geomatics

MDPI

Article

Robust Gross-Error Detection in GNSS Coordinate Time Series by Fitting the Central Residual Distribution

Ming Zhang [1], Han Gao [2,*], Shunqiang Hu [3] and Yufen Niu [4]

[1] Office of Research Technology and Innovation, Yunnan Quality Checking Center for Surveying and Mapping Products, Kunming 650034, China; ming@otofly.com
[2] College of City, Kunming University of Science and Technology, Kunming 650093, China
[3] Key Laboratory of Poyang Lake Wetland and Watershed Research, Ministry of Education, Jiangxi Normal University, Nanchang 330022, China; husq@jxnu.edu.cn
[4] School of Mining and Geomatics Engineering, Hebei University of Engineering, Handan 056038, China; niuyufen@hebeu.edu.cn
* Correspondence: gaohan@kust.edu.cn; Tel.: +86-137-0066-8582

Abstract

Gross errors in Global Navigation Satellite System (GNSS) coordinate time series can bias velocity and deformation estimates. Conventional three-sigma screening is prone to masking because contaminated observations inflate the residual scale. This study evaluates AEC-NDF, which estimates scale from nested central-residual subsets and applies a three-sigma decision rule. The method was compared with conventional three-sigma and interquartile range (IQR) criteria using 2192-epoch synthetic series and 1000 paired Monte Carlo realizations under correct specification, omission of a 30-day periodic term, and earthquake-like deformation. Under correct specification, AEC-NDF achieved a recall of 99.42%, a precision of 96.93%, an F1 of 98.15%, and a false-positive rate of 0.353%, outperforming both baselines in F1. Omitting the 30-day term reduced recall and F1 to 45.09% and 61.69%, respectively. In earthquake-like simulations, AEC-NDF retained recall of 99.32% and F1 of 98.09% when coseismic and postseismic terms were modeled, but its F1 fell to 49.02% when both were omitted, below IQR (49.83%). Applications to YNMH and XIAG showed contrasting flagging, but independent labels were unavailable. AEC-NDF can reduce contamination-induced masking when deterministic signals are adequately modelled and should be used after known offsets and transients have been represented.

Keywords: GNSS coordinate time series; outlier detection; robust scale estimation; normal-distribution fitting; crustal deformation

Check for updates

Academic Editor: Gino Dardanelli

Received: 26 August 2026
Revised: 27 September 2026
Accepted: 5 October 2026
Published: 8 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Continuous Global Navigation Satellite System (GNSS) coordinate time series have become indispensable space-geodetic observations for monitoring the dynamic Earth. Owing to their ability to provide continuous three-dimensional positions with millimeter-level precision in a stable terrestrial reference frame, GNSS measurements are widely used to investigate plate motion, interseismic strain accumulation, coseismic and postseismic deformation, volcanic activity, hydrological loading, sea-level change, and other geophysical processes [1–3]. High-rate GNSS observations have also been applied to structural-health monitoring and the measurement of dynamic structural displacements [4,5]. The reliability of these applications depends directly on the quality of the coordinate time series,

because estimates of station velocity, seasonal amplitude, transient displacement, and their uncertainties can be strongly affected by even a small number of anomalous observations.

A GNSS coordinate time series generally contains a secular trend, annual and semi-annual variations, transient deformation, position discontinuities, and stochastic noise. Seasonal variations may represent genuine deformation caused by atmospheric, hydrological, and nontidal-ocean loading [6], but may also include apparent signals introduced by orbit modeling, tropospheric mismodelling, antenna phase-center errors, multipath, and reference-frame effects [7,8]. Position offsets constitute another important class of irregularity. Unlike isolated gross errors, offsets produce persistent discontinuities and may originate from earthquakes, antenna or receiver replacements, firmware changes, monument disturbances, or changes in processing strategy [9–11]. The distinction between transient gross errors and persistent offsets is therefore essential: the former should normally be detected during data cleaning, whereas the latter should be parameterized or corrected within the functional model.

In addition to these deterministic components, GNSS residuals commonly contain temporally correlated stochastic errors. Previous investigations have demonstrated that white noise alone is generally inadequate to characterize GNSS coordinate time series; combinations of white, flicker, random-walk, power-law, and autoregressive noise are frequently required [12–19]. Recent studies have also proposed computationally efficient approaches for jointly estimating functional and correlated stochastic components [20] and model-averaging strategies for accounting for uncertainty in noise-model selection [21]. Neglecting temporal correlation may substantially underestimate velocity uncertainty, whereas unmodelled seasonal variations and undetected discontinuities may be absorbed into the estimated noise spectrum. Gross errors further complicate this interaction. An anomalous observation can bias the estimated trend or seasonal terms and inflate the residual variance, while an inadequate functional model can leave genuine deformation signals in the residuals and cause them to be misclassified as outliers. Gross-error detection is thus closely coupled with both deterministic-signal modeling and stochastic-noise characterization.

Classical gross-error detection is generally based on standardized residuals, statistical hypothesis tests, or probability thresholds [22,23]. The three-sigma criterion, which is frequently adopted in GNSS time-series processing [15,24], identifies an observation as anomalous when the magnitude of its residual exceeds a specified multiple of the estimated standard deviation. This method is transparent and computationally efficient, but the sample mean and standard deviation are highly sensitive to contamination. Several large outliers can inflate the standard deviation and consequently increase the detection threshold, producing the well-known masking effect. Conversely, iterative rejection may cause normal observations to be incorrectly removed, resulting in swamping [25]. These limitations become particularly serious when gross errors have large magnitudes, occur in clusters, or account for a substantial proportion of the observations.

Resistant location and scale estimators provide an alternative to conventional variance-based thresholds. An IQR-based robust scale criterion uses quartiles rather than moments and is therefore less sensitive to extreme observations [26]. The Hampel identifier and median absolute deviation (MAD) similarly replace the sample mean and standard deviation with the median and a resistant scale estimate [27,28]. More general robust-statistical frameworks, including M-estimation, offer bounded-influence estimates under contaminated distributions [25,29]. Nevertheless, resistance to isolated extreme values does not in itself resolve the signal-noise separation problem in GNSS time series. If trend changes or seasonal signals remain in the residuals, the interquartile range or MAD may increase and mask moderate anomalies. Conversely, asymmetric, heteroscedastic, or multimodal

residual distributions may lead to excessive rejection on one side of the distribution. Temporal correlation also reduces the validity of thresholds derived under independent and identically distributed assumptions.

A variety of GNSS-specific approaches have consequently been developed. The residual-window and IQR procedure introduced in widely used GNSS time-series analysis was designed to reduce the influence of isolated anomalous epochs [24]. Robust M-estimation has been used to jointly analyze the functional and stochastic properties of GPS position series [30], while the Detection, Identification and Adaptation (DIA) framework provides a formal combination of estimation, model testing, and adaptation [31]. Bayesian methods reformulate anomaly detection in terms of posterior probabilities and can identify both isolated outliers and patches of additive anomalies [32]. Joint noise-and-probability analysis has further demonstrated that outlier removal changes the estimated noise spectrum and that the assumed noise model can, in turn, affect the observations identified as outliers [33]. Although these methods improve statistical rigor, their performance may remain sensitive to prior distributions, stochastic models, iterative strategies, or contamination-dependent scale estimates.

Signal-decomposition and hybrid methods have also received increasing attention. Singular spectrum analysis has been used to reconstruct unevenly sampled GPS time series while simultaneously detecting gross errors [34]. Wavelet-based approaches first extract nonlinear or time-varying signals and then apply IQR statistics to the reconstructed residuals [35]. Related studies have shown that seasonal amplitudes and phases can vary with time [36] and that unmodelled periodic components may be incorrectly absorbed into colored-noise estimates [33,37]. More recently, Bayesian change-point analysis has been combined with the Hampel identifier to distinguish distributional changes from isolated outliers [38]. Complementary ensemble empirical mode decomposition combined with IQR has improved the detection of subtle gross errors in GNSS buoy coordinate series [39], while an integrated fuzzy-entropy, successive variational-mode-decomposition, and IQR method has been proposed for nonlinear and strongly seasonal GNSS series [40]. These developments demonstrate the benefit of separating deterministic signals from high-frequency residuals before thresholding. However, decomposition-based methods usually introduce additional choices, such as window length, wavelet basis, decomposition level, mode number, entropy criterion, and reconstruction threshold. Their results may consequently vary with parameter settings, and inappropriate decomposition can attenuate or distort genuine deformation signals.

The preceding studies indicate that the central difficulty is not merely the selection of a threshold multiplier but the reliable estimation of the residual scale used to construct that threshold. A standard deviation computed from the complete residual distribution can be severely inflated by gross errors, whereas purely quantile-based scale estimators do not explicitly exploit the near-Gaussian structure expected for the central population of accidental measurement errors. After the principal deterministic components have been removed, the high-density central part of a GNSS residual distribution is normally less contaminated than its tails and can therefore provide a more representative statistical basis for estimating the dispersion of ordinary observational errors. This observation motivates a method that combines the interpretability of normal error theory with resistance to tail contamination.

Accordingly, this study proposes an Accidental Error Characteristics-based Normal Distribution Fitting method, termed AEC-NDF. Rather than estimating the standard deviation from all residuals, AEC-NDF fits a normal distribution to the central portion of the residual population and uses the resulting scale estimate to construct the gross-error threshold. This strategy limits the influence of large or numerous outliers while retaining

a statistically interpretable, sigma-based decision rule. The method is evaluated using simulated GNSS coordinate time series with systematically varied gross-error magnitudes and contamination ratios and is compared with the conventional 3σ and IQR criteria. It is further applied to the Menghai and Xiaguan continuous GNSS stations in Yunnan, China, over 2011.0014–2019.9603, with particular attention to the preservation of pronounced annual deformation signals. The purpose of AEC-NDF is not to replace explicit modeling of offsets, seasonal deformation, or colored noise, but to provide a robust, computationally efficient quality-control step before subsequent GNSS time-series parameter estimation and geophysical interpretation.

2. Materials and Methods

2.1. GNSS Coordinate Time-Series Model

The decomposition of a GNSS coordinate series into a deterministic trajectory model and a stochastic residual component follows established GNSS time-series practice [18,20,24]. The coordinate time series for one component of a GNSS station is represented by the following functional model:

$$y(t_i) = b + v_0 t_i + \sum_{m=1}^{m_0} [a_m \sin(2\pi f_m t_i) + b_m \cos(2\pi f_m t_i)] + r_{t_i} \quad (1)$$

where $y(t_i)$ denotes the coordinate observation at epoch t_i for a given GNSS component; b is the constant offset; v_0 is the linear velocity; m_0 is the number of periodic components; a_m and b_m are the sine and cosine coefficients of the m-th periodic component, respectively; f_m is the corresponding frequency; and r_{t_i} is the residual at epoch t_i. The time unit of t_i should be specified explicitly. Accordingly, f_m is expressed in inverse time, v_0 in coordinate units per unit time, and b, a_m, b_m, and r_{t_i} in coordinate units. In this study, the fitted functional model contains the annual and semiannual components only; therefore, $m_0 = 2$, with $f_1 = 1/365$ day^{-1} and $f_2 = 1/182$ day^{-1}. These frequencies are fixed a priori rather than estimated from the observations, whereas their sine and cosine coefficients are estimated during model fitting. The specific implementation and parameter settings used in the simulations are described in Section 2.6.

2.2. Three-Sigma Criterion

The three-sigma criterion is a classical Gaussian-based screening rule for detecting potentially anomalous observations and continues to be used in recent GPS/GNSS coordinate time-series processing [22,23,41,42]. It flags an observation when the absolute residual from the fitted model exceeds three times the standard deviation estimated from the complete residual series. Under an ideal zero-mean Gaussian error model, approximately 99.73% of uncontaminated residuals lie within ±3σ. Because the conventional standard deviation is estimated from the complete residual series, large or numerous gross errors can inflate the estimated scale, enlarge the detection threshold, and produce masking.

For GNSS coordinate time series, the criterion is applied to the residual series obtained from Equation (1):

$$|r_{t_i}| > 3\sigma, \quad i = 1, 2 \cdots n \quad (2)$$

Here, n denotes the total number of observations, and hence the total number of residuals, in the analyzed coordinate time series, whereas i denotes the observation epoch. It is distinct from m_0 in Equation (1), which denotes the number of prescribed periodic components in the functional model. The two quantities are independent: n determines the length of the coordinate time series, whereas m_0 determines the number of cyclic frequencies included in the fitted model. In the synthetic experiments, $n = 2192$ and $m_0 = 2$, with

the latter corresponding to the annual and semiannual components. The 99.73% value is a model-based Gaussian coverage probability; it is not the posterior probability that an observation outside the interval is a gross error.

