Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Gaussian Process and PCA-Based Methods for GNSS Data Gap Filling

THI DUNG VU

dung.vu@ingv.it

National Institute of Geophysics and Volcanology

Adriano Gualandi

University of Cambridge

Francesco Pintori

National Institute of Geophysics and Volcanology

Enrico Serpelloni

National Institute of Geophysics and Volcanology

Massimo Nespoli

University of Bologna

Giuseppe Pezzo

National Institute of Geophysics and Volcanology

Research Article

Keywords: GNSS Time Series, Gap Filling, Gaussian Process Regression, Principal Component Analysis, Noise Modelling, Uncertainty Quantification

Posted Date: March 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9091203/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Gaussian Process and PCA-Based Methods for GNSS Data Gap Filling

Dung Thi Vu[1,2*], Adriano Gualandi[3,4], Francesco Pintori[1], Enrico Serpelloni[1], Massimo Nespoli[2], Giuseppe Pezzo[4]

[1]Istituto Nazionale di Geofisica e Vulcanologia, Sezione di Bologna, Viale Carlo Berti Pichat 6/2, Bologna, 40127, BO, Italy.
[2]Department of Physics and Astronomy, University of Bologna, Viale Carlo Berti Pichat 6/2, Bologna, 40127, BO, Italy.
[3]Department of Earth Sciences, Bullard Laboratories, University of Cambridge, Madingley Rise, Madingley Road, Cambridge, CB30EZ, CB, United Kingdom.
[4]Istituto Nazionale di Geofisica e Vulcanologia, Osservatorio Nazionale Terremoti, Via di Vigna Murata 605, Rome, 00143, RM, Italy.

*Corresponding author(s). E-mail(s): dung.vu@ingv.it;
Contributing authors: ag2347@cam.ac.uk; francesco.pintori@ingv.it; enrico.serpelloni@ingv.it; massimo.nespoli2@unibo.it; giuseppe.pezzo@ingv.it;

Abstract

GNSS position time series are essential for monitoring crustal deformation but often contain data gaps due to equipment failures, maintenance, and quality control. Accurate gap filling is vital for reliable velocity estimation, noise characterisation, and tectonic interpretation. This study compares three methodologies for gap imputation: a novel hierarchical Gaussian process regression (GPR) approach using composite kernels to capture multiple temporal scales, weighted principal component analysis (WPCA), and expectation-maximisation PCA (EMPCA). Our GPR framework employs physically interpretable kernels to represent distinct processes, including long-term trends (radial basis functions), short-term correlations (Matérn), seasonal cycles (periodic), and measurement/coloured noise (i.e., white and flicker), alongside station-specific noise modelling and parallel processing for efficiency. Evaluation using synthetic gaps and field data from 70 stations on Italy's North Adriatic coasts shows that GPR achieves superior accuracy, significantly reducing reconstruction error for

irregular gaps while providing robust uncertainty estimates. PCA-based methods, though faster, introduce more spurious correlations and are less reliable for station-specific signals. Based on these findings, we provide clear guidelines for method selection depending on gap duration, network density, and computational constraints.

Keywords: GNSS Time Series, Gap Filling, Gaussian Process Regression, Principal Component Analysis, Noise Modelling, Uncertainty Quantification.

1 Introduction

Global Navigation Satellite Systems (GNSS) have revolutionised our ability to monitor crustal deformation at the millimetre scale, providing unprecedented temporal resolution [1]. In the North Adriatic region of Italy, GNSS time series have been increasingly used to measure surface deformation associated with subsurface fluid extraction and gas storage operations [2–4], where the extraction and injection of fluids can induce measurable displacement [5]. This study focuses on a network of 70 continuous GNSS stations installed there to measure deformation linked to a nearby gas storage field [6]. The resulting data are vital for operational safety, reservoir management, and environmental compliance.

Such anthropogenic activities can significantly impact the Earth's crust, potentially inducing seismicity and altering local stress and strain fields [7, 8]. Extracting gas or oil reduces reservoir pressure, leading to ground compaction and subsidence [9, 10]. In coastal areas, this may result in seawater intrusion, exacerbated flooding, and increased storm impacts [11, 12]. The GNSS time series used to monitor these effects, however, frequently contain data gaps due to equipment failure, scheduled maintenance, power outages, and quality control procedures that remove unreliable measurements [13]. These interruptions make it difficult to model deformation signals accurately, yet continuous observations are essential for applying advanced time series techniques that can reveal subtle changes related to reservoir behaviour and subsurface processes.

A further challenge arises with the growing adoption of machine learning (ML) techniques in geodesy, as these typically require complete, high-quality datasets to function effectively [14, 15]. Missing data disrupts the inherent structure of the signal and noise on which analytical models rely. Conventional gap-filling methods, such as linear interpolation or spline fitting, often fail to capture the complex temporal correlations and noise characteristics of GNSS data [16], introducing spurious trends and biasing key products such as velocity estimates. With the expansion of ML applications, robust gap-filling methods that preserve statistical properties are paramount, because artificial correlations introduced by poor imputation can degrade model performance and generalisation capability [17].

This limitation has motivated the development of more advanced approaches to data reconstruction. Parametric methods (e.g., least-squares fitting with heuristic functions [18] or Kalman filtering [19]) can be limited by predefined functional forms

that may not adapt well to non-stationary industrial signals. Frequency-domain techniques such as singular spectrum analysis [20] and wavelet methods [21] can struggle with irregularly sampled data and introduce problematic edge effects. Recently, principal component analysis (PCA) and its variants have gained popularity for GNSS data reconstruction [22–24].

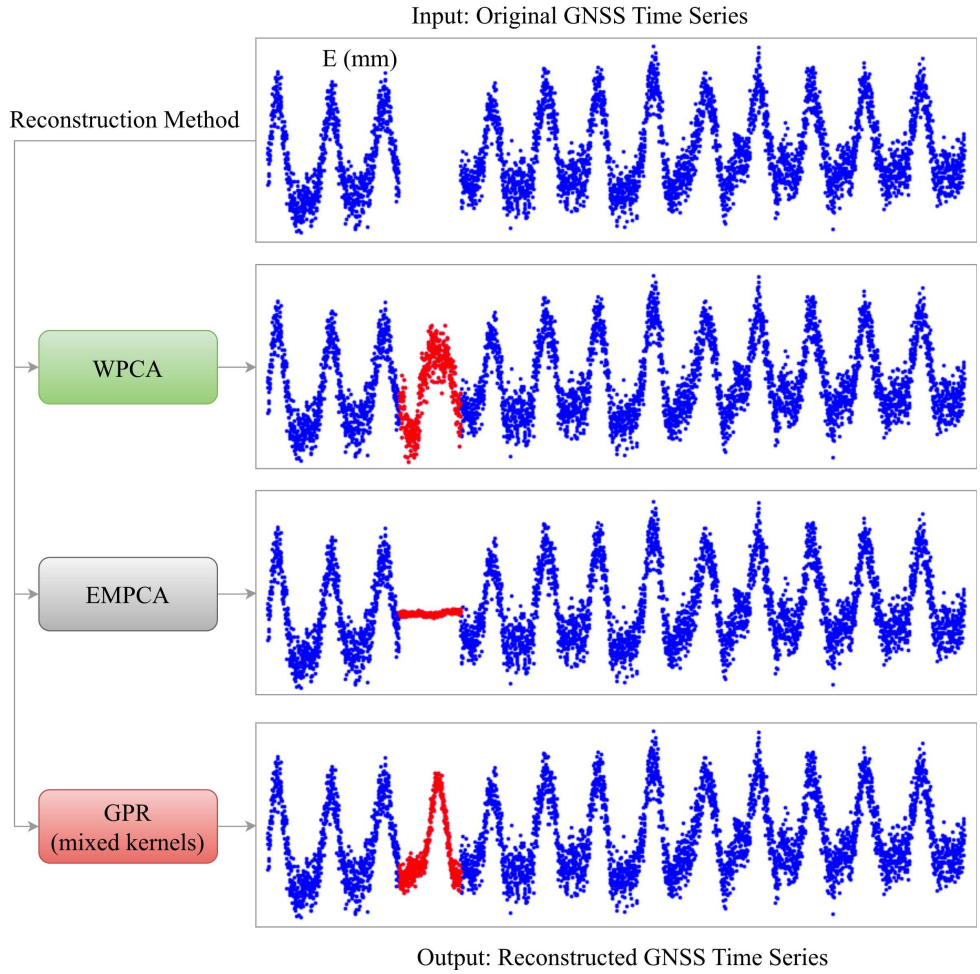


Fig. 1 GNSS east (E) component time series: original data (blue) and reconstructed segments (red) using WPCA, EMPCA, and GPR. Units are in millimetres.

Standard PCA leverages spatial correlation within multi-station networks to fill missing values [25]. Weighted PCA (WPCA) incorporates uncertainty weighting for more reliable estimation [26], and expectation-maximisation PCA (EMPCA) iteratively estimates missing values and principal components [27]. These methods exploit

spatial correlation but have notable limitations: they often assume stationarity and may smooth out station-specific signals; they depend critically on having a sufficient number of stations to estimate the spatial covariance matrix reliably; and their performance degrades with network subsets or temporary outages. This creates a practical paradox, the methods most reliant on network density are least reliable when the network is incomplete, precisely when robust gap-filling is most needed.

The emergence of ML in geodesy has introduced new possibilities [17, 28, 29]. Gaussian process regression (GPR) represents a particularly promising Bayesian non-parametric approach that provides simultaneous prediction and probabilistic uncertainty quantification [30, 31]. A key advantage is that GPR operates independently on each station's time series, eliminating the minimum-station requirement and avoiding the covariance-matrix conditioning issues that affect approaches relying on spatial networks. This makes GPR suitable for applications with variable network density or isolated stations. Its kernel functions can be designed to reflect known geophysical processes, offering interpretability. While GPR has been used in geophysics for scale-level reconstruction [32] and geodetic time-series analysis [33–35], its application to GNSS gap filling in settings characterised by anthropogenic deformation, where transient, often non-stationary signals make data completeness especially critical, remains underexplored. Comprehensive comparisons with established PCA methods in realistic operational settings are also lacking.

Several research gaps motivate this study. First, a systematic comparison between modern GPR and established PCA variants for GNSS position time series is lacking. Second, insufficient attention has been paid to how gap-filling methods affect subsequent ML analyses. Third, methodology evaluation within operational settings characterised by anthropogenic deformation remains underdeveloped. Fourth, there is a lack of clear, evidence-based guidelines for practitioners who must choose a method balancing accuracy, computational cost, and operational constraints.

To address these gaps, we pursue three objectives: i) develop a hierarchical GPR framework using composite kernels tailored to the multi-scale temporal signatures in industrial GNSS data, ii) conduct a comprehensive comparative assessment of this GPR approach against WPCA and EMPCA methods using synthetic tests and field data, and iii) evaluate all methods on criteria critical for ML preprocessing, such as uncertainty quantification and noise preservation, alongside traditional accuracy metrics.

This work therefore provides the first systematic comparison of WPCA, EMPCA, and GPR for GNSS gap-filling within the context of monitoring deformation related to subsurface fluid extraction and storage. We introduce and release an optimised, open-source implementation of the hierarchical GPR model designed for scalability with large networks. A suite of practical evaluation metrics is established and applied, namely RMSE, MAE, R^2, and confidence intervals, to rigorously assess the suitability of gap-filled data for downstream analysis. Based on the findings, we offer evidence-based guidelines to assist practitioners in selecting an appropriate method according to gap duration, network density, and computational requirements. The reconstruction performance of each method is illustrated in Fig. 1, which compares original and gap-filled GNSS east-component time series using WPCA, EMPCA, and

GPR with a mixed kernel of radial basis function (RBF), Matérn, periodic, and noise components.

2 Methodology

This study conducts a comparative analysis of three methods for reconstructing gaps in GNSS time series: WPCA, EMPCA, and a hierarchical GPR framework. Each technique was adapted to GNSS data characteristics and the needs of monitoring deformation linked to subsurface fluid extraction and storage. A common preprocessing stage (detrending) was applied to all methods, with subsequent retrending after reconstruction to ensure fair comparison, as outlined in Fig. 2.

