Russian Journal of Earth Sciences

Machine Learning for GNSS Time Series Analysis in the Time Domain

Yu. V. Gabsatarov[*,1] **and I. S. Vladimirova**[1]

[1] Shirshov Institute of Oceanology, Russian Academy of Sciences, Moscow, Russia
* **Correspondence to:** Yurii Gabsatarov, gabsatarov.yv@ocean.ru

Abstract: The paper presents the results of developing a method for analyzing time series of GNSS measurements based on a machine learning approach. The constructed algorithm was tested on GNSS data from the vicinity of sources of large earthquakes occurred in regions with different tectonic structures: the Japanese islands, Southern California, and the Peruvian-Chilean coast. It is shown that the proposed approach allows one to build an adequate, versatile, interpretable, statistically significant time series model using exclusively statistical data analysis methods, which will further allow one to create automated processing systems operating in a near-real-time mode.

Keywords: Satellite geodesy, recent crustal movements, machine learning, time series analysis.

Citation: Gabsatarov Yu. V. and Vladimirova I. S. (2025), Machine Learning for GNSS Time Series Analysis in the Time Domain, *Russian Journal of Earth Sciences*, 25, ES6022, EDN: UMNEIN, https://doi.org/10.2205/2025es001018

Introduction

Since the late 1980s, data from Global Navigation Satellite Systems (GNSS) have been widely used to solve a wide range of geodynamic problems. The development of theoretical foundations for the application of satellite geodetic methods in solving geodynamic problems determines the increasingly active use of GNSS methods in the study of seismotectonic deformations in addition to classical geological and geophysical methods. Currently, GNSS observation data are actively used to study recent movements of the Earth's surface, detect spatiotemporal variations in deformation fields near active faults and active volcanoes, as well as in a number of other applied geophysical studies.

Regression analysis of GNSS time series is one of the most widely employed methods for analyzing satellite geodetic data. This method enables researchers to collect new data on geodynamical processes with higher accuracy by isolating various components of the observed signal using classical statistical methods. Currently, regression analysis algorithms are extensively employed to derive preliminary data for constructing earthquake source models [*Steblov et al.*, 2008], identifying precursors in contemporary crustal deformation fields [*Gitis et al.*, 2021; *Liu and Kossobokov*, 2021], and analyzing the frequency composition of GNSS data [*Nikolaidis*, 2002], among other applications. The most advanced regression models are employed in the processing of GNSS measurements at the International GNSS Service data analysis centers [*Blewitt et al.*, 2016; *Bock et al.*, 2023] and in the construction of the International Terrestrial Reference Frame, beginning with the ITRF2014 version [*Altamimi et al.*, 2023, 2016].

In the last decade, there has been a sharp increase in the number of studies devoted to the processing of time series of various data using machine learning methods. Active development of the methodology has led to the emergence of a number of new approaches to the analysis of GNSS data time series based on machine learning methods. These algorithms are aimed at solving the main problems arising in the processing of GNSS data time series: detecting outliers, searching and modeling instantaneous shifts, modeling transient deformation processes. In particular, the following have been proposed: 1) new algorithms for studying time series in the frequency domain [*Ji et al.*, 2020]; 2) new methods for modeling transient processes, such as seasonal variations, nonstationary trends, etc. [*Xue and Freymueller*, 2023; *Yamaga and Mitsui*, 2019]; 3) new algorithms for missing data

Research Article

Received: November 13, 2024
Accepted: April 29, 2025
Published: December 30, 2025

recovery [*Zhang et al.*, 2021]; 4) time series modeling methods based on the recurrent neural network algorithm [*Ozbey et al.*, 2024] and many others. New machine learning-based methods can capture complex, nonlinear patterns and interactions that classical regression models might miss.

In this article we consider GNSS measurements not as auxiliary information that merely complements the available geological and geophysical data, but as direct input data that requires direct interpretation. The volumes of the obtained GNSS measurements due to continuous data from dense GNSS networks in tectonically active regions, their variety due to large amount of postprocessing data and accumulation speed have brought GNSS data very close to the ensemble of Big Data sources in Earth Sciences. This approach seems quite justified, since GNSS data themselves contain sufficient information about the nature of recent crustal deformation, and this information can be extracted using specialized quantitative Big Data methods, such as anomaly detection, filtering, interpolation, clustering, etc. [*Gvishiani et al.*, 2022].

The proposed approaches to interpreting time series of GNSS measurements are based on the use of machine learning methods, which corresponds to modern global trends in Earth sciences and allows creating new methods for studying the features of regional fields of recent movements of the Earth's surface and to move on to studying the patterns of their formation. The purpose of this study is to build an adequate, versatile, statistically significant and interpretable regression model to obtain correct estimates of the components of time series in the time domain, which, in turn, can be used to study anomalies in the field of deformations of the Earth's surface and to build models of geodynamic processes.

Data and methods

Satellite geodetic data

Time series of high-precision daily estimates of GNSS station positions were used as initial data for constructing regression models. These time series reflect the variability of positions caused by both tectonic processes and systematic and random errors in GNSS measurements. In order to achieve the requirement of versatility, the regression recovery algorithm was tested and verified on time series of GNSS stations located in regions with different tectonic structure and activity: the Japanese Islands (more than 1300 stations), the coast of Peru-Chile (130 stations), and Southern California (71 stations). The GNSS observation data were provided: 1) on the Japanese Islands – by the Japan Geospatial Information Agency (GSI); 2) on the Peruvian-Chilean coast and in Southern California – by the Nevada Geodetic Laboratory of the University of Reno [*Blewitt et al.*, 2018]. The stations for analysis were selected from the vicinity of the focal zones of strong regional earthquakes in order to obtain the maximum possible amount of data on the action of regional geodynamic processes: on the Japanese islands stations were selected in the vicinity of the focal zone of the Tohoku earthquake, March 11, 2011, with $M_w = 9.0$; on the Peruvian-Chilean coast, in the vicinity of the focal zones of the Maule earthquake, February 27, 2010, with $M_w = 8.8$, the Iquique earthquake, April 1, 2014, with $M_w = 8.1$ and the Illapel earthquake, September 16, 2015, with $M_w = 8.3$; in Southern California, in the vicinity of the focal zone of the Ridgecrest earthquake, July 5, 2019, with $M_w = 7.1$ (Figure 1).

Algorithm for solving the regression recovery problem

In this work we construct an adequate, versatile, statistically significant and interpretable regression model to extract meaningful signal from GNSS time series in tectonically active regions. The model's adequacy is assessed based on its ability to accurately reproduce the time series in the time domain. The versatility of the model means its suitability for describing time series characterizing the total action of geodynamic processes in regions with different dynamics and tectonic structure. The statistical significance of the model is determined by the success of passing statistical tests for the significance of regression coefficients. The interpretability property of the model means the ability