2.3. IQR Criterion

The IQR criterion estimates the location and scale of the residual distribution using the median and interquartile range, respectively. Let Q_1 and Q_3 denote the first and third quartiles of the residual series $\{r_{t_i}\}_{i=1}^{n}$, and let $\tilde{r} = \text{median}(\{r_{t_i}\}_{i=1}^{n})$. The interquartile range and its Gaussian-consistent scale estimate are defined as:

$$IQR = Q_3 - Q_1 \tag{3}$$

which $\hat{\sigma}_{IQR}$ is denoted by 0.7413IQR. The robust standardized residual is then calculated as:

$$Z_i = \frac{r_{t_i} - \tilde{r}}{\hat{\sigma}_{IQR}}, i = 1, 2, \ldots, n \tag{4}$$

An observation at epoch t_i is flagged as a gross error when

$$|Z_i| \geq 3 \tag{5}$$

The factor 0.7413 makes $\hat{\sigma}_{IQR}$ a consistent estimate of the standard deviation under a Gaussian residual model.

The IQR criterion estimates the center and scale of the residual distribution using sample quantiles, which are less sensitive to extreme residuals than the conventional mean and standard deviation. In this study, Q_1 and Q_3 are defined as the 25th and 75th sample percentiles of the residual series, respectively, and the IQR is their difference. The residual median is used as the corresponding robust location estimate.

The criterion implemented here should be distinguished from the conventional Tukey boxplot rule, which standard boxplot commonly identifies observations outside the lower and upper fences defined by $Q_1 - 1.5IQR$ and $Q_3 + 1.5IQR$ [26]. These boxplot fences are not used in the present analysis, and the detection results are not determined through visual inspection of boxplots.

Instead, the IQR is converted into a Gaussian-consistent robust scale estimate using the factor 0.7413. For a Gaussian reference distribution, the difference between the 75th and 25th percentiles is approximately 1.349 times the standard deviation; consequently, 0.7413IQR estimates the standard deviation of the uncontaminated residual population. Each residual is then standardized by its deviation from the residual median relative to this robust scale estimate. An observation is classified as a gross error when the absolute robust standardized residual exceeds three.

The Gaussian reference distribution is used only to calibrate the IQR scale factor and the common threshold multiplier. It does not imply that the complete residual series, including gross errors, must be normally distributed. In the synthetic experiments, the ordinary observational errors are Gaussian by construction. For the real GNSS data, the IQR-based quantity is interpreted as a robust scale estimate, and no exact Gaussian coverage probability is assumed.

2.4. Proposed AEC-NDF Method

The effectiveness of a sigma-based gross-error criterion depends largely on whether the estimated scale represents the dispersion of ordinary observational errors. When the standard deviation is calculated from the complete residual series, gross errors may substantially inflate the scale estimate and consequently enlarge the detection threshold,

resulting in masking. To address this limitation, the AEC-NDF method estimates the residual scale from progressively expanded central subsets of the residual distribution. The method exploits the approximately Gaussian behavior of the high-density residual core after the principal deterministic signals have been removed, without assuming that the entire residual series is normally distributed.

For each coordinate component $y(t_i)$, Equation (1) was fitted using bisquare-weighted regression to estimate the constant term, linear trend, and prescribed periodic components [18,24]. The stochastic errors, any remaining unmodelled signals, and potential gross errors were retained in the resulting residual series r_{t_i}, denoted by r_i. hereafter. The bisquare weighting reduces the influence of anomalous epochs on the estimation of the deterministic components.

$$r_{t_i} = y(t_i) - b - v_0 t_i - \sum_{m=1}^{m_0} [a_m \sin(2\pi f_m t_i) + b_m \cos(2\pi f_m t_i)] \tag{6}$$

The bisquare weighting limits the influence of anomalous epochs on the estimated trend and periodic components. Nevertheless, the resulting residual series r_i may still contain gross errors and is therefore subjected to the following detection procedure.

Let Δ_r denote a prescribed residual interval in the same unit as the coordinate component. A sequence of increasing truncation bounds is defined as

$$\begin{aligned} &b_k = k\Delta_r, k = 1, 2, \ldots, K, \\ &K = \tfrac{\max_i |r_i|}{\Delta_r}. \end{aligned} \tag{7}$$

The residual interval Δr defines the increment between successive truncation bounds, $b_k = k\Delta r$. It controls the resolution of the candidate central subsets rather than directly fixing the width of the final selected subset. At each step, all residuals satisfying $|r_i| < b_k$ are included, and the final subset is determined by the subsequent scale-selection procedure. The value of Δr is selected based on both the measurement precision and empirical considerations. In the experiments conducted in this study, Δr was set to 1 mm. A smaller Δr provides a finer sampling of the scale sequence, but adjacent bounds may produce identical subsets when no additional residuals are included. Consequently, plateaus and local sample fluctuations may affect the first-local-extremum selection rule. A larger Δr provides a coarser sampling and may skip a narrow stabilization region. These effects can change the selected central subset, the estimated scale δ, and hence the detection threshold 3δ. Therefore, decreasing Δr does not necessarily improve detection, and the sensitivity of the results should be assessed empirically.

For each bound b_k, a nested central residual subset is constructed and fitted by a normal distribution

$$\mathcal{R}_k = \{r_i : |r_i| < b_k\}, (\hat{\mu}_k, \hat{\sigma}_k) = \text{NormalFit}(\mathcal{R}_k). \tag{8}$$

Since no censoring or truncation option was used, the estimated parameters are the sample mean and the sample standard deviation

$$\mu_k = \frac{1}{n_k} \sum_{r_i \in \mathcal{R}_k} r_i, \sigma_k = \left[\frac{1}{n_k - 1} \sum_{r_i \in \mathcal{R}_k} (r_i - \mu_k)^2 \right]^{1/2}. \tag{9}$$

The condition $|r_i| < b_k$ is used only to select the central subset of residuals. No truncated-normal correction is applied in the estimation of μ_k and σ_k. Although both the mean and standard deviation are estimated, only the scale sequence $\hat{\sigma}_k$ is used for gross-

error detection because the residuals are expected to be approximately centered after model fitting. The interval Δ_r determines the resolution with which the residual distribution is progressively expanded.

A very narrow central subset does not contain sufficient information to represent the complete random-error distribution and generally produces a small scale estimate. Therefore, AEC-NDF does not directly adopt the first or smallest value of $\hat{\sigma}_k$. Instead, it examines the evolution of the fitted scale as the truncation bound increases. For residuals dominated by ordinary measurement errors, $\hat{\sigma}_k$ increases smoothly and approaches an initial stable region. When observations from the contaminated tails begin to enter the subsets, the evolution of $\hat{\sigma}_k$ changes.

This transition is identified from the first-order differences in the scale sequence

$$d_k = \hat{\sigma}_{k+1} - \hat{\sigma}_k, k^* = \text{FirstLocalMaximum}(-d_k), \delta = \hat{\sigma}_{k^*+1}. \quad (10)$$

Thus, k^* corresponds to the first local minimum in the increment of the fitted scale, representing the initial stabilization point before the inclusion of larger tail residuals produces renewed scale growth. The corresponding value δ is adopted as the representative standard deviation of the uncontaminated residual population. δ is not a user-defined threshold or a fixed value shared among datasets. It is estimated independently from the scale sequence of each residual series. Consequently, different Monte Carlo realizations, stations, and coordinate components may produce different values of δ. Figure 1 illustrates the typical evolution of this scale sequence.

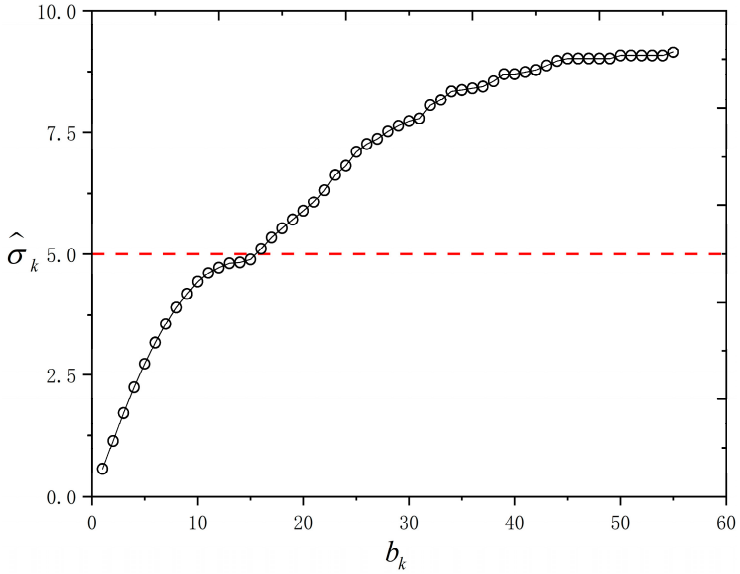


Figure 1. Evolution of the fitted standard deviation $\hat{\sigma}_k$ with increasing residual truncation bound b_k. The first stabilization point is used to determine the adaptive scale δ. The curve illustration of scale selection in a single simulation realization. The curve shows the evolution of the estimated residual scale as the truncation bound increases. The red dashed line indicates the selected scale δ, which is used to construct the gross-error detection threshold.

Finally, an observation is classified as a gross error when

$$\frac{|r_i|}{\delta} > 3, i = 1, 2, \ldots, n. \quad (11)$$

For an ideally centered Gaussian residual population, the interval $\pm 3\delta$ contains approximately 99.73% of uncontaminated observations. This percentage is a model-based coverage probability rather than the posterior probability that an identified observation is a gross

error. Observations satisfying Equation (11) are excluded from subsequent analysis and represented as missing values. If the scale sequence contains fewer than four valid intervals or no local maximum can be identified, the implementation conservatively retains the original observations rather than applying an unsupported threshold. In all experiments reported in this study, this procedure was performed as a single screening pass. The functional model was not refitted after gross-error marking, and the AEC-NDF detection was not repeated.

2.5. Evaluation Metrics

For each simulated dataset, let $\mathcal{G}$ denote the predefined set of epochs at which an additive gross-error term was injected. The reference label of observation i was defined as

$$z_i = \begin{cases} 1, & i \in \mathcal{G}, \\ 0, & i \notin \mathcal{G}, \end{cases} \tag{12}$$

where $z_i = 1$ denotes a contaminated observation and $z_i = 0$ denotes an uncontaminated observation. Thus, the reference classification was determined solely from the known injection record and did not depend on the observed residual magnitude or subsequent detection result.

For a given detection method, let $\hat{\mathcal{G}}$ denote the set of epochs satisfying its prespecified decision rule. Specifically, an observation was classified as a predicted gross error when its detection statistic exceeded the corresponding three-sigma threshold: $|r_i| > 3\hat{\sigma}$ for the conventional three-sigma criterion, $|r_i - \tilde{r}| > 3\hat{\sigma}_{IQR}$ for the IQR criterion, and $|r_i| > 3\delta$ for AEC-NDF. The four classification outcomes were then defined as

$$\begin{aligned} TP &= \left|G \cap \hat{\mathcal{G}}\right|, \\ FN &= \left|G \backslash \hat{\mathcal{G}}\right|, \\ FP &= \left|\hat{\mathcal{G}} \backslash G\right|, \\ TN &= \left|\mathcal{G}^c \cap \hat{\mathcal{G}}^c\right|. \end{aligned} \tag{13}$$

Accordingly, true positives (TP) and false negatives (FN) denote injected gross errors that were detected and missed, respectively, whereas false positives (FP) and true negatives (TN) denote observations without an injected gross-error component that were incorrectly flagged and correctly retained, respectively. The detection rate (Recall), equivalent to recall or sensitivity, was calculated as

$$\text{DetectionRate} = \frac{TP}{TP + FN} \tag{14}$$

The denominator TP + FN represents the total number of injected gross errors. Therefore, the detection rate quantifies the proportion of actual gross errors successfully identified by the method. A higher detection rate indicates a lower frequency of missed detections. However, detection rate alone does not reflect whether an increased number of normal observations has also been incorrectly flagged.

To quantify such incorrect detections, the false-positive rate was defined as

$$FPR = \frac{FP}{FP + TN} \tag{15}$$

where FP + TN is the total number of uncontaminated observations. The false-positive rate therefore represents the proportion of normal observations incorrectly classified as gross errors. This metric is particularly important because an overly aggressive detection

procedure may achieve a high detection rate at the expense of incorrectly removing a substantial number of valid observations.

Precision, also referred to as the positive predictive value, was calculated as

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{16}$$

Unlike detection rate, which is conditioned on the set of actual gross errors, precision is conditioned on the set of observations classified as gross errors. It measures the proportion of reported detections that correspond to truly contaminated observations. Thus, detection rate evaluates the completeness of gross-error identification, whereas precision evaluates the reliability of the reported detections.

