GPS Solutions (2024) 28:79
https://doi.org/10.1007/s10291-024-01616-8

ORIGINAL ARTICLE

Reconstruction of geodetic time series with missing data and time-varying seasonal signals using Gaussian process for machine learning

Keke Xu[1] · Shaobin Hu[1] · Shuanggen Jin[1,2] · Jun Li[1] · Wei Zheng[1] · Jian Wang[1] · Yongzhen Zhu[1] · Kezhao Li[1] · Ankang Ren[1] · Yifu Liu[1]

Received: 21 September 2023 / Accepted: 13 January 2024

Abstract

Seasonal signals in satellite geodesy time series are mainly derived from a number of loading sources, such as atmospheric pressure and hydrological loading. The most common method for modeling the seasonal signal with quasi-period is to use the sine and cosine functions with the constant amplitude for approximation. However, due to the complexity of environmental changes, the time-varying period part is very difficult to model by the geometric or physical method. We present a machine learning method with Gaussian process to capture the quasi-periodic signals in the geodetic time series and optimize the estimation of model parameters by means of maximum likelihood estimation. We test the performance of the method using the synthetic time series by simulating the time-varying and quasi-periodic signals. The results show that the fitting residuals of the new model show a better random fluctuation, while the traditional models still leave the clear periodic systematics signals without being fully modeled. The new model illustrates a higher reliability of linear trend estimation, and a lower uncertainty and model fitting RMSE, even in time series with shorter time span. On the other hand, it shows a strong capacity to restore the missing data and predict the future changes in time series. The method is successfully applied to modeling the real coordinate time series of the GNSS site (BJFS) from IGS network, and the equivalent water height (EWH) time series in North China obtained from gravity satellites. Therefore, it is recommended as an alternative for precise model reconstruction and signals extraction of satellite geodesy time series, especially in modeling the complex time-varying signals, estimating the secular motion velocity, and recovering the large missing data.

Keywords Geodetic time series · Gaussian process · Quasi-periodic signals · GRACE · GNSS

Introduction

With the continuous development of satellite geodesy technology, it has become the most effective space observation technique for monitoring global change and crustal motion and provides valuable basic data for the research of geophysical phenomena at different spatio-temporal scales (Tregoning et al. 2009; Xu et al. 2019). The geodetic time series contain rich information, including tectonic and non-tectonic movement signals, such as ground mass load (e.g., ocean tide, atmosphere, snow, soil water, and nontidal load of the ocean) and model residuals (e.g., physical model residual, nonmodeling error). The seasonal signals may be quasi-periodic due to the complexity of environmental factors, which include not only the period signal of constant amplitude but also the time-varying signal with amplitude variation from year to year (Bogusz and Figurski 2014; Tregoning et al. 2009). According to recent research, the seasonal changes in the various regions worldwide are not consistent in different years, and the response of GNSS stations to environmental change in a seasonal scale is irregular (Kreemer and Blewitt 2021). In the meantime, the noise in geodetic time series is extremely complex, including the white noise, the colored noise, the flicker noise, the power