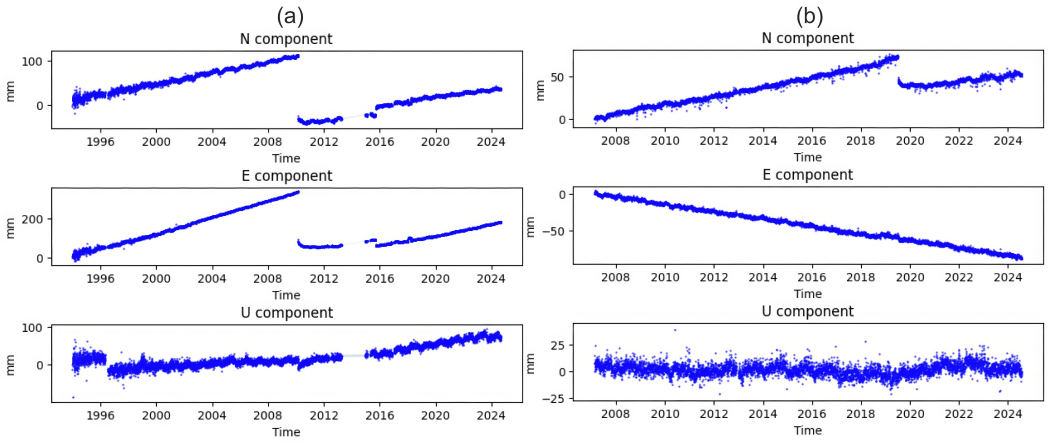

Figure 1. Examples of initial time series used in the analysis. (a) Station SANT (Santiago, Chile) near the source zone of the Maule earthquake; (b) Station P091 (Southern California) near the source zone of the Ridgecrest earthquake.

of further use of the obtained regression coefficients for solving geodynamic problems, in particular, studying recent movements and deformations of the Earth's surface and modeling geodynamic processes.

Within the framework of machine learning methodology, such a regression recovery problem is formulated as the mining of the statistical dependence M between moments of time $t_i \in T$ and the values of the time series $y_i \in Y$, which is usually specified in the form of parametric family of functions:

$$M = \{f(t,\theta)|\theta \in \Theta\}, \tag{1}$$

where $f : T \times \Theta \to Y$ is a fixed function, Θ is a set of admissible values of the parameters θ. In this formulation, the regression recovery problem is a classic supervised learning problem where the set of time moments T determines the set of objects, and the set Y determines the set of labels. The advantage of machine learning methods is the versatility of the algorithm for solving the problem, while its fine-tuning is based on the modification of three key principles: 1) the principle of generating a feature description of objects; 2) the principle of specifying a family of parametric functions that approximate the desired statistical dependence (1); 3) the principle of determining the proximity of the predictive model M and the original time series.

The feature description of objects is their formalized informational description, characterizing some property or aspect of the object:

$$p_j : T \to P_j, \qquad j = 1,\ldots,n, \tag{2}$$

where P_j is the admissible set of values of the j-th feature. In the case of a one-dimensional time series of GNSS observations, the only feature of objects is the time moments $t \in T$, and also, according to (2), any functions of the variable t.

The principle of constructing the family of functions (1) is determined by the type of the analyzed data and the requirements for the resulting model. In this paper, the requirement for interpretability of the regression model determines the need to involve a priori information to form a model characterizing the actions of real geodynamic processes. Due to the additivity of deformation, the parametric family of functions (1) is represented by functions linear with respect to their parameters, reflecting the contribution of various deformation processes to the total displacement of the Earth's surface, recorded by GNSS [*Nikolaidis*, 2002]:

$$M = \sum_{j=1}^{n} \theta_j p_j(T), \theta \in \mathbb{R}^n, \tag{3}$$

where feature vector

$$p(t) = \begin{array}{c}(1, t, \sin(2\pi t), \cos(2\pi t), \sin(4\pi t), \cos(4\pi t), H(t > T_0), \ldots, \\ H(t > T_l), H\left(t > T_0^{\text{post}}\right) \cdot \text{post}_0(t), \ldots, \left(t > T_k^{\text{post}}\right) \cdot \text{post}_k(t))\end{array}. \quad (4)$$

H is a Heaviside function [*Bracewell*, 2000], l and m are the number of instantaneous shifts at time moments $(T_0, \ldots, T_l)$ and episodes of postseismic nonlinear behavior started at time moments $\left(T_0^{\text{post}}, \ldots, T_k^{\text{post}}\right)$, respectively, $\text{post}_i(t)$ is a function, which describes the summary effect of postseismic processes and can be written in different forms [*Yamaga and Mitsui*, 2019]:

$$\text{post}(t) = \begin{cases} \ln\left(1 + \frac{\tau_a}{t}\right) - \exp\left(-\frac{t}{\tau_b}\right) & (I) \\ \ln\left(1 + \frac{\tau_a}{t}\right)\ln\left(1 + \frac{\tau_b}{t}\right) - \exp\left(\frac{t}{\tau_b}\right) & (II) \\ \ln\left(1 + \frac{\tau_c}{t}\right) - \exp\left(-\frac{t}{\tau_b}\right) - \exp\left(-\frac{t}{\tau_c}\right) & (III) \end{cases}, \quad (5)$$

where τ_a, τ_b, τ_c are attenuation constants. The model (3–5) allows us to simulate recent movements and deformations of the Earth's surface both in stable intraplate regions and in areas with high tectonic activity at all stages of the seismic cycle.

The idea of machine learning is to find an optimal set of parameters θ in a certain sense for a sample $X^m = (t_i, y_i)$, $i = 1, \ldots, m$ which determines a certain implementation of the model $a = a(X^m) \in M$ of family (3). This problem can be formalized by setting a loss function $\mathcal{L}(a, t_i)$ on each element of the data sample (t_i, y_i), which characterizes the closeness of the predicted $a(t_i)$ and the true value of the time series (label) y_i or, in other words, the error of the model a on a given element of the sample. Setting the error function allows us to construct a functional of the quality of the model a on the sample X^m, which, in the case of equivalent elements of the sample, is defined as the average value of the error over the entire sample (empirical risk functional):

$$Q(a, X^m) = \frac{1}{m}\sum_{i=1}^{m} \mathcal{L}(a, t_i). \quad (6)$$

The choice of the type of loss function and empirical risk functional depends on the initial problem formulation and the characteristics of the data. The further solution of the machine learning problem consists of finding the optimum of the empirical risk functional. When specifying the empirical risk functional in the form (6), the solution is found using the empirical risk minimization method (ERM):

$$a(X^m) = \underset{a \in M}{\text{argmin}} Q(a, X^m) \quad (7)$$

The obtained solution minimizes the functional (6), but depends significantly on the data sample on which the training took place, the selected families of functions (3) and the methods for specifying the loss function and the empirical risk functional. The common way to check the performance of a learning algorithm (7) is to use cross-validation techniques, such as Leave-one-out, K-fold, etc. The key of this approach is to split the original data set X^m into two sub-sets – training set and test set. The training set is used to build the model and adjust the hyperparameters of the algorithm. The test set is used to check the generalizing ability of the model and assess its quality. Sequential methods (such as Leave-one-out) are extremely computationally intensive for long-term GNSS time series. On the other hand, the use of batch cross-validation methods (e.g., K-fold) in the analysis of GNSS measurement time series is not possible due to the presence of short-term and instantaneous shifts in the time series. Thus, in this paper we construct training and test sets by dividing the original data set into even and odd elements. This approach allows us to evaluate the quality of the constructed model and minimize data loss in the training sample.

Machine learning methods are sensitive to the quality of the initial data and the presence of random errors that distort the statistical dependence existing in the data. Thus, solving the regression recovery problem requires a preliminary analysis of GNSS displacement time series in order to clean them, as well as collecting a priori information necessary for constructing a regression model. In tectonically active regions, GNSS displacement time series can be significantly complicated by the presence of instantaneous shifts caused by earthquakes and volcanic eruptions, long-term transient processes occurring near source zones of strong earthquakes and areas of preparation for volcanic eruptions, as well as due to changes in the stress-strain state of the Earth's crust and lithosphere. The effects of these geodynamic processes of different spatiotemporal scales cannot always be correctly separated, which requires the development and application of special methods and criteria. In addition, the analyzed time series contain outliers and seasonal variations. Incorrect consideration of all the above-described complicating factors can significantly distort the resulting estimates of the geodynamic processes' effects, which, in turn, will affect the correctness of the subsequent interpretation of the time series.