To provide a single measure balancing detection rate and precision, the F1 score was calculated as their harmonic mean:

$$\text{F1} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \tag{17}$$

The F1 score approaches one only when both precision and recall are high. It therefore penalizes methods that achieve a high detection rate through excessive false-positive detections, as well as methods that achieve high precision by detecting only a small subset of the injected gross errors. Because the number of uncontaminated observations can be substantially larger than the number of injected gross errors, the F1 score is generally more informative than overall classification accuracy for evaluating gross-error detection under class imbalance.

For the earthquake-like simulation, we additionally calculated the false-positive rate and recall within the prespecified event window $W = [t_e, t_e + 3\tau]$, together with recall outside this window. The event-window FPR uses only uncontaminated observations in W as its denominator. Its complement, $1 - \text{FPR}_W$, describes the proportion of valid event-window observations retained; it is not an independent measure of physical-signal preservation. To assess downstream parameter estimation, the known earthquake-aware functional model was fitted by ordinary least squares to the observations before and after removal of flagged epochs. The absolute errors in the estimated coseismic and postseismic amplitudes were then compared with the injected values. This refitting served only as an evaluation diagnostic and was not fed back into the detection procedure.

In the present tables, detection performance is summarized using the detection rate and the absolute number of false positives. The false-positive count provides a direct interpretation of the number of valid observations incorrectly removed. However, because the absolute FP count depends on the number of uncontaminated observations, comparisons between datasets of different sizes should preferably be based on the false-positive rate or should report both FP and FPR. When all compared datasets contain the same number of uncontaminated observations, the FP count and FPR lead to identical rankings, although the FPR remains more readily comparable across experimental settings.

Unless otherwise stated, all metrics were calculated at the observation level using the same predefined detection criterion or decision threshold for all methods. Higher values of detection rate, precision, and F1 score indicate better performance, whereas lower values of FP and FPR indicate better control of false detections.

For the baseline Monte Carlo comparison, 1000 paired replicates were generated. The replicates were generated sequentially from this stream without reseeding. Different random draws were therefore used across replicate pairs. Within each pair, the correctly specified and misspecified scenarios shared the same white-noise realization, gross-error increments, and contaminated epochs. The reported performance metrics were calculated

separately for each realization and then summarized across realizations. For the baseline comparison, performance metrics were calculated separately for each of the 1000 paired Monte Carlo realizations and then summarized across realizations. Table 1 reports the arithmetic mean and sample standard deviation across realizations; approximate 95% confidence intervals for the Monte Carlo means are presented with the results in Section 3.1. For a given method, scenario, and performance metric, let M_j denote the result from the j-th valid realization. The Monte Carlo mean, sample standard deviation, and standard error were calculated as

$$\overline{M} = \frac{1}{n}\sum_{j=1}^{n} M_j, s_M = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(M_j - \overline{M})^2}, MCSE = \frac{s_M}{\sqrt{n}} \quad (18)$$

where n is the number of valid realizations for the corresponding metric. The approximate 95% confidence interval was calculated as $\overline{M} \pm 1.96$ MCSE.

Table 1. Gross-error detection performance over 1000 paired Monte Carlo realizations. Values are arithmetic means ± between-realization sample standard deviations.

Scenario	Method	Recall (%)	FPR (%)	Precision (%)	F1 (%)
Correctly specified	Three-sigma	24.88 ± 2.04	0.000 ± 0.000	100.00 ± 0.00	39.80 ± 2.62
Correctly specified	IQR	83.64 ± 4.14	0.074 ± 0.063	99.22 ± 0.65	90.71 ± 2.40
Correctly specified	AEC-NDF	99.42 ± 0.97	0.353 ± 0.193	96.93 ± 1.60	98.15 ± 0.80
Misspecified	Three-sigma	15.61 ± 1.84	<0.001	100.00 ± 0.10	26.96 ± 2.76
Misspecified	IQR	25.24 ± 3.47	<0.001	99.97 ± 0.22	40.18 ± 4.43
Misspecified	AEC-NDF	45.09 ± 6.77	0.040 ± 0.056	99.28 ± 0.96	61.69 ± 6.40

These standard deviations describe performance variability among simulated datasets and are distinct from the Monte Carlo standard errors of the reported means.

For metrics bounded between 0 and 1, the interval endpoints were restricted to this range and then converted to percentages for plotting. These intervals quantify the Monte Carlo uncertainty in the estimated mean performance under the specified simulation design. In contrast, the standard deviations reported in Table 1 describe variability across individual realizations. The between-realization sample standard deviation describes how detection performance varied among independently generated datasets. It should not be interpreted as the numerical uncertainty of the estimated Monte Carlo mean.

The Monte Carlo standard error (MCSE), calculated as the between-realization sample standard deviation divided by the square root of the number of valid realizations, quantifies the uncertainty arising from the finite number of simulation realizations. Table 1 reports the arithmetic mean and between-realization sample standard deviation, whereas the error bars in the Monte Carlo performance figure represent approximate 95% confidence intervals for the corresponding mean values.

The approximate confidence intervals were based on the sampling distribution of the realization-level mean. The central limit theorem was therefore applied to the independent realization-level performance metrics, rather than to the individual GNSS observations or to the contaminated residual distribution. With 1000 realizations, the distribution of the Monte Carlo mean was approximated by a normal distribution. These intervals quantify numerical uncertainty in the estimated mean performance under the specified data-generating process; they are not prediction intervals for the performance of an individual GNSS coordinate series and do not imply that the contaminated observations themselves are Gaussian.

Pairwise comparisons between scenarios and methods were based on within-realization metric differences. The corresponding confidence intervals were calculated

using the standard deviation of these paired differences rather than by treating results obtained from the same simulated observations as independent.

2.6. Simulation Design

Simulation defines a controlled data-generating process with known reference labels rather than an empirical model calibrated to a particular GNSS station. The deterministic component follows the conventional representation of a GNSS coordinate time series by an initial position, a linear trend, and annual and semi-annual harmonic terms [18,22,24]. The numerical coefficients, sampling interval, record length, noise level, and contamination settings specified below were prescribed to construct a reproducible benchmark and should not be interpreted as universal parameters of operational GNSS coordinate series.

A synthetic GNSS coordinate time series containing n = 2192 epochs was generated at a sampling interval of 4 h. The resulting record covered approximately 365.17 days. The four-hour interval and approximately one-year duration were selected as controlled experimental settings to provide a sufficiently dense series for the Monte Carlo comparisons; they are not proposed as a generally applicable GNSS sampling configuration. The observation times were defined as

$$t_i = (i-1)\frac{4}{24}\text{days}, i = 1, \ldots, 2192 \tag{19}$$

covering approximately one year. Two simulation scenarios were considered: a correctly specified functional model and a misspecified functional model.

For the correctly specified scenario, the deterministic signal was generated as

$$\begin{aligned} x_i^{(C)} = 5 + 0.5t_i + 5\sin\left(\tfrac{2\pi t_i}{365}\right) + 5\cos\left(\tfrac{2\pi t_i}{365}\right) \\ +3\sin\left(\tfrac{2\pi t_i}{182}\right) + 3\cos\left(\tfrac{2\pi t_i}{182}\right) \end{aligned} \tag{20}$$

Thus, the equation follows the conventional GNSS functional-model structure comprising an intercept, a linear trend, and annual and semi-annual sine–cosine terms. The intercept was 5 mm, the linear-trend coefficient was 0.5 mm day^{-1}, and the annual and semi-annual sine–cosine coefficient pairs were (5, 5) and (3, 3) mm, respectively. Although the structure of the model is literature-based, these numerical coefficients were fixed to construct a reproducible numerical benchmark and were not estimated from, or intended to represent, a particular GNSS station.

For the misspecified scenario, an additional 30-day periodic component was included in the data-generating process:

$$x_i^{(M)} = x_i^{(C)} + 6\sin\left(\frac{2\pi t_i}{30}\right) + 6\cos\left(\frac{2\pi t_i}{30}\right) \tag{21}$$

To construct a known case of functional-model misspecification, an additional 30-day periodic component was included in the data-generating process but deliberately omitted from the fitted model. Subseasonal and technique-related periodic signals have been documented in GNSS coordinate products [8]. Nevertheless, the 30-day period and the corresponding 6 mm sine–cosine coefficients were selected specifically for this controlled stress test and are not asserted to represent a universal GNSS signal or a particular geophysical process.

Independent Gaussian white noise was first generated as

$$\varepsilon_i \sim \mathcal{N}\left(0, \delta^2\right), \delta = 5 \tag{22}$$

Independent Gaussian white noise was adopted as a deliberately simplified baseline so that the effects of gross-error contamination and functional-model misspecification could

be isolated. This assumption is not intended as a complete stochastic representation of operational GNSS coordinate series, which commonly contain both white and temporally correlated noise components [8,18]. In the present experiment, the white-noise standard deviation was fixed at $\delta = 5$ mm as a study-specific simulation setting.

Controlled GNSS simulations commonly evaluate detection methods by injecting synthetic perturbations at known epochs. For example, Wang et al. generated high-variance normally distributed candidate errors, retained values exceeding a three-sigma reference, and inserted them into a simulated GNSS coordinate series [39]. Following this general controlled-injection strategy, candidate additive gross-error increments were generated in the present study as follows:

$$g_i \sim \mathcal{N}\left(0, (3\delta)^2\right) = \mathcal{N}\left(0, 15^2\right) \tag{23}$$

The distribution of g_i, its scale of 3δ, and the subsequent eligibility rule were prespecified experimental choices. Although g_i was initially generated without direct truncation, the condition $|\varepsilon_i + g_i| > 3\delta$ conditioned the perturbations ultimately injected into the observations. Therefore, the realized injected-error distribution was conditional rather than Gaussian. This rule produced a clearly separated gross-error benchmark relative to the nominal ordinary-noise level; it did not evaluate small perturbations that remained within the ordinary three-sigma noise range. Instead, an epoch was included in the candidate contamination set when the magnitude of the combined random and gross error exceeded three times the standard deviation of the ordinary noise:

$$\mathcal{C} = \{i : |\varepsilon_i + g_i| > 3\delta\} = \{i : |\varepsilon_i + g_i| > 15\} \tag{24}$$

A total of 219 epochs were then sampled uniformly without replacement from $\mathcal{C}$. The resulting contamination rate was $219/2192 \approx 9.99\%$. Gross errors were added to, rather than substituted for, the noise-contaminated observations. Therefore, the final observations were

$$y_i^{(s)} = x_i^{(s)} + \varepsilon_i + I(i \in \mathcal{S}) g_i, s \in \{C, M\} \tag{25}$$

where $\mathcal{S} \subset \mathcal{C}$ denotes the selected set of contaminated epochs and $I(\cdot)$ is the indicator function. Equivalently,

$$y_i^{(s)} = \begin{cases} x_i^{(s)} + \varepsilon_i + g_i, & i \in \mathcal{S}, \\ x_i^{(s)} + \varepsilon_i, & i \notin \mathcal{S}. \end{cases} \tag{26}$$

The same realization of Gaussian noise, the same gross-error values, and the same contaminated epochs were used in the correctly specified and misspecified scenarios, enabling a paired comparison in which the unmodeled 30-day periodic component was the principal difference between the two datasets.