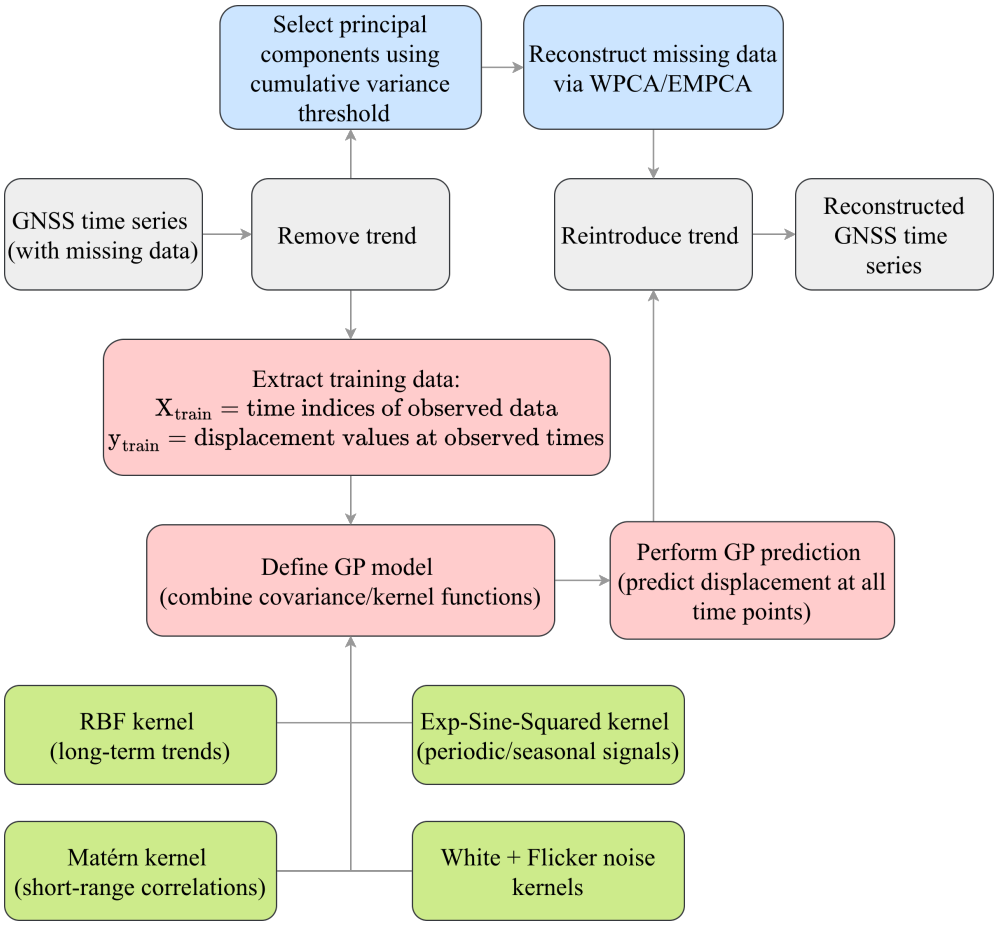


Fig. 2 Comparative workflow of three gap-filling approaches: WPCA, EMPCA, and GPR.

2.1 Weighted Principal Component Analysis (WPCA)

WPCA extends standard PCA to handle missing data by assigning binary weights, excluding missing entries from the decomposition [27]. For a data matrix $\mathbf{X} \in \mathbb{R}^{T \times N}$ with T temporal samples and N stations, the method minimizes::

$$\min_{\mathbf{U},\mathbf{V}} \sum_{i=1}^{T} \sum_{j=1}^{N} w_{ij} \left(x_{ij} - \mathbf{u}_i^\top \mathbf{v}_j \right)^2 \tag{1}$$

where $\mathbf{U} \in \mathbb{R}^{T \times h}$ contains the temporal patterns (scores), $\mathbf{V} \in \mathbb{R}^{N \times h}$ represents spatial loadings (with $\mathbf{V}^\top \mathbf{V} = \mathbf{I}$), x_{ij} is the recorded displacement at station j at time i, and w_{ij} is a binary weight (1 for observed data, 0 for missing). The reconstructed matrix is:

$$\hat{\mathbf{X}} = \mathbf{U}\mathbf{V}^\top \tag{2}$$

The number of principal components h was selected using the cumulative explained variance ratio (CEVR) criterion [22, 36]:

$$\text{CEVR}(h) = \frac{\sum_{i=1}^{h} \lambda_i}{\sum_{i=1}^{N} \lambda_i} \tag{3}$$

where λ_i denotes the eigenvalues. A threshold of 85% was used for all displacement components unless otherwise noted.

2.2 Expectation-Maximization PCA (EMPCA)

EMPCA formulates PCA probabilistically, handling missing values via an iterative expectation-maximization (EM) algorithm [26]. The generative model is:

$$\mathbf{x}_i = \mathbf{W}\mathbf{z}_i + \boldsymbol{\mu} + \boldsymbol{\epsilon}_i \tag{4}$$

where $\mathbf{z}_i \sim \mathcal{N}(0, \mathbf{I})$ are the latent variables, $\mathbf{W}$ contains the principal axes, $\boldsymbol{\mu}$ is the mean vector, and $\boldsymbol{\epsilon}_i \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$ is Gaussian noise. The EM algorithm iterates between:

- E-step: Compute $\mathbb{E}[\mathbf{z}_i]$ and $\mathbb{E}[\mathbf{z}_i \mathbf{z}_i^\top]$ given current parameters.
- M-step: Update $\mathbf{W}$, $\boldsymbol{\mu}$, and σ^2 to maximize the expected complete-data log-likelihood.

The number of components was determined using the same CEVR criterion as in WPCA (Eq. 3).

2.3 Gaussian Process Regression (GPR) Framework

We developed a hierarchical GPR model that operates independently on each station's detrended time series. The approach follows three stages: i) data preprocessing and training set construction; ii) model fitting with a physically motivated composite kernel; and iii) prediction with uncertainty quantification.

After detrending (subsection 2.4), the training set for each station and component is constructed from available, non-missing data:

$$\mathcal{D} = \{(t_i, y_{\text{detrend}}(t_i)) \mid y_{\text{detrend}}(t_i) \text{ is non-NaN}\} \tag{5}$$

where t_i denotes the time coordinate (in days) of the i-th epoch, and $y_{\text{detrend}}(t_i)$ is the corresponding detrended displacement value. This yields training inputs $\mathbf{X}_{\text{train}} \in \mathbb{R}^{n\times 1}$ (time indices of observed points) and training targets $\mathbf{y}_{\text{train}} \in \mathbb{R}^n$ (detrended displacements), with $n < T$ being the number of observed data points.

A Gaussian process (GP) is defined as a collection of random variables, any finite number of which follow a joint Gaussian distribution [30]. It is fully characterized by a mean function $m(t)$ and a covariance (kernel) function $k(t, t')$:

$$f(t) \sim \mathcal{GP}\left(m(t), k(t, t')\right) \tag{6}$$

After detrending and normalization, we set $m(t) = 0$. The observed data are then modelled as:

$$y_{\text{detrend}}(t) = g(t) + \epsilon(t) \tag{7}$$

where $\epsilon(t) \sim \mathcal{N}(0, \sigma^2_{\text{obs}})$ is independent Gaussian measurement noise.

To accurately model the multi-scale temporal signatures in GNSS data, we designed a composite kernel combining five distinct components:

$$k(\mathbf{t}, \mathbf{t}') = k_{\text{RBF}}(\mathbf{t}, \mathbf{t}') + k_{\text{Matérn}}(\mathbf{t}, \mathbf{t}') + k_{\text{periodic}}(\mathbf{t}, \mathbf{t}') + k_{\text{white}}(\mathbf{t}, \mathbf{t}') + k_{\text{flicker}}(\mathbf{t}, \mathbf{t}') \tag{8}$$

Each component captures specific geophysical or noise processes:

- The RBF kernel captures long-term, smooth deformations such as secular tectonic motion [37]:

$$k_{\text{RBF}}(\mathbf{t}, \mathbf{t}') = \sigma_g^2 \exp\left(-\frac{\|\mathbf{t} - \mathbf{t}'\|^2}{2\ell^2_{\text{RBF}}}\right), \ \ell_{\text{RBF}} \in [50, 1000] \text{ days} \tag{9}$$

- Short-term correlations and transient deformations are modelled with a Matérn kernel ($\nu = 1.5$) [38]:

$$k_{\text{Matérn}}(\mathbf{t}, \mathbf{t}') = \sigma_g^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}\|\mathbf{t} - \mathbf{t}'\|}{\ell_{\text{Matérn}}}\right)^{\nu} K_\nu \left(\frac{\sqrt{2\nu}\|\mathbf{t} - \mathbf{t}'\|}{\ell_{\text{Matérn}}}\right) \tag{10}$$

  with $\ell_{\text{Matérn}} \in [5, 150]$ days.
- A periodic kernel represents annual cyclical patterns (e.g., seasonal loading) [39]:

$$k_{\text{periodic}}(\mathbf{t}, \mathbf{t}') = \sigma_g^2 \exp\left(-\frac{2\sin^2(\pi\|\mathbf{t} - \mathbf{t}'\|/P)}{\ell^2_{\text{periodic}}}\right) \tag{11}$$

  Here, the period P is fixed at 365.25 days, with bounds between 350 and 380 days to allow for minor variation, while $\ell_{\text{periodic}} \in [1, 50]$ days controls harmonic smoothness.

- Uncorrelated measurement uncertainty is represented by a white noise kernel [40]:

$$k_{\text{white}}(\mathbf{t}, \mathbf{t}') = \sigma_w^2 \delta(\mathbf{t}, \mathbf{t}') \tag{12}$$

  The noise amplitude σ_w^2 is station-specific, based on an empirical pre-analysis of the GNSS network's residual time series. For stations with zero or negative estimates, σ_w^2 was fixed at $10^{-3} mm^2$; otherwise, bounds of $[0.1\sigma_w^2, 10\sigma_w^2]$ were applied.
- Flicker (or pink) noise, characterized by a $1/f$ power spectrum, has been widely documented in GNSS position time series and significantly influences velocity uncertainty estimates [41, 42]. Since our observed GNSS data from the North Adriatic network confirmed the presence of this noise component, we implemented a pink noise kernel using a Matérn covariance function with smoothness parameter $\nu = 0.5$, a configuration mathematically equivalent to an exponential kernel. While a true $1/f$ spectrum is more complex, the Matérn family with $\nu = 0.5$ provides a computationally convenient approximation that reproduces the essential power-law behaviour and is commonly used to model flicker noise in geodetic time series [42]. This choice directly incorporates the characteristic pink noise spectrum into the GP's covariance structure, ensuring predictive variances correctly account for this stochastic process:

$$k_{\text{flicker}}(\mathbf{t}, \mathbf{t}') = \sigma_p^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}\|\mathbf{t} - \mathbf{t}'\|}{\ell_{\text{flicker}}} \right)^{\nu} K_\nu \left(\frac{\sqrt{2\nu}\|\mathbf{t} - \mathbf{t}'\|}{\ell_{\text{flicker}}} \right) \tag{13}$$

  with $\nu = 0.5$, $\ell_{\text{flicker}} \in [0.1, 5]$ days, and σ_p^2 derived from station-specific pink noise amplitudes.

The hyperparameters of the composite kernel, collectively denoted as $\boldsymbol{\theta} = \{\sigma_g^2,\ \ell_{\text{RBF}},\ \ell_{\text{Matérn}},\ \ell_{\text{periodic}},\ \ell_{\text{flicker}},\ \sigma_w^2,\ \sigma_p^2\}$, were optimised by maximising the log marginal likelihood [35]:

$$\log p(\mathbf{y}|\mathbf{t}, \boldsymbol{\theta}) = -\frac{1}{2}\mathbf{y}^\top \mathbf{K}_{\boldsymbol{\theta}}^{-1}\mathbf{y} - \frac{1}{2}\log|\mathbf{K}_{\boldsymbol{\theta}}| - \frac{n}{2}\log 2\pi \tag{14}$$

where $\mathbf{K}_{\boldsymbol{\theta}}$ represents the full covariance matrix including both signal and noise components. Optimization was performed with five random restarts to avoid local minima.

Once trained, the model predicts the mean $\boldsymbol{\mu}_*$ and covariance $\boldsymbol{\Sigma}_*$ at all time points $\mathbf{t}_*$ using standard GP equations:

$$\begin{aligned} \boldsymbol{\mu}_* &= \mathbf{K}(\mathbf{t}_*, \mathbf{X}_{\text{train}}) \left[\mathbf{K}(\mathbf{X}_{\text{train}}, \mathbf{X}_{\text{train}}) + \sigma_{\text{obs}}^2 \mathbf{I}\right]^{-1} \mathbf{y}_{\text{train}} \\ \boldsymbol{\Sigma}_* &= \mathbf{K}(\mathbf{t}_*, \mathbf{t}_*) - \mathbf{K}(\mathbf{t}_*, \mathbf{X}_{\text{train}}) \left[\mathbf{K}(\mathbf{X}_{\text{train}}, \mathbf{X}_{\text{train}}) + \sigma_{\text{obs}}^2 \mathbf{I}\right]^{-1} \mathbf{K}(\mathbf{X}_{\text{train}}, \mathbf{t}_*) \end{aligned} \tag{15}$$

where $\mathbf{K}(\cdot, \cdot)$ denotes the kernel matrix and $\mathbf{I}$ is the identity matrix.

Predictions at all daily epochs $\mathbf{t}_{\text{all}} = 1, 2, \ldots, T$ yield the imputed mean $\boldsymbol{\mu}_{\text{all}}$, full covariance $\boldsymbol{\Sigma}_{\text{all}}$, and predictive standard deviations $\boldsymbol{\sigma}_{\text{all}} = \sqrt{\text{diag}(\boldsymbol{\Sigma}_{\text{all}})}$. Gap-filled

values are then obtained by adding scaled Gaussian noise to the mean predictions and restoring the polynomial trend (subsection 2.4).

2.4 Implementation Details for GNSS Data

All time series were detrended prior to analysis. For each station, an F-test ($\alpha = 0.05$) selected between linear and quadratic polynomial trends [43]. The detrended residual matrix $\mathbf{X}_{\text{detrend}}$ was used as input for WPCA and EMPCA, with weights defined as the inverse of formal uncertainties: $w_{ij} = \sigma_{ij}^{-2}$ for observed data [44], and 0 otherwise. After reconstruction, the polynomial trend was restored.