The algorithm proposed in this work for solving the regression recovery problem involves processing the original time series through two main phases, each comprising several stages (Figure 2):

- data preparation phase;
- modeling phase.

The data preparation phase is aimed at primary processing and analysis of time series data and consists of the following stages

1. determination of the trend using robust optimization algorithms in order to obtain a modified time series for subsequent analysis;
2. pre-processing of GNSS time series in order to remove outliers and fill in the gaps that arose due to a malfunction of the station or after removing outliers;
3. detecting the instantaneous shifts in time series.

An important feature of GNSS observation time series is their nonstationarity, both in terms of sample mean and sample variance. Such properties of time series do not allow us to use classic methods of statistical analysis of time series. In this regard, it is necessary to make a preliminary analysis and modification of the original time series in order to bring them to a quasi-stationary form, which will allow further use of statistical analysis methods to solve problems of cleaning the time series from statistically unlikely events (outliers), determining the times of instantaneous shifts, etc.

The first stage of such preliminary analysis consists in correct estimation of the trend, which usually makes the largest contribution to the sample variance. The problem of trend removal in GNSS time series is complicated by the presence of a seasonal component (the fundamental harmonic has a period of 1 year), outliers, instantaneous and nonlinear shifts, which requires the use of robust optimization methods for estimating the linear trend. In the proposed algorithm, a modified robust Theil-Sen algorithm presented in [*Blewitt et al.*, 2016] is used for these purposes. This approach allows obtaining a robust estimate of the linear trend and modifying the original series by removing the linear trend model calculated for each point of the series. In this case, the modified time series can be considered conditionally stationary within a short time window under the assumption of an uncorrelated error model.

In the original version of the Theil-Sen algorithm, the slope of a straight line is calculated as the median of the distribution of slopes of straight lines passing through all possible pairs of points in the time series. In a modified version, pairs of data separated by 1 year are selected to account for seasonal component, the fundamental harmonic of which has a period of 1 year [*Blewitt et al.*, 2016]. It is assumed that the measurement errors are statistically independent and follow a normal distribution. The results of numerical experiments showed that the distribution of the obtained slope estimates can differ significantly from the normal one (in particular, multimodal distributions (Figure 3b)

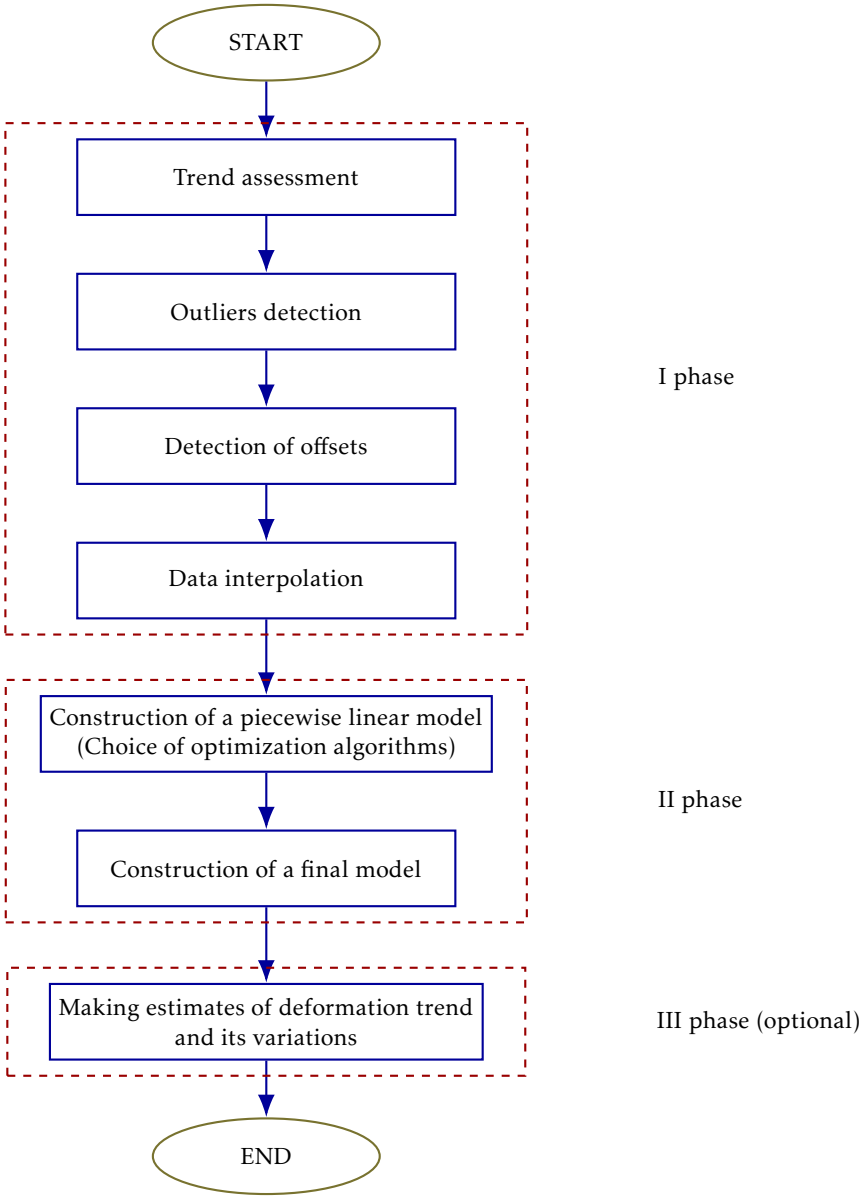

Figure 2. The algorithm operating scheme.

associated with possible trend changes, heavy-tailed distributions (Figure 3c) caused by unlikely trend values obtained due to the presence of seasonality, instantaneous shifts and nonlinear shifts in the time series). In order to reduce the error in determining the median trend, we modified the algorithm described in [*Blewitt et al.*, 2016] by determining the trend based on the maximum mode estimate calculated from the histogram, and by discarding slope values that are outside 2 standard deviations from the median (Figure 3). The median absolute deviation (MAD) estimate [*Blewitt et al.*, 2016], calculated based on the median, was used as the standard deviation.

In the next stage of the preprocessing phase of the algorithm, we clean the modified detrended time series to exclude outliers (Figure 4). In the following, we define an outlier exclusively as an undesired point anomaly resulting from random measurement errors [*Blazquez-Garciá et al.*, 2021]. Thus, cleaning outliers from a time series is a crucial preprocessing step since measurement errors can seriously distort model fitting. At the same time, a thorough analysis of the detected outliers must be conducted to ensure that meaningful signal components are not mistakenly removed. We evaluated several widely used algorithms for outlier detection, including classical sliding window methods such

