predictability
2. Outlier & Anomaly Behavior	Residual Anomaly Rate (Std Dev-based)	Fraction of residuals deemed anomalous based on standard deviation.	Frequency of Large Forecast Errors	Identifying Forecast "Busts"
2. Outlier & Anomaly Behavior	Mean Anomalous Residual Magnitude	Average magnitude of anomalous residuals.	Severity of Large Forecast Errors	Assessing Impact of Forecast "Busts"
2. Outlier & Anomaly Behavior	Average Time of Error Occurrence	Average index of non-zero residuals.	Temporal Location of Errors	Evaluating Error Timing/Horizon Degradation
2. Outlier & Anomaly Behavior	Error Persistence (Longest Run)	Longest consecutive run of non-zero residuals.	Duration of Error Streaks	Assessing Error Autocorrelation/ Self-Correction
3. Direction & Velocity Dynamics	Directional Accuracy	Fraction of correct up/down/flat movement predictions (first	Directional Correctness	Evaluating Trend Capture

		differences).		
3. Direction & Velocity Dynamics	Velocity Error (MAE of 1st Diff)	MAE of the first differences (error in rate of change).	Momentum Accuracy (1st Derivative)	Evaluating Prediction of Change Magnitude
3. Direction & Velocity Dynamics	Acceleration Error (MAE of 2nd Diff)	MAE of the second differences (error in rate of change of rate of change).	Curvature Accuracy (2nd Derivative)	Evaluating Prediction of Momentum Shifts
3. Direction & Velocity Dynamics	Turning Point F1 Score	F1 score for detecting local peaks and valleys.	Turning Point Detection Accuracy	Assessing Ability to Predict Reversals
3. Direction & Velocity Dynamics	Trend Change Detection Delay	Average lag between actual and predicted turning points.	Lag in Trend Reversal Detection	Quantifying Forecast Reaction Time at Reversals
4. Probabilistic Calibration	Prediction Interval Coverage Probability (PICP)	Percentage of actuals falling within a nominal prediction interval.	Interval Reliability/Calibration	Validating Stated Confidence Levels
4. Probabilistic Calibration	Average Prediction Interval Width	Average width of the nominal prediction interval.	Interval Sharpness/Precision	Assessing Forecast Uncertainty Magnitude
4. Probabilistic Calibration	Winkler Interval Score	Winkler score combining interval width and penalty for misses.	Combined Calibration & Sharpness	Overall Evaluation of Prediction Intervals
4. Probabilistic Calibration	Continuous Ranked	Continuous Ranked	Overall Probabilistic	Evaluating Full Predictive

	Probability Score (CRPS)	Probability Score assuming Gaussian predictive distribution.	Accuracy	Distribution (Gaussian)
5. Distributional Drift/Stability	Sliding Window Jensen-Shannon Distance (JSD)	Average Jensen-Shannon Distance between actuals/preds in sliding windows.	Local Distribution Similarity	Detecting Dynamic/Local Distribution Drift
5. Distributional Drift/Stability	Population Stability Index (PSI)	Population Stability Index comparing actuals and predictions distributions.	Overall Distribution Shift	Quantifying Static Output Distribution Mismatch
5. Distributional Drift/Stability	Kullback-Leibler (KL) Divergence	Kullback-Leibler Divergence from predictions distribution to actuals distribution.	Information Loss / Asymmetric Shift	Measuring Divergence (Actuals from Preds)
5. Distributional Drift/Stability	Rolling Error Variance	Mean variance of residuals in sliding windows.	Error Variance Stability	Monitoring Consistency of Error Magnitude
5. Distributional Drift/Stability	Maximum Mean Discrepancy (MMD)	Maximum Mean Discrepancy between actuals and predictions using RBF kernel.	Non-parametric Distribution Distance	Detecting General Distributional Differences

---

Section 1: Systematic Bias Assessment

Overview

Systematic bias in forecasting refers to a consistent, non-random tendency for predictions to deviate from actual outcomes in a particular direction – either

consistently overestimating or consistently underestimating the true value. This is distinct from random error, which should average out over time. Persistent bias is detrimental because it distorts expectations and leads to suboptimal decisions regarding planning, inventory management, resource allocation, and financial budgeting.[5] For instance, overly optimistic forecasts (positive bias) can result in excess inventory, wasted resources, and inflated revenue expectations, while overly conservative forecasts (negative bias) can lead to stockouts, missed sales opportunities, and under-allocation of resources.[7] Identifying and quantifying systematic bias is therefore a fundamental step in forecast evaluation, helping to pinpoint potential flaws in the modeling process, assumptions, or even external influences like misaligned incentives.[12] This section details metrics designed to measure different aspects of systematic forecast bias.

1.1 Mean Bias (Mean Forecast Error)

- **Technical Definition:** The Mean Bias, also commonly referred to as Mean Forecast Error (MFE), is the arithmetic average of the forecast errors (residuals) over a given period. The residual is calculated as the predicted value minus the actual value.
- **Mathematical Computation:** Given actual values yt and predicted values y^t for t=1,...,N, the residuals are et=y^t−yt. The Mean Bias is: Mean Bias=N1t=1ΣN(y^t−yt)=N1t=1ΣNet
- **Business Understanding & Interpretation:**
    - This metric provides a direct measure of the average direction and magnitude of forecast errors.[5] A positive Mean Bias indicates that, on average, the forecasts are higher than the actual outcomes (over-forecasting). Conversely, a negative Mean Bias signifies an average tendency to under-forecast (actuals are higher than predictions).[5]
    - In an ideal scenario, the Mean Bias should be close to zero, suggesting a balanced forecasting process where over-predictions and under-predictions cancel each other out on average.[5] A consistently non-zero bias, whether positive or negative, points towards systemic issues that require investigation. These could stem from model misspecification, incorrect assumptions about trends or seasonality, biased input data, or even organizational factors like sales targets influencing forecast inputs.[8]
    - The business impact of bias is significant. Consistent over-forecasting can lead to excessive inventory holding costs, capital tied up unnecessarily, increased risk of obsolescence or waste (especially for perishable goods), and misallocation of production or staffing resources.[7] Consistent under-forecasting can result in stockouts, lost sales revenue, unmet customer

demand leading to dissatisfaction, and potentially emergency measures (like expedited shipping) that increase costs.[7]
    - While related to forecast *attainment* (often defined as Actual / Forecast), Mean Bias offers a different perspective. Attainment measures performance against the target (the forecast), while bias measures the systematic error direction relative to the actual outcome. A consistent attainment below 100% often corresponds to a negative bias (under-forecasting), and attainment above 100% to a positive bias (over-forecasting), but bias quantifies the average error magnitude directly.[1]
    - Mean Bias is widely used due to its simplicity in calculation and interpretation.[2] However, a major drawback is that it averages errors over the entire period. Large positive errors can be cancelled out by large negative errors, resulting in a Mean Bias close to zero even when the forecast is consistently inaccurate in opposite directions for different periods or conditions.[17]
- **Deeper Implications & Connections:**
    - The observation that Mean Bias can mask offsetting errors highlights its insufficiency as a sole measure of systematic error. A forecast might accurately predict the average level but exhibit conditional bias – for example, consistently over-predicting high-demand periods and under-predicting low-demand periods. While the average bias might be near zero, the forecast fails systematically under specific conditions. This necessitates the use of complementary metrics like the Tracking Signal, which assesses persistence, or graphical residual analysis to uncover such patterns.[17]
    - Even a small but persistent bias can have a substantial cumulative impact over longer planning horizons. In inventory management, a consistent 5% under-forecast might seem minor in any single period but translates into chronic under-stocking, leading to accumulating lost sales, potential erosion of customer loyalty, and ultimately reduced market share over months or years.[7] This underscores the operational significance of bias detection, which may outweigh concerns about minor improvements in scale-dependent accuracy metrics like MAE or RMSE if those improvements come at the cost of introducing bias.
- **Limitations & Considerations:** Averages positive and negative errors, potentially masking significant inaccuracies. Provides a static view and doesn't indicate if the bias is consistent over time or varies. Does not measure the magnitude of typical errors (addressed by MAE/RMSE).

1.2 Tracking Signal

- **Technical Definition:** The Tracking Signal is a monitoring metric that assesses

the consistency of forecast bias over time. It is typically calculated as the ratio of the cumulative sum of forecast errors (the running sum of residuals) to the mean absolute deviation (MAD) of the forecast errors.

- **Mathematical Computation:** Let et=y^t−yt be the residual at time t. The cumulative error up to time N is CEN=Σt=1Net. The Mean Absolute Deviation of the errors is MADN=N1Σt=1N∣et∣. The Tracking Signal at time N is: $$ \text{Tracking Signal}*N = \frac{CE_N}{MAD_N} = \frac{\sum*{t=1}^{N} e_t}{\frac{1}{N} \sum_{t=1}^{N} |e_t|} $$ (Note: Some definitions use a smoothed MAD, like MADt=β∣et∣+(1−β)MADt−1.[21] The denominator MADN is prevented from being zero by using a minimum value if the MAD calculation results in zero).
- **Business Understanding & Interpretation:**
  - The primary purpose of the Tracking Signal is to detect *persistent* or *systematic* bias that accumulates over time.[20] It flags situations where the forecast is consistently erring in one direction for a sustained period.
  - Interpretation relies on comparing the Tracking Signal value against predefined control limits or thresholds. Commonly cited thresholds are around ±3.75 or ±4.[5] A Tracking Signal value consistently falling outside this range is considered statistically unlikely under the assumption of random, unbiased errors, thus signaling a probable systematic bias in the forecast model.[16]
  - A sustained positive Tracking Signal indicates persistent under-forecasting (actuals tend to be higher than predictions), while a sustained negative value indicates persistent over-forecasting (predictions tend to be higher than actuals).[20]
  - It serves as an essential *early warning system* in operational forecasting environments.[21] A breach of the threshold suggests that the forecast model's validity might be compromised, perhaps due to changes in underlying demand patterns, market conditions, or model degradation, prompting a review and potential recalibration or adjustment.[21]
  - The common threshold value (e.g., 3.75) is often justified by assuming forecast errors are approximately normally distributed. Under normality, the standard deviation is roughly 1.25 times the MAD (specifically, σ≈π/2×MAD≈1.253×MAD). A threshold of 3.75 MAD then corresponds approximately to a 3-sigma limit (3.75/1.25=3) for the cumulative error, implying a low probability (around 0.3%) of exceeding this limit by chance if the forecast is truly unbiased.[20] However, the metric's reliability decreases significantly when the number of observations (N) is very small (e.g., N≤3), as the maximum possible Tracking Signal value is N.[25]
- **Deeper Implications & Connections:**
  - The Tracking Signal framework borrows directly from Statistical Process

Control (SPC) methodologies used in quality management. It essentially treats the forecasting process as a system whose output (the forecast error) should ideally be centered around zero with random variation. The Tracking Signal monitors the cumulative deviation from this target (zero bias), scaled by the typical variation (MAD), analogous to how control charts monitor process means.[20] A signal outside the limits suggests the process is "out of control" and requires intervention.

    - Compared to the Mean Bias, which provides a static average over a period, the Tracking Signal offers a dynamic perspective on bias. It is sensitive to the *persistence* and *accumulation* of errors. A forecast could have a near-zero Mean Bias over a long horizon but experience significant periods of sustained positive bias followed by sustained negative bias. The Tracking Signal, especially if calculated over rolling windows, would likely detect these persistent deviations that the overall Mean Bias might obscure.[21] This makes it particularly valuable for ongoing monitoring where the stability and consistency of bias are critical.

- **Limitations & Considerations:** Threshold values are heuristic and rely on distributional assumptions (often normality) that may not hold. Less reliable for very short time series.[25] The choice of MAD calculation (simple average vs. exponentially smoothed) can affect responsiveness.[21] It signals bias presence but not the root cause.