✉ Keke Xu
xkk@hpu.edu.cn

✉ Shuanggen Jin
sgjin@hpu.edu.cn

1 Present Address: Henan Polytechnic University, Jiaozuo 454000, China

2 Shanghai Astronomical Observatory, Chinese Academy of Sciences, Shanghai 200030, China

law noise, and the random walking. If any seasonal signal or residual periodicity is not properly modeled and removed, it will move the stochastic part to much more correlated noise causing the uncertainties to be artificially overestimated (Bogusz and Klos 2016; Ren et al. 2023). Previous research has shown that ignoring the colored noise will overestimate the velocity error by 2–3 times (Kreemer and Blewitt 2021; Williams 2003). The above complex nonlinear and time-varying characteristics in the satellite observation series are very difficult to model, whether using geophysical or geometric models. Furthermore, due to the various irresistible reasons such as satellite signals disturbance, instrument antenna damage, equipment failure, and replacement upgrading or updating of satellite sensors as well as data loss, most geodetic time series contain a lot of missing data, which may destroy the evenly spaced symmetry and thereby the nature of the covariance matrix (Shen et al. 2014). For example, the aging of Gravity Recovery and Climate Recovery (GRACE) satellite components led to its retirement in 2017 and the launch of the next generation gravity satellite GRACE-FO in 2018, resulting in a data gap for about one year between GRACE and GRACE-FO observations. Therefore, it is important to find an alternative method to fill the data gap between GRACE and GRACE-FO. Although a small number of missing data can be compensated easily by data interpolation, large data gaps are difficult to interpolate across, which brings certain difficulties to the interpretation and extraction of subsequent signals in time series.

Classic modeling methods usually regard seasonal signals as having constant amplitude (Bevis and Brown 2014; Wu et al. 2015), which can no longer satisfy the nonstationary behavior of practical geophysical phenomena. A number of methods have been proposed to detect the quasi-periodic variability in geodetic time series.

(1) Time–frequency analyses methods, such as Jumps Upon Spectrum and Trend JUST (Ghaderpour and Vujadinovic 2020), Least-Squares Wavelet Analysis (LSWA) (Ghaderpour and Pagiatakis 2019), Anti-Leakage Least-Squares Spectral Analysis (ALLSSA), or Least-Squares Spectral Analysis (LSSA) (Ghaderpour and Ghaderpour 2020). LSSA and ALLSSA can accurately estimate the periodic signals but cannot explain the nature of the estimated signals and how the frequencies and amplitudes of components of interest change over time. LSWA can determine periodic and aperiodic signals and show how the signal amplitudes and frequencies change over time. However, aliasing remains a critical issue when estimating signals at high frequencies in coarsely sampled time series (Ghaderpour and Ghaderpour 2020). In JUST, its shortcoming is its sensitivity to the segment size. In certain applications, when there is significant variability of frequency and amplitude within the seasonal component over time, the window size may be defined to have variable sizes for different frequencies to account for all the irregularities.

(2) Filtering methods, such as the Kalman filter (Didova et al. 2016), wavelet decomposition (Bogusz 2015; Ghaderpour and Pagiatakis 2019), and semi-parametric model-based Chebyshev polynomials (Bennett 2008) and Wiener filter-based approaches (Klos et al. 2019), have excellent performance for high signal-to-noise ratios in capturing the varying seasonal signal, but the precision of SSA deteriorates for higher noise levels (Klos et al. 2018). The spatiotemporal filtering methods, considering the spatio-temporal correlation among stations, such as Empirical Orthogonal Function (EOF)/Principal Component Analysis (PCA) (Dong et al. 2006; Shen et al. 2014) and Independent Component Analysis (ICA) (Hyvärinen and Oja 1997), multichannel singular spectrum analysis (MSSA) (Chen et al. 2013; Kondrashov and Ghil 2006), Singular Value Thresholding (SVT) (Bao et al. 2021), Kriged Kalman Filter (Liu et al. 2017), mainly emphasize the smoothing or extracting of common mode errors (CME) or filling of the missing data in time series; little attention is paid to the separation of the varying seasonal signal and subtle deformation in geodetic time series. The result is that the filtered residuals may still contain an artificial signal driven by colored noise and un- or mismodeled geophysical signals. Xu and Yue (2015) emphasize that the seasonal signals filtered may contain an artificial signal. Therefore, some of the power may be artificially removed from power spectra of the residuals, leading to imprecise estimates. Only recently, Koulali and Clarke (2021) use the Gaussian processes to capture the quasi-periodic signals in the time series, but no special attention is paid to the missing data in time series.