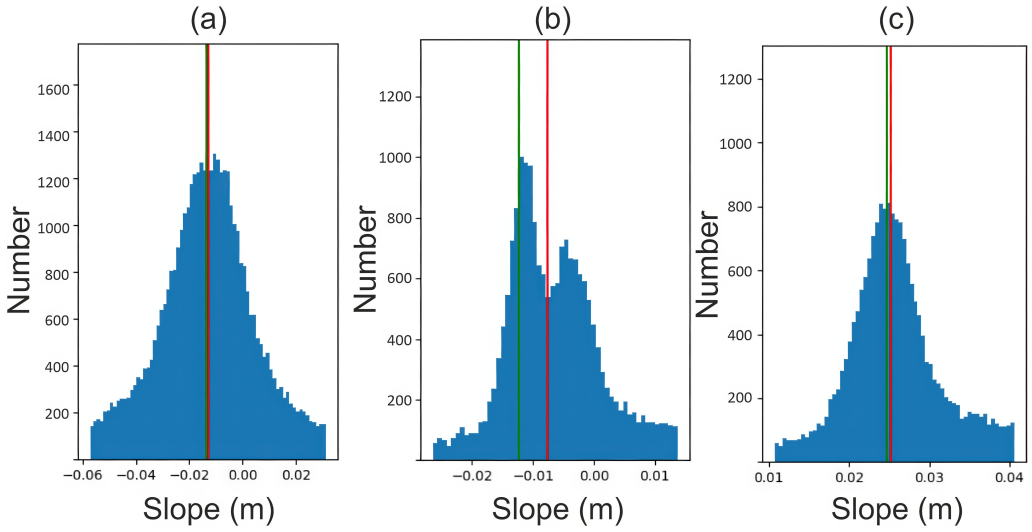

Figure 3. Possible type of histograms for calculated slopes. (a) – quasi-normal distribution (median (red line) and mode (green line) coincide), (b) – bi-modal distribution (mode shifted to the left related to the median), (c) – heavy-tailed distribution (mode slightly shifted to the left related to the median).

as Z-score and Interquartile Range (IQR), as well as modern machine learning-based approaches, specifically Local Outlier Factor (LOF) [*Alghushairy et al.*, 2020] and Isolation Forest [*Liu et al.*, 2008].

The IQR and Z-score methods operate by comparing each point in the time series to a local statistical threshold, computed within a sliding window centered at the point of interest. Specifically, the IQR method uses the interquartile range, while the Z-score method relies on the sample mean (μ) and standard deviation (σ):

$$\text{outlier}(Z\text{-score}) = \left| \frac{y_i^{\text{detrend}} - \mu}{\sigma} \right| > 3.$$

The interquartile range is the difference between the estimates of the first quartile ($Q1$) and the third quartile ($Q3$) of the distribution. In this case, the values of the time series that are outside $1.5 \cdot \text{IQR}$ from $Q1$ to the left and $Q3$ to the right are defined as outliers:

$$\text{outlier}(\text{IQR}) = \begin{cases} y_i^{\text{detrend}} < Q1 - 1.5 \cdot \text{IQR} \\ y_i^{\text{detrend}} > Q3 + 1.5 \cdot \text{IQR} \end{cases}.$$

The size of the sliding window for IQR and Z-score is chosen to be small enough (namely, 30 days) to consider the time series segment as quasi-stationary, and the measurement errors as conditionally independent and normally distributed.

The results of applying the outlier detection algorithms are shown in Table 1. Analysis of the detected outliers shows that their removal improves the performance metric of the resulting model. At the same time, classical methods are more reliable, since the tested machine learning methods, if the hyperparameters are chosen incorrectly, can lead to the loss of a significant amount of initial data.

One of the most challenging aspects of GNSS time series processing is the identification of instantaneous shifts caused mainly by seismic and volcanic processes or changes in GNSS station equipment. The solution to the problem of determining the moments of coseismic displacements is necessary for further modeling of postseismic processes and construction of the final regression model. Displacements in time series of coordinates are defined as an instantaneous change in the sample mean, leading to a long-term effect on the estimated parameters. Depending on their presence in the time series, unmodeled displacements can seriously affect the coefficients of the resulting model.

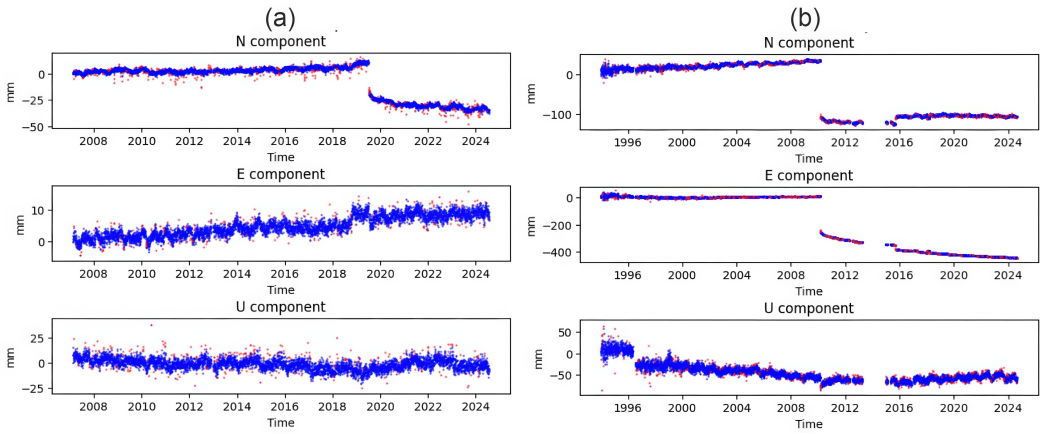

Figure 4. An example of cleaning outlier procedure results for P091 GNSS station (Southern California) (a) and SANT GNSS station (Santiago, Chile) (b). Blue dotes denote daily estimates for displacements, red dots denote identified outliers.

Table 1. Comparative analysis of the performance of various outlier detection algorithms and the impact of outlier cleaning procedures on model fitting for J001 time series (Japanese islands)

Method	Window size, days	Number of outliers			Performance metric, mm			Coefficient of determination, %			Determined offsets		
		N	*E*	*U*	*N*	*E*	*U*	*N*	*E*	*U*	*N*	*E*	*U*
IQR	30	482	445	545	1.78	2.58	5.51	99.44	99.98	79.74	1997.2041 2011.1932 2012.9631	1997.2232 1998.8918 2011.1932 2012.2746	2003.1000 2011.1932 2012.9631
Z-score	30	482	445	545	1.78	2.58	5.51	99.44	99.98	79.74	1997.2041 2011.1932 2012.9631	1997.2233 1998.8918 2011.1932 2012.2746	2003.1000 2011.1932 2012.9631
Isolation forest	1000	1968	1838	884	1.82	2.29	4.48	99.02	99.98	80.4	1997.2014 2013.6700 2015.4343	1997.2096 1999.6370 2013.6699	2003.1027 2013.6700
Local Outlier Factor	500	368	769	880	1.72	4.08	4.58	99.21	99.95	87.30	1999.3712 2011.1904 2012.9631	1999.3712 2011.1904	1999.3767 2000.6243 2003.1027 2011.1904 2012.9631
Raw data	–	–	–	–	2.44	5.11	7.09	99.22	99.94	78.23	1997.2014 2011.1904 2012.9631	1997.2095 2011.1904	2003.1027 2011.1904 2012.1762 2012.9631

The algorithm proposed in the article allows us to use any methods originally developed for analyzing data of various natures, such as financial, meteorological, sociological, etc., to solve this complex problem [*Crocetti et al.*, 2021; *Londschien et al.*, 2023; *Truong et al.*, 2020]. The analysis of specialized methods for automated offset detection in synthetic GNSS time series was carried out as part of the Detection of Offset in GPS Experiment (DOGEx) [*Gazeaux et al.*, 2013]. During our experiments, we tested several methods: the sequential JPL_STP1 analysis method, which showed good results on synthetic tests [*Gazeaux et al.*, 2013]; the changeForest method [*Londschien et al.*, 2023], based on the use of Random Forest technology, and several methods implementing the Change Point Detection (CPD) approach [*Truong et al.*, 2020].