For GPR, each station's detrended series was processed independently. Following prediction, scaled Gaussian noise was added to gap-filled values to match observed noise characteristics:

$$\mathbf{y}_{\text{filled}}[t_{\text{gap}}] = \mathbf{y}_{\text{GP}}[t_{\text{gap}}] + \eta \cdot \boldsymbol{\sigma}_{\text{GP}}[t_{\text{gap}}] \cdot \xi, \quad \xi \sim \mathcal{N}(0,1) \tag{16}$$

where $\eta = 0.2$ is a scaling factor that controls the noise level added to the imputed values. This value was determined empirically by analysing the average ratio between the predictive standard deviation and the residuals in the training data, ensuring that the filled gaps exhibit realistic noise characteristics consistent with the observed data. Further testing of the η value is presented in subsection 4.2. The polynomial trend was then added back to obtain the final reconstructed series.

Reconstruction accuracy was assessed using three standard metrics calculated across all data points: root mean square error (RMSE), mean absolute error (MAE), and the coefficient of determination (R^2) [31]:

$$\text{RMSE} = \sqrt{\frac{1}{M}\sum_{i=1}^{M}(y_i - \hat{y}_i)^2}, \ \text{MAE} = \frac{1}{M}\sum_{i=1}^{M}|y_i - \hat{y}_i|, \ R^2 = 1 - \frac{\sum_{i=1}^{M}(y_i - \hat{y}i)^2}{\sum_{i=1}^{M}(y_i - \bar{y})^2} \tag{17}$$

where M is the total number of data points assessed, y_i is the true values, $\hat{y}_i$ is the reconstructed values, and $\bar{y}$ the mean of the observed data.

For GPR, 95% confidence intervals were computed as $\hat{y}_i \pm 1.96\sigma_i$ [30, 35], and empirical coverage probability was evaluated. Computational efficiency was compared via average runtime per station.

3 Synthetic Data Generation

To rigorously test the gap-filling methods outlined in section 2, we generated a synthetic dataset designed to replicate the statistical and geophysical characteristics of the observed GNSS time series from the 70-station North Adriatic region (Fig. 3). While applying the three methods directly to real time series would allow qualitative comparison, synthetic data offer a critical advantage: they provide a known ground truth against which reconstruction accuracy can be quantitatively evaluated. This is particularly important for assessing performance on transient deformation signals associated with human activities, precisely the type of signals that motivate this study, as

real-world gaps obscure the very data needed for validation. Synthetic data enable us to introduce controlled gaps and measure exactly how well each method recovers the original signal, including the anthropogenic components that are of primary interest for operational monitoring.

The study area (Fig. 3) is geodynamically complex, characterized not only by active hydrocarbon operations (five gas storage sites and eighteen production sites) but also by natural subsidence due to sediment compaction, ongoing tectonic activity, as well as transient signals from events such as the 2012 Emilia earthquake sequence [45]. This combination of natural and anthropogenic deformation sources produces a rich mixture of signals at different temporal scales, making the region an ideal testbed for evaluating gap-filling methods.

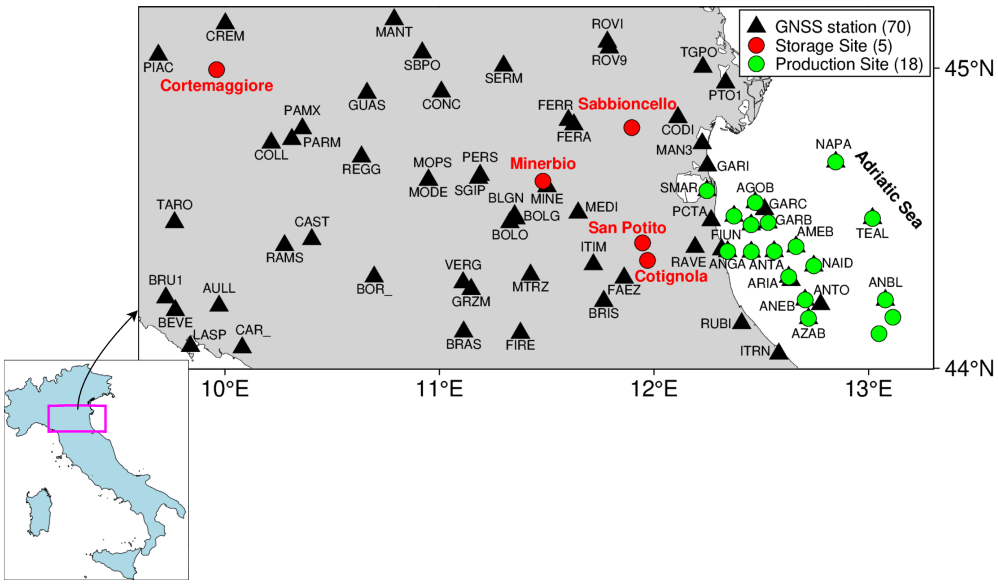


Fig. 3 Map of the study area in the North Adriatic Sea, showing the GNSS network alongside gas storage and hydrocarbon production sites.

The synthetic time series were constructed to mirror field GNSS data processed via a standard methodology [44]. The functional model incorporates the key components identified in the processed field data:

$$\begin{aligned} y(t) = y_0 + b(t - t_0) + \alpha_1 \cdot \sin[2\pi(t - t_0) + \varphi_1] \\ + \alpha_2 \cdot \sin[4\pi(t - t_0) + \varphi_2] + D(t) + \epsilon_{\mathrm{w}}(t) + \epsilon_{\mathrm{p}}(t) \end{aligned} \tag{18}$$

where $y(t)$ is the synthetic station position at time t. The parameters include an initial position bias y_0, a linear rate b that varies spatially across the network, and annual and semi-annual seasonal signals with amplitudes α_1, α_2 and phases φ_1, φ_2.

These parameters were randomly drawn from normal distributions defined by the median and inter-quartile range of the corresponding field data estimates. The rate b was assigned based on the geographic location of each synthetic station, with higher subsidence rates in areas subject to greater sediment compaction and fluid extraction, following the spatial patterns observed in the field data. The noise terms, white noise ϵ_{w} and pink noise ϵ_{p}, were generated as Gaussian noise with a power-law spectrum [46]. Their amplitudes were randomly selected from site-specific uniform distributions derived from the field data.

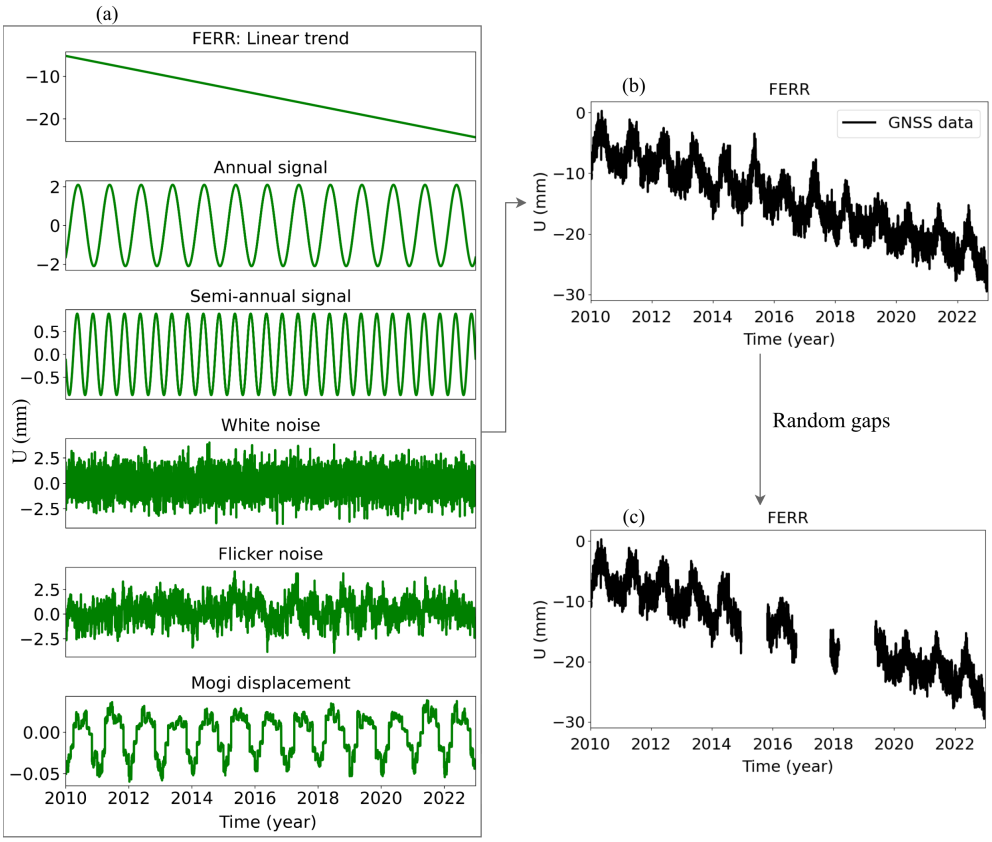


Fig. 4 Synthetic GNSS time series for the up component at the Sabbioncello gas storage site (near FERR station). (a) Decomposition into individual components. (b) Complete synthetic series. (c) Series with realistic random data gaps.

The anthropogenic deformation component $D(t)$, resulting from reservoir volume changes, was simulated using a Mogi model [47]. This model provides a simple analytical relationship between volume change and surface displacement, making it suitable for approximating ground movement from gas storage or hydrocarbon extraction [48]. Surface displacements were calculated based on realistic volume-change time series

derived from operational data at the actual storage and production sites. Further details for the GNSS station locations, noise amplitudes, and volume conversions are provided in supplementary Appendices A and B.

To simulate realistic data imperfections, 30% of the data points were randomly set to missing, with gap lengths ranging from 100 to 500 days. Fig. 4 shows an example of the complete synthetic series for the vertical component at the Sabbioncello storage site (near the FERR station). This synthetic dataset (see Data availability statement) provides a controlled, realistic testbed for evaluating the performance of the WPCA, EMPCA, and GPR gap-filling methods.

4 Results

4.1 Results of Synthetic GNSS Data

We first evaluate the three gap-filling methods using synthetic GNSS data from 70 stations in Fig. 3, where the true signal is known and can serve as ground truth for quantitative comparison. The aggregate performance across the entire network, summarized in Fig. 5, establishes a clear performance hierarchy: GPR achieves the highest accuracy and consistency, followed by WPCA, with EMPCA showing the largest errors and variability, particularly for horizontal components. To illustrate the distinct signal morphologies underlying these aggregate trends, we examine two representative synthetic stations: AMEB (east) for oscillatory, complex signals and ANTA (up) for trend-dominated vertical motion. Validation on real GNSS data from the same network follows in section 4.3.

Fig. 5 presents the distribution of RMSE, MAE, and R^2 across all stations for the east (E), north (N), and up (U) components. For both RMSE and MAE (panels a and b), GPR exhibits the lowest median errors and narrowest interquartile ranges (IQRs) across all three components. Its median RMSE remains below 1 mm for the horizontal components, with few outliers. On the other hand, EMPCA shows the highest median errors for E and N (RMSE $>$ 1.5 mm), with broad IQRs and numerous outliers. WPCA occupies an intermediate position, with errors consistently higher than GPR but lower than EMPCA for the horizontal components.

The R^2 distributions (Fig. 5c) reinforce this pattern: GPR maintains the highest median values (0.8–0.9) for E and N, while EMPCA's R^2 drops sharply for these components (median $\approx$ 0.4). Notably, for the up component, EMPCA's performance improves substantially, nearly matching GPR in R^2 (median $\approx$ 0.9), reflecting its suitability for trend-dominated signals.