1.3 Residual Sign Count Difference

- **Technical Definition:** This metric calculates the difference between the number of positive residuals (where the prediction exceeded the actual value) and the number of negative residuals (where the prediction was less than the actual value). Residuals exactly equal to zero are typically ignored.
- **Mathematical Computation:** Let $e_t = \hat{y}_t - y_t$ be the residual at time t. The residual count difference is: $\text{Residual Counts} = \sum_{t=1} I(e_t > 0) - \sum_{t=1} I(e_t < 0)$ where $I(\cdot)$ is the indicator function (1 if the condition is true, 0 otherwise).
- **Business Understanding & Interpretation:**
    - This metric offers a simple, non-parametric check for directional bias by focusing on the *frequency* of over- versus under-forecasting, irrespective of the error magnitudes.[19]
    - A large positive value suggests the forecast predicts higher than the actual value more often than it predicts lower. A large negative value indicates the opposite – under-forecasting is more frequent.[19]
    - It serves as a useful complement to the Mean Bias. A Mean Bias near zero doesn't guarantee an equal number of positive and negative residuals. For

instance, numerous small positive errors could be balanced by a few large negative errors, yielding a low Mean Bias but a large positive residual count difference. This situation would indicate an asymmetric error distribution.
    - For a truly unbiased forecast, one would expect the number of positive and negative residuals to be roughly equal, resulting in a residual count difference close to zero.[19] This relates to the desirable property that forecast residuals should ideally have a mean (and median) of zero.[19]
    - Its non-parametric nature makes it a useful quick diagnostic, particularly when the distributional assumptions required for interpreting other metrics (like the normality assumption for Tracking Signal thresholds) might be questionable.[19]
- **Deeper Implications & Connections:**
    - Fundamentally, this metric performs a sign test on the forecast residuals. It assesses whether the median of the error distribution is significantly different from zero by comparing the counts of positive and negative signs. A substantial deviation from a 50/50 split (which could be formally tested using a binomial test, though here it's assessed informally by the magnitude of the difference) implies a bias in the forecast's central tendency.[13]
    - Comparing the result of this metric with the Mean Bias can reveal important characteristics of the error distribution's symmetry. If the residual count difference is large (e.g., strongly positive, indicating many over-forecasts) but the Mean Bias is small, it implies that the less frequent under-forecasts (negative residuals) must be significantly larger in magnitude on average to bring the overall sum of errors close to zero.[2] This points towards a skewed error distribution, where the model might be consistently making small errors in one direction and occasional large errors in the other.
- **Limitations & Considerations:** Completely ignores the magnitude of the errors; a large count difference might arise from many tiny errors. Statistical significance is not assessed; a large difference might occur by chance with few data points. Does not account for potential autocorrelation in the residuals, where errors might cluster together.[19]

1.4 Area Under Sparsification Curve

- **Technical Definition:** This metric computes the area under a "sparsification curve." This curve plots the cumulative coverage (fraction of data points included) against an increasing threshold based on the absolute magnitude of the forecast residual. The area is calculated numerically using the trapezoidal rule.
- **Mathematical Computation:**
    1. Calculate absolute residuals: $|e_t| = |\hat{y}_t - y_t|$.

2. Sort the absolute residuals in ascending order: $|e|_{(1)} \le |e|_{(2)} \le \dots \le |e|_{(N)}$. Let these sorted values be $r_i = |e|_{(i)}$.
3. Calculate the corresponding cumulative coverage values: $c_i = i/N$ for $i = 1, \dots, N$.
4. Approximate the integral of coverage with respect to the residual threshold using the trapezoidal rule: $$ \text{Area} = \int_{r_1}^{r_N} \text{coverage}(r) dr \approx \sum_{i=1}^{N-1} \frac{(c_i + c_{i+1})}{2} (r_{i+1} - r_i) $$

- **Business Understanding & Interpretation:**
  - This metric provides an assessment of how well the magnitude of forecast errors aligns with their frequency, drawing conceptual parallels to methods used in evaluating confidence or uncertainty estimates in fields like computer vision.[30] Here, the absolute error magnitude itself acts as the ranking criterion.
  - Interpretation centers on the shape of the curve being integrated. A curve that rises very steeply at low residual values indicates that a large fraction of the errors are small. Conversely, a curve that rises slowly implies a significant proportion of large errors. The area under this curve summarizes this distributional characteristic. A smaller area generally suggests a better-performing forecast in the sense that its errors are more concentrated at lower magnitudes.
  - When comparing different forecasting models applied to the same dataset, the model yielding a lower Area Under Sparsification curve might be preferred, as it indicates its errors tend to be smaller overall, or at least less spread out towards large values.
  - The concept is related to the Area Under the Sparsification Error (AUSE) curve sometimes used in uncertainty quantification.[30] In AUSE, data points are typically removed based on a separate confidence score, and the error on the remaining points is plotted. Here, points are effectively "removed" (or accumulated) based on the error magnitude itself.
- **Deeper Implications & Connections:**
  - This metric offers a non-parametric summary of the entire distribution of absolute error magnitudes, going beyond simple averages like MAE or RMSE. While MAE and RMSE provide single points summarizing the central tendency of errors, the sparsification curve visualizes the cumulative distribution, and the area under it provides a scalar summary of this distribution's shape, particularly its concentration towards zero. A model with a low MAE but a heavy tail of large errors would likely have a larger area compared to a model with a slightly higher MAE but fewer extreme errors.[30]
  - Monitoring changes in the Area Under Sparsification over time could serve as an indicator of shifts in the error distribution's characteristics. For example, if a model starts producing more frequent large errors (outliers), the

sparsification curve will tend to shift to the right, increasing the area under it. This might occur even if the mean error (MAE/RMSE) remains relatively stable due to offsetting changes in smaller errors. This offers a different perspective for detecting drift compared to metrics focused solely on the mean or variance of errors.[35]

- **Limitations & Considerations:** The interpretation is less direct than metrics like MAE or RMSE. The absolute value of the area depends on the scale of the forecast errors, making direct comparisons across datasets with different units difficult. It is primarily useful as a comparative metric between models on the same data or for tracking changes in a single model's error distribution over time. Its application in standard forecast evaluation practice is less common compared to its use in uncertainty calibration for perception models.

Bias Metrics Comparison

The following table summarizes the key characteristics of the bias metrics discussed in this section, highlighting their complementary roles in assessing systematic forecast errors.

Metric	What it Measures	Sensitivity	Primary Use
Mean Bias (MFE)	Average error magnitude and direction	Magnitude & Direction (Average)	Overall static bias check
Tracking Signal	Persistence of cumulative error relative to MAD	Error Persistence, Accumulation	Monitoring bias stability over time
Residual Sign Count Difference	Frequency difference between over/under forecasts	Frequency of Error Direction	Non-parametric directional bias check
Area Under Sparsification Curve	Cumulative distribution shape of absolute errors	Overall Distribution of Error Magnitudes	Comparing error concentration / magnitude

---

Section 2: Outlier & Anomaly Behavior Analysis

Overview

This section delves into metrics designed to identify, quantify, and characterize

anomalous or outlier behavior within the forecasting context. Anomalies can occur either in the actual observed data (representing unexpected real-world events, data errors, or high volatility) or in the forecast residuals (representing unusually large prediction errors). Understanding and monitoring anomalies is crucial for several reasons: it helps assess the inherent predictability of the data, identifies periods or events that challenge the model, evaluates the model's robustness to extreme values, informs risk management, and can signal data quality problems or shifts in the underlying process being forecast.[37] The metrics in this section examine the frequency (rate), size (magnitude), and duration (persistence) of such anomalies.

2.1 Data Anomaly Rate (MAD-based)

- **Technical Definition:** This metric calculates the fraction of data points in the actuals time series that are classified as anomalous. The classification is based on the deviation of each point from the series median, scaled by the Median Absolute Deviation (MAD).
- **Mathematical Computation:**
  1. Calculate the median of the actuals: med=median(actuals).
  2. Calculate the Median Absolute Deviation (MAD) of the actuals from their median: mad=median(|actuals−med|). Handle the case where mad=0.
  3. Calculate the modified Z-score for each point: Zmad,t=mad|actualst−med|.
  4. Determine the anomaly rate as the fraction of points where the modified Z-score exceeds a threshold k: Data Anomaly Rate=N1t=1ΣNI(Zmad,t>k) The default threshold is k=3.0.
- **Business Understanding & Interpretation:**
  - This metric quantifies the inherent "spikiness" or frequency of outliers within the historical actuals data, independent of any forecasting model. It utilizes the median and MAD, which are robust statistics, meaning they are less influenced by extreme values compared to the mean and standard deviation.[37]
  - A higher data anomaly rate indicates that the historical process being measured is more volatile, prone to extreme events, or potentially contains data quality issues. The threshold k controls the sensitivity; a common choice is k=3, which roughly corresponds to a 3-sigma deviation if the data were normally distributed, but the MAD-based approach is more robust for non-normal data.[37]
  - Understanding the baseline anomaly rate in the actuals is crucial for setting realistic expectations about forecast accuracy. Data with a high intrinsic anomaly rate is fundamentally harder to predict using standard time series models that primarily capture regular patterns.[7] Identifying these anomalies can also trigger investigations into their causes (e.g., promotions, holidays,

system outages, sensor errors) and inform decisions about data cleaning, preprocessing, or the need for causal models that explicitly incorporate such events.[38]

- **Deeper Implications & Connections:**
  - The data anomaly rate provides a measure of the inherent unpredictability or noise level in the target series. A high rate suggests that a significant portion of the data's variance is driven by irregular events rather than systematic patterns like trend or seasonality. This inherently limits the performance ceiling for any forecasting model that does not have access to information about the causes of these anomalies.[7] Models might smooth over these anomalies, leading to large errors at those points, or attempt to fit them, potentially distorting the forecasts for normal periods.
  - A valuable diagnostic step involves comparing the Data Anomaly Rate with the Residual Anomaly Rate (detailed next). This comparison helps distinguish between inherent data volatility and model-induced errors. If the data anomaly rate is low but the residual anomaly rate is high, it suggests the model itself might be unstable or poorly specified, generating large errors even for relatively smooth data. Conversely, if both rates are high, the model is likely struggling to cope with inherently challenging, volatile data. A scenario where the data rate is high but the residual rate is relatively low might indicate the model is effectively smoothing or dampening the anomalies, which could be desirable or undesirable depending on the application's goals.
- **Limitations & Considerations:** The choice of the threshold k is somewhat arbitrary, though values around 3 are common. While MAD is robust, the interpretation of the k threshold as equivalent to standard deviations relies on an assumption of approximate symmetry around the median. The metric counts anomalies but doesn't differentiate their types (e.g., point anomalies vs. contextual anomalies vs. collective anomalies).

2.2 Residual Anomaly Rate (Std Dev-based)

- **Technical Definition:** This metric calculates the fraction of forecast residuals (errors) that are considered anomalous. Anomalies are identified based on the deviation of each residual from the mean residual, scaled by the standard deviation of the residuals.
- **Mathematical Computation:**
  1. Calculate the residuals: $e_t = \hat{y}_t - y_t$.
  2. Calculate the mean of the residuals: $\bar{e} = \frac{1}{N}\sum_{t=1}^{N} e_t$.
  3. Calculate the standard deviation of the residuals: $s_e = \sqrt{\frac{1}{N-1}\sum_{t=1}^{N}(e_t - \bar{e})^2}$. Handle the case where $s_e = 0$.

4. Calculate the standard Z-score for each residual: $Z_{std,t} = se|e_t - \bar{e}|$.
5. Determine the anomaly rate as the fraction of residuals where the Z-score exceeds a threshold k: $\text{Residual Anomaly Rate} = \frac{1}{N}\sum_{t=1}^{N} I(Z_{std,t} > k)$ The default threshold is k=3.0.

- **Business Understanding & Interpretation:**
  - This metric identifies the frequency of "surprisingly large" forecast errors, meaning errors that fall far outside the typical range of the model's prediction errors.[48] It uses the mean and standard deviation, making it sensitive to large deviations and implicitly assuming a somewhat unimodal, symmetric error distribution for the standard interpretation of the k-sigma rule (e.g., k=3 corresponds to approximately 99.7% coverage under normality).[48]
  - A high residual anomaly rate signifies that the forecasting model frequently produces errors that are extreme relative to its average performance. This can indicate that the model fails to capture certain dynamics (e.g., volatility spikes, sudden shifts), is overly sensitive to noise or specific input patterns, or its underlying assumptions are being violated.[45]
  - From a business perspective, frequent large errors undermine confidence in the forecast and can lead to poor operational or financial decisions. Even if the average error (MAE or RMSE) seems acceptable, a high residual anomaly rate highlights potential unreliability under specific, possibly recurring, conditions.[48] This metric is crucial for risk management, as it flags models prone to significant failures, helping to quantify the likelihood of encountering substantially incorrect forecasts.[43]
- **Deeper Implications & Connections:**
  - Unlike MAE or RMSE which summarize the central tendency of the error distribution, the residual anomaly rate specifically probes the *tails* of this distribution. It directly measures how often the forecast "busts" or produces errors considered extreme relative to its own typical variability.[48] The use of standard deviation (which squares errors in its calculation) makes it particularly sensitive to these large deviations compared to MAD-based approaches.
  - An increasing trend in the residual anomaly rate over time serves as a strong indicator of model drift or concept drift. If a model was initially well-calibrated, its residuals should behave like random noise around zero. A rising rate of anomalous residuals suggests that the model's representation of the underlying process is becoming less accurate, possibly due to changes in data patterns, seasonality, or external factors that the model fails to capture.[36] This connects directly to the need for continuous model monitoring in production.