The JPL_STP1 method is a sequential method in which each point in the series is tested for possible shift. Identification of shift is based on constructing a linear regression using data from two sliding windows before and after the time point under study. The potential

shift is estimated as the difference between the absolute terms of the two constructed regression models. To improve the robustness of the regression, a series of time window sizes (from 10 days to 90 days) is considered, and the median of all the resulting shift estimates is used as the final estimate of shift. The significance of the resulting shift was tested using the F-test. We also modified the JPL_STP1 method by changing the optimization method from the non-robust Least Squares to the more robust Theil-Sen method. We conducted a series of experiments using the modified JPL_STP1 method on various time series and found that it has a fairly high computational complexity due to the sequential checking of all points in the series, as well as a fairly large number of "False Positive" detections due to wrong estimates of linear regression coefficients on points surrounding the real shift.

The changeForest algorithm [*Londschien et al.*, 2023] is based on the use of Random Forest to find time series segments with different sample statistical characteristics, while on each segment, the data are assumed to be distributed equally. As a result of applying the method, a tree-like structure is obtained by the division of the original series into segments. Numerical experiments have shown the method produces an excessive number of "False Positive" detections when applied to the analysis of GNSS time series (Figure 5, left panel).

The CPD approach is a general approach to finding changes in statistical quantities (mean, variance, etc.) in time series. The essence of the method is to find the minimum of the functional characterizing the total misfit between the model and the initial data [*Truong et al.*, 2020]:

$$\min_{\mathcal{T}} V(\mathcal{T}) + \mathrm{pen}(\mathcal{T}), \tag{8}$$

where $\mathcal{T}$ is the best possible segmentation, $V(\mathcal{T}) = \sum_{k=0}^{K} c(y_{t_k..t_{k+1}})$ is the summary of cost functions $c(y)$ defined on all the sub-segments $y_{t_k..t_{k+1}}$ of the initial time series, $\mathrm{pen}(\mathcal{T})$ is the penalty function introduced to avoid large number of shifts.

We tested several CPD algorithms solving (8) and choose the fast approximate sliding-window method Win [*Truong et al.*, 2020]. The advantages of this method are its very low computational cost and high possibilities of fine-tuning via variation of hyperparameters which allowed us to get better shift detections for very different time series used in the experiment (Figure 5, right panel).

To ensure that the chosen CPD algorithm can detect the actual displacements, we compared the moments and magnitudes of coseismic displacements for the time series of the P091 GNSS station located near the epicentral zone of the 2019 Ridgecrest earthquake. We compared the data provided by the leading GNSS data analysis centers: SOPAC – Scripps Orbit and Permanent Array Center and NGL – Nevada Geodetic Laboratory, and the results of two our models (Table 2). All models including the selected CPD model, correctly detected the time moment of coseismic slip caused by the 2019 Ridgecrest earthquake. We calculated direct estimates of coseismic shift as the difference between linear models built from 10-day pre-earthquake and 10-day postearthquake time series segments. The obtained displacement magnitudes are close to the obtained direct estimates ($N_{\mathrm{slip}} = -25.24\,\mathrm{mm}$, $E_{\mathrm{slip}} = -2.53\,\mathrm{mm}$, $U_{\mathrm{slip}} = 2.22\,\mathrm{mm}$), and the difference in estimates is due to different postseismic displacement models used.

The instantaneous shifts obtained at this stage of the preprocessing phase allow us to determine the feature vector (4). It is known that significant postseismic displacements are observed only during large earthquakes and within only a few hundred kilometers from the earthquake source, that is why we used the magnitude of the observed shift as a threshold value when determining the moments of the onset of postseismic processes. In particular, postseismic processes were modeled only for those displacements that exceeded the threshold of 10 standard deviations.

The final optional stage of the time series pre-processing phase is to interpolate the time series to fill in any gaps in the data that exist or have arisen due to the outlier cleaning procedure, using the monotonic piecewise cubic interpolation algorithm [*Fritsch and Carlson*, 1980].

Table 2. Comparison of times and magnitudes of coseismic shifts obtained using different algorithms (N – North, E – East, U – Vertical components of time series)

Coseismic offset source	NGL offsets, mm			SOPAC offsets, mm			[*Gabsatarov*, 2012] offsets, mm			This work offsets, mm		
	N	*E*	*U*	*N*	*E*	*U*	*N*	*E*	*U*	*N*	*E*	*U*
2019 Ridgecrest earthquake $M_w = 7.1$ [USGS] 2019-07-06 03:19:53 (UTC)	−29.11	−4.46	3.91	−25.72	−3.07	5.32	−29.62	1.51	3.85	−35.25	−0.1	–
Determined offset date	2019.5099 [2019-07-06]			2019.5110 [2019-07-06]			2019.5110 [2019-07-06]			2019.5096 [2019-07-06]		

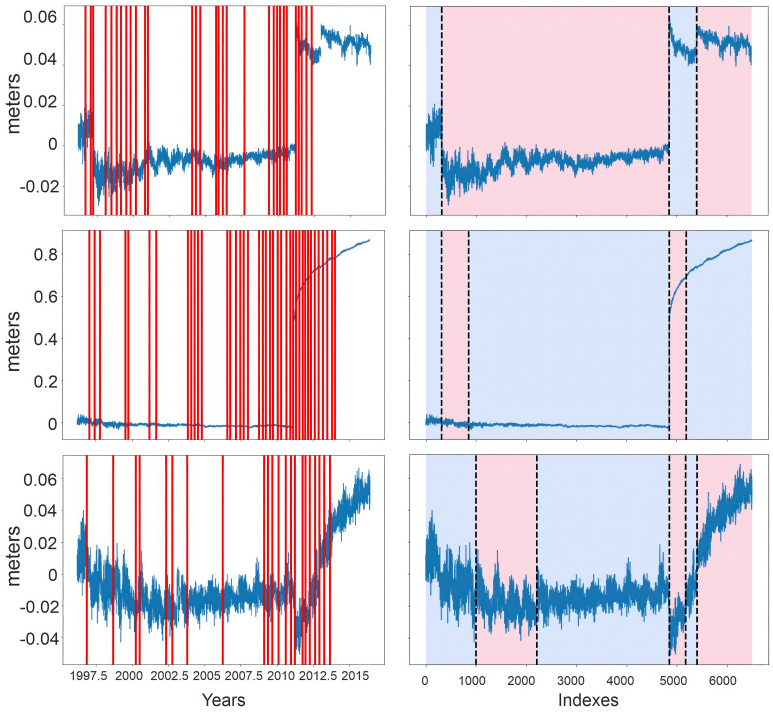

Figure 5. (Left panel) An example of instantaneous shift detection using `changeForest` algorithm [*Londschien et al.*, 2023] for J001 GNSS station (Japanese islands). (Right panel) An example of CPD using `Win` algorithm [*Truong et al.*, 2020] for J001 GNSS station (Japanese islands). From up to down: North, East and U component of GNSS time series. Red lines for left panel and black dashed lines for right panel denote moments of slips. Colors for right panel indicate the stable parts of the time series.