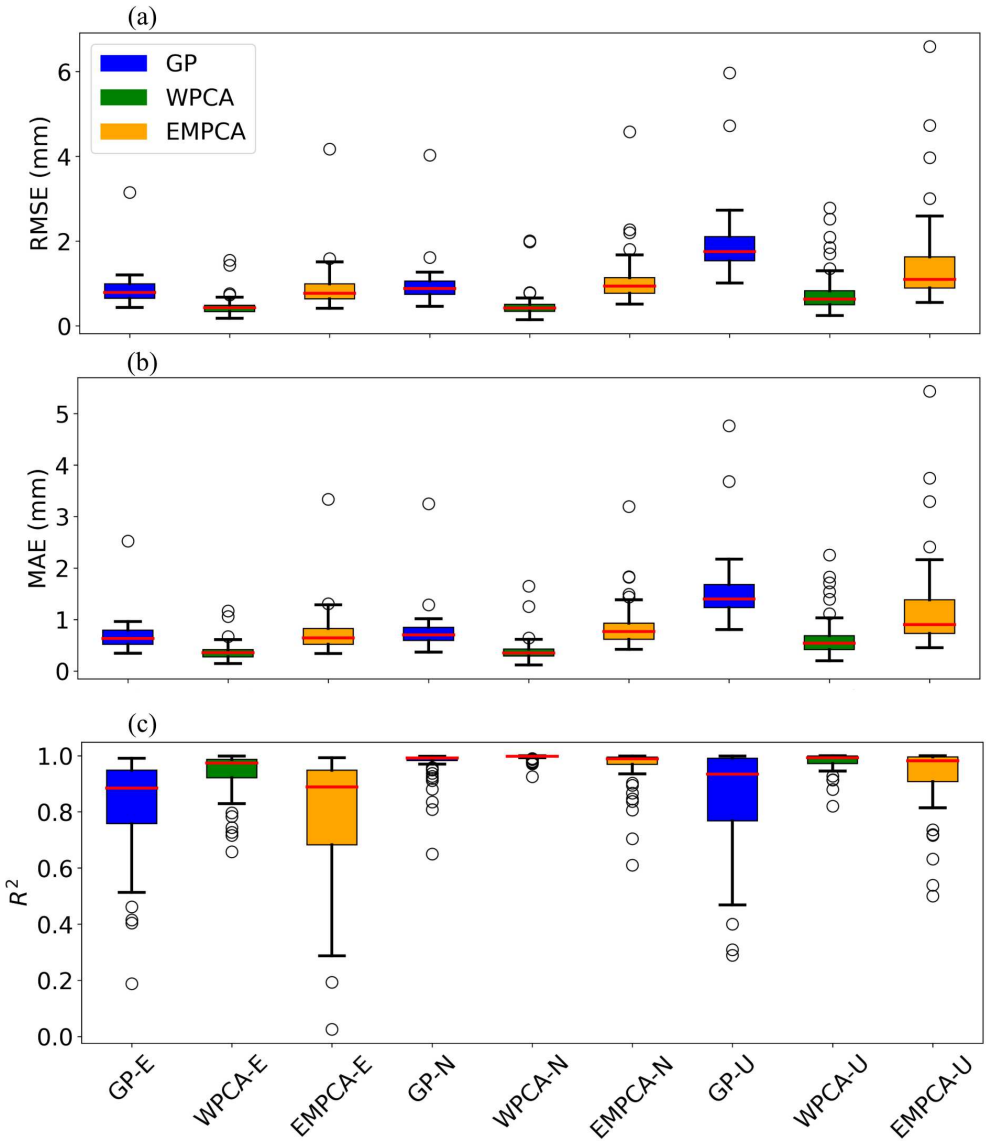


Fig. 5 Distribution of performance metrics across 70 synthetic GNSS stations. (a) RMSE, (b) MAE, and (c) R^2 for the GP, WPCA, and EMPCA reconstruction methods, shown for the east (E), north (N), and up (U) components. Red line stands for median values.

Fig. 6 presents a systematic evaluation of methods to reconstruct synthetic GNSS signals and mitigate data gaps, focuses on the east direction component at the AMEB station. Panel (a) exhibits pronounced temporal variability with irregular gaps between 2012 and 2014. GPR reconstruction (Fig. 6b) closely follows the original signal, achieving an RMSE of 0.507 mm, MAE of 0.407 mm, R^2 of 0.830 (Table 1). WPCA requires

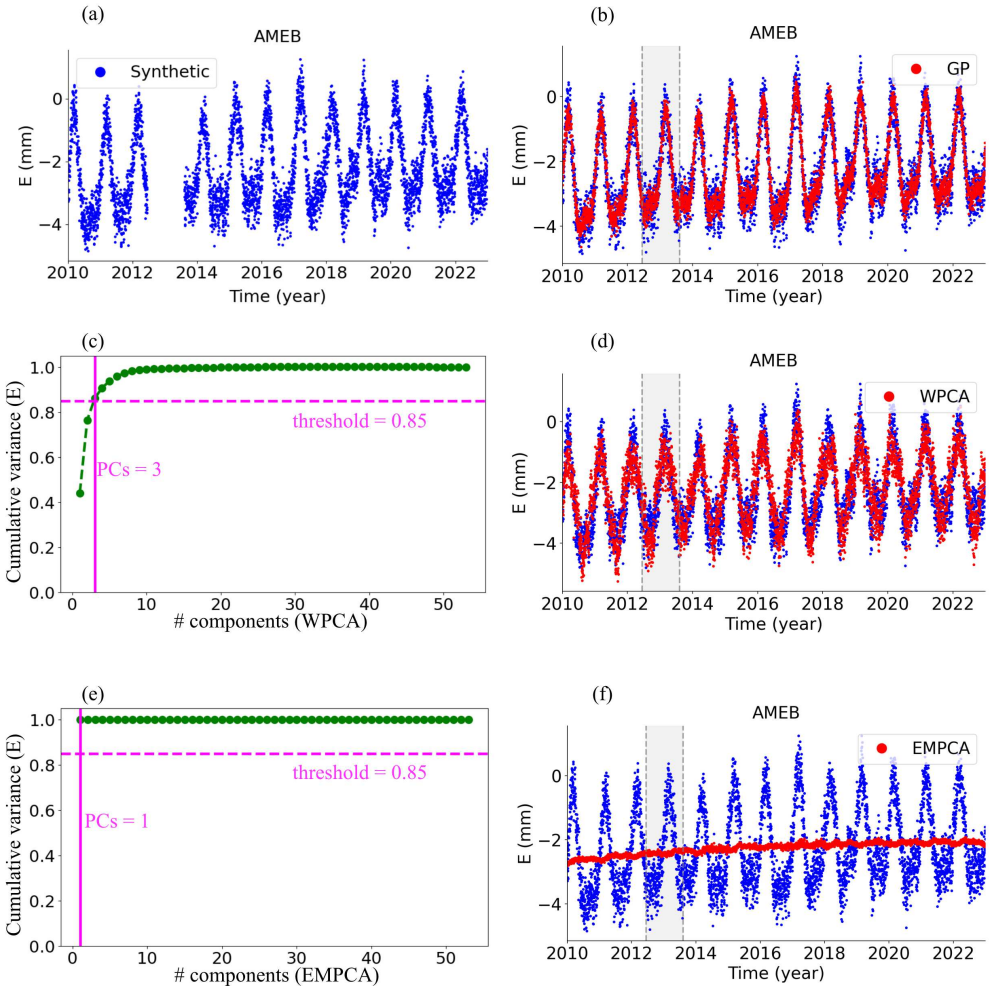


Fig. 6 Reconstruction of the synthetic East-component signal at station AMEB. (a) Original gapped signal. (b) GPR reconstruction. (c, d) WPCA cumulative variance and reconstruction using three PCs. (e, f) EMPCA cumulative variance and reconstruction using one PC. Grey bands mark data gaps; blue points and red lines show original and reconstructed data, respectively.

three principal components to explain 85% of the variance (Fig. 6c), but its reconstruction (Fig. 6d) retains residual high-frequency noise (RMSE = 0.555 mm, $R^2 = 0.797$). EMPCA explains the same variance with only one component (Fig. 6e), resulting in an overly smooth reconstruction (Fig. 6f) that poorly fits the oscillatory data (RMSE = 1.215 mm, $R^2 = 0.025$). This case highlights EMPCA's inability to capture complex, and non-stationary dynamics.

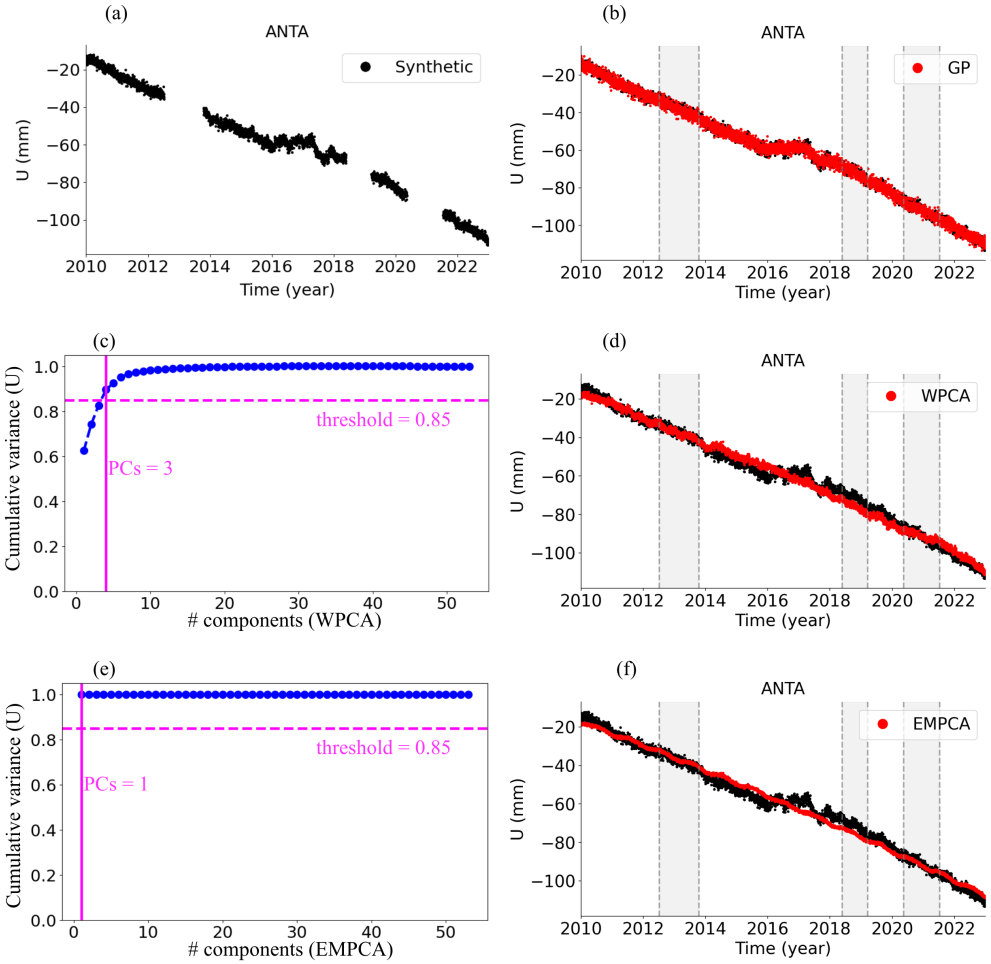


Fig. 7 Gap-filling performance for the synthetic up-component signal at station ANTA. (a) Original gapped signal. (b) GPR-imputed signal. (c, d) WPCA cumulative variance and imputed signal. (e, f) EMPCA cumulative variance and imputed signal. Format follows Fig. 6.

The synthetic vertical signal at ANTA (Fig. 7a) shows a strong, consistent linear trend from -20 mm to -100 mm between 2010 and 2022. GPR again provides the most accurate reconstruction (Fig. 7b), with an RMSE of 2.373 mm and R^2 of 0.992 (as shonw in Table 1). WPCA (using three PCs, Fig. 7d) and EMPCA (using one PC, Fig. 7f) both perform well on this trend-dominated signal, with closely matched results (WPCA: RMSE = 2.783 mm, R^2 = 0.989; EMPCA: RMSE = 3.006 mm, R^2 = 0.987). The minimal performance gap here contrasts sharply with the AMEB case, demonstrating that EMPCA can be effective when the signal is dominated by a simple, and low-dimensional structure.

Table 1 Performance metrics for the reconstruction of synthetic GNSS data at stations AMEB (east) and ANTA (up).

Station (Component)	Method	RMSE (mm)	MAE (mm)	R^2
AMEB (E)	GP	**0.507**	**0.407**	**0.830**
	WPCA	0.555	0.476	0.797
	EMPCA	1.215	1.040	0.025
ANTA (U)	GP	**2.373**	**1.894**	**0.992**
	WPCA	2.783	2.254	0.989
	EMPCA	3.006	2.409	0.987

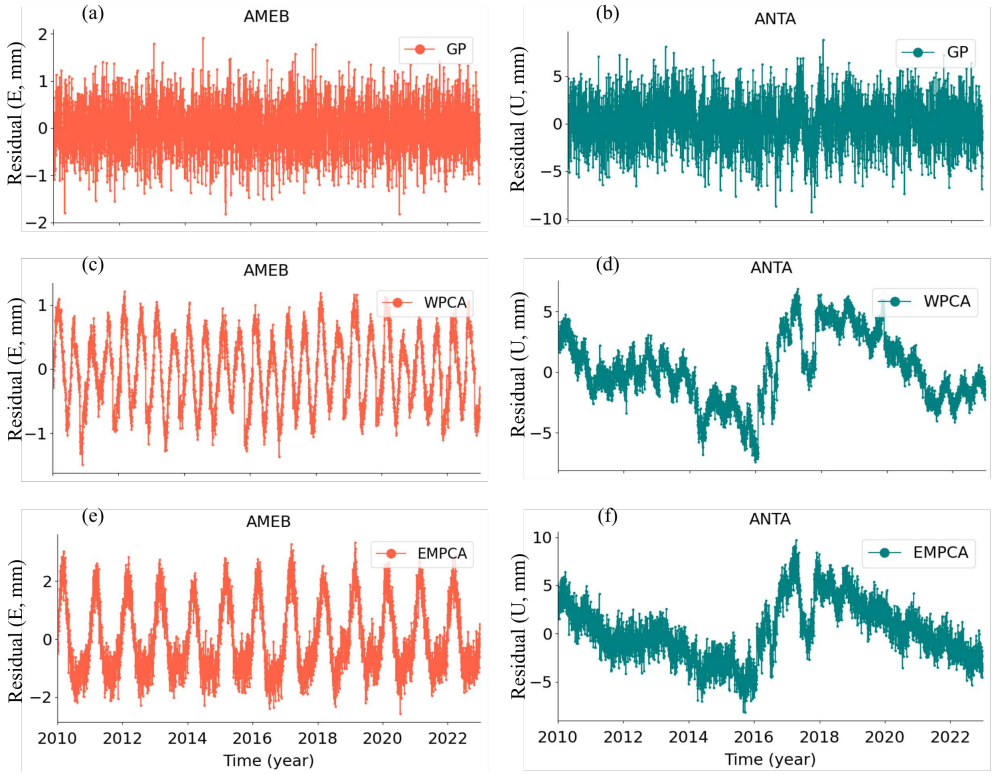


Fig. 8 Residual plots (true minus reconstructed) for (a–c) AMEB east and (d–f) ANTA up, for GP, WPCA, and EMPCA.