In this study, "correctly specified" and "misspecified" refer to whether the fitted functional model includes all deterministic components present in the data-generating process. In both scenarios, the fitted model contains an intercept, a linear trend, and periodic terms with periods of 365 and 182 days. Its coefficients are estimated from the observations rather than fixed at their true simulation values. In the correctly specified scenario, these terms match the deterministic components used to generate the data. In the misspecified scenario, the data additionally contain the component $6\sin(2\pi t/30) + 6\cos(2\pi t/30)$, which is deliberately excluded from the fitted model. The correctly specified scenario provides a baseline for evaluating gross-error detection when the deterministic signal is adequately modeled. The misspecified scenario examines the effect of an omitted periodic signal on residual-based detection and identifies a limitation of the method. Both scenarios

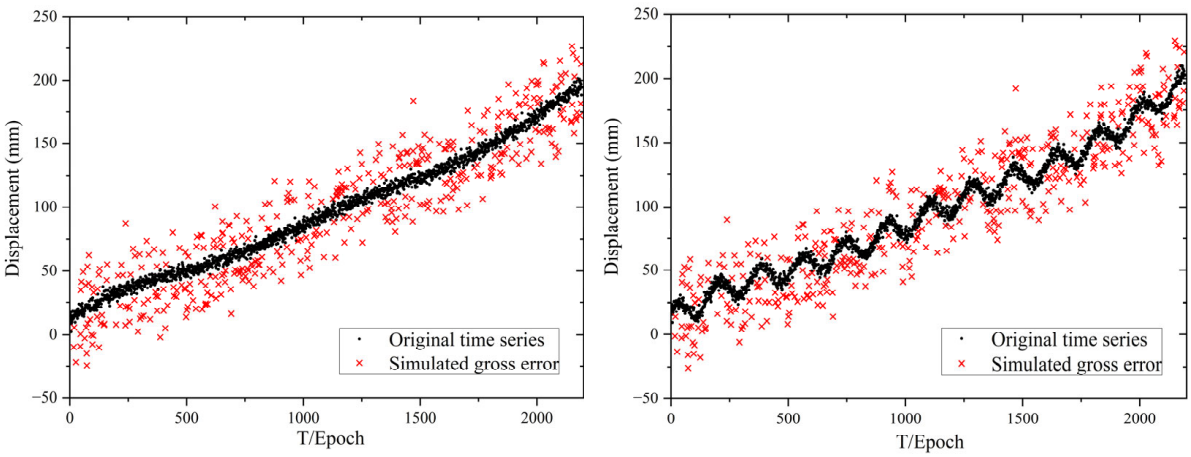
were processed using the same functional-model fitting procedure, residual calculation method, and fixed detection settings. Model coefficients and residual scales were estimated separately for each dataset; therefore, identical processing settings did not imply identical residuals or numerical detection thresholds. Each realization contained a deterministic coordinate signal, an independently generated Gaussian background-error sequence, and additive gross-error increments at known epochs. The Gaussian assumption applied only to the ordinary observational errors. After gross errors were introduced, the complete simulated series followed a contaminated and generally non-Gaussian distribution; the full dataset, including the injected gross errors, was not assumed to follow a single normal distribution. Candidate gross-error increments were initially generated from the distribution specified above. However, because only candidates satisfying the prescribed magnitude condition were eligible for injection, the increments actually injected into the observations followed the corresponding conditionally selected distribution rather than the unrestricted generating distribution. Moreover, exactly 219 contaminated epochs were sampled without replacement in each baseline realization. Thus, the contamination indicators within an individual realization were not mutually independent, whereas all stochastic components were newly generated across realization pairs. These features formed part of the prespecified simulation design and were independent of the subsequent detection results.

All simulated coordinate values are expressed in millimeters (mm), and time t is measured in days. Accordingly, the intercept, the sine and cosine coefficients, and the noise standard deviation are expressed in millimeters; the linear-trend coefficient is expressed in mm day^{-1}; and the frequencies are expressed in day^{-1}. Specifically, the frequencies $1/365$, $1/182$, and $1/30$ day^{-1} represent annual, semiannual, and approximately monthly variations, respectively. The parameter values were selected to construct a controlled synthetic coordinate series comprising an initial offset, secular motion, periodic variations, and Gaussian random noise. The intercept defines the reference level, the trend and periodic terms represent the deterministic temporal variations, and the noise standard deviation controls the background stochastic variability. All simulations and detection procedures were implemented in MATLAB (R2026b) with the Statistics and Machine Learning Toolbox. Random numbers were generated using the Mersenne Twister generator initialized by rng(20260730, 'twister'). Gaussian random values were generated using randn, contaminated epochs were selected using randperm, and robust fitting was performed using robustfit with bisquare weighting. The AEC-NDF procedure was implemented using custom MATLAB functions. The representative realization shown in Figure 2 was the first realization generated using this seeded stream. The bisquare-weighted fitting is a common preprocessing step for all three methods.

An additional earthquake-like synthetic signal was constructed as

$$x_i^{(E)} = x_i^{(C)} + A_c H(t_i - t_e) + A_p H(t_i - t_e)\left[1 - \exp\left(-\frac{t_i - t_e}{\tau}\right)\right] \tag{27}$$

We specified $t_e = 180$ days, $A_c = 50$ mm, $A_p = 20$ mm, and $\tau = 30$ days. The same earthquake-affected observations were processed using three fitted models: (i) a model containing both the coseismic step and postseismic term, (ii) a model containing the step only, and (iii) the original trend-and-periodic model containing neither term. Within each of the 1000 paired realizations, the three fits used identical observations, noise, gross-error increments, and contaminated epochs. Thus, differences among them reflect functional-model specification rather than different random realizations. The event epoch and relaxation time were assumed known in this controlled experiment; the chosen signal does not represent a specific earthquake.



(**a**) Correctly specified model (**b**) Misspecified model

Figure 2. Illustrative realizations of the synthetic GNSS coordinate series under the correctly specified and misspecified functional-model scenarios. Both series contain 2192 observations, including 219 observations contaminated by additive gross errors: (**a**) correctly specified functional model; (**b**) misspecified functional model with an unmodelled high-frequency periodic component.

2.7. Real GNSS Data

Daily coordinate time series from the Menghai (YNMH) and Xiaguan (XIAG) continuous GNSS stations in Yunnan, China, were obtained from the Crustal Movement Observation Network of China (CMONOC). The analyzed records span 2011.0014–2019.9603. The station observations were processed using GAMIT/GLOBK, together with the International GNSS Service (IGS) precise ephemerides and IGS global daily unconstrained solutions provided by the Scripps Orbit and Permanent Array Center (SOPAC), to generate daily station-coordinate products in the International Terrestrial Reference Frame 2014 (ITRF2014). YNMH contains several large excursions, whereas XIAG exhibits pronounced recurring seasonal variability. No missing epochs were interpolated, and no undocumented offsets were manually corrected before the analysis. Only available epochs were used in the functional-model fitting.

The daily positions were expressed in a local topocentric north-east-up (NEU) system relative to the station-specific reference coordinates at epoch 2011.0014 adopted in the GAMIT/GLOBK solution. All plotted coordinate offsets are expressed in meters; positive N, E, and U values denote northward, eastward, and upward displacements, respectively.

3. Results

3.1. Synthetic Results: Representative Realization and Monte Carlo Evaluation

Figure 3 illustrates the gross-error detection results obtained from one fixed-seed baseline realization. The purpose of these figures is to provide an intuitive visualization of the gross errors detected and missed by each method under the correctly specified and misspecified coordinate models. The black dots denote the original time series, whereas the red crosses and green triangles denote undetected and correctly detected injected gross errors, respectively.

For the correctly specified case shown in the left-hand panels, the recall rates in this baseline realization were 24.7% for the three-sigma criterion, 82.7% for IQR, and 99.7% for AEC-NDF. The conventional three-sigma criterion missed many injected gross errors because the full-series standard deviation was inflated by contamination. The IQR criterion substantially improved gross-error recovery, while AEC-NDF detected nearly all injected gross errors. These results are presented as a visual example rather than as estimates of the methods' expected performance.

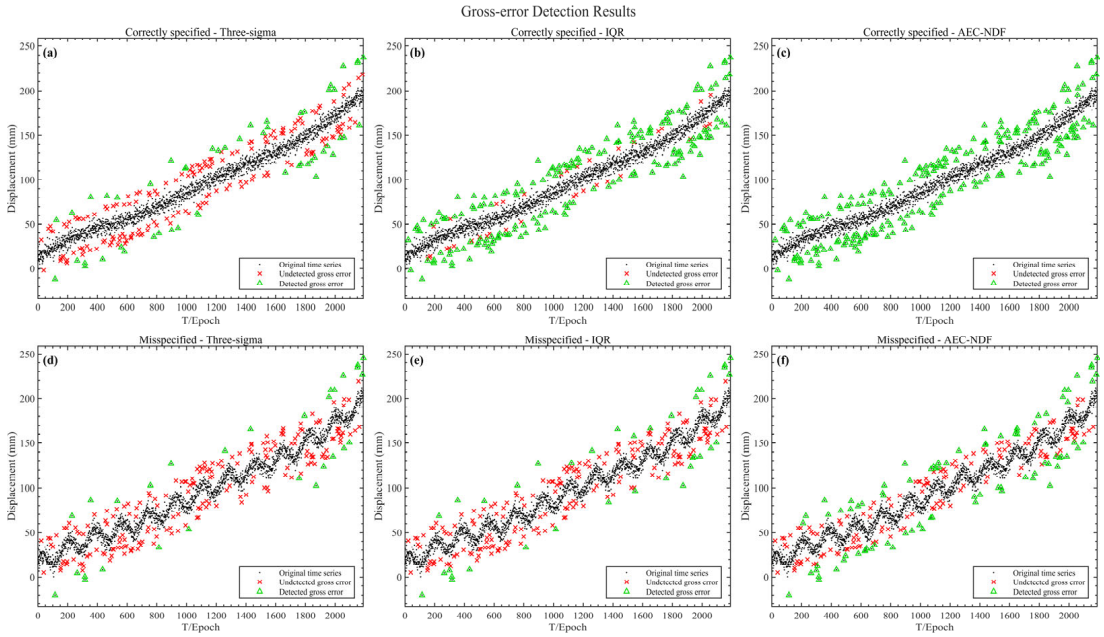


Figure 3. Gross-error detection results for one fixed-seed representative realization: (**a**) correctly specified model using the three-sigma criterion; (**b**) correctly specified model using the IQR criterion; (**c**) correctly specified model using AEC-NDF; (**d**) misspecified model using the three-sigma criterion; (**e**) misspecified model using the IQR criterion; and (**f**) misspecified model using AEC-NDF. Black dots denote the original time series, red crosses denote undetected injected gross errors, and green triangles denote correctly detected gross errors.

To evaluate whether the single-realization results were reproducible, the three methods were subsequently assessed using 1000 paired Monte Carlo realizations. Each realization contained 2192 epochs, including 219 epochs contaminated by gross errors. Within each paired realization, the correctly specified and misspecified scenarios shared the same ordinary noise, gross-error values, and contaminated epochs. Consequently, differences between the two scenarios primarily reflected the effect of model misspecification rather than random variation between simulations.

Under the correctly specified model, the mean recall rates were 24.88% (95% confidence interval (CI): 24.75–25.01%) for the three-sigma criterion, 83.64% (83.38–83.89%) for IQR, and 99.42% (99.36–99.48%) for AEC-NDF. These averages were close to the corresponding baseline-realization values of 24.7%, 82.7%, and 99.7%, indicating that the visual results in Figure 4 were broadly representative of the correctly specified simulations.

The corresponding mean F1 scores were 39.80% for the three-sigma criterion, 90.71% for IQR, and 98.15% for AEC-NDF. AEC-NDF therefore provided the best overall balance between gross-error recovery and false detections. Its improved recall was accompanied by a modest increase in the mean false-positive rate to 0.353%, compared with 0.074% for IQR and 0% for the three-sigma criterion. Thus, the almost complete gross-error recovery achieved by AEC-NDF did not result from an excessive number of false detections.

Under model misspecification, the mean recall rates decreased to 15.61% for the three-sigma criterion, 25.24% for IQR, and 45.09% for AEC-NDF. Based on the paired realizations, these values represented mean reductions of 9.27, 58.40, and 54.33 percentage points, respectively, relative to the correctly specified scenario. Nevertheless, AEC-NDF retained the highest recall and F1 score, with a mean F1 score of 61.69%, compared with 40.18% for IQR and 26.96% for the three-sigma criterion.

False-positive rates remained very low under model misspecification, and mean precision exceeded 99% for all three methods. Therefore, the deterioration in performance

was primarily attributable to an increase in false negatives rather than excessive false detections. The unmodelled 30-day periodic component broadened the residual distribution and increased the estimated detection thresholds, causing genuine gross errors to be missed. Overall, the paired Monte Carlo results confirm the superior performance of AEC-NDF under a correctly specified coordinate model, while also demonstrating that adequate modeling of deterministic coordinate variations remains important for reliable gross-error detection.

Within the deliberately conditioned white-noise simulations used here, AEC-NDF yielded higher recall and F1 scores than the two simple scale-based baselines. These results do not yet establish superiority under realistic GNSS noise or independently generated contamination.

Under the earthquake-aware fitted model, AEC-NDF achieved a mean recall of 99.32% and an F1 score of 98.09%, compared with 83.71% and 90.75%, respectively, for IQR. When the postseismic term was omitted but the coseismic step was retained, the AEC-NDF recall and F1 score decreased modestly to 96.58% and 96.63%. When both earthquake terms were omitted, its recall and F1 score fell to 44.90% and 49.02%, while its FPR rose from 0.355% to 3.381%. In this last condition, IQR achieved a slightly higher mean F1 score (49.83%) and a lower FPR (2.989%) than AEC-NDF. Thus, AEC-NDF did not retain an unconditional advantage under severe earthquake-signal misspecification.