- **Limitations & Considerations:** Sensitivity to outliers in residuals can be a double-edged sword – it highlights large errors but can be heavily influenced by a few extreme values. The interpretation of the k-sigma threshold relies on the assumption of approximate normality or at least symmetry of the residual distribution. The choice of k remains subjective.

2.3 Mean Anomalous Residual Magnitude

- **Technical Definition:** This metric calculates the average absolute magnitude of only those forecast residuals that were identified as anomalous by the Residual Anomaly Rate method (i.e., residuals whose standardized Z-score exceeds the threshold k).
- **Mathematical Computation:**
  1. Calculate residuals et=y^t−yt.
  2. Calculate standardized Z-scores Zstd,t=se∣et−e¯∣.
  3. Identify the set of anomalous residuals Eanom={et∣Zstd,t>k}.
  4. Calculate the mean absolute magnitude of these anomalous residuals: Mean Anomaly Magnitude=∣Eanom∣1et∈EanomΣ∣et∣ (Returns NaN if no anomalies are found, i.e., ∣Eanom∣=0). Default k=3.0.
- **Business Understanding & Interpretation:**
  - This metric quantifies the typical *size* or *severity* of the large forecast errors, complementing the anomaly rate which measures their frequency.[18] It answers the question: "When the forecast is significantly wrong, how wrong does it tend to be?"
  - A high value indicates that the anomalous errors, when they occur, are typically very large in magnitude. A lower value suggests that even the "anomalous" errors are not drastically larger than typical errors.
  - This metric is critical for risk assessment. A low anomaly *rate* might seem acceptable, but if the *magnitude* of those few anomalies is extremely high, the potential business impact (e.g., cost of a major stockout, financial loss from a bad trade, system failure) could still be severe.[18] It helps stakeholders understand the potential downside risk associated with relying on the forecast, particularly for setting safety margins, contingency plans, or capital reserves.
- **Deeper Implications & Connections:**
  - This metric provides an estimate of the *conditional expectation* of the absolute error, given that the error falls into the tail of the distribution (as defined by the k-sigma rule). It specifically characterizes the expected severity of forecast failures or "busts," offering a more targeted risk measure than overall error metrics.[18]

    - Comparing the Mean Anomalous Residual Magnitude to the overall MAE or RMSE provides valuable context about the shape of the error distribution. If the mean anomaly magnitude is substantially larger (e.g., several times larger) than the overall MAE or RMSE, it strongly suggests that the error distribution is heavy-tailed. This implies that the average error metrics might be misleadingly optimistic, as they are influenced by many small errors, while the rare large errors (anomalies) are significantly more impactful than the average suggests.[2]
- **Limitations & Considerations:** The value is highly dependent on the chosen threshold k. If the number of anomalies detected is very small, the average magnitude can be unstable or highly influenced by a single event. It provides an average magnitude, which might still mask variability in the size of the anomalies themselves.

2.4 Average Time of Error Occurrence

- **Technical Definition:** This metric calculates the average index (or time step) at which non-zero forecast residuals occur within the evaluation period. (Note: This reflects the code's calculation, which averages the indices of *all* non-zero residuals, not just the first one).
- **Mathematical Computation:**
    1. Calculate residuals et=y^t−yt.
    2. Identify the set of indices where the residual is non-zero: Idxnonzero={t∣et優=0}.
    3. If Idxnonzero is empty (all residuals are zero), return NaN.
    4. Otherwise, return the mean of the indices in Idxnonzero: Avg Time of Error=∣Idxnonzero∣1t∈IdxnonzeroΣt
- **Business Understanding & Interpretation:**
    - Measures, on average, how far into the forecast period errors tend to appear.
    - Interpretation: A low average index suggests errors occur frequently throughout the period, including early on. A high average index might suggest errors are concentrated later in the forecast horizon. NaN indicates a perfect forecast within the evaluation set.
    - Business Relevance: Can provide insights into how forecast accuracy degrades over the horizon. If errors cluster early (low average index), it might point to issues with model initialization or capturing immediate dynamics. If errors cluster late (high average index), it might reflect challenges in long-term trend or seasonality prediction. This relates to understanding lead time accuracy decay.[1]
- **Deeper Implications & Connections:**

    - This metric gives a sense of the temporal center of mass of the forecast errors. It implicitly weights time points by the presence of an error.
    - If interpreted as the average time of the *first* error (which differs from the code), a consistently low value suggests the model struggles with capturing the initial state or immediate short-term movements accurately. It might indicate poor initialization, immediate lag, or failure to model high-frequency components present from the start.[54] Conversely, a high time-to-first-error might be desirable if short-term accuracy is paramount, but could be detrimental in systems relying on forecast errors for rapid anomaly detection, as it implies a delay in signaling deviations.[1]
- **Limitations & Considerations:** The name "Time to Detect" might be misleading given the code calculates the average index of all errors. Sensitive to the definition of "non-zero" (floating point precision). Does not consider error magnitude. The average index interpretation can be difficult to translate directly into business action without further context.

2.5 Error Persistence (Longest Run)

- **Technical Definition:** This metric measures the length of the longest consecutive sequence (run) of time steps where the forecast residual is non-zero.
- **Mathematical Computation:**
    1. Create a boolean series indicating non-zero residuals: flagt=(et僾=0), where et=y^t−yt.
    2. Iterate through the flags t=1,...,N. Maintain a current run length run and maximum run length max_run, both initialized to 0.
    3. If flagt is True, increment run. Update max_run = max(max_run, run).
    4. If flagt is False, reset run to 0.
    5. Return max_run.
- **Business Understanding & Interpretation:**
    - Quantifies the tendency for forecast errors, once they occur, to continue occurring in subsequent periods without interruption.[19] It measures the longest period during which the model remained "off track."
    - A high persistence value suggests that when the model deviates from the actuals, it takes a long time to recover or get back on course. This could indicate problems such as uncaptured autocorrelation in the errors (meaning errors predict future errors), unmodeled shifts in the underlying data level or trend, slow adaptation to changes, or model misspecification.[19]
    - High error persistence is often problematic in operational settings. For example, in inventory control, a long run of under-forecasting errors can lead to prolonged stockouts, while a long run of over-forecasting errors results in

sustained excess inventory. Similarly, in financial planning, persistent errors lead to extended deviations from budgets or targets.[19] The metric highlights a lack of self-correction or rapid adaptation in the forecasting model.

- **Deeper Implications & Connections:**
  - A long persistence of non-zero residuals is a strong qualitative indicator of autocorrelation remaining in the residuals. Ideally, residuals from a well-specified model should resemble white noise, meaning they are uncorrelated over time.[19] Long runs of non-zero errors (particularly if they also tend to have the same sign, although this metric doesn't check sign) are statistically unlikely under the white noise assumption. This suggests that the error at time t contains information that could be used to predict the error at time t+1, implying the model has failed to capture all predictable temporal structure in the data.[19]
  - While standard autocorrelation function (ACF) plots provide a quantitative measure of residual correlation averaged over the entire series, the persistence metric can be more effective at identifying specific, potentially isolated, periods of prolonged model failure. This is particularly relevant in non-stationary time series or situations involving structural breaks or regime shifts, where the model might perform well overall but fail consistently during specific intervals. Persistence pinpoints the length of the worst such failure period.[19]
- **Limitations & Considerations:** Only captures the single *longest* run of consecutive errors, potentially ignoring other periods that had significant, albeit shorter, persistence. Does not consider the magnitude or the sign pattern of the errors within the run. Sensitive to the definition of "non-zero."

Section 3: Direction & Velocity Dynamics Evaluation

Overview

This section focuses on evaluating how well a forecast captures the dynamic behavior of the time series, moving beyond static point accuracy. It assesses the model's ability to predict the direction of change, the speed (velocity) of change, the rate of change in speed (acceleration), and critical moments like turning points (peaks and valleys). These dynamic aspects are often crucial for practical applications. For instance, in financial markets, predicting the direction of price movement might be more important than predicting the exact price level for timing trades.[59] In supply chain management, anticipating an acceleration or deceleration in demand (velocity/acceleration) is key for adjusting production and inventory levels effectively.[3]

Similarly, correctly identifying turning points can signal important shifts in trends or cycles.[60] Standard error metrics like MAE or RMSE can sometimes be low even when a forecast consistently misses these crucial dynamic features, highlighting the need for specialized metrics like those presented here.[23]

3.1 Directional Accuracy

- **Technical Definition:** This metric calculates the proportion of time steps for which the direction of change (increase, decrease, or no change) in the forecasted value matches the direction of change in the actual value. It is based on comparing the signs of the first differences of the prediction and actual series.
- **Mathematical Computation:** Let $\Delta y_t = y_t - y_{t-1}$ and $\Delta \hat{y}_t = \hat{y}_t - \hat{y}_{t-1}$ be the first differences (changes) of the actual and predicted series, respectively (handling the first element appropriately, e.g., assuming $\Delta y_1 = 0$). The direction accuracy is: $$ \text{Direction Accuracy} = \frac{1}{N} \sum_{t=1}^{N} \mathbb{I}(\text{sign}(\Delta \hat{y}_t) = \text{sign}(\Delta y_t)) $$ where sign(x) returns +1 if $x>0$, −1 if $x<0$, and 0 if $x=0$. The comparison handles the case where both are zero as a match.
- **Business Understanding & Interpretation:**
  - This metric specifically assesses the forecast's ability to predict the *direction* of movement (up, down, or flat) from one period to the next, ignoring the magnitude of the change.[23]
  - The score ranges from 0 to 1 (or 0% to 100%). A score of 1 indicates perfect directional prediction. For a series with predominantly up/down movements, a score around 0.5 suggests the forecast is no better than random chance at predicting direction. Scores significantly below 0.5 would imply the model is systematically getting the direction wrong.
  - Correctly predicting the direction of change is fundamental for many business decisions. In finance, it's crucial for market timing strategies.[59] In demand planning, knowing whether demand is expected to increase or decrease informs decisions about inventory adjustments, production scheduling, and resource allocation.[3] Assessing directional accuracy helps determine if the model captures the basic underlying trend and momentum, even if point accuracy varies.[7] A model might have a low MAE by consistently predicting values close to the previous actual but fail directionally if it lags behind changes.
- **Deeper Implications & Connections:**
  - Direction accuracy serves as a fundamental check on the model's grasp of the series' basic generative process. If a model cannot reliably predict whether the series will go up or down next, its ability to predict the magnitude

of that change (velocity) or subsequent changes (acceleration, turning points) is inherently compromised.[23] It's a necessary, though not sufficient, condition for accurately capturing dynamics.
    - A common reason for poor direction accuracy, even with acceptable point accuracy (low MAE/RMSE), is forecast lag. If a model consistently lags the actual series by one or more periods (e.g., $\hat{y}_t \approx y_{t-1}$), it will predict the correct direction when the trend persists but will predict the *old* direction whenever the actual series changes direction. This lag leads to directional errors specifically at and immediately after turning points, negatively impacting the overall direction accuracy score, especially in volatile series.
- **Limitations & Considerations:** Ignores the magnitude of the change; a correct prediction of a tiny uptick is treated the same as correctly predicting a major trend reversal. Can be sensitive to noise, as small random fluctuations can change the sign of the difference.