Residual analyses for both stations (Fig. 8) further clarify these patterns. For AMEB, GPR residuals are tightly clustered within ±1 mm, while WPCA shows oscillatory residuals and EMPCA exhibits large, systematic deviations. For ANTA, all three methods produce small residuals, consistent with the strong trend that is easily captured even by simpler methods. Collectively, the synthetic data results demonstrate that GPR provides the most robust and accurate gap-filling across diverse signal

types, while EMPCA is only reliable for strongly trend-dominated vertical signals, and WPCA offers a middle-ground compromise. The following sections examine the sensitivity of these methods to parameter choices and their performance on real field data.

4.2 Sensitivity to Model Parameters

The robustness of any gap-filling method depends not only on its core algorithm but also on the appropriate configuration of its key parameters. We examine how reconstruction accuracy is influenced by two critical aspects: i) kernel composition in GPR and the variance threshold in PCA-based methods, and ii) hyperparameter bounds within the GPR framework. This sensitivity analysis reveals fundamental characteristics of each method's flexibility and practical tuning requirements.

Regarding kernel composition and PCA variance threshold, Fig. 9 presents a systematic sensitivity analysis for the east component at station FIUN. For WPCA, increasing the cumulative variance threshold (and thus the number of principal components) improves reconstruction accuracy. Raising the threshold from 0.75 (2 PCs) to 0.95 (6 PCs) reduces RMSE from 0.734 mm to 0.658 mm and increases R^2 from 0.656 to 0.724 (Fig. 9a–c). This confirms that capturing more signal variance leads to better fit for oscillatory data. In contrast, EMPCA exhibits a structural limitation: even with a high variance threshold of 0.95, it utilizes only one principal component (Fig. 9d), resulting in poorer reconstruction (RMSE = 0.893 mm, $R^2 = 0.491$). This demonstrates that EMPCA's constrained dimensionality prevents it from matching WPCA's accuracy on complex signals, regardless of threshold setting. Extended cumulative variance plots for both methods across the entire network are available in supplementary Appendix C.

For GP regression, the choice of kernel functions is critical. The best performance is achieved by the full mixed kernel (Fig. 9e), which combines RBF, Matérn, periodic, and noise components, yielding an RMSE of 0.618 mm and an R^2 of 0.756. Removing key elements from this mixture degrades performance. For instance, omitting the Matérn kernel (Fig. 9f) or the periodic kernel (Fig. 9g) increases the RMSE. The simplest kernel, using only RBF and noise components (Fig. 9h), performs the worst among the GP variants tested. These results highlight distinct tuning considerations for each method. WPCA's accuracy can be improved by increasing the variance threshold to include more components. EMPCA's performance is inherently limited by its use of a single component. GP delivers optimal results only when its kernel composition is sufficiently complex to model all relevant signal characteristics. Therefore, careful parameter selection is essential, with GP offering the most robust performance when properly configured.

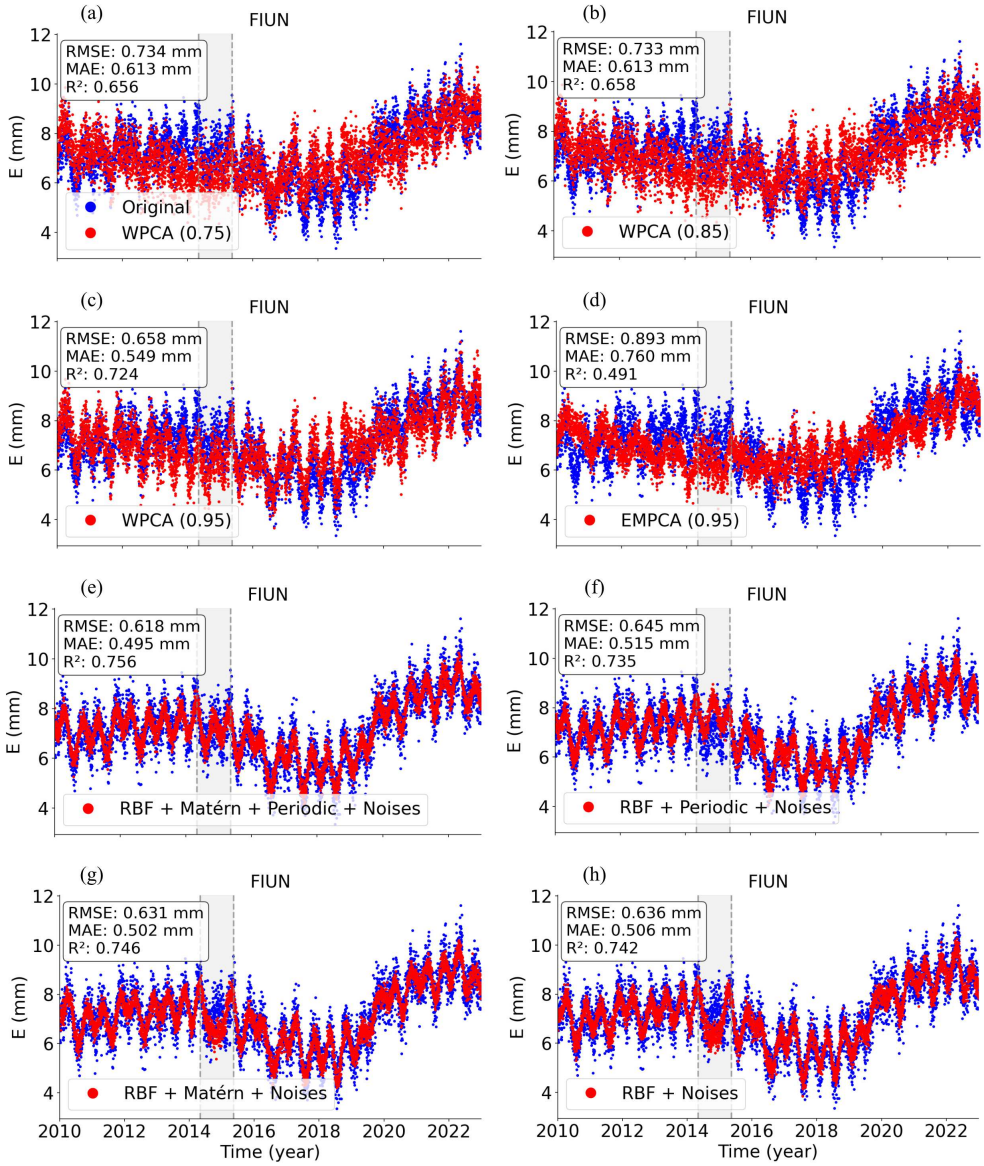


Fig. 9 Parameter selection analysis for synthetic GNSS data at station FIUN (east/E). WPCA reconstruction with variance thresholds of (a) 0.75, (b) 0.85, and (c) 0.95. (d) EMPCA reconstruction with a 0.95 threshold. GPR reconstruction using (e) a full mixed kernel, (f) RBF+periodic+noise kernels, (g) RBF+Matérn+noise kernels, and (h) RBF+noise kernels only. Grey bands mark data gaps; original and reconstructed data are shown as blue points and red lines, respectively.

We further investigate the sensitivity of GPR to specific hyperparameter bounds using the east component at station SBPO (Fig. 10). The baseline reconstruction

(GP-Fixed) with an optimized composite kernel achieves RMSE = 0.443 mm and R^2 = 0.847 (Fig. 10a). Altering individual hyperparameters while keeping others fixed reveals a clear hierarchy of sensitivity.

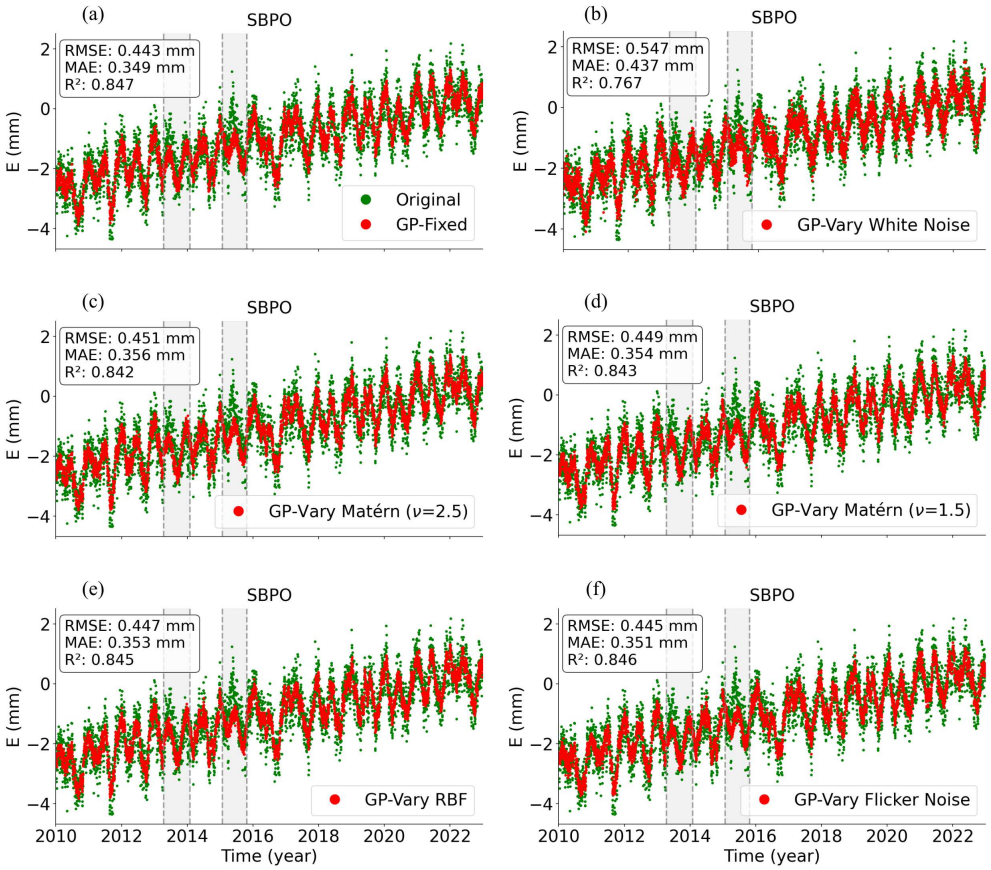


Fig. 10 Effects of individual kernel hyperparameter variations on GP reconstruction for the east component at station SBPO. (a) Baseline (GP-Fixed) with optimised composite kernel. (b) Modified white noise bounds. (c, d) Adjusted Matérn smoothness η. (e) Altered RBF length scale. (f) Modified flicker noise parameters. Performance metrics indicate that white noise calibration is most critical. Grey bands indicate missing data; green and red represent original and reconstructed data, respectively. Residual plots for all these hyperparameter tests (Appendix D) are provided in the supplementary file.

The most significant performance drop occurs when varying the white noise kernel parameters (Fig. 10b), where RMSE increases to 0.547 mm and R^2 falls to 0.767. This indicates that accurately modeling the noise level is critical for reconstruction fidelity. In contrast, changes to other kernel components have minimal impact. Adjusting the smoothness parameter (ν) of the Matérn kernel (Fig. 10c,d) or the length scale of the

RBF kernel (Fig. 10e) results in nearly identical metrics to the baseline. Variations to the flicker noise parameters (Fig. 10f) also cause negligible change. This demonstrates that while a comprehensive kernel mixture is essential, not all parameters require equally precise tuning; those governing noise modelling are the most sensitive. The specific parameter configurations tested are detailed in Table 2.

Table 2 Parameter configurations for GP kernel sensitivity tests with fixed $\eta = 0.2$. Key: ℓ=length scale, P=period, ν=smoothness parameter, σ_n^2=noise variance, bounds=parameter bounds. P_b=periodicity bounds.

Configuration	**Kernel Parameters**				
	RBF	Matérn	periodic	white noise	flicker noise
GP-fixed	$\ell = 365$; bounds: (50, 1000)	$\ell = 30.0$, $\nu = 1.5$; bounds: (5.0, 150.0)	$\ell = 10.0$, $P = 365.25$; ℓ_b: (1.0, 50.0), P_b: (350.0, 380.0)	bounds: $(0.1\sigma_n^2, 10\sigma_n^2)$	$\ell = 1.0$, $\nu = 0.5$; bounds: (0.1, 5.0)
GP-vary RBF	$\ell = 50.0$; bounds: (100.0, 3000.0)	–	–	–	–
GP-vary Matérn ($\nu = 2.5$)	–	$\ell = 45.0$, $\nu = 2.5$; bounds: (10.0, 200.0)	–	–	–
GP-vary Matérn ($\nu = 1.5$)	–	$\ell = 45.0$, $\nu = 1.5$; bounds: (10.0, 200.0)	–	–	–
GP-vary white noise	–	–	–	bounds: $(0.5\sigma_n^2, 2\sigma_n^2)$	–
GP-wary flicker noise	–	–	–	–	$\ell = 2.0$, $\nu = 0.5$; bounds: (0.5, 10.0)

Additionally, we test the influence of the noise scaling hyperparameter η (which controls the amplitude of added noise in the gap-filled values) on the up component at station FERR (Fig. 11). As η (in Eq. 16) increases from 0.2 to 0.5, reconstruction quality systematically declines: RMSE rises from 1.992 mm to 3.312 mm, and R^2 drops from 0.888 to 0.692. Visually, the fit becomes poorer, with the reconstructed line diverging more noticeably from the original data within gaps. This confirms that even with an optimal kernel, inappropriate noise scaling can degrade performance; for our data, $\eta = 0.2$ provides the best balance.