The second phase of time series processing consists of directly constructing the regression model (3–5). This phase is divided into two stages: 1) constructing a piecewise linear model without taking into account the effect of postseismic processes (Figure 6a); 2) constructing a complete time series model using the feature vector generated based on the results of the previous modeling stage (Figure 6c). The first stage is necessary to select the best optimization method and determine the onset of postseismic processes. Modeling is performed for the feature vector (4) with removed components (5) responsible for modeling postseismic processes. The problem of selecting the best optimization method is solved by performing calculations with various optimization methods and selecting the best one based on the maximum of the determination coefficient (Figure 6b). The following are considered as candidate methods: a) Least Squares method – a classical non-robust optimization method, in which the solution is defined

as $a(T,Y) = \min_{\theta\in\Theta}\sum_{i=1}^{m}(p(t_i)\theta - y_i)^2$; b) Lasso method is an optimization method that works well with models with a large number of zero coefficients, the solution is specified as $a(T,Y) = \min_{\theta\in\Theta}\left[\frac{1}{2m}\sum_{i=1}^{m}(p(t_i)\theta - y_i)^2 + \sum_{j=1}^{n}\theta_j\right]$; c) Bayesian Ridge method, which is more robust than Least squares due to defining additional prior distribution for θ, the solution is given in the form $a(T,Y) = \min_{\theta\in\Theta}\left[\frac{1}{2m}\sum_{i=1}^{m}(p(t_i)\theta - y_i)^2 + \sum_{j=1}^{n}\theta_j^2\right]$; d) The Theil-Sen method is a robust optimization method based on the estimation of the median of the distribution of model and initial data residuals. The determining of the onset of postseismic processes is performed using the threshold method described above for the resulting estimates of shifts.

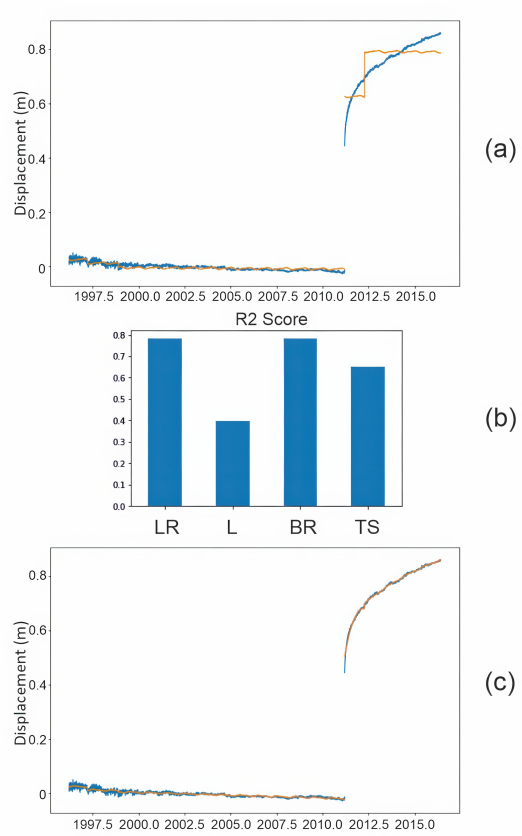

Figure 6. Results of regression modeling for J001 GNSS station (Japanese islands) (a) piecewise linear model (blue line is a time series, orange line denote regression model); (b) The estimates of coefficient of determination for different optimization algorithms (LR – Least Squares, L – Lasso, BR – Bayesian Ridge, TS – Thiel-Sen); (c) resulting regression model accounting for postseismic process after 2011 Tohoku earthquake.

The construction of a complete nonlinear model is based on an iterative approach in which the attenuation constants of postseismic processes (5) are considered as a hyper-parameter and vary from 1 to 300 days. The best model is determined by the minimum of the mean absolute error for all models maximizing the coefficient of determination. This approach is quite similar to that used in the model [*Gabsatarov*, 2012]. The statistical significance of the obtained regression models was tested using the standard F-test.

During the third phase of the algorithm operation, the residuals obtained during the construction of the final regression model are used to study the rate of accumulation of elastic deformations and their variations. The initial data for modeling the variations in the process of deformation accumulation are obtained by analyzing the time series of residuals of the final model in a 1 year-long sliding window, modified by adding to each of its points a deformation trend model calculated as the difference between the estimated final trend

and its modeled value, based on the MORVEL plate tectonic model [*Argus et al.*, 2011]. The deformation trend in each time window is estimated using the modification of the Theil-Sen algorithm proposed in the work, taking into account all possible pairs of points.

The proposed algorithm for analyzing GNSS time series is based on a combination of the machine learning approach described above and classical methods of statistical analysis of time series. This approach allows us to significantly reduce the amount of a priori geological and geophysical data used and completely avoid the direct modeling of the action of geodynamic processes, which accelerates the analysis of time series by reducing the computational complexity of algorithms, and will allow further the designing of fully automated systems for analyzing GNSS data.

Results

In order to test the proposed algorithm, we created its software implementation in Python using open-source modules (`NumPy`, `SciPy`, `Scikit-learn`, `Pandas`, `Matplotlib`, `Ruptures`). The architecture of the created software is a modular scheme, which allows us to easily improve and replace individual blocks of the algorithm. Using an object-oriented approach also simplifies the support and further improvement of software by combining data and processing methods in one data structure. Further, these software features will help us to expand the capabilities of the algorithm for interpreting data, increase the accuracy of estimates and reduce the time to build a model.

The algorithm presented in the work is the first stage of creating an automated system for statistical analysis of GNSS time series. In this regard, our task was to create a fully operational version of the algorithm and test it on the most diverse data sets in order to obtain data on the limits of applicability of the algorithm, determine its shortcomings and outline promising areas of development. The algorithm was tested on a cloud server (CPU: 2×Xeon E5-2670 2.6 GHz (8 cores) RAM: 32 GB). Results of applying the algorithm for several time series in the stable part of the lithospheric plate (ARTU station) and in different tectonically active regions (subduction zones: PETS and J001 stations, shear zone: P091 station) are shown in Table 3.

The piecewise linear model is quite good for analyzing data from stable inner plate regions but absence of postseismic features in feature vector for piecewise model can cause mismodeling for short-term and small nonlinear effects (Figure 8b) and even ruin the solution for prominent nonlinear postseismic motions (Figure 7b). The complete model demonstrated adequacy for approximately 85% of the tested time series, which includes observations from various tectonically active regions, such as the immediate vicinity of megathrust zones (Figure 7) and significant strike-slip events (Figure 8). Model adequacy is defined here as the ability to accurately reproduce the GNSS time series based on established machine learning performance metrics. Specifically, this is characterized by a low sample variance of the residuals—reflected by a mean absolute error of approximately 2 mm to 3 mm for the horizontal components and about 5 mm for the vertical component—as well as high values of the coefficient of determination, averaging around 0.95 for horizontal components and 0.85 for the vertical component (Figures 7d and 8d). In all studied cases the constructed regression models were found to be statistically significant.

The obtained results (Table 3) indicate that even at this early stage of development, the constructed regression model achieves performance that is comparable to or exceeds that of existing classical regression analysis models [*Blewitt et al.*, 2016; *Bock et al.*, 2023; *Gabsatarov*, 2012] in terms of the performance metric. Testing further revealed that, in the absence of postseismic effects in the time series—modeling of which remains computationally intensive—the new algorithm operates significantly faster than its predecessor [*Gabsatarov*, 2012]. This enhanced efficiency is primarily attributed to the utilization of vectorized computations and the elimination of the need to gather a priori information or perform direct calculations of coseismic displacements.

However, we found that despite the high values of the coefficient of determination and small values of mean absolute error, the final model didn't account for all the deformation effects. The complete model residuals (Figures 7d–8d) show prominent long-term nonlinear effects possibly caused by slow-slip events.