3.2 Velocity Error (MAE of First Differences)

- **Technical Definition:** This metric calculates the Mean Absolute Error (MAE) between the first differences of the predicted time series and the first differences of the actual time series. The first difference represents the change or 'velocity' from one time step to the next.
- **Mathematical Computation:** Let $\Delta y_t = y_t - y_{t-1}$ and $\Delta \hat{y}_t = \hat{y}_t - y_{t-1}$ be the first differences. The velocity error is: $\text{Velocity Error} = \frac{1}{N'} \sum_{t=2}^{N} |\Delta \hat{y}_t - \Delta y_t|$ where N' is the number of differences calculated (typically $N-1$).
- **Business Understanding & Interpretation:**
    - This metric measures the average error in the forecast's prediction of the *rate of change* or *velocity* of the time series.[2] It assesses how well the model predicts the magnitude of the increase or decrease from one period to the next.
    - A lower velocity error indicates that the forecast more accurately captures the size of the period-to-period movements. A high velocity error suggests the model struggles to predict *how much* the series will change, even if it sometimes gets the direction correct (as measured by Directional Accuracy).
    - Velocity error is important in contexts where the magnitude of change is critical for decision-making. Examples include estimating the required adjustment in inventory levels based on predicted demand change, planning changes in production rates, or allocating resources based on the expected increase or decrease in workload.[2] It complements directional accuracy by evaluating the accuracy of the predicted *step size*.
- **Deeper Implications & Connections:**

    - The velocity error directly evaluates the accuracy of the forecast's momentum prediction. The first difference, Δy_t, represents the velocity or instantaneous rate of change (in discrete time) of the series. This metric computes the MAE specifically on these velocity values, thus assessing the model's ability to predict this period-to-period delta accurately.[2]
    - Models that exhibit good performance on standard point error metrics (like MAE or RMSE) but perform poorly on velocity error might be overly smooth. Such models might capture the overall level or long-term trend effectively but fail to represent short-term fluctuations or the correct speed of change.[2] Techniques like heavy smoothing or models focusing solely on long-term components can lead to this discrepancy.
- **Limitations & Considerations:** The metric is scale-dependent, with units representing the change in the original variable's units (e.g., change in dollars per period). It can be sensitive to outliers in the differenced series, as differencing can amplify noise. It measures the absolute error in velocity, not the bias (i.e., whether the predicted changes are consistently too large or too small).

3.3 Acceleration Error (MAE of Second Differences)

- **Technical Definition:** This metric calculates the Mean Absolute Error (MAE) between the second differences of the predicted time series and the second differences of the actual time series. The second difference represents the change in velocity, or 'acceleration'.
- **Mathematical Computation:** Let $\Delta 2y_t = \Delta y_t - \Delta y_{t-1} = (y_t - y_{t-1}) - (y_{t-1} - y_{t-2})$ and $\Delta 2\hat{y}_t = \Delta \hat{y}_t - \Delta \hat{y}_{t-1}$ be the second differences. The acceleration error is: $\text{Acceleration Error} = N'' \sum_{t=3}^{N} |\Delta 2\hat{y}_t - \Delta 2y_t|$ where N'' is the number of second differences calculated (typically $N-2$).
- **Business Understanding & Interpretation:**
    - This metric assesses the forecast's ability to predict the *rate of change of the rate of change*, or the acceleration, of the time series.[65] It evaluates how well the model captures curvature or shifts in the momentum of the series.
    - A lower acceleration error indicates the forecast is better at capturing changes in the speed of the trend (e.g., identifying when growth is accelerating or decline is decelerating). High values suggest the model fails to anticipate these shifts in momentum or struggles with non-linear patterns.
    - While less commonly used than velocity error, acceleration error is relevant for identifying potential inflection points or anticipating when a current trend might be strengthening or weakening. This could be useful in strategic forecasting contexts like predicting market saturation points, changes in technology adoption rates, or the sharpness of economic cycle peaks and

troughs.

- **Deeper Implications & Connections:**
  - Mathematically, the acceleration error evaluates the model's accuracy in predicting the second derivative of the time series (approximated by the second difference). The second derivative relates to the convexity or concavity of the series' path.[65] Models with low acceleration error are better equipped to handle non-linear trends and anticipate changes in the slope. Linear forecasting models, by definition, have zero acceleration and will thus likely exhibit high acceleration error when applied to series with significant curvature.
  - There is a strong connection between acceleration error and the ability to detect turning points. Turning points (local maxima or minima) are characterized by a change in the sign of the velocity (first difference) and typically involve significant changes in acceleration around the peak or trough. A model that struggles to accurately predict acceleration is consequently likely to misplace, miss, or smooth over these critical turning points, leading to poor performance on metrics like the Turning Point F1 Score.
- **Limitations & Considerations:** Scale-dependent. Taking the second difference can significantly amplify noise present in the original series, making this metric potentially unstable or difficult to interpret unless the underlying series is relatively smooth. Its practical interpretation can be less intuitive than velocity error. Requires a longer time series for reliable calculation.

3.4 Turning Point F1 Score

- **Technical Definition:** This metric calculates the F1 score, a standard measure from classification, to evaluate the forecast's performance in detecting turning points (local peaks and valleys) in the time series. Turning points are identified as locations where the direction of movement (sign of the first difference) changes.
- **Mathematical Computation:**
  1. Calculate first differences for actuals (Δy_t) and predictions ($\Delta \hat{y}_t$).
  2. Identify actual turning points (TP_a) where $\text{sign}(\Delta y_t) \neq \text{sign}(\Delta y_{t-1})$ (and signs are non-zero).
  3. Identify predicted turning points (TP_p) where $\text{sign}(\Delta \hat{y}_t) \neq \text{sign}(\Delta \hat{y}_{t-1})$ (and signs are non-zero).
  4. Calculate True Positives (TP): Number of points present in both TP_a and TP_p (exact index match). $TP = |TP_a \cap TP_p|$.
  5. Calculate False Positives (FP): Number of points in TP_p but not in TP_a. $FP = |TP_p| - TP$.
  6. Calculate False Negatives (FN): Number of points in TP_a but not in TP_p.

$FN = |TPa| - TP$.

7. Calculate Precision: $P = \frac{TP}{TP+FP}$ (if $TP+FP>0$, else NaN or 0).
8. Calculate Recall: $R = \frac{TP}{TP+FN}$ (if $TP+FN>0$, else NaN or 0).
9. Calculate F1 Score: $F1 = \frac{2 \times P \times R}{P+R}$ (if $P+R>0$, else NaN or 0).

- **Business Understanding & Interpretation:**
  - Measures the model's effectiveness in correctly identifying local peaks (where the series changes from increasing to decreasing) and troughs (where it changes from decreasing to increasing).[69]
  - The F1 score ranges from 0 to 1, with 1 being the best. It represents the harmonic mean of precision and recall, providing a balanced assessment.[69] Precision measures the accuracy of the predicted turning points (what fraction of predicted peaks/troughs were actual peaks/troughs?), while Recall measures the completeness (what fraction of actual peaks/troughs did the model find?).[69]
  - This metric is particularly relevant in applications where anticipating reversals or cyclical extremes is paramount. Examples include financial market timing (predicting tops and bottoms), capacity planning (identifying peak demand periods), inventory management (predicting troughs in sales before restocking), or economic forecasting (identifying business cycle peaks and troughs).[71] Standard accuracy metrics might remain high even if a model completely smooths over or misplaces these critical turning points.
- **Deeper Implications & Connections:**
  - The use of the F1 score provides a more robust evaluation of turning point detection than simple accuracy, especially when turning points are relatively rare events in the time series (an imbalanced classification problem). Accuracy can be misleadingly high if a model simply predicts no turning points in a series that has few of them. F1 balances the cost of missing actual turning points (False Negatives, impacting Recall) and the cost of predicting non-existent ones (False Positives, impacting Precision).[69]
  - Poor performance on the turning point F1 score often points to issues with either model complexity or forecast lag. Overly simplistic or heavily smoothed models tend to miss turning points entirely, resulting in high False Negatives and low Recall. Overly complex or reactive models might capture noise as turning points, leading to high False Positives and low Precision. Models that lag the actual series will predict turning points later than they occur; since this metric requires an exact index match for a True Positive, lag hurts both Precision (predicted point doesn't match an actual point at that index) and Recall (actual point isn't matched by a prediction at that index). This connects turning point detection performance closely to the Trend Change Detection

Delay metric.

- **Limitations & Considerations:** Highly sensitive to the precise definition and identification of turning points (based strictly on sign changes of the first difference). Noise in the data can create spurious turning points, affecting both actual and predicted sets. It treats all turning points equally, regardless of their magnitude or economic significance. Requires careful handling of edge cases (start/end of series, consecutive zero differences).

3.5 Trend Change Detection Delay

- **Technical Definition:** This metric calculates the average time lag (delay) between the occurrence of a turning point (trend reversal) in the actual time series and the corresponding turning point identified in the predicted time series.
- **Mathematical Computation:**
  1. Identify the indices of turning points in the actual series (tpa) and the predicted series (tpp) as in the Turning Point F1 Score calculation.
  2. For each actual turning point index $ta \in tpa$, find the index tp of the first predicted turning point such that $tp \geq ta$.
  3. If such a tp exists, calculate the delay $d = tp - ta$. If no such tp exists (the forecast misses the turning point or predicts it too early relative to subsequent actual turning points), the delay is considered undefined (NaN).
  4. Calculate the average of all defined (non-NaN) delays: Trend Change Delay=mean($\{d \mid d \text{ is defined}\}$) Returns NaN if no corresponding turning points are found.
- **Business Understanding & Interpretation:**
  - Measures how promptly, on average, the forecast detects changes in the trend's direction compared to when they actually occur.[76] It quantifies the forecast's reaction lag specifically at moments of reversal.
  - A lower value, ideally close to zero, indicates that the forecast recognizes trend changes quickly. A large positive value signifies a substantial delay between the real-world reversal and its reflection in the forecast.
  - This lag is critical for timely decision-making. If a forecast significantly delays identifying a downturn, businesses might continue investing or holding excess inventory too long. Conversely, delaying the identification of an upturn can lead to missed sales opportunities or insufficient capacity.[1] The metric essentially quantifies the forecast's reaction time when it matters most – at trend shifts.
- **Deeper Implications & Connections:**
  - While the Turning Point F1 Score assesses *whether* turning points are detected around the correct time, Trend Change Detection Delay specifically

measures the *temporal displacement* or timing error of those detected turning points. It isolates the lag component of turning point forecasting performance.[76] A model could have a decent F1 score by detecting most turning points but still exhibit a large delay, making it less useful for rapid response.
    - A consistently positive trend change delay observed across different forecast horizons or models likely points to a systemic issue within the forecasting methodology. This could be due to excessive smoothing, high-order autoregressive components that rely too heavily on past data, inappropriate model structure, or a failure to incorporate relevant leading indicators that could signal impending changes earlier.[1] The metric helps quantify this inherent lag.
- **Limitations & Considerations:** Relies on the accurate identification and matching of corresponding turning points between the actual and predicted series, which can be ambiguous or difficult in noisy or complex series. The average delay might obscure significant variability (e.g., some turning points detected early, others very late). Missed turning points in the forecast are excluded from the average (as NaN), potentially biasing the result if missed turning points are common.

Section 4: Probabilistic Calibration & Sharpness Assessment

Overview

This section transitions from evaluating point forecasts to assessing the quality of probabilistic forecasts. Probabilistic forecasts provide richer information than single point estimates by quantifying the uncertainty associated with the prediction, often expressed through prediction intervals (a range within which the actual value is expected to fall with a certain probability) or predictive quantiles. Evaluating such forecasts requires assessing two distinct but equally important properties [80]:

1. **Calibration (or Reliability):** This refers to the statistical consistency between the probabilistic forecast and the observed outcomes. For example, if a model generates 90% prediction intervals, are the actual values truly contained within these intervals 90% of the time?.[80]
2. **Sharpness:** This refers to the concentration or narrowness of the predictive distribution. Given that a forecast is well-calibrated, narrower prediction intervals are preferred as they indicate lower uncertainty and provide more precise information for decision-making.[80]

Metrics like the Winkler Interval Score and the Continuous Ranked Probability Score (CRPS) are designed to evaluate both calibration and sharpness simultaneously. This section explores metrics designed to assess these crucial aspects of probabilistic forecast quality.

4.1 Prediction Interval Coverage Probability (PICP)

- **Technical Definition:** The Prediction Interval Coverage Probability (PICP) measures the empirical frequency with which the actual observed values fall within a specified nominal prediction interval. In this implementation, the interval is constructed symmetrically around the point forecast ($\hat{y}_t$) using a multiple (z) of the standard deviation of the residuals (se).
- **Mathematical Computation:**
  1. Calculate residuals: $e_t = \hat{y}_t - y_t$.
  2. Calculate the standard deviation of residuals: $se = \text{std}(e_t)$.
  3. Determine the interval half-width: $h = z \times se$. The value z corresponds to the desired nominal coverage under a Gaussian assumption (e.g., $z = 1.96$ for a nominal 95% interval).
  4. Calculate the lower bound $L_t = \hat{y}_t - h$ and upper bound $U_t = \hat{y}_t + h$.
  5. Compute the PICP as the percentage of actuals y_t falling within $[L_t, U_t]$: $PICP = \left(\frac{1}{N}\sum_{t=1}^{N} I(L_t \le y_t \le U_t)\right) \times 100$
- **Business Understanding & Interpretation:**
  - PICP directly assesses the *reliability* or *calibration* of the prediction intervals generated by the forecast method.[85] It answers the fundamental question: "How often did the predicted range actually contain the true outcome?"
  - The interpretation involves comparing the calculated PICP to the *nominal* coverage level implied by the choice of z. For instance, if $z = 1.96$ (nominal 95%), the PICP should ideally be close to 95%.
    - If PICP < Nominal Level (e.g., PICP is 70% for a 95% nominal interval): The intervals are too narrow, meaning the forecast is overconfident and underestimates the true uncertainty. The actual value falls outside the predicted range more often than expected.[88]
    - If PICP > Nominal Level (e.g., PICP is 99% for a 95% nominal interval): The intervals are too wide, meaning the forecast is underconfident or overly conservative. While capturing the actual value frequently, the intervals might be too broad to be useful.[88]
  - PICP is crucial for risk management and decision-making under uncertainty. If a business relies on a 95% prediction interval for setting safety stock or financial reserves, but the interval's actual PICP is only 70%, then stockouts or

budget shortfalls will occur much more frequently than anticipated, leading to increased operational risk and potential losses.[87] PICP validates whether the uncertainty estimates provided by the forecast can be trusted.