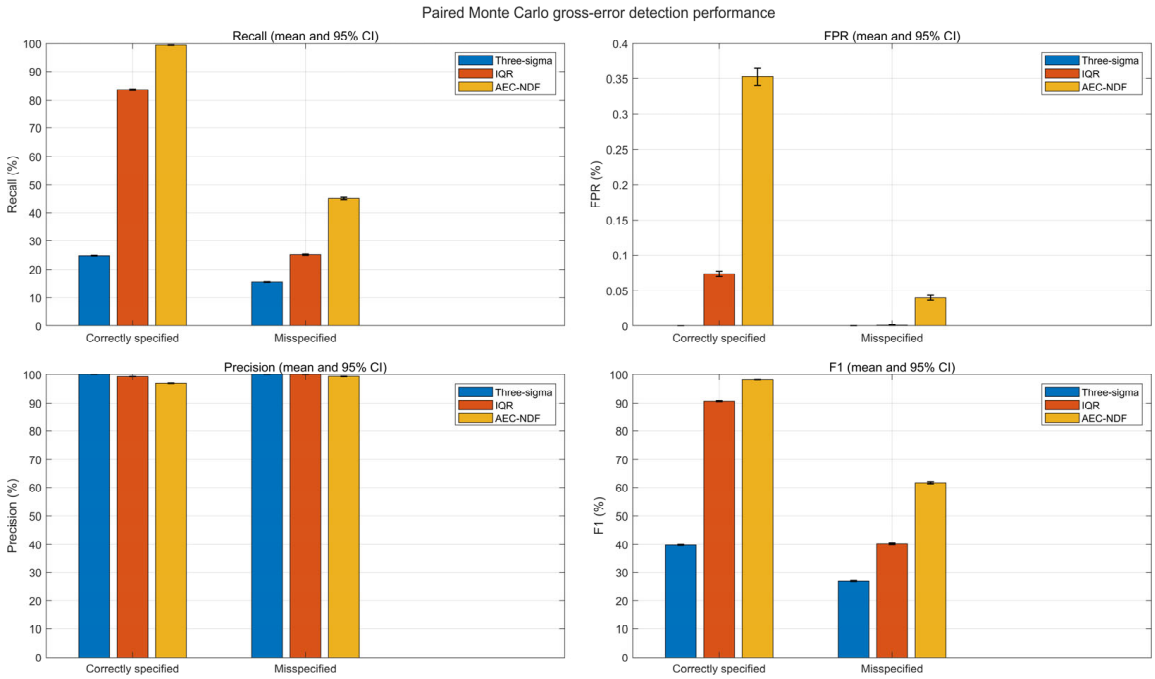


Figure 4. Gross-error detection performance over 1000 paired Monte Carlo realizations under correctly specified and misspecified coordinate models. Bars represent the mean recall, false-positive rate, precision, and F1 score. Error bars indicate approximate 95% confidence intervals for the Monte Carlo means, calculated as $\overline{M} \pm 1.96 s_M / \sqrt{n}$, where n is the number of valid realizations. Interval endpoints are restricted to 0–100%. The intervals quantify uncertainty arising from the finite number of realizations and are not prediction intervals for individual coordinate series. Within each realization, all methods were evaluated using the same simulated observations. Figure 5 shows the detection patterns for the first paired earthquake-like realization.

For AEC-NDF, the event-window FPR increased from 0.349% under the earthquake-aware model to 0.464% under the step-only model and 1.999% when both terms were omitted. The corresponding event-window recall decreased from 99.29% to 95.05% and 46.54%. The mean absolute error in the estimated coseismic amplitude after AEC-NDF

cleaning increased from 0.73 mm to 0.81 mm and 8.11 mm across these three conditions; the corresponding postseismic-amplitude errors were 1.94, 2.56, and 6.04 mm. These parameter errors are controlled-simulation diagnostics obtained with a common earthquake-aware evaluation model. The aggregate performance over 1000 paired realizations is summarized in Table 2, and the corresponding mean metrics are compared graphically in Figure 6.

Table 2. Gross-error detection performance under different earthquake-signal model specifications over 1000 paired Monte Carlo realizations.

Fitted Earthquake-Signal Model	Detection Method	Recall (%)	FPR (%)	Precision (%)	F1 Score (%)
Earthquake-aware	Three-sigma	24.91 ± 2.03	0.000 ± 0.000	100.00 ± 0.00	39.84 ± 2.61
	IQR	83.71 ± 4.12	0.075 ± 0.065	99.21 ± 0.66	90.75 ± 2.38
	AEC-NDF	99.32 ± 1.01	0.355 ± 0.193	96.91 ± 1.60	98.09 ± 0.80
Earthquake-partially specified	Three-sigma	24.57 ± 2.04	0.000 ± 0.000	100.00 ± 0.00	39.41 ± 2.63
	IQR	81.23 ± 4.06	0.077 ± 0.066	99.17 ± 0.69	89.24 ± 2.42
	AEC-NDF	96.58 ± 2.12	0.367 ± 0.210	96.74 ± 1.73	96.63 ± 0.97
Earthquake-unaware models	Three-sigma	11.98 ± 1.80	0.275 ± 0.132	82.96 ± 7.56	20.90 ± 2.88
	IQR	42.21 ± 4.06	2.989 ± 0.459	61.17 ± 3.47	49.83 ± 3.32
	AEC-NDF	44.90 ± 18.65	3.381 ± 2.021	62.02 ± 6.56	49.02 ± 11.24

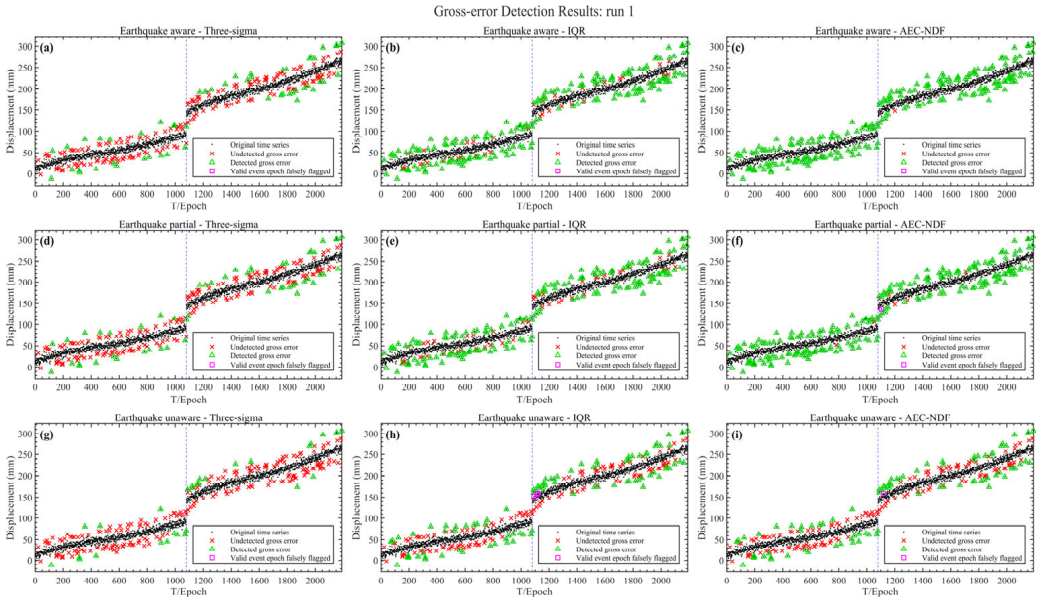


Figure 5. Gross-error detection results for the first paired earthquake-like simulation. Rows represent the earthquake-aware model, the partially specified model containing only the coseismic step, and the earthquake-unaware model, respectively. Columns represent the three-sigma, IQR, and AEC-NDF methods. Black dots show all simulated observations; red crosses indicate injected gross errors that were missed (false negatives); green triangles indicate correctly detected injected gross errors (true positives); and magenta squares indicate uncontaminated observations within the prescribed event window that were incorrectly flagged (false positives). The blue dashed line marks the simulated earthquake epoch. All panels contain the same observations and injected gross errors; only the fitted functional model and detection method differ.

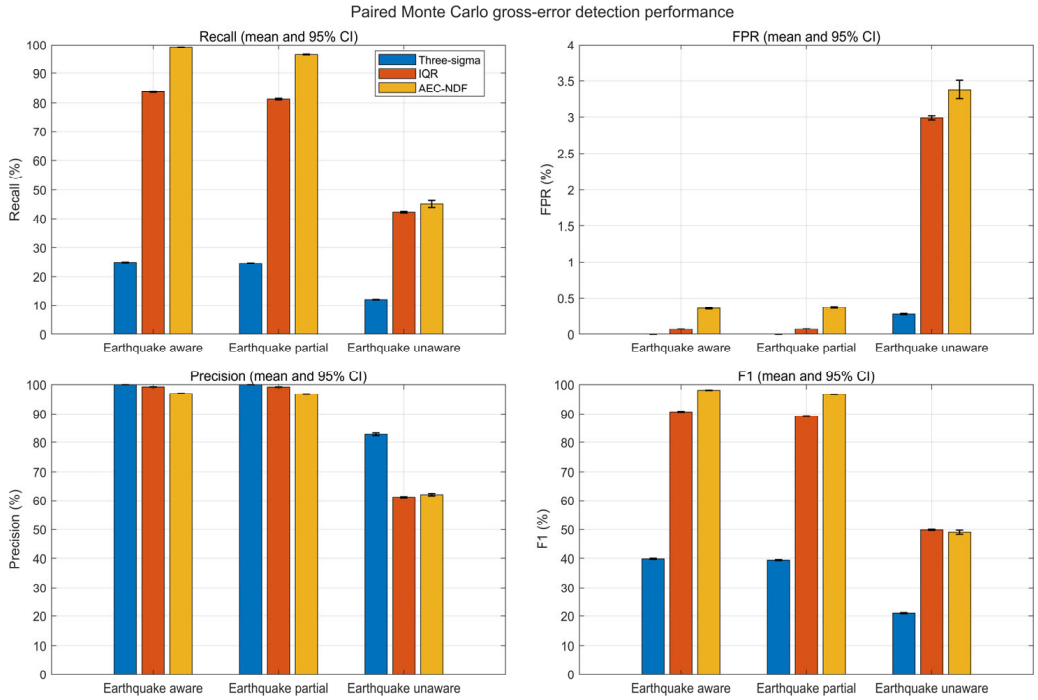


Figure 6. Gross-error detection performance over 1000 paired Monte Carlo realizations under earthquake-aware, earthquake-partially specified and earthquake-unaware models. Bars represent the mean recall, false-positive rate, precision, and F1 score. Error bars indicate approximate 95% confidence intervals for the Monte Carlo means, calculated as $\overline{M} \pm 1.96 s_M / \sqrt{n}$, where n is the number of valid realizations. Interval endpoints are restricted to 0–100%.

3.2. Sensitivity to Gross-Error Magnitude

Table 3 compares the three detection methods for gross-error generation scale multipliers of 3, 4, 5, 6, 7, 8, and 9. With the ordinary-noise standard deviation fixed at 5 mm, these settings correspond to candidate gross-error generating standard deviations of 15, 20, 25, 30, 35, 40, and 45 mm, respectively. Candidate contaminated epochs were identified using $|\varepsilon_i + g_i| > 3\sigma_0$, and 219 epochs were randomly selected for gross-error injection under each setting. The table therefore evaluates sensitivity to the scale of the generating distribution. The mean detection rates across the reported intervals were 24.7% for the three-sigma criterion, 84.6% for IQR, and 99.7% for AEC-NDF. AEC-NDF produced no more than four false positives in any interval (approximately 0.2% of the uncontaminated epochs in the 10% contamination experiment). These values describe the reported simulations; repeated Monte Carlo experiments are needed to quantify variability across random realizations.

Table 3. Detection performance for different gross-error generation scale multipliers.

Method	Performance Metric	3σ	4σ	5σ	6σ	7σ	8σ	9σ
Three-sigma	Detection Rate (%)	27.85	21.46	23.74	27.40	22.83	23.74	25.57
	False Positives (count)	0	0	0	0	0	0	0
IQR	Detection Rate (%)	89.50	85.39	84.93	78.08	86.30	84.93	83.11
	False Positives (count)	2	1	1	2	2	1	1
AEC-NDF	Detection Rate (%)	100.00	99.54	99.54	99.09	100.00	100.00	100.00
	False Positives (count)	4	3	3	1	4	3	4

3.3. Sensitivity to Contamination Ratio

Table 4 compares the three methods as the contamination ratio increases from 5% to 30%. The detection rates of the conventional three-sigma and IQR criteria decreased to 2.44% and 38.97%, respectively, at 30% contamination. In the reported realizations, AEC-NDF maintained detection rates between 98.63% and 100%, with no more than four false positives (below 0.3% of the uncontaminated epochs). Because these results appear to be based on individual random realizations, they should be interpreted as scenario-specific until repeated simulations are reported.

Table 4. Detection performance of the three methods for contamination ratios from 5% to 30%.