Therefore, we apply machine learning method with Gaussian process (GP) to modeling the complex time-varying characteristics, simultaneously, recovering the missing data in the satellite geodesy time series. We first introduce the implementation process of the methodology of GP for machine learning. Then, a lot of simulation experiments are performed to demonstrate the abilities of the approach in modeling the time-variable and quasi-periodic signals from simulated GNSS time series with different time spans, emphasizing the secular velocity and its uncertainty estimation. We also demonstrate the performance in recovering and predicting the missing data. Finally, we apply the method to the real GNSS coordinate time series and GRACE gravity time series in North China.

Methodology

Gaussian process is a machine learning method developed based on statistical learning theory and Bayesian theory. It is a nonparametric modeling method generally used for modeling nonlinear functions. The difference between the Gaussian process and neural network is that a lot of Bayesian regression models based on neural networks can converge to the Gaussian process, in the case of infinite networks. Therefore, the Gaussian process can solve nonlinear problems (Matthias 2008). The key to the Gaussian process is constructing the mean function and covariance kernel function. The mean function is used to depict the relatively long-term changes, while the kernel function is used to capture the quasi-periodic signals in time series. The model parameters are estimated based on Bayesian theory.

It is assumed that the collected observation data $Data = [t, Y].t = [t_1, t_2, \cdots, t_n]^T$, representing the observation epoch vector; $Y = [y_1, y_2, \cdots, y_n]^T$, representing the observation value corresponding to each observation epoch. The observation and random model of time series are written as:

$$Y_i = f(t_i) + \varepsilon_i, i = 1, \cdots, n \quad \varepsilon_i\ N(0, \sigma_n^2) \tag{1}$$

where ε_i is the observation noise; σ_n represents the uncertainty of the position solution of the time series.

The time series is regarded as the Gaussian process; then, the observation value Y satisfies the following multivariate Gaussian distribution:

$$p(Y|t, \beta, \theta) = GP(\mu(t, \beta), \Sigma(t, \theta)) \tag{2}$$

where β and θ is the hyperparameter of the mean function μ and the covariance matrix Σ, respectively.

The mean function is expressed with the seasonal signals with constant amplitude:

$$\mu(t, \beta) = h(t)\beta \tag{3}$$

where $h(t)$ is basis matrix, $h(t) = [1 \quad t \quad \sin(2\pi t) \quad \cos(2\pi t)]$.β is basis coefficient (hyperparameter vector), $\beta = [m \quad n \quad a \quad b]^T$. m is intercept; n is the linear rate; a and b is the signal amplitude of the annual sine and cosine functions. The above parameters together constitute the hyperparameters of the mean function.

Each element of the covariance matrix Σ can be obtained by means of the covariance kernel function κ. The kernel function is the core of a Gaussian process, which determines the properties of a GP. There are various kernel functions, such as RBF kernel, Matern kernel, and exponential kernel. We here employ Matern 3/2 as the kernel function (Zhang et al. 2018):

$$\kappa(t_i, t_j) = \sigma^2\left(1 + \frac{\sqrt{3}r}{\ell}\right)exp\left(-\frac{\sqrt{3}r}{\ell}\right) \tag{4}$$

where $r = \sqrt{(t_i - t_j)^T(t_i - t_j)}$, $\kappa_y(t_i, t_j) = \kappa(t_i, t_j) + \sigma_{\mathrm{n}}^2\delta_{ij}$, δ_{ij} represents Dirac function and its value is 1 when $i = j$, or is 0. Parameters σ^2,ℓ and σ_n form the hyperparameter vector θ .

The predicted epochs are set as t^*, and its corresponding predicted set function values $f(t^*)$ still conform to GP. Then the joint probability distribution of the observation set function values Y and $f(t^*)$ can be represented as:

$$\begin{bmatrix} Y \\ f(t^*) \end{bmatrix} GP\left(\begin{bmatrix} \mu(t) \\ \mu(t^*) \end{bmatrix}, \begin{bmatrix} \kappa(t, t) + \sigma_n^2 I_n & \kappa(t, t^*) \\ \kappa(t^*, t) & \kappa(t^*, t^*) \end{bmatrix}\right) \tag{5}$$

where I_n is the identity matrix; $\kappa(t, t)$ is the GP covariance matrix of $n \times n$ dimensions.