- **Deeper Implications & Connections:**
  - PICP provides a direct empirical check on the validity of the confidence level associated with the forecast's uncertainty quantification (here, derived from residual standard deviation). Probabilistic forecasts essentially make a "probabilistic contract" with the user (e.g., "the value will be in this range 95% of the time").[85] PICP verifies if this contract holds true on the test data. Significant deviations between PICP and the nominal level indicate miscalibration, meaning the model's assessment of its own uncertainty is flawed.[80]
  - Achieving perfect calibration (PICP matching the nominal level) is a necessary condition for a good probabilistic forecast, but it is not sufficient. A model could achieve 95% coverage by producing extremely wide, uninformative intervals (e.g., predicting demand between 0 and 1,000,000). Such intervals, while calibrated, lack *sharpness* and offer little practical value. Therefore, calibration must always be assessed in conjunction with interval width (sharpness).[80] This motivates the use of metrics like the Winkler score or evaluating PICP conditional on interval width.[93]
- **Limitations & Considerations:** This implementation assumes the prediction interval is symmetric around the point forecast and its width is determined solely by the overall standard deviation of residuals (implying a Gaussian assumption for uncertainty). Real-world prediction intervals might be asymmetric (e.g., from quantile regression) or have widths that vary based on forecast inputs (heteroscedasticity). The accuracy of the PICP depends on the reliability of the residual standard deviation estimate. It only evaluates the coverage for a single, specific nominal level defined by z.

4.2 Average Prediction Interval Width

- **Technical Definition:** This metric calculates the average width of the prediction intervals, where the intervals are constructed as $\hat{y}_t \pm z \times se$.
- **Mathematical Computation:** Let $se = \text{std}(\hat{y}_t - y_t)$ be the standard deviation of residuals. The interval width for a given z is $W = 2 \times z \times se$. Since this width is constant across all time points in this formulation, the average width is simply W. Average Interval Width $= 2 \times z \times se$ The default is $z = 1.96$.
- **Business Understanding & Interpretation:**
  - This metric measures the *sharpness* or *precision* of the probabilistic forecast.[80] It quantifies the average range spanned by the prediction interval,

representing the degree of uncertainty expressed by the forecast.
    - Interpretation requires context, specifically the corresponding PICP. Generally, narrower intervals (lower width) are preferred, as they provide more precise guidance, *but only if they maintain the desired level of calibration (coverage)*.[81] An extremely narrow interval that frequently fails to contain the actual value (low PICP) is misleading and not useful.
    - The width of the prediction interval directly impacts its utility for business decisions. Wide intervals signify high uncertainty, which may necessitate larger safety stocks, more conservative financial planning, wider operational buffers, or indicate lower confidence in the central point forecast.[90] When comparing models that achieve similar calibration (PICP), the one with the smaller average interval width is generally considered superior as it provides a more informative and actionable forecast.
- **Deeper Implications & Connections:**
    - Interval width is purely a characteristic of the forecast's uncertainty component (here, estimated by residual standard deviation) and the chosen confidence level (z). Unlike PICP, its calculation does not involve the actual observed values.[80] It measures the forecast's *assertiveness* or claimed precision about the future, independent of whether that claim is accurate (which is measured by PICP).
    - There exists an inherent trade-off between achieving high calibration (PICP close to nominal) and high sharpness (narrow interval width). Models can often artificially improve PICP by simply increasing the estimated uncertainty (e.g., inflating the residual standard deviation or choosing a larger z), which directly increases the interval width but reduces sharpness and practical utility.[81] The goal of probabilistic forecasting is to find models that are *inherently* less uncertain (low residual variance) while remaining well-calibrated, thus achieving both sharpness and reliability. Metrics like the Winkler score are designed to explicitly evaluate this trade-off.
- **Limitations & Considerations:** Assumes symmetric intervals based on a single standard deviation value. The width calculated here is constant across all forecasts; more sophisticated methods produce intervals whose width varies depending on forecast horizon or input features. Meaningless without considering the corresponding PICP. Scale-dependent, with units matching the forecast variable.

4.3 Winkler Interval Score

- **Technical Definition:** This function computes the Winkler Interval Score, a proper scoring rule designed to evaluate the quality of a prediction interval. It combines

the interval's width (sharpness) with a penalty term that is applied only when the observed actual value falls outside the interval (calibration failure). The penalty increases proportionally to the distance the observation lies outside the interval bounds.

- **Mathematical Computation:**
  1. Calculate residuals et=y^t−yt and their standard deviation se=std(et).
  2. Determine the interval bounds based on the nominal confidence level 1−α. Using the normal distribution assumption, the half-width is h=Φ−1(1−α/2)×se, where Φ−1 is the inverse CDF (quantile function) of the standard normal distribution.
  3. Lower bound Lt=y^t−h, Upper bound Ut=y^t+h.
  4. The Winkler score for a single observation yt is: $$ W_{\alpha, t} = (U_t - L_t) + \frac{2}{\alpha} \max(0, L_t - y_t) + \frac{2}{\alpha} \max(0, y_t - U_t) $$
  5. The overall score is the average over all observations: Winkler Score=N1t=1ΣNWα,t The default is α=0.1, corresponding to a 90% nominal interval.
- **Business Understanding & Interpretation:**
  - The Winkler score provides a single, integrated measure for evaluating the overall quality of prediction intervals, simultaneously considering both their width (sharpness) and their coverage performance (calibration).[94]
  - Lower scores indicate better performance. The score consists of the interval width (Ut−Lt) plus a penalty term that is zero if the actual value yt falls within the interval [Lt,Ut]. If yt falls outside, a penalty is added, proportional to the distance yt is from the nearest interval bound, and scaled by 2/α.[95] This scaling ensures that, on average, narrower intervals are rewarded, but only if they maintain the target coverage level α.
  - This score is directly useful for comparing different probabilistic forecasting models or different methods for generating prediction intervals (e.g., comparing intervals from bootstrapping vs. quantile regression vs. this residual-based method). A model achieving a lower average Winkler score produces intervals that are considered better overall – either narrower for the same level of coverage, or achieving better coverage for the same width, or a superior combination of both.[94] It aids in selecting models that provide uncertainty estimates that are both reliable (calibrated) and precise (sharp), which is crucial for informed risk assessment, resource planning, and setting operational parameters like safety stock.[91]
  - As a *proper scoring rule*, the Winkler score incentivizes the forecaster to report their true belief about the quantiles defining the interval. It cannot be "gamed" by strategically misreporting the interval bounds.[96]

- **Deeper Implications & Connections:**
  - The Winkler score provides an elegant mathematical formulation that explicitly balances the competing objectives of sharpness (narrow intervals) and calibration (correct coverage). Its structure ensures that the expected score is minimized when the interval bounds Lt and Ut correspond to the true α/2 and 1−α/2 quantiles of the conditional distribution of yt given the information available at the time of forecasting.[95] The score penalizes deviations from both optimal sharpness and optimal calibration.
  - There is a direct mathematical relationship between the Winkler score and the quantile score (also known as pinball loss). The Winkler score for a 1−α interval is equivalent to the average of the quantile scores for the lower bound (quantile α/2) and the upper bound (quantile 1−α/2), scaled by 1/α: Wα,t=(Qα/2,t+Q1−α/2,t)/α.[95] This highlights that evaluating the interval's quality via the Winkler score is fundamentally tied to evaluating the accuracy of the quantile forecasts that define its boundaries.
- **Limitations & Considerations:** This implementation assumes symmetric prediction intervals derived from the overall residual standard deviation using a normal distribution's quantile function. It requires the user to specify the nominal coverage level α. The score is scale-dependent (in the same units as the forecast variable).

4.4 Continuous Ranked Probability Score (CRPS)

- **Technical Definition:** The Continuous Ranked Probability Score (CRPS) is a proper scoring rule that evaluates the accuracy of a full probabilistic forecast (represented by a predictive Cumulative Distribution Function, CDF) against a single observed outcome. It generalizes the Mean Absolute Error (MAE) to the probabilistic setting. This implementation calculates the CRPS assuming the predictive distribution is Gaussian, with its mean given by the point forecast (y^t) and its standard deviation estimated from the overall standard deviation of the residuals (se).
- **Mathematical Computation:** Let F be the predicted CDF (assumed Gaussian with mean μ=y^t and standard deviation σ=se=std(y^t−yt)) and let yt be the observed outcome. The CRPS is defined as: CRPS(F,yt)=∫−∞∞(F(x)−I(x≥yt))2dx where I(·) is the indicator function (1 if condition true, 0 otherwise). For a Gaussian forecast F~N(μ,σ2) and observation yt, this integral has a closed-form solution: $$ \text{CRPS}(N(\mu, \sigma^2), y_t) = \sigma \left[Z (2 \Phi(Z) - 1) + 2 \phi(Z) - \frac{1}{\sqrt{\pi}} \right] $$ where Z=σyt−μ, Φ(Z) is the standard normal CDF, and ϕ(Z) is the standard normal PDF. The function computes the average CRPS over all observations: CRPS=N1t=1ΣNCRPS(N(y^t,se2),yt)

- **Business Understanding & Interpretation:**
  - CRPS provides a single score that measures the overall accuracy of a probabilistic forecast (here, approximated as a Gaussian distribution) compared to the actual observed value.[102]
  - Lower CRPS values indicate better forecasts. The score is expressed in the same units as the observed variable, making it somewhat interpretable in magnitude.[105] Conceptually, it represents the integrated squared difference between the predicted CDF and the empirical CDF of the observation (which is a step function jumping from 0 to 1 at the actual value yt).[102] It penalizes forecasts where the predicted distribution is far from the actual outcome.
  - CRPS is considered a comprehensive metric for evaluating probabilistic forecasts because it implicitly assesses both calibration and sharpness across the entire range of the predictive distribution, not just at specific quantiles or intervals.[82] It is useful for comparing different models that produce probabilistic outputs (even if approximated, as done here). Selecting models based on lower CRPS helps identify those that provide the most accurate overall representation of future uncertainty.[103]
  - A key property is that CRPS generalizes MAE. If the predictive distribution collapses to a single point forecast (i.e., $\sigma \to 0$), the CRPS converges to the MAE between the point forecast and the actual value.[102]
- **Deeper Implications & Connections:**
  - Unlike metrics focused on specific intervals (PICP, Interval Width, Winkler Score), CRPS evaluates the *entire* predicted distribution against the outcome. This provides a more holistic assessment, as it considers discrepancies across all possible thresholds or quantiles simultaneously.[102] A forecast might have good 90% interval calibration but perform poorly in the extreme tails; CRPS would reflect the overall distributional mismatch more effectively.
  - The implementation's reliance on a Gaussian assumption (N(y^t,se2)) is a significant simplification. The true underlying predictive distribution from the forecasting model might be non-Gaussian (e.g., skewed, multimodal, or possess time-varying variance). The calculated CRPS score evaluates how well this *assumed Gaussian distribution* performs as a forecast, which might differ from the performance of the model's potentially more complex (but unarticulated) true predictive distribution. This limitation arises from using only the point forecast and the overall residual standard deviation to characterize uncertainty, rather than a more direct probabilistic output from the model itself.
- **Limitations & Considerations:** The current implementation strongly assumes a Gaussian predictive distribution with constant variance estimated from residuals.

This may not be appropriate for many real-world scenarios. Requires estimating the standard deviation of residuals. The score is scale-dependent.

Probabilistic Metrics Overview

The table below summarizes the focus of the probabilistic metrics discussed, clarifying their roles in evaluating calibration and sharpness.