Method	Performance Metric	5%	10%	15%	20%	25%	30%
Three-sigma	Detection Rate (%)	51.38	28.77	14.94	6.62	4.20	2.44
	False Positives (count)	0	0	0	0	0	0
IQR	Detection Rate (%)	90.83	79.91	77.74	63.93	56.93	38.97
	False Positives (count)	1	0	0	2	2	1
AEC-NDF	Detection Rate (%)	100.00	99.54	99.78	98.86	99.09	98.63
	False Positives (count)	4	3	3	4	4	3

3.4. Real GNSS Results

3.4.1. YNMH Station

Figures 7–9 compare the observations flagged by the three-sigma, IQR, and AEC-NDF methods for the north, east, and up components of YNMH. Red crosses indicate observations flagged by the respective method.

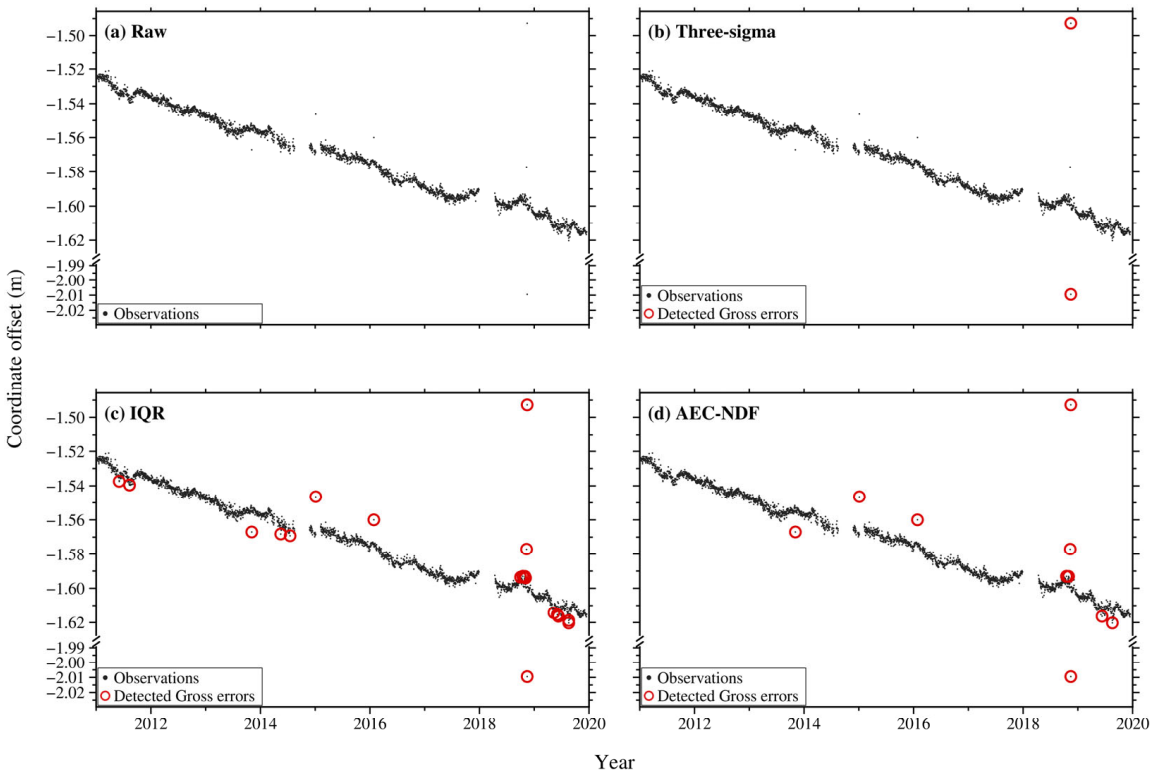


Figure 7. Observations flagged by the three methods for the north component of the YNMH station.

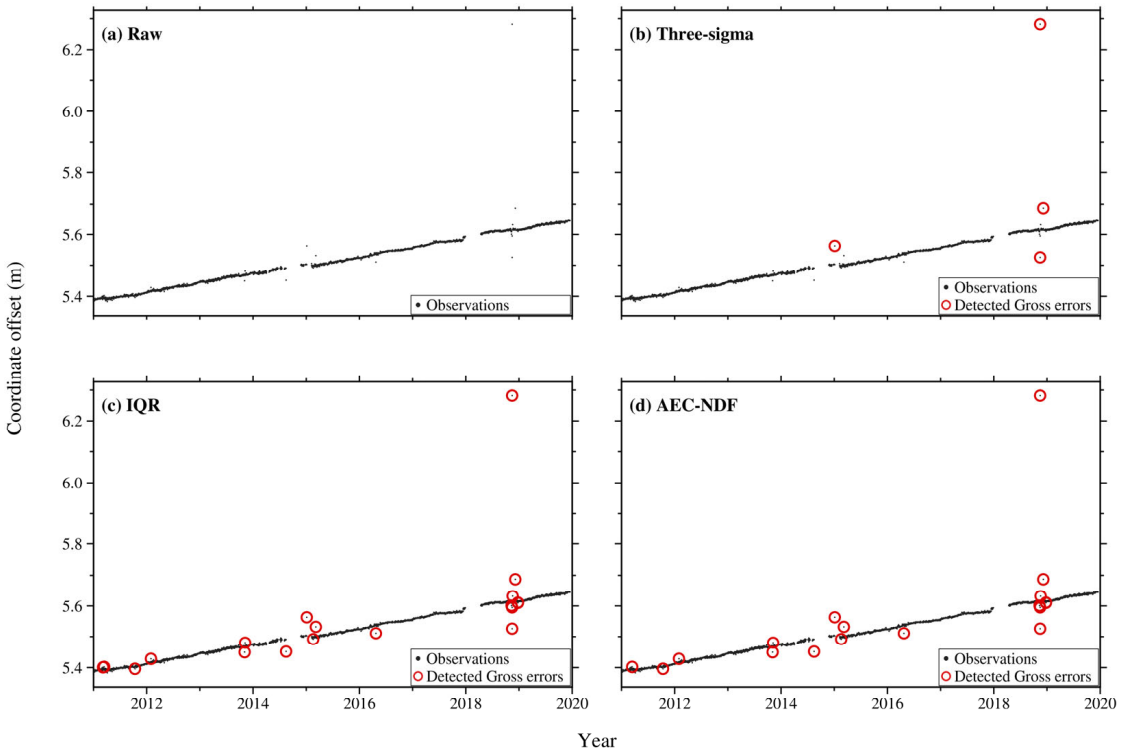


Figure 8. Observations flagged by the three methods for the east component of the YNMH station.

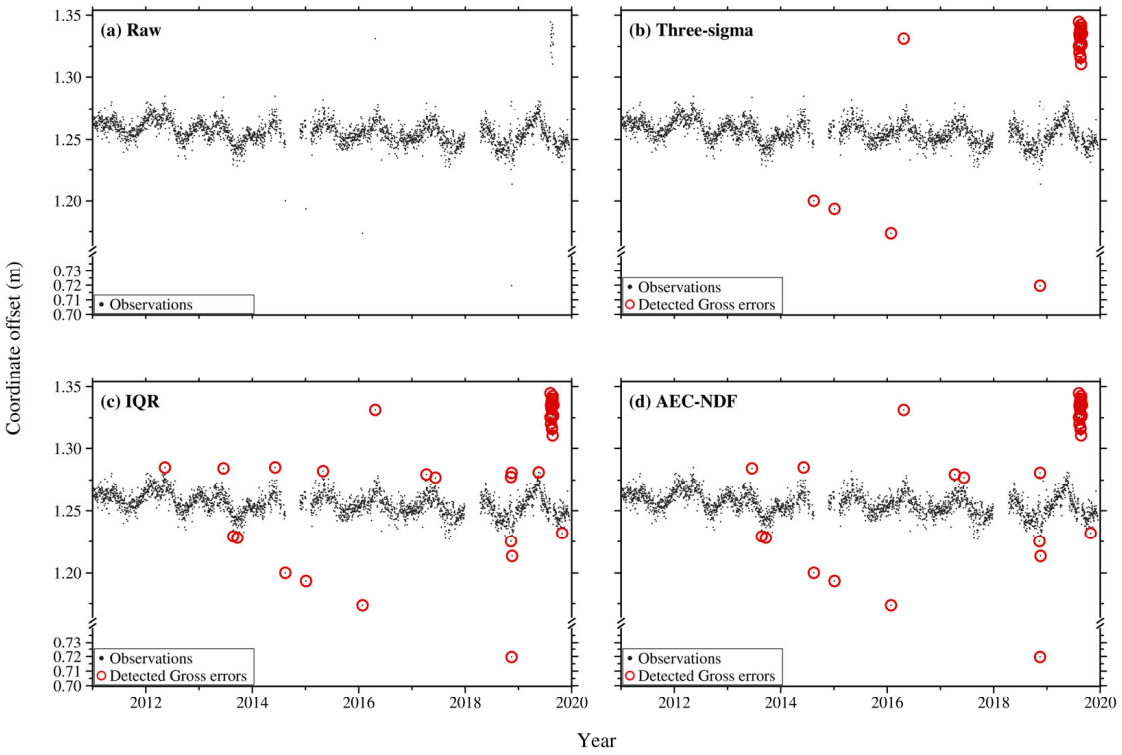


Figure 9. Observations flagged by the three methods for the up component of the YNMH station.

At YNMH, all three methods flagged the largest excursions. In the north and east components, the conventional three-sigma criterion flagged fewer secondary excursions than IQR and AEC-NDF, consistent with threshold inflation caused by large residuals. Because independent labels are unavailable, these plots compare flagging behavior rather than provide a direct estimate of real-data detection accuracy.

3.4.2. XIAG Station

Figures 10–12 compare the observations flagged by the three methods for the north, east, and up components of XIAG.

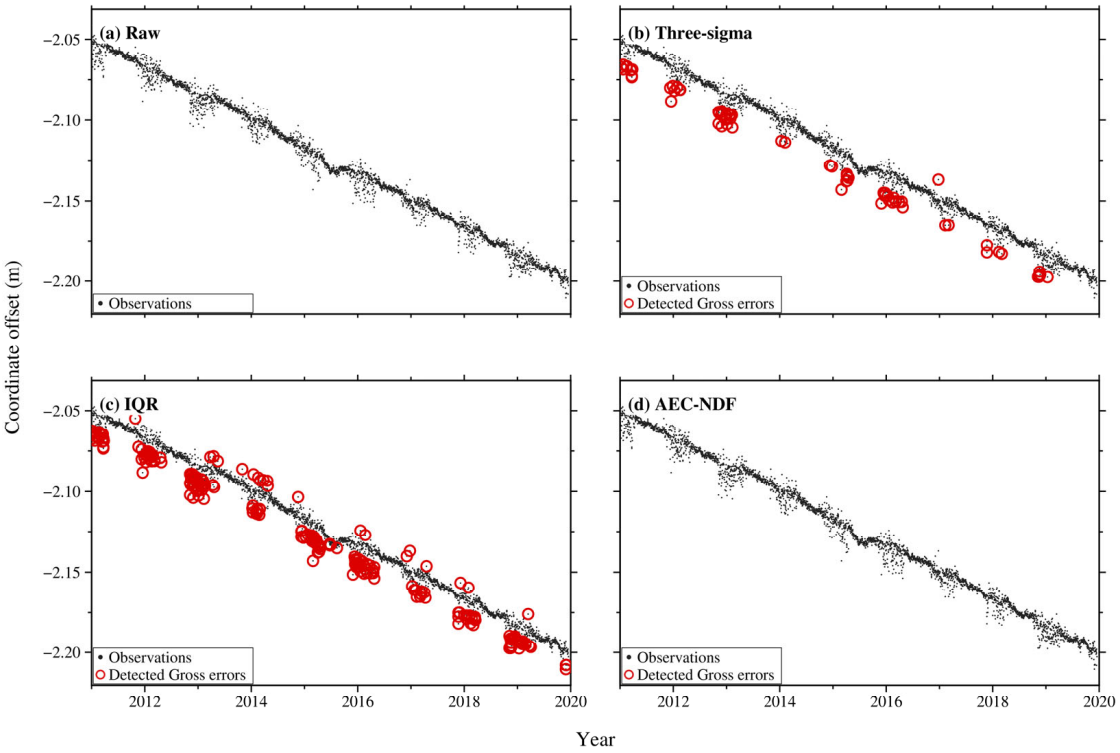


Figure 10. Observations flagged by the three methods for the north component of the XIAG station.