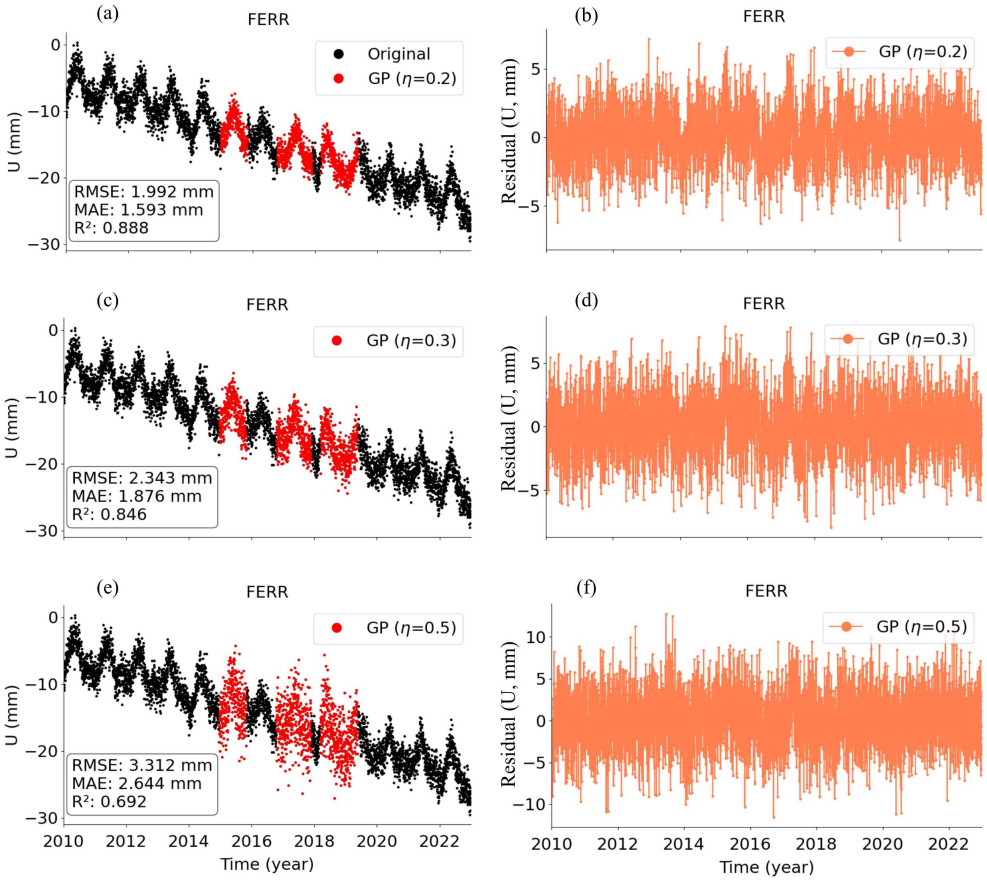


Fig. 11 Influence of noise scaling on GP reconstruction for the up (U) component at station FERR. (a, c, e) reconstructed signals (red) against original data (black) for $\eta = 0.2, 0.3$, and 0.5. (b, d, f) corresponding residual plots. Accuracy systematically declines with higher η values.

4.3 Results of Real GNSS Data

We now apply the three reconstruction methods to real GNSS observations from the North Adriatic network (in Fig. 3). The results from field data validate and extend the synthetic benchmarks, confirming the general performance hierarchy in an operational setting.

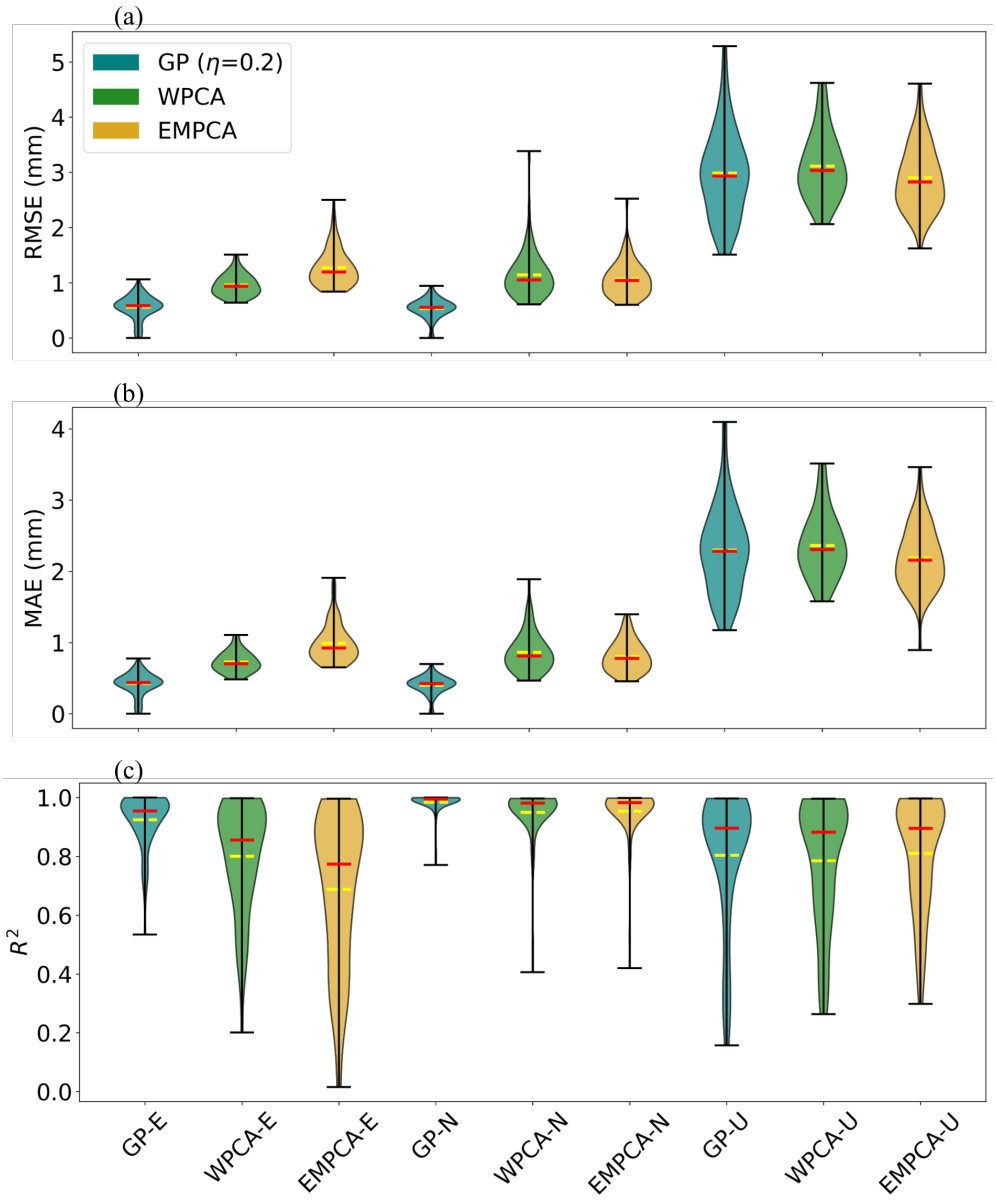


Fig. 12 Performance distributions across the 70-station field network for (a) RMSE, (b) MAE, and (c) R^2 comparing GP, WPCA, and EMPCA for east (E), north (N), and up (U) components. The red and yellow lines indicate the median and mean values, respectively.

Aggregate performance across the entire network is shown in Fig. 12. For both RMSE and MAE (panels a and b), GPR maintains the lowest median error and narrowest spread for all displacement components (east, north, and up). Its performance

is robust, with a median RMSE between 0.5 and 1.0 mm for the horizontal components and few outliers. WPCA shows intermediate results with higher median errors and greater outcome variability, indicating station-dependent sensitivity. EMPCA exhibits the highest errors and greatest spread, particularly for the east and north components.

The R^2 distributions (Fig. 12c) corroborate this pattern. GPR achieves the highest median values (0.8–0.9), confirming its superior signal recovery. EMPCA yields very low R^2 scores for horizontal components but performs notably better for the vertical component, a trend that directly echoes the synthetic data finding that its single-component model is adequate for simple, trend-dominated signals but fails for complex dynamics.

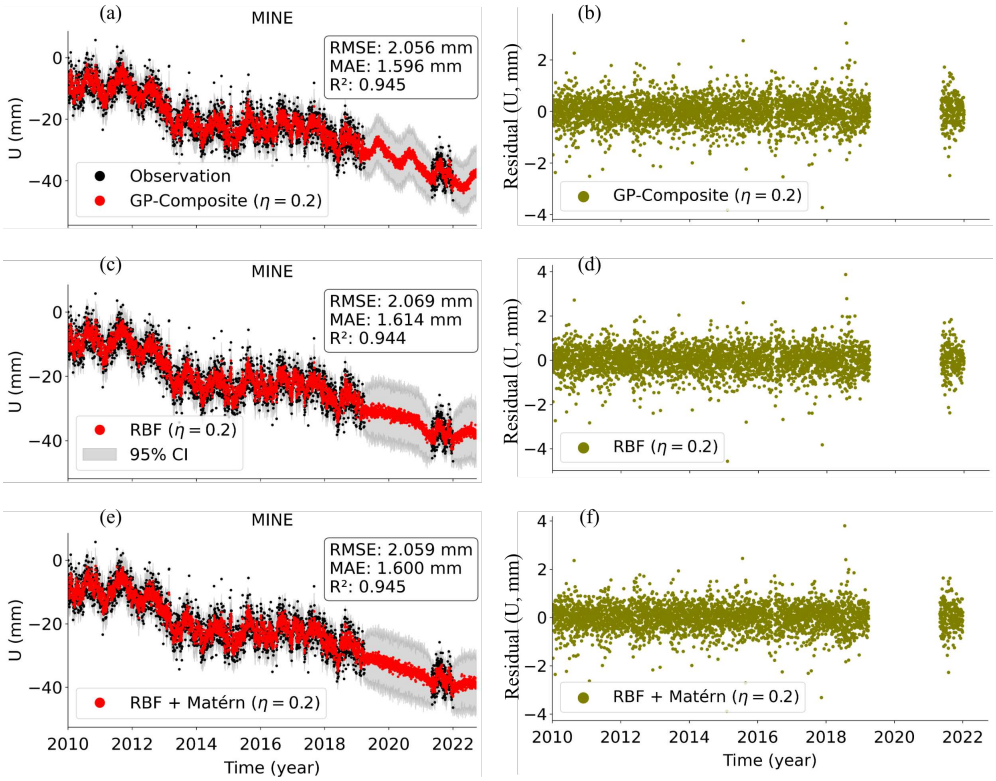


Fig. 13 Sensitivity of GP reconstruction to kernel composition for the vertical component at station MINE. Rows: (a, b) Full composite kernel (RBF + Matérn + periodic + white + flicker noise); (c, d) Simplified RBF + noise kernels; (e, f) Simplified RBF + Matérn + noise kernels. Left column shows reconstructions with 95% confidence intervals (CI, grey bands); right column shows corresponding residuals. Performance metrics indicate that simplified kernels achieve equivalent accuracy and uncertainty estimation for this real dataset, highlighting the flexibility of the GPR framework.

Fig. 13 examines whether simplified GPR kernel configurations can match the performance of the full composite kernel when reconstructing real GNSS data. Using

the vertical component at station MINE, we compare reconstructions from three kernel sets: i) the full composite kernel (RBF + Matérn + periodic + white + flicker noise), ii) a simplified RBF + noise kernel, and iii) RBF + Matérn + noise. All three yield nearly identical accuracy (RMSE $\approx$ 2.06 mm, $R^2 \approx 0.945$) and uncertainty bounds. This result critically indicates that for real vertical signals dominated by long-term trends, simpler kernels can maintain fidelity while substantially reducing computational cost, a key practical advantage for large-scale operational processing where efficiency is paramount.