Metric	Aspect Measured	Input Required	Output Interpretation
PICP	Calibration / Reliability	Interval Definition (z, std)	Empirical Coverage % vs. Nominal %
Average Prediction Interval Width	Sharpness / Precision	Interval Definition (z, std)	Average Interval Size (Units of Y)
Winkler Interval Score	Combined Calibration & Sharpness	Interval Definition (alpha, std, norm.ppf)	Combined Score (Lower is better)
CRPS	Overall Distributional Accuracy	Predictive Distribution (Assumed Gaussian)	Generalized MAE (Units of Y, Lower better)

Section 5: Distributional Drift & Stability Monitoring

Overview

This final section explores metrics designed to detect changes or instability in the statistical distributions associated with the forecasting process. This includes monitoring the distribution of the actual data, the distribution of the predictions, or the distribution of the forecast errors over time. Such monitoring is crucial, particularly when models are deployed in production environments. The underlying data generating process can change due to seasonality, market shifts, policy changes, or other external factors, causing the statistical properties of the incoming data to deviate from those observed during model training. This phenomenon, known as data drift or concept drift, can lead to significant degradation in forecast performance if not detected and addressed.[36] The metrics in this section provide tools to quantify distributional shifts and assess the stability of the forecast errors, helping to identify when model retraining, recalibration, or investigation might be warranted.

5.1 Sliding Window Jensen-Shannon Distance (JSD)

- **Technical Definition:** This metric calculates the average Jensen-Shannon Distance (JSD) between the empirical probability distributions of the actual values and the predicted values, computed over sliding windows of a specified size w. JSD is a method for measuring the similarity between two probability distributions.
- **Mathematical Computation:**
  1. Define a window size w.
  2. Iterate through the time series from t=1 to N−w+1.
  3. For each window i (from time t to t+w−1): a. Extract actuals Yi={yt,...,yt+w−1} and predictions Y^i={y^t,...,y^t+w−1}. b. Estimate the probability distributions Pi (from Yi) and Qi (from Y^i) by creating histograms with a fixed number of bins (e.g., 10 bins). Let pi(k) and qi(k) be the probabilities in bin k. c. Calculate the Jensen-Shannon Divergence between Pi and Qi. Let Mi=21(Pi+Qi). $$ D_{JS}(P_i | | Q_i) = \frac{1}{2} D_{KL}(P_i | | M_i) + \frac{1}{2} D_{KL}(Q_i | | M_i) $$ where DKL(P||Q)=Σkp(k)log2q(k)p(k) is the Kullback-Leibler divergence. d. The Jensen-Shannon Distance is the square root of the divergence: JSDi=DJS(Pi||Qi). (Note: scipy.spatial.distance.jensenshannon computes this square root version, bounded between 0 and 1).
  4. Return the average JSD over all windows: Sliding JSD=N−w+11i=1ΣN−w+1JSDi
- **Business Understanding & Interpretation:**
  - This metric assesses the average similarity between the distribution of predictions and the distribution of actuals within recent, localized time windows.[35] It checks if the forecast's distributional characteristics (shape, spread, location, as captured by the histogram) locally match those of the actual data as the series progresses.
  - JSD values range from 0 to 1 (for the square root version). Values closer to 0 indicate high similarity between the predicted and actual distributions within the windows. Values closer to 1 signify substantial divergence.[35] A low average sliding JSD suggests the forecast distribution consistently tracks the actual distribution locally.
  - This metric is useful for detecting *local* or *dynamic* distributional drift. It can identify specific periods or regimes where the model's assumptions about the data distribution (implicitly learned during training) no longer hold, leading to a mismatch between the predicted and actual distributions in those windows.[108] It monitors the stability of the relationship between the prediction and actual distributions over time.
- **Deeper Implications & Connections:**

    - The use of sliding windows explicitly focuses this metric on the *temporal evolution* of distributional similarity, making it suitable for detecting drift as it happens, rather than just providing a static comparison over the entire dataset.[35] Averaging the JSD across windows provides a summary measure of how consistently the predicted distribution tracks the actual distribution locally throughout the evaluation period.
    - Compared to metrics based only on moments (like comparing rolling means or variances), JSD offers a more comprehensive comparison because it considers the entire shape of the distributions (as discretized by the histograms). Two distributions might share the same mean and variance but differ significantly in skewness, modality, or tail behavior. JSD, being derived from KL divergence which compares probability mass across all bins, is sensitive to these finer distributional differences, potentially detecting drifts that moment-based methods might miss.[35]
- **Limitations & Considerations:** Requires careful selection of the window size w and the number of histogram bins; results can be sensitive to these choices. The quality of the JSD calculation depends on the accuracy of the histogram estimation, which can be poor with small window sizes or inappropriate binning. It measures the degree of divergence but doesn't inherently specify *how* the distributions differ (e.g., shift in mean vs. change in variance).

5.2 Population Stability Index (PSI)

- **Technical Definition:** The Population Stability Index (PSI) is a metric used to quantify the difference between the probability distribution of a variable in a current sample (here, preds) compared to its distribution in a baseline or reference sample (here, actuals).
- **Mathematical Computation:**
    1. Determine bin edges based on the distribution of the reference sample (actuals). Typically uses deciles or a fixed number of bins (e.g., 10 bins).
    2. Calculate the percentage (or fraction) of observations falling into each bin k for both the reference sample (actuals, let this be $P_{a,k}$) and the current sample (preds, let this be $P_{p,k}$). Ensure $\Sigma_k P_{a,k}=1$ and $\Sigma_k P_{p,k}=1$.
    3. Add a small epsilon to probabilities to avoid division by zero or ln(0).
    4. Calculate the PSI using the formula: $PSI=\Sigma_k(P_{a,k}-P_{p,k})\ln(P_{p,k}P_{a,k})$
- **Business Understanding & Interpretation:**
    - PSI measures the magnitude of the shift in the overall distribution of predictions compared to the distribution of the actuals over the entire evaluation period.[109] While traditionally used to compare a variable's distribution between training and production datasets to detect input drift,

this application compares the output distribution (preds) against the target distribution (actuals).
    - Interpretation often relies on established rules of thumb [109]:
        - PSI < 0.1: Indicates no significant difference between the distributions. The forecast distribution is considered stable relative to the actuals distribution.
        - 0.1 <= PSI < 0.25: Suggests a moderate shift or minor divergence between the distributions. This might warrant closer monitoring or investigation.
        - PSI >= 0.25: Signals a significant difference between the distributions. The forecast distribution is considerably different from the actuals distribution, potentially indicating systematic issues, model misspecification, or significant drift if applied over time.
    - PSI is widely used in industries like finance (especially credit risk modeling) and increasingly in general machine learning operations (MLOps) for monitoring distributional drift.[109] In this context (comparing preds to actuals), a high PSI suggests that the model is generating predictions whose overall distributional shape, location, or spread differs substantially from the target variable it's trying to predict. This could point to calibration issues across the range of predicted values or systematic biases.
- **Deeper Implications & Connections:**
    - PSI is mathematically related to, and can be seen as a symmetric version of, the Kullback-Leibler (KL) Divergence.[118] The formula involves terms related to both $D_{KL}(Actuals||Preds)$ and $D_{KL}(Preds||Actuals)$. Specifically, $PSI = \Sigma(Pa - Pp)(\ln Pa - \ln Pp) = D_{KL}(Pa||Pp) + D_{KL}(Pp||Pa)$. This symmetry makes it easier to interpret as a distance-like measure between the two distributions compared to the asymmetric KL divergence.[118]
    - When applied to compare the marginal distribution of predictions against the marginal distribution of actuals, PSI serves as a measure of *output distribution calibration*. Ideally, a well-calibrated forecasting system should produce predictions whose overall distribution mirrors that of the actual observed data. A low PSI indicates good alignment in these marginal distributions.[109] This complements metrics like PICP, Winkler Score, or CRPS, which assess *conditional* calibration (i.e., the accuracy of the uncertainty estimate given a specific prediction or input).
- **Limitations & Considerations:** Requires binning for continuous variables, and the resulting PSI value can be sensitive to the number and definition of bins. The interpretation thresholds (0.1, 0.25) are heuristics and may need adjustment based on the specific context and business tolerance for drift. PSI measures the overall shift but doesn't pinpoint which parts of the distribution have changed

most (though examining bin-level contributions can provide this detail).

5.3 Kullback-Leibler (KL) Divergence

- **Technical Definition:** The Kullback-Leibler (KL) Divergence, also known as relative entropy, measures the difference between two probability distributions. Specifically, DKL(P||Q) quantifies the information lost when distribution Q is used to approximate the true distribution P. It measures the inefficiency of assuming the distribution is Q when it is actually P. This implementation calculates DKL(Actuals||Preds).
- **Mathematical Computation:**
  1. Estimate probability distributions Pa (from actuals) and Pp (from preds) using histograms with shared bins derived from actuals. Let pa(k) and pp(k) be the probabilities in bin k.
  2. Add a small epsilon to probabilities to avoid ln(0).
  3. Calculate the KL divergence of Pa from Pp: $$ D_{KL}(P_a || P_p) = \sum_{k} p_a(k) \ln\left(\frac{p_a(k)}{p_p(k)}\right) $$
- **Business Understanding & Interpretation:**
  - KL Divergence quantifies how much the actual data distribution (Pa) diverges from the distribution generated by the predictions (Pp), measured in units of information (nats or bits).[123]
  - KL Divergence is always non-negative, and DKL(Pa||Pp)=0 if and only if Pa and Pp are identical. Larger values indicate greater divergence between the two distributions. A key property is its asymmetry: DKL(Pa||Pp) is generally not equal to DKL(Pp||Pa).[123] The implementation calculates the divergence *of* the actual distribution *from* the predicted distribution.
  - In the context of drift detection, KL divergence is often used to compare the distribution of data in a current window against a reference window (e.g., training data). A significant increase in KL divergence over time signals that the data distribution is changing.[36] When comparing actuals vs. preds directly, it assesses how well the overall shape and characteristics of the predictions match the reality over the evaluation period.
- **Deeper Implications & Connections:**
  - The interpretation of KL divergence as information loss or "surprise" is central. DKL(Pa||Pp) represents the expected additional information (average extra bits or nats) required to encode samples from the true distribution Pa when using a code optimized for the predicted distribution Pp, compared to using a code optimized for Pa itself.[122] It measures the inefficiency imposed by using the model's distribution (Pp) as a proxy for reality (Pa).
  - The asymmetry of KL divergence carries meaning. DKL(Pa||Pp) (calculated

here) is particularly sensitive to situations where the actual distribution Pa assigns probability mass to regions where the predicted distribution Pp assigns very low probability (i.e., the model fails to predict events that actually occur). This is because the term ln(pa(k)/pp(k)) becomes very large if pp(k) is close to zero while pa(k) is not. Conversely, DKL(Pp||Pa) would be more sensitive to the model predicting events (pp(k)>0) that never actually occur (pa(k)≈0).[123] The choice of which divergence to calculate depends on which type of mismatch is considered more critical for the application.

- **Limitations & Considerations:** Requires binning for continuous data, and results are sensitive to the binning strategy. KL divergence is asymmetric. It can be infinite if the predicted distribution assigns zero probability to an outcome that occurs in the actual data (though adding a small epsilon mitigates this in practice). There are no universal thresholds for significance; KL divergence is typically used for relative comparisons or monitored for changes over time.