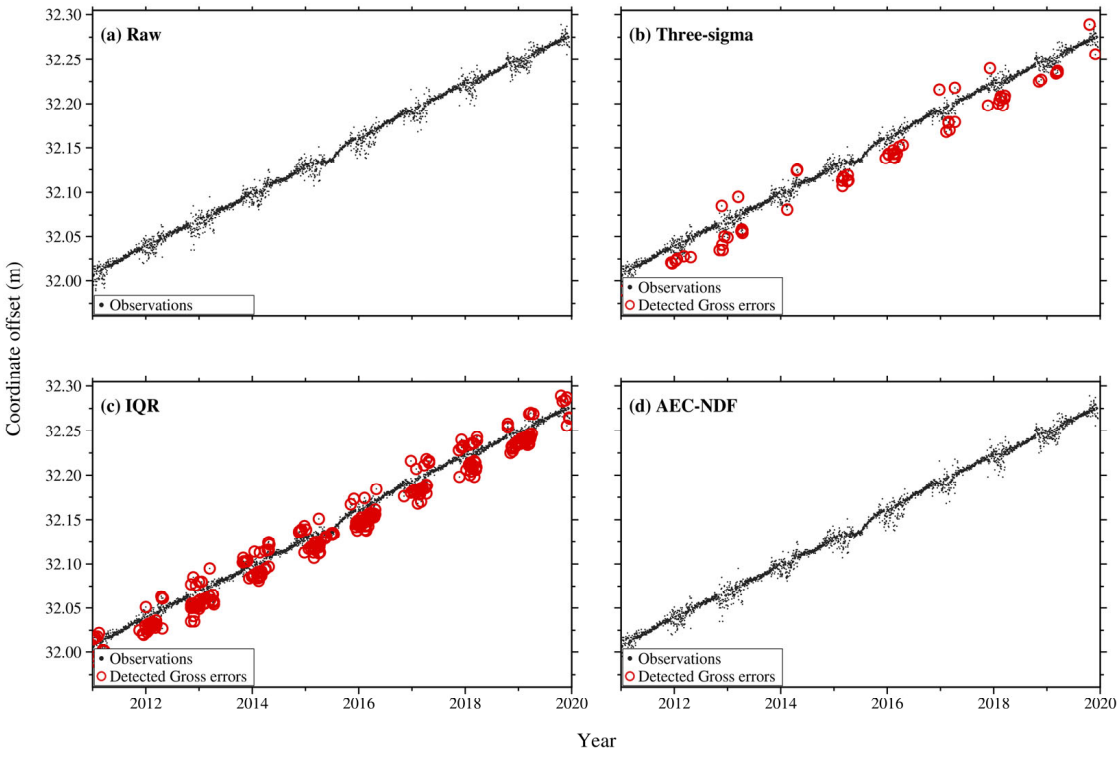


Figure 11. Observations flagged by the three methods for the east component of the XIAG station.

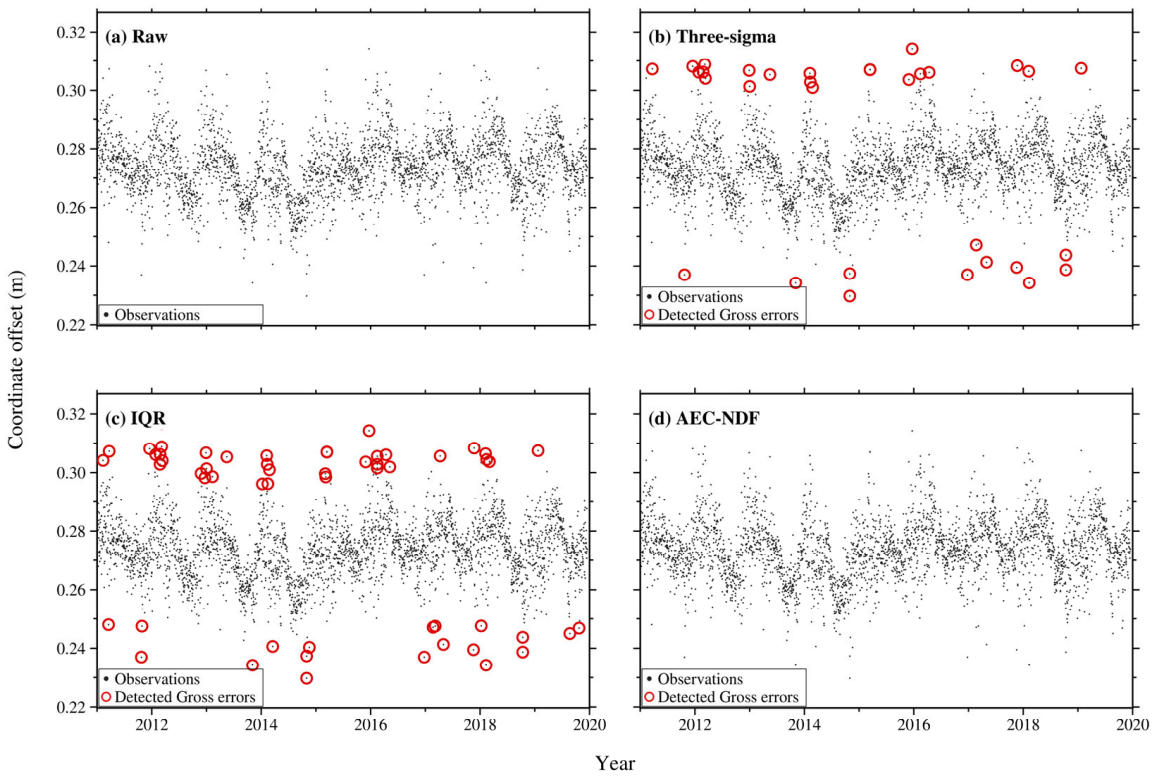


Figure 12. Observations flagged by the three methods for the up component of the XIAG station.

The XIAG series shows a clear difference in flagging behavior. The three-sigma and especially the IQR criterion flagged many observations aligned with recurring seasonal excursions, whereas AEC-NDF retained most of these observations. This pattern is consistent with less aggressive removal of seasonally coherent variability by AEC-NDF. However, because no independent outlier labels are available, the result does not prove that every retained observation is valid or that every IQR-flagged observation is a true seasonal signal.

4. Discussion

The principal finding of this study is that estimating the residual scale from the high-density central population substantially reduced masking under the correctly specified synthetic model. Across 1000 paired Monte Carlo realizations, AEC-NDF achieved a mean recall of 99.42 $\pm$ 0.97%, a mean precision of 96.93 $\pm$ 1.60%, and a mean F1 score of 98.15 $\pm$ 0.80%. The corresponding F1 scores were 90.71 $\pm$ 2.40% for IQR and 39.80 $\pm$ 2.62% for the conventional three-sigma criterion. AEC-NDF did incur a modestly higher false-positive rate than IQR under this scenario (0.353 $\pm$ 0.193% versus 0.074 $\pm$ 0.063%), but the gain in recall was sufficiently large to produce the best overall balance between missed and false detections. Conversely, the zero false-positive rate of the conventional three-sigma method was accompanied by a recall of only 24.88 $\pm$ 2.04%; it therefore reflected an overly conservative, contamination-inflated threshold rather than superior classification performance.

This behavior is consistent with the statistical mechanism that motivated AEC-NDF. A scale estimate computed from the complete residual series has an unbounded response to large residuals; as contamination increases, the threshold expands and genuine gross errors become masked [25]. AEC-NDF instead evaluates a sequence of nested central-residual subsets and selects the scale near the change in the fitted-scale trajectory before the gross-error tail dominates the estimate. The improvement observed here therefore results primarily from scale estimation rather than from a more aggressive threshold

multiplier, because all decisions are ultimately expressed through an interpretable three-sigma rule. IQR is more resistant than the conventional standard deviation to isolated extremes, which explains its intermediate performance, but it does not explicitly use the expected convergence pattern of the near-Gaussian central residual population. This distinction is important when the contaminated tail is sufficiently large to influence a global scale summary.

The magnitude- and contamination-sensitivity experiments provide additional evidence for this interpretation. Across the reported gross-error intervals from generation-scale multipliers from $3\sigma_0$ to $9\sigma_0$, AEC-NDF detected 99.09–100% of the injected errors and produced only one to four false positives, whereas the detection rates of the other two methods remained substantially lower. Similarly, as the contamination ratio increased from 5% to 30%, the reported AEC-NDF detection rate remained between 98.63% and 100%, while the three-sigma and IQR rates decreased to 2.44% and 38.97%, respectively, at 30% contamination. These patterns support the proposed resistance to threshold inflation over the investigated range. Nevertheless, unlike the baseline Monte Carlo comparison, Tables 3 and 4 report individual scenario realizations without uncertainty intervals. Their numerical ranges should therefore be interpreted as descriptive evidence for the tested realizations, not as population-level performance bounds.

The misspecified-model experiment also identifies an important boundary of the method. When the 30-day periodic component was omitted from the fitted functional model, the mean recall of AEC-NDF decreased from 99.42% to 45.09%, and its mean F1 score decreased from 98.15% to 61.69%. AEC-NDF still outperformed IQR and the conventional three-sigma criterion, whose mean F1 scores were 40.18% and 26.96%, respectively, but it was not immune to inadequate signal modeling. Because mean precision remained above 99% and false-positive rates remained very low for all three methods, the loss of performance was dominated by false negatives. The omitted periodic signal broadened the residual distribution and increased the estimated scale, thereby masking contaminated epochs. This result reinforces the established coupling between outlier treatment, functional modelling, and stochastic-noise characterization [33,36,37,43]. In practice, AEC-NDF should therefore be applied after known offsets and major deterministic terms have been modeled, and preferably within an iterative procedure in which the coordinate model and the set of flagged observations are re-estimated until stable.

The earthquake-like stress test identifies a practical boundary of central-residual scale fitting. AEC-NDF provided the strongest F1 performance among the tested methods when the coseismic step and postseismic relaxation were represented in the fitted model and retained an advantage when only the relaxation term was omitted. However, omitting both terms left structured geophysical deformation in the residuals. In that condition, AEC-NDF showed substantially lower and more variable gross-error recovery, higher false-positive rates, and a mean F1 score slightly below that of IQR. Robust scale estimation therefore cannot substitute for explicit modeling of major geophysical transients.

The station examples illustrate two potentially useful operational behaviors. At YNMH, all three methods identified the largest excursions, whereas AEC-NDF and IQR flagged secondary excursions that were missed by the conventional three-sigma rule. This pattern is consistent with reduced masking in the presence of several very large residuals. At XIAG, AEC-NDF flagged markedly fewer observations aligned with recurring annual variability than either IQR or, in some components, the conventional rule. Such restraint is desirable if these observations represent genuine hydrological, atmospheric, or thermal loading rather than measurement faults. However, the same visual pattern could also arise from reduced sensitivity to moderate anomalies; without independent labels, it cannot by itself demonstrate better signal preservation. The real-data results should

therefore be understood as a comparison of flagging behavior, not as a measurement of detection accuracy.

A stronger real-data validation should evaluate the downstream geodetic consequences of each flagging strategy. Relevant tests include changes in estimated velocity, annual and semi-annual amplitude and phase, residual spectral power, and colored-noise parameters before and after cleaning. Temporal coherence across coordinate components, agreement with nearby stations, and station logs documenting equipment changes or environmental disturbances could provide additional evidence for whether a flagged epoch is anomalous. Semi-synthetic experiments, in which controlled gross errors are injected into cleaned real residuals, would also retain realistic temporal dependence while providing known labels. These evaluations are particularly important because removing valid seasonal extremes can bias geophysical interpretation, whereas retaining true gross errors can bias both parameter estimates and uncertainty assessments [24,33,37,43].

Several limitations constrain the generality of the present results. The simulations assume Gaussian white noise, uniformly located additive gross errors, and a relatively simple deterministic signal. They do not represent flicker or random-walk noise, heteroscedasticity, long data gaps, clustered or persistent anomalies, asymmetric contamination, offsets, coseismic steps, or equipment changes that are common in operational GNSS series. Only two stations were examined, neither with independently verified gross-error labels. The comparison was also restricted to conventional three-sigma and IQR rules; broader benchmarking against MAD/Hampel procedures, M-estimation, DIA, Bayesian methods, and signal-decomposition approaches would clarify the conditions under which central-residual fitting offers a practical advantage [29–32,34,38–40]. Finally, reproducibility depends on complete specification of the residual-bin width, candidate subset sequence and detrending procedures, local-peak criterion, minimum subset size, and fallback rule when no stable change point is found.

Future work should therefore combine factorial Monte Carlo experiments with network-scale real-data validation. The synthetic design should vary colored-noise amplitude, temporal correlation, sampling irregularity, gross-error clustering, contamination sign and magnitude, and functional-model misspecification, while reporting uncertainty for every sensitivity scenario. Parameter-sensitivity and ablation analyses are needed to determine how the selected central subset and change-point procedure affect the estimated scale. Extensions could include component-specific or adaptive thresholds and an iterative refitting strategy, but any additional flexibility should be evaluated against false-positive control and preservation of genuine deformation signals. These steps would establish whether the strong performance observed under the correctly specified simulations transfers to heterogeneous GNSS networks and long operational records.