Table 3. Comparison of the results of different regression analysis algorithms. (*N* – North, *E* – East, *U* – Vertical components of time series)

Station name	Location	Model	Program running time	Number of instantaneous shifts			Number of modeled postseismic processes			Performance metric, mm			Length of time series, years
				N	E	U	N	E	U	N	E	U	
PETS	Kamchatka Peninsula, Kuril-Kamchatka subduction zone	[*Gabsatarov*, 2012]	6 min 30 s	1	1	1	1	1	1	9.55	13.38	14.13	25.15
		This work	9 min 05 s	7	11	6	1	1	0	1.05	1.52	3.87	
		SOPAC	–	3	3	3	2	2	2	1.91	3.44	5.39	17.93
		NGL	–	33	33	33	7	7	0	1.81	1.76	5.95	17.93
ARTU	Ural mountains, stable part of the Eurasian lithospheric plate	[*Gabsatarov*, 2012]	3 min 50 s	0	0	0	0	0	0	1.22	3.42	5.67	22.81
		This work	2 min 30 s	4	3	0	0	0	0	0.99	0.94	4.58	
		SOPAC	–	0	0	0	0	0	0	1.41	1.50	6.59	22.81
		NGL	–	1	1	1	0	0	0	1.55	1.58	7.27	22.81
I001	Japanese islands, Japan subduction zone	[*Gabsatarov*, 2012]	6 min 38 s	1	1	1	1	1	1	3.25	27.58	8.18	15.72
		This work	53 min 16 s	7	6	5	2	2	0	1.81	3.33	5.94	20.18
		SOPAC	–	–	–	–	–	–	–	–	–	–	–
		NGL	–	40	40	40	7	7	7	4.45	1.89	5.35	15.72
P091	California, Eastern California Shear Zone	[*Gabsatarov*, 2012]	4 min 17 s	1	1	1	1	1	1	1.71	1.31	4.31	17.52
		This work	3 min 01 s	7	5	1	1	0	0	0.96	1.00	3.18	
		SOPAC	–	2	2	1	1	1	0	1.82	1.18	4.22	17.96
		NGL	–	7	7	7	–	1	1	1.53	1.88	4.75	17.52

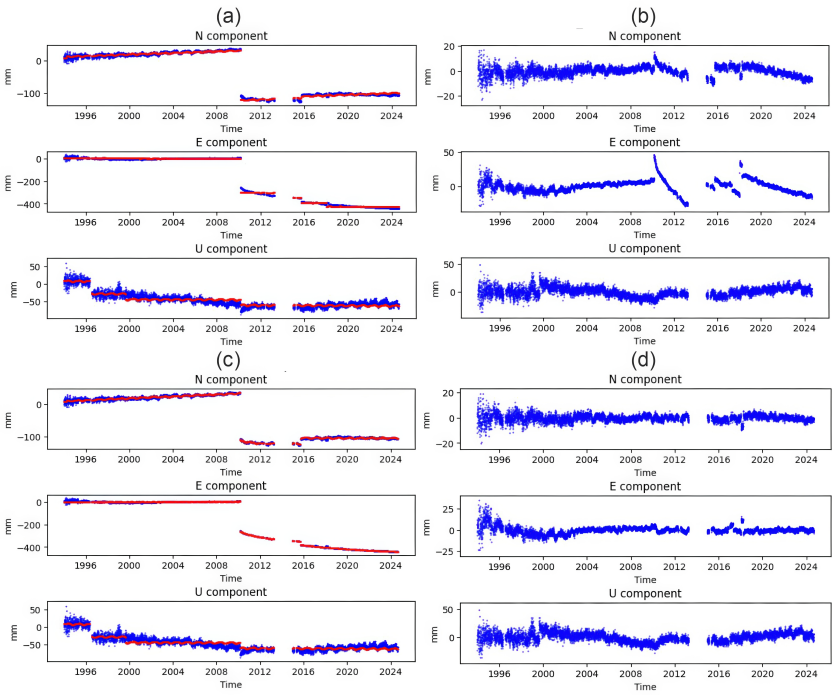

Figure 7. An example of regression modeling for SANT GNSS station (Santiago, Chile). Blue dotes denote daily estimates for displacements, red line denotes constructed regression model. (a) – piecewise linear regression model, (b) – residuals of piecewise linear regression model, (c) – final regression model, (d) – residuals of final regression model.

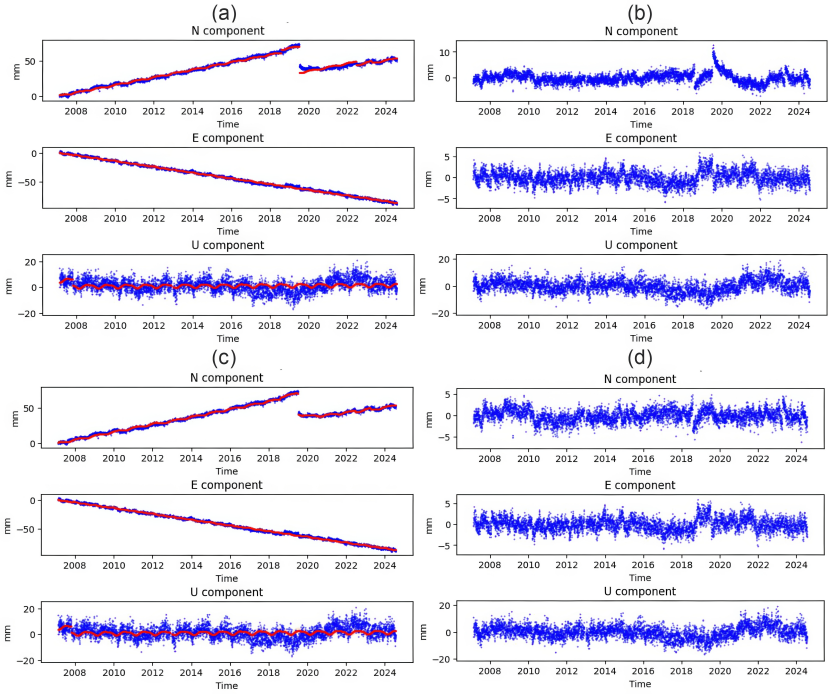

Figure 8. An example of regression modeling for P091 GNSS station (Southern California). The designations in the figure are similar to Figure 6.

The disadvantages of the presented algorithm revealed during the tests include the excessively simple CPD method, which does not cope well enough with the determination of instantaneous shifts in the case of long and intensive nonlinear processes, a rough method for selecting the onsets of postseismic processes, the need to further reduce the algorithm's operating time for long time series and more complex models of postseismic processes. Due to the modular structure of the algorithm and the software implementing it, these problems can be overcome by using more complex models taking into account shifts and postseismic effects such as application of a recurrent neural network to extract nonlinear transient displacements [*Xue and Freymueller*, 2023] and mark the onset of postseismic processes, new neural network-based methods to extract instantaneous shifts [*Ozbey et al.*, 2024].

Conclusion

In this paper, we present the theoretical foundations of the algorithm for processing GNSS time series using machine learning methods in solving the problem of regression recovery. This algorithm allows us to solve the problem of constructing a linear model of crustal deformation at the location of GNSS station without using a priori information and direct modeling.

The versatility and adequacy of the algorithm were tested by analyzing time series of GNSS stations located in various tectonically active regions (the Japanese islands, Northern California, the coast of Peru-Chile) both near the sources of large earthquakes with different types of mechanisms (thrust and strike-slip) and magnitudes from $M_w = 7.1$ to $M_w = 9.0$ and at distances of up to several hundred kilometers across and along the azimuths of the strike of the main rupture plane. According to the results of statistical tests for the significance of the regression coefficients all the resulting regression models were recognized as statistically significant.