5.4 Rolling Error Variance

- **Technical Definition:** This metric calculates the variance of the forecast errors (residuals) computed over sliding windows of a specified size w, and then averages these variance values across all windows.
- **Mathematical Computation:**
  1. Calculate residuals: et=y^t−yt.
  2. Define a window size w.
  3. For each possible window starting at time i (from i=1 to N−w+1): a. Calculate the variance of the residuals within that window: Vari=Var(ei,...,ei+w−1).
  4. Return the mean of these windowed variances: Rolling Error Variance=N−w+11i=1ΣN−w+1Vari
- **Business Understanding & Interpretation:**
  - This metric measures the average stability or consistency of the forecast error variance over time.[1] It assesses whether the typical magnitude (spread) of forecast errors remains relatively constant or fluctuates significantly

Works cited

1. Measuring forecast accuracy: The complete guide - RELEX Solutions, accessed April 20, 2025, https://www.relexsolutions.com/resources/measuring-forecast-accuracy/
2. Error Metrics: How to Evaluate Your Forecasting Models - Jedox, accessed April 20, 2025, https://www.jedox.com/en/blog/error-metrics-how-to-evaluate-forecasts/
3. Unlocking Business Success: The Crucial Role of Forecasting Accuracy - C-Suite

Support, accessed April 20, 2025, https://www.c-suitesupport.com/post/the-role-of-forecasting-accuracy
4. Understanding Performance Metrics: Making Sense of Forecast Errors | Blogs - Soon, accessed April 20, 2025, https://www.soon.works/blog/understanding-performance-metrics-making-sense-of-forecast-errors
5. 4 Demand Forecast Accuracy KPIs You'll Actually Use | Farseer, accessed April 20, 2025, https://www.farseer.com/blog/4-demand-forecast-accuracy-kpis-youll-actually-use/
6. A Comprehensive Guide to Mean Absolute Percentage Error (MAPE) - Aporia, accessed April 20, 2025, https://www.aporia.com/learn/a-comprehensive-guide-to-mean-absolute-percentage-error-mape/
7. Forecast Accuracy: The Ultimate Guide from Data to Decisions - Manhattan Associates, accessed April 20, 2025, https://www.manh.com/our-insights/resources/blog/from-data-to-decision-enhancing-forecast-accuracy-with-best-practices
8. The Monthly Metric: Demand Forecast Error Percentage, accessed April 20, 2025, https://www.ismworld.org/supply-management-news-and-reports/news-publications/inside-supply-management-magazine/blog/2024/2024-01/the-monthly-metric-demand-forecast-error-percentage/
9. Measuring forecast accuracy: Keeping score on keeping score | NTT DATA, accessed April 20, 2025, https://us.nttdata.com/en/blog/2022/march/measuring-forecast-accuracy-keeping-score-on-keeping-score
10. Forecast KPIs: Business success through forecast accuracy - numi, accessed April 20, 2025, https://numi.digital/blog/forecast-kpis
11. What is Forecasting Accuracy? Full Guide | Revenue Grid, accessed April 20, 2025, https://revenuegrid.com/blog/forecasting-accuracy/
12. How to Solve Forecast Attainment Challenges & Meet Sales Targets - BoostUp.ai, accessed April 20, 2025, https://www.boostup.ai/blog/forecast-attainment
13. Tackling Forecast Bias: Signals and Noise | FP&A Trends, accessed April 20, 2025, https://fpa-trends.com/article/tackling-forecast-bias-signals-and-noise
14. 7 Key Demand Metrics for Supply Chain Forecasting, accessed April 20, 2025, https://demandplanning.net/demand-metrics-wmape-wape/
15. Measuring Effectiveness of Demand Forecasts - Planalytics, Inc., accessed April 20, 2025, https://www.planalytics.com/forecast-bias-accuracy-difference-in-differences/
16. A Critical Look at Measuring and Calculating Forecast Bias - Demand Planning, accessed April 20, 2025, https://demand-planning.com/2021/08/06/a-critical-look-at-measuring-and-calculating-forecast-bias/
17. How to Report Forecast Accuracy to Management, accessed April 20, 2025, https://blog.arkieva.com/how-to-report-forecast-accuracy-to-management/

18. Section 12.A Statistical Concepts - Deterministic Data - ECMWF Confluence Wiki, accessed April 20, 2025, https://confluence.ecmwf.int/display/FUG/Section+12.A+Statistical+Concepts+-+Deterministic+Data
19. 3.3 Residual diagnostics | Forecasting: Principles and Practice (2nd ed) - OTexts, accessed April 20, 2025, https://otexts.com/fpp2/residuals.html
20. Tracking Signal - Example & Formula - Valtitude, accessed April 20, 2025, https://valuechainplanning.com/blog-details/7
21. Tracking signal - Wikipedia, accessed April 20, 2025, https://en.wikipedia.org/wiki/Tracking_signal
22. What Is Tracking Signal? Calculation and Example, accessed April 20, 2025, https://efex.vn/en/blog/tracking-signal
23. Forecast Accuracy Metrics to Know for Business Forecasting - Fiveable, accessed April 20, 2025, https://fiveable.me/lists/forecast-accuracy-metrics
24. What Does Tracking Signal Mean? - Bizmanualz, accessed April 20, 2025, https://www.bizmanualz.com/library/what-does-tracking-signal-mean
25. Usage of tracking signal as a metric when number of forecasts is low - Cross Validated, accessed April 20, 2025, https://stats.stackexchange.com/questions/544863/usage-of-tracking-signal-as-a-metric-when-number-of-forecasts-is-low
26. 5.4 Residual diagnostics | Forecasting: Principles and Practice (3rd ed) - OTexts, accessed April 20, 2025, https://otexts.com/fpp3/diagnostics.html
27. Forecast Accuracy and Evaluation - Rob J Hyndman, accessed April 20, 2025, https://robjhyndman.com/files/1-ForecastEvaluation.pdf
28. How To Analyse Your Time Series Model Using Residuals | Towards Data Science, accessed April 20, 2025, https://towardsdatascience.com/how-to-analyse-your-time-series-model-using-residuals-f980f597332e/
29. Residual autocorrelation and forecasting - Cross Validated - Stack Exchange, accessed April 20, 2025, https://stats.stackexchange.com/questions/188217/residual-autocorrelation-and-forecasting
30. On the calibration of underrepresented classes in LiDAR-based semantic segmentation - OpenReview, accessed April 20, 2025, https://openreview.net/pdf?id=YzRNccnAdd6
31. What does AUSE metric mean in uncertainty estimation - AI Stack Exchange, accessed April 20, 2025, https://ai.stackexchange.com/questions/36912/what-does-ause-metric-mean-in-uncertainty-estimation
32. Uncertainty Quantification Metrics for Deep Regression - arXiv, accessed April 20, 2025, https://arxiv.org/html/2405.04278v2
33. On the Calibration of Uncertainty Estimation in LiDAR-based Semantic Segmentation - arXiv, accessed April 20, 2025, https://arxiv.org/pdf/2308.02248
34. Uncertainty Quantification of Neural Reflectance Fields for Underwater Scenes - MDPI, accessed April 20, 2025, https://www.mdpi.com/2077-1312/12/2/349

35. Presenting Univariate Drift Detection Methods - NannyML's documentation! - Read the Docs, accessed April 20, 2025, https://nannyml.readthedocs.io/en/stable/how_it_works/univariate_drift_detection.html
36. Detecting & Handling Data Drift in Production - MachineLearningMastery.com, accessed April 20, 2025, https://machinelearningmastery.com/detecting-handling-data-drift-in-production/
37. 7 Surprising Stats on MAD in Modern Data Analytics, accessed April 20, 2025, https://www.numberanalytics.com/blog/mad-modern-data-analytics-stats
38. Anomaly Detection for Influx Telegraf Explained - Eyer.ai, accessed April 20, 2025, https://www.eyer.ai/blog/anomaly-detection-for-influx-telegraf-explained/
39. Forecasting Anomalies: How to Identify and Forecast Anomalies for Investment Forecasting - FasterCapital, accessed April 20, 2025, https://fastercapital.com/content/Forecasting-Anomalies--How-to-Identify-and-Forecast-Anomalies-for-Investment-Forecasting.html
40. Large Language Models for Forecasting and Anomaly Detection: A Systematic Literature Review - arXiv, accessed April 20, 2025, https://arxiv.org/html/2402.10350v1
41. Anomaly detection - DataRobot docs, accessed April 20, 2025, https://docs.datarobot.com/en/docs/modeling/special-workflows/unsupervised/anomaly-detection.html
42. Developing the Splunk App for Anomaly Detection, accessed April 20, 2025, https://www.splunk.com/en_us/blog/platform/developing-the-splunk-app-for-anomaly-detection.html
43. Anomaly Detection Techniques Summary - Kaggle, accessed April 20, 2025, https://www.kaggle.com/code/praxitelisk/anomaly-detection-techniques-summary
44. What is Anomaly Detection? Benefits, Challenges & Real-World Examples - Atlan, accessed April 20, 2025, https://atlan.com/what-is-anomaly-detection/
45. Evaluating Statistical Models for Network Traffic Anomaly Detection, accessed April 20, 2025, https://par.nsf.gov/servlets/purl/10144014
46. Anomaly Detection with Median Absolute Deviation | InfluxData, accessed April 20, 2025, https://www.influxdata.com/blog/anomaly-detection-with-median-absolute-deviation/
47. Anomaly Detection for Time Series Data: Part 2 - CleverTap Tech Blog, accessed April 20, 2025, https://tech.clevertap.com/anomaly-detection-for-time-series-data-part-2/
48. Residual Standard Deviation: Definition, Formula, and Examples - Investopedia, accessed April 20, 2025, https://www.investopedia.com/terms/r/residual-standard-deviation.asp
49. 5 Key Techniques in Residual Analysis for Better Models, accessed April 20, 2025, https://www.numberanalytics.com/blog/5-key-techniques-residual-analysis-better-models

50. Top 8 Most Useful Anomaly Detection Algorithms For Time Series - Spot Intelligence, accessed April 20, 2025, https://spotintelligence.com/2023/03/18/anomaly-detection-for-time-series/
51. Effective Anomaly Detection in Time-Series Using Basic Statistics - RisingWave, accessed April 20, 2025, https://risingwave.com/blog/effective-anomaly-detection-in-time-series-using-basic-statistics/
52. Understanding RMSE Use in Forecasting Models for Accurate Decision Making, accessed April 20, 2025, https://www.numberanalytics.com/blog/understanding-rmse-use-in-forecasting-models-accurate-decision-making
53. Initial-Value vs. Model-Induced Forecast Error: A New Perspective - the NOAA Institutional Repository, accessed April 20, 2025, https://repository.library.noaa.gov/view/noaa/58671/noaa_58671_DS1.pdf
54. 3.4 Evaluating forecast accuracy | Forecasting: Principles and Practice (2nd ed) - OTexts, accessed April 20, 2025, https://otexts.com/fpp2/accuracy.html
55. Time Series Forecast Error - Real Statistics Using Excel, accessed April 20, 2025, https://real-statistics.com/time-series-analysis/forecasting-accuracy/time-series-forecast-error/
56. Time Series Analysis for Business Forecasting, accessed April 20, 2025, http://home.ubalt.edu/ntsbarsh/stat-data/forecast.htm
57. How to Model Residual Errors to Correct Time Series Forecasts with Python, accessed April 20, 2025, https://machinelearningmastery.com/model-residual-errors-correct-time-series-forecasts-python/
58. Lumpy Forecasts - Isaac Baley, accessed April 20, 2025, https://www.isaacbaley.com/uploads/6/7/3/5/6735245/lumpyforecasts_baleyturen_final.pdf
59. Full article: Assessing the accuracy of directional forecasts - Taylor & Francis Online, accessed April 20, 2025, https://www.tandfonline.com/doi/full/10.1080/00036846.2024.2393902
60. Make Informed Decisions With Business Forecasting - Radford University Online, accessed April 20, 2025, https://online.radford.edu/degrees/business/mba/business-analytics/informed-decisions-business-forecasting/
61. What is mean absolute error (MAE) in time series forecasting? - Milvus Blog, accessed April 20, 2025, https://blog.milvus.io/ai-quick-reference/what-is-mean-absolute-error-mae-in-time-series-forecasting
62. Introducing WAPE: A Key Metric for Time Series Forecasting | Zams - Obviously AI, accessed April 20, 2025, https://www.zams.com/blog/introducing-wape
63. Mean Absolute Relative Difference Error (MARDE) Metric for Improved Forecasting Evaluation - Jeremy Whittaker, accessed April 20, 2025, https://jeremywhittaker.com/index.php/2023/04/07/mean-absolute-relative-difference-error-marde-metric-for-improved-forecasting-evaluation/