5. Conclusions

This study developed AEC-NDF as an interpretable gross-error detection method for GNSS coordinate time series. By fitting the scale to nested subsets of the central residual population rather than to the complete residual series, the method limits the influence of contaminated tails and reduces the masking that affects a conventional three-sigma threshold.

The strongest quantitative evidence was obtained from 1000 paired Monte Carlo realizations under a correctly specified coordinate model. AEC-NDF achieved a mean recall of 99.42%, a mean F1 score of 98.15%, and a mean false-positive rate of 0.353%, outperforming both IQR and the conventional three-sigma criterion in the balance between error recovery and false detections. The reported single-realization sensitivity experiments were consistent with this result: detection remained at 99.09–100% across the tested gross-

error magnitude intervals and at 98.63–100% across contamination ratios of 5–30%, with no more than four false positives in any reported realization.

The experiments also show that performance depends on the adequacy of the fitted coordinate model. Omitting a 30-day periodic component reduced the mean AEC-NDF recall to 45.09% and the mean F1 score to 61.69%, although both remained higher than those of the comparison methods. The YNMH and XIAG applications further demonstrated distinct flagging behavior in operational series, including less removal of seasonally coherent XIAG observations by AEC-NDF. Because independent labels were not available, these station results provide qualitative evidence only and should not be interpreted as real-data accuracy estimates.

In an illustrative earthquake-like simulation, AEC-NDF retained high recall and F1 scores when the major coseismic and postseismic terms were included in the fitted coordinate model. Its performance deteriorated markedly when both terms were omitted, and its mean F1 score was then slightly lower than that of IQR. The method is therefore recommended as a gross-error screening step after known offsets and major transient deformation have been modeled, rather than as a means of distinguishing all unmodelled earthquake signals from gross errors.

AEC-NDF is therefore best regarded as a promising preprocessing tool rather than a substitute for explicit modeling of offsets, deterministic deformation, and temporally correlated noise. Its use is most defensible after adequate functional modelling and with iterative re-estimation of the coordinate model. Validation under realistic colored-noise conditions, clustered anomalies, data gaps, and larger labeled or semi-synthetic GNSS datasets is required before the method can be generalized to routine network-wide quality control.

Author Contributions: Conceptualization, M.Z. and H.G.; methodology, M.Z.; software, M.Z.; validation, H.G., S.H. and Y.N.; formal analysis, M.Z. and H.G.; investigation, M.Z.; resources, H.G.; data curation, M.Z.; writing—original draft preparation, M.Z.; writing—review and editing, H.G., S.H. and Y.N.; visualization, M.Z.; supervision, H.G.; funding acquisition, H.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant number 42404055.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The GNSS coordinate time series used in this study are available from the corresponding author upon reasonable request, subject to permission from the data providers.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Bock, Y.; Melgar, D. Physical applications of GPS geodesy: A review. *Rep. Prog. Phys.* **2016**, *79*, 106801. [CrossRef] [PubMed]
2. Stamps, D.S.; Kreemer, C. Open access GNSS data for studies of the lithosphere. *Geochem. Geophys. Geosyst.* **2024**, *25*, e2024GC011567. [CrossRef]
3. He, X.; Montillet, J.-P.; Li, Z.; Kermarrec, G.; Fernandes, R.; Zhou, F. Recent Advances in Modelling Geodetic Time Series and Applications for Earth Science and Environmental Monitoring. *Remote Sens.* **2022**, *14*, 6164. [CrossRef]
4. Yu, J.; Meng, X.; Yan, B.; Xu, B.; Fan, Q.; Xie, Y. Global Navigation Satellite System-based positioning technology for structural health monitoring: A review. *Struct. Control Health Monit.* **2020**, *27*, e2467. [CrossRef]
5. Moschas, F.; Stiros, S. Measurement of the dynamic displacements and of the modal frequencies of a short-span pedestrian bridge using GPS and an accelerometer. *Eng. Struct.* **2011**, *33*, 10–17. [CrossRef]

6. Li, Z.; Jiang, W.; van Dam, T.; Zou, X.; Chen, Q.; Chen, H. A Review on Modeling Environmental Loading Effects and Their Contributions to Nonlinear Variations of Global Navigation Satellite System Coordinate Time Series. *Engineering* **2025**, *47*, 26–37. [CrossRef]
7. Dong, D.; Fang, P.; Bock, Y.; Cheng, M.K.; Miyazaki, S. Anatomy of apparent seasonal variations from GPS-derived site position time series. *J. Geophys. Res. Solid Earth* **2002**, *107*, ETG 9-1–ETG 9-16. [CrossRef]
8. Ray, J.; Altamimi, Z.; Collilieux, X.; van Dam, T. Anomalous harmonics in the spectra of GPS position estimates. *GPS Solut.* **2008**, *12*, 55–64. [CrossRef]
9. Gazeaux, J.; Williams, S.; King, M.; Bos, M.; Dach, R.; Deo, M.; Moore, A.W.; Ostini, L.; Petrie, E.; Roggero, M.; et al. Detecting offsets in GPS time series: First results from the detection of offsets in GPS experiment. *J. Geophys. Res. Solid Earth* **2013**, *118*, 2397–2407. [CrossRef]
10. Perfetti, N. Detection of station coordinate discontinuities within the Italian GPS Fiducial Network. *J. Geod.* **2006**, *80*, 381–396. [CrossRef]
11. Williams, S.D.P. Offsets in Global Positioning System time series. *J. Geophys. Res. Solid Earth* **2003**, *108*, 2310. [CrossRef]
12. Amiri-Simkooei, A.R.; Tiberius, C.C.J.M.; Teunissen, P.J.G. Assessment of noise in GPS coordinate time series: Methodology and results. *J. Geophys. Res. Solid Earth* **2007**, *112*, B07413. [CrossRef]
13. Bos, M.S.; Fernandes, R.M.S.; Williams, S.D.P.; Bastos, L. Fast error analysis of continuous GNSS observations with missing data. *J. Geod.* **2013**, *87*, 351–360. [CrossRef]
14. Langbein, J. Noise in GPS displacement measurements from Southern California and Southern Nevada. *J. Geophys. Res. Solid Earth* **2008**, *113*, B05405. [CrossRef]
15. Mao, A.; Harrison, C.G.A.; Dixon, T.H. Noise in GPS coordinate time series. *J. Geophys. Res. Solid Earth* **1999**, *104*, 2797–2816. [CrossRef]
16. Santamaría-Gómez, A.; Ray, J. Chameleonic Noise in GPS Position Time Series. *J. Geophys. Res. Solid Earth* **2021**, *126*, e2020JB019541. [CrossRef]
17. Williams, S.D.P. The effect of coloured noise on the uncertainties of rates estimated from geodetic time series. *J. Geod.* **2003**, *76*, 483–494. [CrossRef]
18. Williams, S.D.P.; Bock, Y.; Fang, P.; Jamason, P.; Nikolaidis, R.M.; Prawirodirdjo, L.; Miller, M.; Johnson, D.J. Error analysis of continuous GPS position time series. *J. Geophys. Res. Solid Earth* **2004**, *109*, B03412. [CrossRef]
19. Wang, L.; Wu, Q.; Wu, F.; He, X. Noise content assessment in GNSS coordinate time-series with autoregressive and heteroscedastic random errors. *Geophys. J. Int.* **2022**, *231*, 856–876. [CrossRef]
20. Cucci, D.A.; Voirol, L.; Kermarrec, G.; Montillet, J.-P.; Guerrier, S. The Generalized Method of Wavelet Moments with eXogenous inputs: A fast approach for the analysis of GNSS position time series. *J. Geod.* **2023**, *97*, 14. [CrossRef] [PubMed]
21. Huan, Y.; Chang, G.; Huang, Y.; Feng, Y.; Zhu, Y.; Yang, S. Selection of noise models for GNSS coordinate time series based on model averaging algorithm. *Meas. Sci. Technol.* **2024**, *35*, 076305. [CrossRef]
22. Grubbs, F.E. Procedures for Detecting Outlying Observations in Samples. *Technometrics* **1969**, *11*, 1–21. [CrossRef]
23. Koch, K.-R. *Parameter Estimation and Hypothesis Testing in Linear Models*, 2nd ed.; Springer: Berlin/Heidelberg, Germany, 1999. [CrossRef]
24. Nikolaidis, R.M. Observation of Geodetic and Seismic Deformation with the Global Positioning System. Ph.D. Thesis, University of California, San Diego, CA, USA, 2002.
25. Barnett, V.; Lewis, T. *Outliers in Statistical Data*, 3rd ed.; Wiley: Chichester, UK, 1994.
26. Tukey, J.W. *Exploratory Data Analysis*; Addison-Wesley: Reading, MA, USA, 1977.
27. Hampel, F.R. The Influence Curve and its Role in Robust Estimation. *J. Am. Stat. Assoc.* **1974**, *69*, 383–393. [CrossRef]
28. Rousseeuw, P.J.; Croux, C. Alternatives to the Median Absolute Deviation. *J. Am. Stat. Assoc.* **1993**, *88*, 1273–1283. [CrossRef]
29. Huber, P.J.; Ronchetti, E.M. *Robust Statistics*, 2nd ed.; Wiley: Hoboken, NJ, USA, 2009. [CrossRef]
30. Khodabandeh, A.; Amiri-Simkooei, A.R.; Sharifi, M.A. GPS position time-series analysis based on asymptotic normality of M-estimation. *J. Geod.* **2012**, *86*, 15–33. [CrossRef]
31. Teunissen, P.J.G. Distributional theory for the DIA method. *J. Geod.* **2018**, *92*, 59–80. [CrossRef]
32. Zhang, Q.; Gui, Q. Bayesian methods for outliers detection in GNSS time series. *J. Geod.* **2013**, *87*, 609–627. [CrossRef]
33. Klos, A.; Olivares, G.; Teferle, F.N.; Hunegnaw, A.; Bogusz, J. On the combined effect of periodic signals and colored noise on velocity uncertainties. *GPS Solut.* **2018**, *22*, 1. [CrossRef]
34. Wang, X.; Cheng, Y.; Wu, S.; Zhang, K. An effective toolkit for the interpolation and gross error detection of GPS time series. *Surv. Rev.* **2016**, *48*, 202–211. [CrossRef]
35. Ji, K.; Shen, Y. A Wavelet-Based Outlier Detection and Noise Component Analysis for GNSS Position Time Series. In *Beyond 100: The Next Century in Geodesy*; Freymueller, J.T., Sánchez, L., Eds.; International Association of Geodesy Symposia; Springer: Cham, Switzerland, 2020; Volume 152, pp. 45–53. [CrossRef]

36. Klos, A.; Bos, M.S.; Bogusz, J. Detecting time-varying seasonal signal in GPS position time series with different noise levels. *GPS Solut.* **2018**, *22*, 21. [CrossRef]
37. Bogusz, J.; Klos, A. On the significance of periodic signals in noise analysis of GPS station coordinates time series. *GPS Solut.* **2016**, *20*, 655–664. [CrossRef]
38. Pehlivan, H. A novel outlier detection method based on Bayesian change point analysis and Hampel identifier for GNSS coordinate time series. *EURASIP J. Adv. Signal Process.* **2024**, *2024*, 44. [CrossRef]
39. Wang, J.; Yan, S.; Tu, R.; Zhang, P. Detection and Mitigation of GNSS Gross Errors Utilizing the CEEMD and IQR Methods to Determine Sea Surface Height Using GNSS Buoys. *Sensors* **2025**, *25*, 2863. [CrossRef] [PubMed]
40. Chen, J.; Wang, Y.; Zhao, X.; Liu, C. An integrated FE-SVMD-IQR method for improved anomaly detection in GNSS coordinate time series. *Adv. Space Res.* **2026**, *77*, 11994–12012. [CrossRef]
41. Lenzano, M.G.; Rivera, A.; Durand, M.; Vacaflor, P.; Carbonetti, M.; Lannutti, E.; Gende, M.; Lenzano, L. Detection of Crustal Uplift Deformation in Response to Glacier Wastage in Southern Patagonia. *Remote Sens.* **2023**, *15*, 584. [CrossRef]
42. Li, X.; Wang, L.; Qu, X.; Zhang, S.; Shu, B.; Xu, H. A GPS multipath mitigation method in coordinate-domain considering the effects of gross errors and missing data. *Measurement* **2024**, *225*, 114035. [CrossRef]
43. Klos, A.; Bogusz, J.; Figurski, M.; Kosek, W. On the Handling of Outliers in the GNSS Time Series by Means of the Noise and Probability Analysis. In *IAG 150 Years*; Rizos, C., Willis, P., Eds.; International Association of Geodesy Symposia; Springer: Cham, Switzerland, 2015; Volume 143, pp. 657–664. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.