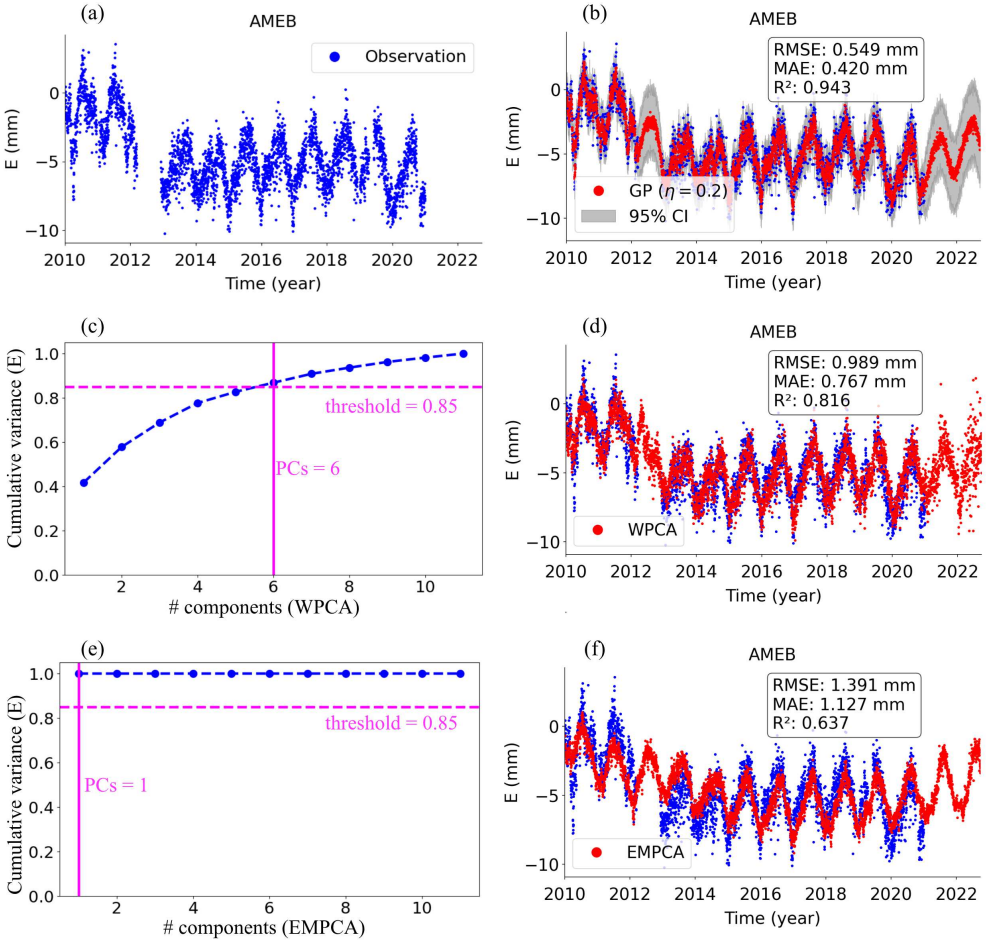


Fig. 14 Field data imputation for the GNSS east component at station AMEB. (a) Original series with gaps. (b) GP-filled series using $\eta = 0.2$ to control added noise (see subsection 2.4), with 95% confidence intervals shown in grey. (c) WPCA cumulative variance plot. (d) WPCA reconstruction using six principal components. (e) EMPCA cumulative variance plot. (f) EMPCA reconstruction using a single principal component. GP achieves the most accurate result (RMSE = 0.549 mm, $R^2 = 0.943$), confirming the conclusions drawn from synthetic tests.

Fig. 14 compares time series modelling techniques for the east component of the AMEB station from 2010 to 2022. Panel (a) shows the observed GNSS data, characterised by strong seasonal oscillations, a notable gap between 2012 and 2013, and missing data from 2021 onwards. Panel (b) applies a GP model with $\eta = 0.2$, which achieves the highest accuracy, with an R^2 of 0.943 and a low RMSE of 0.549 mm, and effectively captures both the trend and the 95% confidence intervals.

We then turn to dimensionality reduction techniques. Panels (c) and (d) of Fig. 14 show the WPCA approach. The cumulative variance plot in panel (c) shows that six components are required to explain 85% of the variance. The reconstruction in panel (d) is reasonably accurate but noisier than the GP model, with an R^2 of 0.816 and RMSE of 0.989 mm. Finally, panels (e) and (f) present the EMPCA results. The variance plot in panel (e) suggests that a single component meets the threshold, leading to a highly smoothed reconstruction in panel (f). This oversimplification yields the poorest performance, with the lowest R^2 of 0.637 and the highest RMSE of 1.391 mm, indicating that one component cannot capture the complex variability of the original signal. Additional examples for real GNSS data with vertical trends at the ANTA station are provided in Appendix E of the supplementary materials.

Overall, the field data analysis strongly reinforces our synthetic test conclusions. GPR provides the most accurate and reliable gap filling for real GNSS time series, delivering well quantified uncertainties. The PCA based methods are less consistent, with EMPCA particularly unsuitable for complex horizontal signals. GPR also offers flexibility in kernel design, allowing simplified configurations that retain accuracy and uncertainty quantification for trend dominated vertical components.

5 Discussion

5.1 Interpretation of Core Findings

The comparative analysis confirms that the performance of each gap-filling method is fundamentally shaped by its underlying modelling paradigm. The hierarchical GPR framework achieved superior accuracy, particularly for long gaps and station-specific signals, by directly modelling multi-scale temporal processes. Its composite kernel decomposes the covariance structure into physically interpretable components namely secular trends, seasonal cycles, transient deformations, and noise, providing a principled, station-adaptive interpolation that preserves local anomalies critical for operational monitoring. This temporal focus enables reliable operation even on isolated stations or during network outages, addressing a key operational vulnerability.

Conversely, the PCA-derived algorithms (WPCA and EMPCA) are fundamentally constrained by their reliance on spatial covariance. While efficient for reconstructing network-wide common-mode signals during periods of high station availability, they inevitably smooth out station-specific deformation, a significant drawback near individual extraction wells where local signals dominate. Their performance degrades with reduced network density, creating a practical paradox: they are least reliable when robust gap-filling is most needed. EMPCA's further simplification to a single dominant component, while computationally attractive, sacrifices too much fidelity for all but the simplest vertical trends. Therefore, method selection involves a trade-off between

statistical rigour (GPR) and computational efficiency (WPCA/EMPCA), mediated by network density and signal complexity.

5.2 Relation to Existing Literature

Our findings align with and extend the established understanding of GNSS time series analysis. The robust performance of hierarchical GPR supports its growing recognition in geodesy for tasks such as sea-level reconstruction and signal decomposition [31, 32, 35]. This study advances the application by developing a composite, physics-informed kernel specifically for the gap-filling challenges of monitoring deformation in anthropogenic settings, where preserving non-stationary, multi-scale signals is essential.

Several methodological approaches for GNSS gap filling have emerged in recent years. First, Liu et al. [49] developed a Matlab-based Kriged Kalman filter that combines least squares prediction with Kalman filtering to interpolate missing data. This hybrid approach leverages spatial correlation between neighbouring stations, similar in spirit to PCA-based methods, but within a stochastic filtering framework. While effective for networks with strong spatial coherence, its performance naturally degrades when stations operate in isolation or during widespread network outages, a limitation shared with PCA methods and one that our station-wise GPR framework circumvents. Second, the reliable performance of WPCA and EMPCA during periods of strong spatial correlation reinforces well-documented principles about PCA-based reconstruction in dense networks [22, 25]. However, we provide critical empirical evidence of their practical limitations, which were previously acknowledged mainly in theory [26]. Specifically, we demonstrate how the smoothing of localised deformation and acute sensitivity to network sparsity can bias results in heterogeneous industrial settings, a vital caveat for practitioners.

Our work builds most directly on recent Gaussian process applications in geodesy. Xu et al. [35] proposed a GPR framework for reconstructing geodetic time series with missing data and time-varying seasonal signals. While their study shares our focus on GP-based imputation, several key distinctions set our work apart. First, Xu et al. [35] emphasise the estimation of time-varying seasonal signals through quasi-periodic kernels, whereas our composite kernel is explicitly designed to capture the multi-scale processes characteristic of anthropogenic deformation, including a dedicated flicker noise component informed by observed noise characteristics in our study area. Second, their validation focuses on simulated and real GNSS data from tectonically active regions, while we target the specific challenges of monitoring human-induced deformation, a context where transient, non-stationary signals demand particularly robust gap filling. Third, we provide a systematic comparison with PCA-based methods (WPCA and EMPCA) under both synthetic and operational conditions, offering practitioners clear, evidence-based guidance that extends beyond a single methodological proposal.

Furthermore, our results underscore the substantial advantage of both GPR and advanced PCA variants over traditional interpolation (e.g., linear splines, averaging). Such simplistic methods fail to model the coloured noise and seasonal components inherent in GNSS data, often introducing artefacts that compromise velocity estimates and noise characterisation [16]. By conducting a direct, systematic comparison using

shared metrics and realistic data, this study moves beyond isolated case studies to provide evidence-based guidance, responding to a recognised need in recent literature [14, 29].

5.3 Guidelines for Application

Based on our comparative findings, we propose the following practical guidelines for operational monitoring and modern data-processing workflows.

- Operational monitoring (e.g., hydrocarbon fields, gas storage sites): The hierarchical GPR framework provides critical advantages in resilience and risk awareness. Its station-independent nature ensures reliable assessment continues at safety-critical locations even during wider network outages. Its inherent uncertainty quantification delivers confidence-bounded data, enabling risk-informed decision-making for tasks such as subsidence forecasting or regulatory compliance.
- Network-wide processing of large datasets with predominantly short, random gaps: For short gaps and when uncertainty quantification is not a primary concern, WPCA remains an efficient and effective choice, especially when a strong common-mode signal is present. EMPCA can be a suitable probabilistic alternative for similar scenarios, offering a balance between simplicity and spatial inference.
- Data preprocessing in ML pipelines: Many ML algorithms require continuous time series as input and cannot handle missing data natively. The choice of gap-filling method directly influences downstream model performance. Techniques that preserve the statistical and noise properties of the data, such as GPR or weighted PCA, create a more robust foundation, reducing the risk of algorithms learning artefacts introduced by poorer imputation methods.
- Simplified decision framework can be based on three key factors: i) For long gaps (exceeding one year), GPR is generally preferable when the signal follows patterns consistent with surrounding data. If long gaps coincide with periods where unmodelled transient activity (e.g., post-seismic deformation, volcanic unrest, anthropogenic events) is anticipated, spatial methods may offer complementary value, though their reliance on neighbouring stations must be weighed against network density; ii) For sparse or isolated stations, GPR is the only viable option, whereas for dense networks, WPCA or EMPCA provide efficient alternatives, and iii) If uncertainty quantification is required, GPR is essential; if computational speed is the priority, WPCA or EMPCA are more suitable.

5.4 Limitations of the Study

Several limitations contextualise the findings of this study and warrant careful consideration when interpreting our results and recommendations.

First, the computational demand of the hierarchical GPR framework, primarily from optimising its composite kernel's hyperparameters, is inherently higher than the linear algebra underlying PCA-based methods. Despite our use of parallel processing, this presents a scalability challenge for real-time processing of continental-scale networks with hundreds of stations [30, 35].

Second, our station-wise GPR fills gaps using only temporal patterns learned from the target station. It cannot detect events occurring entirely within a data gap: if an unmodelled transient (e.g., post-seismic deformation or anthropogenic signal) takes place when a station is offline, GPR will be unaware of it. This model dependency extends to kernel design. Our composite kernel suits signals with secular trends, annual cycles, and specific noise types observed in the North Adriatic, but applying it to different geophysical transients would require kernel adaptation [29]. GPR can be extended to incorporate spatial information [31], but multi-station GPR falls outside this study's scope. WPCA and EMPCA, by contrast, leverage spatial correlations. During localised gaps, they can draw on neighbouring stations to potentially reconstruct anomalous signals. However, this advantage depends on reliable spatial structure, PCA methods tend to smooth away station, specific details, and the benefit vanishes during widespread network outages, precisely when robust gap filling is most needed.

Third, the method assumes GP priors and Gaussian residuals [50]. While flexible, this assumption may not capture extreme non-Gaussian anomalies present in real GNSS data, particularly during instrumental issues or rapid deformation events.

Lastly, while synthetic validation provides controlled tests, it relies on a simplified analytical displacement model (Mogi). This model does not capture the full complexity of poroelastic reservoir deformation and heterogeneous subsurface properties [8, 12], underscoring that our field data validation remains an essential component of the assessment.

5.5 Future Research Directions

Building on this work, future research should address the following interconnected challenges:

- Scalability of GPR: Developing reduced-order, sparse, or variational approximations of the GPR framework is necessary to improve its computational efficiency for real-time, and large-scale applications [35].
- Physics-informed kernels: Integrating more sophisticated geomechanical models of reservoir deformation (e.g., poroelastic, fault-driven) directly into the GP covariance structure would move beyond the simplified Mogi source, enhancing physical interpretability [8, 12, 51].
- Benchmarking with emerging methods: The evaluation framework should be expanded to include deep-learning-based imputation models (e.g., transformers, neural processes), establishing a comprehensive benchmark for next-generation, and data-driven techniques in geoscience [14, 17].
- Cross-dataset generalisability: Applying the developed GPR framework to other geodetic time series with analogous gap and noise characteristics, such as InSAR displacement fields or GRACE/GRACE-FO gravity data [35, 52], would test its generalisability and potentially unify processing approaches across geophysical datasets.
- Quantitative runtime benchmarking: While our study establishes the accuracy advantages of GPR, a systematic runtime comparison across varying network sizes

and gap configurations would provide valuable guidance for practitioners. Preliminary tests suggest GPR is approximately 5–10 times slower per station than WPCA, but a dedicated benchmarking study could quantify this trade-off more rigorously and explore pathways for computational optimisation.