In addition, we created a software implementation of the presented algorithm, allowing to obtain initial data for constructing models of geodynamic processes and studying variations in the field of recent movements and deformations of the Earth's surface. The obtained variations of the studied fields can be used for a feature description of points on

the Earth's surface when solving the clustering problem in order to identify stable domains in the recent crustal movements field and its spatiotemporal variations.

The presented results will allow creating a theoretical and practical basis for a versatile tool for analyzing variations in the fields of recent movements and deformations of the Earth's surface in order to identify a regional fault-block structure and localize areas of increased geodynamic hazard based on the use of new data analysis methods.

Acknowledgments. The study was supported by a grant from the Russian Science Foundation No. 24-27-00176, https://rscf.ru/en/project/24-27-00176/.

References

Alghushairy O., Alsini R., Soule T., et al. A Review of Local Outlier Factor Algorithms for Outlier Detection in Big Data Streams // Big Data and Cognitive Computing. — 2020. — Vol. 5, no. 1. — https://doi.org/10.3390/bdcc5010001.

Altamimi Z., Rebischung P., Collilieux X., et al. ITRF2020: an augmented reference frame refining the modeling of nonlinear station motions // Journal of Geodesy. — 2023. — Vol. 97, no. 5. — https://doi.org/10.1007/s00190-023-01738-w.

Altamimi Z., Rebischung P., Metivier L., et al. ITRF2014: A new release of the International Terrestrial Reference Frame modeling nonlinear station motions // Journal of Geophysical Research: Solid Earth. — 2016. — Vol. 121, no. 8. — P. 6109–6131. — https://doi.org/10.1002/2016jb013098.

Argus D. F., Gordon R. G. and DeMets C. Geologically current motion of 56 plates relative to the no-net-rotation reference frame: NNR-MORVEL56 // Geochemistry, Geophysics, Geosystems. — 2011. — Vol. 12, no. 11. — P. 1–13. — https://doi.org/10.1029/2011gc003751.

Blazquez-Garciá A., Conde A., Mori U., et al. A Review on Outlier/Anomaly Detection in Time Series Data // ACM Computing Surveys. — 2021. — Vol. 54, no. 3. — P. 1–33. — https://doi.org/10.1145/3444690.

Blewitt G., Hammond W. C. and Kreemer C. Harnessing the GPS Data Explosion for Interdisciplinary Science // Eos. — 2018. — Vol. 99. — https://doi.org/10.1029/2018eo104623.

Blewitt G., Kreemer C., Hammond W. C., et al. MIDAS robust trend estimator for accurate GPS station velocities without step detection // Journal of Geophysical Research: Solid Earth. — 2016. — Vol. 121, no. 3. — P. 2054–2068. — https://doi.org/10.1002/2015jb012552.

Bock Y., Moore A. W., Argus D., et al. Extended Solid Earth Science ESDR System (ES3): Algorithm Theoretical Basis Document, NASA MEaSUREs Project. — SOPAC/CSRC Archive, 2023.

Bracewell R. Heaviside's Unit Step Function, H(x) // The Fourier Transform and Its Applications. — New York : McGraw-Hill, 2000. — P. 61–65.

Crocetti L., Schartner M. and Soja B. Discontinuity Detection in GNSS Station Coordinate Time Series Using Machine Learning // Remote Sensing. — 2021. — Vol. 13, no. 19. — P. 3906. — https://doi.org/10.3390/rs13193906.

Fritsch F. N. and Carlson R. E. Monotone Piecewise Cubic Interpolation // SIAM Journal on Numerical Analysis. — 1980. — Vol. 17, no. 2. — P. 238–246.

Gabsatarov Yu. V. Analysis of deformation processes in the lithosphere from geodetic measurements based on the example of the San Andreas fault // Geodynamics & Tectonophysics. — 2012. — Vol. 3, no. 3. — P. 275–287. — https://doi.org/10.5800/gt-2012-3-3-0074.

Gazeaux J., Williams S., King M., et al. Detecting offsets in GPS time series: First results from the detection of offsets in GPS experiment // Journal of Geophysical Research: Solid Earth. — 2013. — Vol. 118, no. 5. — P. 2397–2407. — https://doi.org/10.1002/jgrb.50152.

Gitis V., Derendyaev A. and Petrov K. Analyzing the Performance of GPS Data for Earthquake Prediction // Remote Sensing. — 2021. — Vol. 13, no. 9. — P. 1842. — https://doi.org/10.3390/rs13091842.

Gvishiani A. D., Dobrovolsky M. N., Dzeranov B. V., et al. Big Data in Geophysics and Other Earth Sciences // Izvestiya, Physics of the Solid Earth. — 2022. — Vol. 58. — P. 1–29. — https://doi.org/10.1134/s1069351322010037.

Ji K., Shen Y. and Wang F. Signal Extraction from GNSS Position Time Series Using Weighted Wavelet Analysis // Remote Sensing. — 2020. — Vol. 12, no. 6. — P. 992. — https://doi.org/10.3390/rs12060992.

Liu F. T., Ting K. M. and Zhou Z.-H. Isolation Forest // 2008 Eighth IEEE International Conference on Data Mining. — Pisa, Italy : IEEE, 2008. — P. 413–422. — https://doi.org/10.1109/icdm.2008.17.

Liu T. and Kossobokov V. G. Displacements Before and After Great Earthquakes: Geodetic and Seismic Viewpoints // Pure and Applied Geophysics. — 2021. — Vol. 178, no. 4. — P. 1135–1155. — https://doi.org/10.1007/s00024-021-02694-2.

Londschien M., Bühlmann P. and Kovács S. Random Forests for Change Point Detection // Journal of Machine Learning Research. — 2023. — Vol. 24. — P. 1–45.

Nikolaidis R. Observation of Geodetic and Seismic Deformation with the Global Positioning System: Ph.D. Thesis. — San Diego : University of California, 2002. — 265 p.

Ozbey V., Ergintav S. and Tari E. GNSS Time Series Analysis with Machine Learning Algorithms: A Case Study for Anatolia // Remote Sensing. — 2024. — Vol. 16, no. 17. — P. 3309. — https://doi.org/10.3390/rs16173309.

Steblov G. M., Kogan M. G., Levin B. V., et al. Spatially linked asperities of the 2006-2007 great Kuril earthquakes revealed by GPS // Geophysical Research Letters. — 2008. — Vol. 35, no. 22. — https://doi.org/10.1029/2008gl035572.

Truong C., Oudre L. and Vayatis N. Selective review of offline change point detection methods // Signal Processing. — 2020. — Vol. 167. — P. 107299. — https://doi.org/10.1016/j.sigpro.2019.107299.

Xue X. and Freymueller J. T. Machine Learning for Single-Station Detection of Transient Deformation in GPS Time Series With a Case Study of Cascadia Slow Slip // Journal of Geophysical Research: Solid Earth. — 2023. — Vol. 128, no. 2. — https://doi.org/10.1029/2022jb024859.

Yamaga N. and Mitsui Y. Machine Learning Approach to Characterize the Postseismic Deformation of the 2011 Tohoku-Oki Earthquake Based on Recurrent Neural Network // Geophysical Research Letters. — 2019. — Vol. 46, no. 21. — P. 11886–11892. — https://doi.org/10.1029/2019gl084578.

Zhang S., Gong L., Zeng Q., et al. Imputation of GPS Coordinate Time Series Using missForest // Remote Sensing. — 2021. — Vol. 13, no. 12. — P. 2312. — https://doi.org/10.3390/rs13122312.