64. Comparisons of RMSE and MAE in velocity prediction between numerical and experimental models at different cross-sections in 90° bend - ResearchGate, accessed April 20, 2025, https://www.researchgate.net/figure/Comparisons-of-RMSE-and-MAE-in-velocity-prediction-between-numerical-and-experimental_tbl2_312131265
65. MAE, MAPE, MASE and the Scaled RMSE - Paul Morgan, accessed April 20, 2025, https://www.pmorgan.com.au/tutorials/mae%2C-mape%2C-mase-and-the-scaled-rmse/
66. Forecast error: How to measure and minimize it - FasterCapital, accessed April 20, 2025, https://fastercapital.com/content/Forecast-error--How-to-measure-and-minimize-it.html
67. Dynamic mean absolute error as new measure for assessing forecasting errors | Request PDF - ResearchGate, accessed April 20, 2025, https://www.researchgate.net/publication/324141962_Dynamic_mean_absolute_error_as_new_measure_for_assessing_forecasting_errors
68. Mean absolute error (MAE) of different correction methods. - ResearchGate, accessed April 20, 2025, https://www.researchgate.net/figure/Mean-absolute-error-MAE-of-different-correction-methods_fig6_354756025
69. F1 Score in Machine Learning: How to Calculate, Apply, and Use It Effectively - Grammarly, accessed April 20, 2025, https://www.grammarly.com/blog/ai/what-is-f1-score/
70. F1 Score in Machine Learning Explained - Encord, accessed April 20, 2025, https://encord.com/blog/f1-score-in-machine-learning/
71. F1 Score in Machine Learning - Lyzr AI, accessed April 20, 2025, https://www.lyzr.ai/glossaries/f1-score/
72. F1 Score in Machine Learning: Intro & Calculation - V7 Labs, accessed April 20, 2025, https://www.v7labs.com/blog/f1-score-guide
73. F1 Score Demystified: Practical Applications in Machine Learning - Number Analytics, accessed April 20, 2025, https://www.numberanalytics.com/blog/f1-score-demystified-practical-applications-machine-learning
74. Improving Model Performance with F1 Score Insights and Best Practices - Number Analytics, accessed April 20, 2025, https://www.numberanalytics.com/blog/improving-model-performance-f1-score-insights-best-practices
75. F1 Score vs ROC AUC vs Accuracy vs PR AUC: Which Evaluation Metric Should You Choose? - Neptune.ai, accessed April 20, 2025, https://neptune.ai/blog/f1-score-accuracy-roc-auc-pr-auc
76. Metrics For Forecast Evaluation - FasterCapital, accessed April 20, 2025, https://fastercapital.com/topics/metrics-for-forecast-evaluation.html
77. Addressing Prediction Delays in Time Series Forecasting: A Continuous GRU Approach with Derivative Regularization - arXiv, accessed April 20, 2025, https://arxiv.org/pdf/2407.01622

78. The Ultimate Guide to Inventory Forecasting, accessed April 20, 2025, https://www.inventory-planner.com/ultimate-guide-to-inventory-forecasting/
79. Forecasting Time Series - Evaluation Metrics - AutoGluon 1.2.0 documentation, accessed April 20, 2025, https://auto.gluon.ai/stable/tutorials/timeseries/forecasting-metrics.html
80. Probabilistic Forecasts, Calibration and Sharpness | Journal of the Royal Statistical Society Series B - Oxford Academic, accessed April 20, 2025, https://academic.oup.com/jrsssb/article/69/2/243/7109375
81. Probabilistic Forecasts, Calibration and Sharpness - DTIC, accessed April 20, 2025, https://apps.dtic.mil/sti/tr/pdf/ADA454827.pdf
82. Probabilistic Forecasts, Calibration and Sharpness | Journal of the Royal Statistical Society Series B - Oxford Academic, accessed April 20, 2025, https://academic.oup.com/jrsssb/article-abstract/69/2/243/7109375
83. Probabilistic Forecasts, Calibration and Sharpness | Request PDF - ResearchGate, accessed April 20, 2025, https://www.researchgate.net/publication/227621176_Probabilistic_forecasts_calibration_and_sharpness
84. Proper Scoring Rules for Multivariate Probabilistic Forecasts based on Aggregation and Transformation - arXiv, accessed April 20, 2025, https://arxiv.org/html/2407.00650v1
85. Module 8 – Verification of probabilistic forecasts - Pierre Pinson, accessed April 20, 2025, http://pierrepinson.com/31761/Slides/31761lecture8p3.pdf
86. Evaluating Probabilistic Predictions: Proper Scoring Rules - Sophia Sun, accessed April 20, 2025, https://huiwenn.github.io/predictive-distributions
87. The overarching methodology. PIPC: Prediction Interval Coverage... - ResearchGate, accessed April 20, 2025, https://www.researchgate.net/figure/The-overarching-methodology-PIPC-Prediction-Interval-Coverage-Probability-DDS_fig1_337739212
88. Probabilistic forecasting: prediction intervals and prediction distribution - Skforecast, accessed April 20, 2025, https://skforecast.org/0.12.0/user_guides/probabilistic-forecasting
89. Optimize the Coverage Probability of Prediction Interval for Anomaly Detection of Sensor-Based Monitoring Series - MDPI, accessed April 20, 2025, https://www.mdpi.com/1424-8220/18/4/967
90. Quantifying Uncertainty in Survival Forecasts - IDA, accessed April 20, 2025, https://www.ida.org/-/media/f7ccb9ab568e43f8aa5f4f57473fe517.ashx
91. Quantifying the Uncertainty of Reservoir Computing: Confidence Intervals for Time-Series Forecasting - MDPI, accessed April 20, 2025, https://www.mdpi.com/2227-7390/12/19/3078
92. Quantifying and Visualizing Forecast Uncertainty with the FIFE - IDA, accessed April 20, 2025, https://www.ida.org/-/media/feature/publications/q/qu/quantifying-and-visualizing-forecast-uncertainty-with-the-fife/p-31857.ashx
93. Metrics for prediction intervals · Issue #20162 - GitHub, accessed April 20, 2025, https://github.com/scikit-learn/scikit-learn/issues/20162

94. Theoretical Description Metrics : contents — MAPIE 0.9.2 documentation, accessed April 20, 2025, https://mapie.readthedocs.io/en/latest/theoretical_description_metrics.html
95. 5.9 Evaluating distributional forecast accuracy | Forecasting: Principles and Practice (3rd ed), accessed April 20, 2025, https://otexts.com/fpp3/distaccuracy.html
96. Evaluating epidemic forecasts in an interval format - PMC - PubMed Central, accessed April 20, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC7880475/
97. Winkler Interval score metric - Kaggle, accessed April 20, 2025, https://www.kaggle.com/datasets/carlmcbrideellis/winkler-interval-score-metric
98. Forecast Scoring and Calibration - UC Berkeley Statistics, accessed April 20, 2025, https://www.stat.berkeley.edu/~ryantibs/statlearn-s23/lectures/calibration.pdf
99. Winkler scores for single and combined interval forecasts... - ResearchGate, accessed April 20, 2025, https://www.researchgate.net/figure/Winkler-scores-for-single-and-combined-interval-forecasts-four-step-ahead-forecasts_tbl5_332865482
100. Scoring Interval Forecasts: Equal-Tailed, Shortest, and Modal Interval - arXiv, accessed April 20, 2025, https://arxiv.org/pdf/2007.05709
101. Proper Scoring Rules for Interval Probabilistic Forecasts - University of Exeter, accessed April 20, 2025, https://ore.exeter.ac.uk/rest/bitstreams/113928/retrieve
102. CRPS - A Scoring Function for Bayesian Machine Learning Models | Towards Data Science, accessed April 20, 2025, https://towardsdatascience.com/crps-a-scoring-function-for-bayesian-machine-learning-models-dd55a7a337a8/
103. Essential Guide to Continuous Ranked Probability Score (CRPS) for Forecasting, accessed April 20, 2025, https://towardsdatascience.com/essential-guide-to-continuous-ranked-probability-score-crps-for-forecasting-ac0a55dcb30d/
104. Scoring epidemiological forecasts on transformed scales - PMC - PubMed Central, accessed April 20, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC10495027/
105. Continuous Ranked Probability Score (CRPS) - Lokad, accessed April 20, 2025, https://www.lokad.com/continuous-ranked-probability-score/
106. (PDF) Online Learning with Continuous Ranked Probability Score - ResearchGate, accessed April 20, 2025, https://www.researchgate.net/publication/331396989_Online_Learning_with_Continuous_Ranked_Probability_Score
107. Continuous Ranked Probability Score (CRPS) — scores 2.0.0 documentation, accessed April 20, 2025, https://scores.readthedocs.io/en/stable/tutorials/CRPS_for_CDFs.html
108. Learning under Concept Drift: ML during interesting times - Stefano Meschiari, accessed April 20, 2025, https://www.stefanom.io/notes/2021/02/25/concept_drift.html
109. A Practical Introduction to Population Stability Index (PSI) - Coralogix,

accessed April 20, 2025, https://coralogix.com/ai-blog/a-practical-introduction-to-population-stability-index-psi/
110. Population Stability Index (PSI) - Radicalbit MLOps Platform, accessed April 20, 2025, https://radicalbit.ai/resources/glossary/population-stability-index/
111. Population Stability Index and feature selection in Python - Train in Data's Blog, accessed April 20, 2025, https://www.blog.trainindata.com/population-stability-index-and-feature-selection-python/
112. A Comprehensive Guide to Univariate Drift Detection Methods - NannyML, accessed April 20, 2025, https://www.nannyml.com/blog/comprehensive-guide-univariate-methods
113. Anomaly Detection in Network Traffic using Jensen-Shannon Divergence - CiteSeerX, accessed April 20, 2025, https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=91266cdc16719707dcdee8f3dbec436ea37b2b51
114. Detecting dynamical changes in time series by using the Jensen Shannon divergence - CONICET, accessed April 20, 2025, https://ri.conicet.gov.ar/bitstream/handle/11336/74669/CONICET_Digital_Nro.ce542c0f-5dcf-425d-bca0-12436f568a25_A.pdf?sequence=2
115. Anomaly detection in network traffic using Jensen-Shannon divergence - ResearchGate, accessed April 20, 2025, https://www.researchgate.net/publication/261336781_Anomaly_detection_in_network_traffic_using_Jensen-Shannon_divergence
116. Jensen-Shannon divergence and Hubert space embedding | Request PDF - ResearchGate, accessed April 20, 2025, https://www.researchgate.net/publication/285698047_Jensen-Shannon_divergence_and_Hubert_space_embedding
117. Understanding Data Drift and Model Drift: Drift Detection in Python - DataCamp, accessed April 20, 2025, https://www.datacamp.com/tutorial/understanding-data-drift-model-drift
118. Population Stability Index (PSI): What You Need To Know - Arize AI, accessed April 20, 2025, https://arize.com/blog-course/population-stability-index-psi/
119. Population Stability Index and Characteristic Analysis - ListenData, accessed April 20, 2025, https://www.listendata.com/2015/05/population-stability-index.html
120. Concepts: Performance Monitoring - SAS Help Center, accessed April 20, 2025, https://documentation.sas.com/doc/en/mdlmgrcdc/v_050/mdlmgrug/p1c6xm7tthdajkn1t6esm4n3kwnq.htm
121. Population Stability Index (PSI) - Machine Learning Plus, accessed April 20, 2025, https://www.machinelearningplus.com/deployment/population-stability-index-psi/
122. How to test categorical data drift in Machine Learning systems - Giskard,

accessed April 20, 2025, https://www.giskard.ai/knowledge/how-to-test-ml-models-2-n-categorical-data-drift
123. Kullback–Leibler divergence - Wikipedia, accessed April 20, 2025, https://en.wikipedia.org/wiki/Kullback%E2%80%93Leibler_divergence
124. KL Divergence: Understanding the impact of AI - Innovatiana, accessed April 20, 2025, https://en.innovatiana.com/post/kl-divergence
125. 8 Concept Drift Detection Methods - AI Infrastructure Alliance, accessed April 20, 2025, https://ai-infrastructure.org/8-concept-drift-detection-methods/
126. Measuring Data Drift with Kullback-Leibler Divergence - YouTube, accessed April 20, 2025, https://www.youtube.com/watch?v=lYLg4e6lxr8
127. Data Drift Detection, from First Principles - Alex Abraham, accessed April 20, 2025, https://alextabraham.com/post/2022-03-28-driftdetection/
128. How to Choose the Right Metrics to Analyze Model Data Drift - Deepchecks, accessed April 20, 2025, https://www.deepchecks.com/how-to-choose-the-right-metrics-to-analyze-model-data-drift/
129. Detecting Drifts in Data Streams Using Kullback-Leibler (KL) Divergence Measure for Data Engineering Applications, accessed April 20, 2025, https://digitalcommons.chapman.edu/cgi/viewcontent.cgi?article=1199&context=engineering_articles
130. Rolling forecasts: Definition, how to make one, and when to use them - Cube Software, accessed April 20, 2025, https://www.cubesoftware.com/blog/rolling-forecast
131. Rolling Forecasts vs Static Budgets Which Works Best for Tech Industries? - Techfunnel, accessed April 20, 2025, https://www.techfunnel.com/fintech/rolling-forecasts-vs-static-budgets-which-is-better/
132. The Importance of Rolling Forecasts in Chaotic Market Conditions - CRMT, accessed April 20, 2025, https://www.crmt.com/resources/blog/the-importance-of-rolling-forecasts-in-chaotic-market-conditions/