6 Conclusion

Methodological trade-offs in reconstructing GNSS data gaps for monitoring deformation caused by subsurface fluid extraction and storage have been clarified through a direct, evidence-based comparison between spatially correlated two PCA variants (WPCA and EMPCA) and a temporally focused hierarchical GPR framework.

The analysis demonstrates that the choice of method is fundamentally a choice between modelling paradigms. Extensions of PCA are powerful for leveraging spatial correlation in dense networks but are inherently limited by it, struggling with station-specific signals and sparse data scenarios. Conversely, the GPR framework circumvents this spatial dependency by using a physics-informed temporal model to provide accurate, station-specific reconstructions with principled uncertainty, a critical capability for operational risk assessment.

This work contributes an optimised, open-source implementation of the GPR framework and establishes that advanced gap-filling is a crucial preprocessing step that safeguards the integrity of downstream analyses, from velocity estimation to machine-learning applications. For practitioners, the decision can be guided by a simple rule: select GPR for critical stations, long gaps, or when uncertainties matter; apply WPCA or EMPCA for efficient, network-wide processing of short gaps in dense arrays. Moving forward, the integration of such probabilistic, physically interpretable models into operational processing chains represents a key step toward more resilient and transparent monitoring of human-induced deformation.

Data availability statement. The field GNSS dataset is available upon request from Serpelloni et al. [44]. The Python code developed for this study, along with the complete synthetic dataset (including both complete and gapped time series for all 70 stations), has been made publicly available on Zenodo (https://doi.org/10.5281/zenodo.18617705).

Acknowledgments. This work was supported by funding from the Istituto Nazionale di Geofisica e Vulcanologia (INGV).

Conflicts of Interest. The authors declare that they have no known competing financial interests or personal relationships that could have influenced the work reported in this paper.

Author's contribution. DTV and AG: conceptualisation, methodology, and numerical computation; DTV and FP: data analysis, experimentation, coding, and original draft preparation; ES: data curation; ES and MN: review and editing of the manuscript; AG and GP: method validation and final manuscript editing; GP: funding acquisition; DTV and all authors participated in discussions throughout the project.

References

[1] Blewitt, G., Hammond, W., Kreemer, C.: Harnessing the gps data explosion for interdisciplinary science. Eos **99**(2), 2020943118 (2018)

[2] Baldi, P., Casula, G., Cenni, N., Loddo, F., Pesci, A.: Gps-based monitoring of land subsidence in the po plain (northern italy). Earth and Planetary Science Letters **288**(1-2), 204–212 (2009)

[3] Devoti, R., Pietrantonio, G., Riguzzi, F.: Gnss networks for geodynamics in italy. Física de la Tierra **26**(2014), 11–24 (2014)

[4] Gandolfi, S., Macini, P., Poluzzi, L., Tavasci, L.: Gnss measurements for ground deformations detection around offshore natural gas fields in the northern adriatic region. Proceedings of the International Association of Hydrological Sciences **382**, 89–93 (2020)

[5] Teatini, P., Castelletto, N., Ferronato, M., Gambolati, G., Janna, C., Cairo, E., Marzorati, D., Colombo, D., Ferretti, A., Bagliani, A., et al.: Geomechanical response to seasonal gas storage in depleted reservoirs: A case study in the po river basin, italy. Journal of Geophysical Research: Earth Surface **116**(F2) (2011)

[6] Pezzo, G., Petracchini, L., Devoti, R., Maffucci, R., Anderlini, L., Antoncecchi, I., Billi, A., Carminati, E., Ciccone, F., Cuffaro, M., *et al.*: Active fold-thrust belt to foreland transition in northern adria, italy, tracked by seismic reflection profiles and gps offshore data. Tectonics **39**(11), 2020–006425 (2020)

[7] Geertsma, J.: Land subsidence above compacting oil and gas reservoirs. Journal of petroleum technology **25**(06), 734–744 (1973)

[8] Segall, P.: Induced stresses due to fluid extraction from axisymmetric reservoirs. pure and applied geophysics **139**(3), 535–560 (1992)

[9] Hoffmann, J., Galloway, D.L., Zebker, H.A.: Inverse modeling of interbed storage parameters using land subsidence observations, antelope valley, california. Water Resources Research **39**(2) (2003)

[10] Xuejun, Q., Wei, C., Dijin, W., Zhaosheng, N., Zhengsong, C., Jie, L., Xiaoqiang, W., Yu, L., Tan, W., Guangcai, F.: Crustal deformation in the hutubi underground gas storage site in china observed by gps and insar measurements. Seismological Research Letters **89**(4), 1467–1477 (2018)

[11] Teatini, P., Gambolati, G., Ferronato, M., Settari, A.T., Walters, D.: Land uplift due to subsurface fluid injection. Journal of Geodynamics **51**(1), 1–16 (2011)

[12] Codegone, G., Rocca, V., Verga, F., Coti, C.: Subsidence modeling validation through back analysis for an italian gas storage field. Geotechnical and Geological

Engineering **34**(6), 1749–1763 (2016)

[13] Böhm, J., Schuh, H.: Atmospheric Effects in Space Geodesy vol. 5. Springer, Berlin (2013)

[14] Gao, W., Li, Z., Chen, Q., Jiang, W., Feng, Y.: Modelling and prediction of gnss time series using gbdt, lstm and svm machine learning approaches: W. gao et al. Journal of Geodesy **96**(10), 71 (2022)

[15] Haji-Aghajany, S., Tasan, M., Izanlou, S., Rohm, W.: Tropodeep: a deep learning-based model for insar tropospheric correction on large-scale interferograms using gnss and wrf outputs. Journal of Geodesy **99**(10), 76 (2025)

[16] Langbein, J.: Noise in gps displacement measurements from southern california and southern nevada. Journal of Geophysical Research: Solid Earth **113**(B5) (2008)

[17] Del Castillo, E., Colosimo, B.M., Tajbakhsh, S.D.: Geodesic gaussian processes for the parametric reconstruction of a free-form surface. Technometrics **57**(1), 87–99 (2015)

[18] Bos, M., Fernandes, R., Williams, S., Bastos, L.: Fast error analysis of continuous gnss observations with missing data. Journal of Geodesy **87**(4), 351–360 (2013)

[19] Dong, D., Fang, P., Bock, Y., Cheng, M., Miyazaki, S.: Anatomy of apparent seasonal variations from gps-derived site position time series. Journal of Geophysical Research: Solid Earth **107**(B4), 9 (2002)

[20] Zhigljavsky, A.A.: Singular spectrum analysis for time series: Introduction to this special issue. Statistics and its Interface **3**(3), 255–258 (2010)

[21] Mallat, S.: A Wavelet Tour of Signal Processing, Academic Press (1999)

[22] Gualandi, A., Serpelloni, E., Belardinelli, M.E.: Blind source separation problem in gps time series. Journal of Geodesy **90**(4), 323–341 (2016)

[23] Mandler, E., Pintori, F., Gualandi, A., Anderlini, L., Serpelloni, E., Belardinelli, M.E.: Post-seismic deformation related to the 2016 central italy seismic sequence from gps displacement time-series. Journal of Geophysical Research: Solid Earth **126**(9), 2021–022200 (2021)

[24] Pintori, F., Serpelloni, E., Gualandi, A.: Common-mode signals and vertical velocities in the greater alpine area from gnss data. Solid Earth **13**(10), 1541–1567 (2022)

[25] Dong, D., Fang, P., Bock, Y., Webb, F., Prawirodirdjo, L., Kedar, S., Jamason, P.: Spatiotemporal filtering using principal component analysis and karhunen-loeve expansion approaches for regional gps network analysis. Journal of geophysical

research: solid earth **111**(B3) (2006)

[26] Delchambre, L.: Weighted principal component analysis: a weighted covariance eigendecomposition approach. Monthly Notices of the Royal Astronomical Society **446**(4), 3545–3555 (2015)

[27] Bailey, S.: Principal component analysis with noisy and/or missing data. Publications of the Astronomical Society of the Pacific **124**(919), 1015 (2012)

[28] Hines, T., Hetland, E.: Revealing transient strain in geodetic data with gaussian process regression. Geophysical Journal International **212**(3), 2116–2130 (2018)

[29] Zarrinnegar, K., Sormayli, J., Mosavi, M.: Robust gnss spoofing detector using optimized machine learning model on embedded platforms. GPS Solutions **30**(1), 1–21 (2026)

[30] Williams, C.K., Rasmussen, C.E.: Gaussian Processes for Machine Learning vol. 2. MIT press Cambridge, MA (2006)

[31] Qiu, X., Wang, F., Zhang, Q., Tao, G., Zhou, S.: An improved gaussian process for filling the missing data in gnss position time series considering the influence of adjacent stations. Scientific Reports **14**(1), 19268 (2024)

[32] Kopp, R.E., Horton, B.P., Kemp, A.C., Tebaldi, C.: Past and future sea-level rise along the coast of north carolina, usa. Climatic Change **132**(4), 693–707 (2015)

[33] Khelifa, S., Gourine, B., Taibi, H., Dekkiche, H.: Filling gaps in time series of space-geodetic positioning. Arabian Journal of Geosciences **11**(12), 329 (2018)

[34] Bao, Z., Chang, G., Zhang, L., Chen, G., Zhang, S.: Filling missing values of multi-station gnss coordinate time series based on matrix completion. Measurement **183**, 109862 (2021)

[35] Xu, K., Hu, S., Jin, S., Li, J., Zheng, W., Wang, J., Zhu, Y., Li, K., Ren, A., Liu, Y.: Reconstruction of geodetic time series with missing data and time-varying seasonal signals using gaussian process for machine learning. GPS Solutions **28**(2), 79 (2024)

[36] Li, W., Jiang, W., Li, Z., Chen, H., Chen, Q., Wang, J., Zhu, G.: Extracting common mode errors of regional gnss position time series in the presence of missing data by variational bayesian principal component analysis. Sensors **20**(8), 2298 (2020)

[37] Ding, X., Liu, J., Yang, F., Cao, J.: Random radial basis function kernel-based support vector machine. Journal of the Franklin Institute **358**(18), 10121–10140 (2021)

[38] Borovitskiy, V., Terenin, A., Mostowsky, P., *et al.*: Matérn gaussian processes on

riemannian manifolds. Advances in Neural Information Processing Systems **33**, 12426–12437 (2020)

[39] Pearce, T., Tsuchida, R., Zaki, M., Brintrup, A., Neely, A.: Expressive priors in bayesian neural networks: Kernel combinations and periodic functions. In: Uncertainty in Artificial Intelligence, pp. 134–144 (2020). PMLR

[40] Shileche, E.A., Weke, P., Achia, T.: Kernel density estimation of white noise for non-diversifiable risk in decision making. Journal of Risk Analysis and Crisis Response **10**(1), 6–11 (2020)

[41] Williams, S.: The effect of coloured noise on the uncertainties of rates estimated from geodetic time series. Journal of Geodesy **76**(9), 483–494 (2003)

[42] Jeong, J., Jun, M.: A class of matérn-like covariance functions for smooth processes on a sphere. Spatial Statistics **11**, 1–18 (2015)

[43] Sureiman, O., Mangera, C.M.: F-test of overall significance in regression analysis simplified. Journal of the Practice of Cardiovascular Sciences **6**(2), 116–122 (2020)

[44] Serpelloni, E., Cavaliere, A., Martelli, L., Pintori, F., Anderlini, L., Borghi, A., Randazzo, D., Bruni, S., Devoti, R., Perfetti, P., *et al.*: Surface velocities and strain-rates in the euro-mediterranean region from massive gps data processing. Frontiers in Earth Science **10**, 907897 (2022)

[45] Serpelloni, E., Anderlini, L., Avallone, A., Cannelli, V., Cavaliere, A., Cheloni, D., D'Ambrosio, C., D'Anastasio, E., Esposito, A., Pietrantonio, G., et al.: Gps observations of coseismic deformation following the may 20 and 29, 2012, emilia seismic events (northern italy): data, analysis and preliminary models. Annals of Geophysics **55**(4) (2012)

[46] Timmer, J., Koenig, M.: On generating power law noise. Astronomy and Astrophysics, v. 300, p. 707 **300**, 707 (1995)

[47] Kiyoo, M.: Relations between the eruptions of various volcanoes and the deformations of the ground surfaces around them. Earthq Res Inst **36**(99), 134 (1958)

[48] Wei, L., Yinping, L., Chunhe, Y., Deyi, J., Daemen, J., Jie, C., Junfeng, K.: A new method of surface subsidence prediction for natural gas storage cavern in bedded rock salts. Environmental Earth Sciences **75**(9), 800 (2016)

[49] Liu, N., Dai, W., Santerre, R., Kuang, C.: A matlab-based kriged kalman filter software for interpolating missing data in gnss coordinate time series. Gps Solutions **22**(1), 25 (2018)

[50] Wang, C., Neal, R.M.: Gaussian process regression with heteroscedastic or non-gaussian residuals. arXiv preprint arXiv:1212.6246 (2012)

[51] Dong, D., Herring, T., King, R.W.: Estimating regional deformation from a combination of space and terrestrial geodetic data. Journal of Geodesy **72**(4), 200–214 (1998)

[52] Lecomte, H., Rosat, S., Mandea, M.: Gap filling between grace and grace-fo missions: assessment of interpolation techniques. Journal of Geodesy **98**(12), 107 (2024)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterial.docx