The posterior probability density of $f(t^*)$ still obeys the Gaussian distribution:

$$p(f(t^*)|t, Y) = GP(\mu_{t^*}, \Sigma_{t^*}) \tag{6}$$

$$\begin{aligned} \mu_{t^*} &= \kappa(t^*, t)^T \kappa_y^{-1}(Y - \mu(t)) + \mu(t^*) \\ \Sigma_{t^*} &= \kappa(t^*, t^*) - \kappa(t^*, t)\kappa_y^{-1}\kappa(t, t^*) \end{aligned} \tag{7}$$

where $\kappa_y = \kappa(t, t) + \sigma_n^2 I_n$,$\mu_{t^*}$ is regarded as the predicted results. It can be found that the mean value μ_{t^*} is actually a linear function of the known observation vector. The first part of the covariance term Σ_{t^*} is a priori covariance, and the last part represents the reduction of the uncertainty of the function distribution.

We use the method of maximum likelihood estimation to optimize the hyperparameters. The marginal log-likelihood function of GP model can be represented as

$$\log(p(R|t, \theta, \beta)) = -\frac{1}{2}R^T\kappa_y^{-1}R - \frac{1}{2}\log\left|\kappa_y\right| - \frac{N}{2}\log(2\pi) \tag{8}$$

where N is the number of the training data points;$R = Y - \mu(t, \beta)$, representing the difference between the observed value and the mean function.

Test by synthetic time series

To test the performance of the GP model proposed in the previous section, we produced a synthetic position time series including a long-term linear rate, a constant annual signal, and a sinusoidal variation representing the time-variable part of the quasi-periodic signal. The comparisons are made between the modeled value from of three different model, including

Standard (St) model, the Bennett model, and GP model, and the simulated value.

Data generation

Based on the trajectory motion model as Eq. (9), the synthetic time series with quasi-periodic signals is simulated, as shown in Fig. 1.

$$y(t) = vt + A\sin\left(\frac{2\pi}{T}t\right) + B\cos\left(\frac{2\pi}{T}t\right) + m(t)\sin\left(\frac{2\pi}{T}t + q\right) \tag{9}$$

where A,B represents the amplitudes of the annual signal; $m(t)\sin\left(\frac{2\pi}{T}t + q\right)$ represents the time-varying part of the quasi-periodic signal, which is defined as a sine function with a period of 5 years and an amplitude of 2.4 mm. The long-term linear rate v is set as $5\,mm/yr$. It is generally believed that the optimal random model of the noise characteristics of GNSS position time series is white noise + flicker noise (Jiang et al. 2014; Mao 1999; Williams and Simon 2004); therefore, in the synthetic time series, we add the noise composed of white noise and flicker noise with the amplitude of 0.9 mm and 2.0 mm/yr$^{1/4}$, respectively.

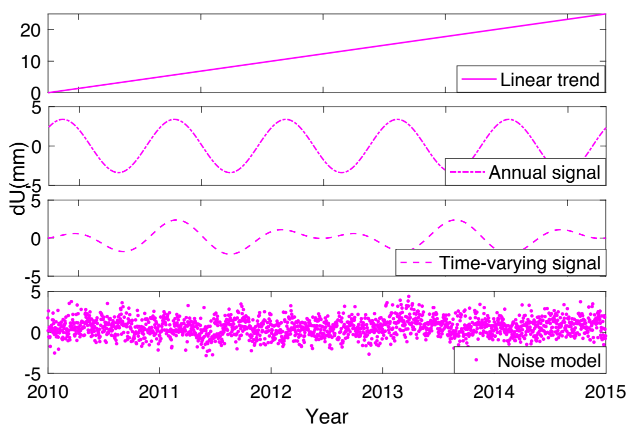


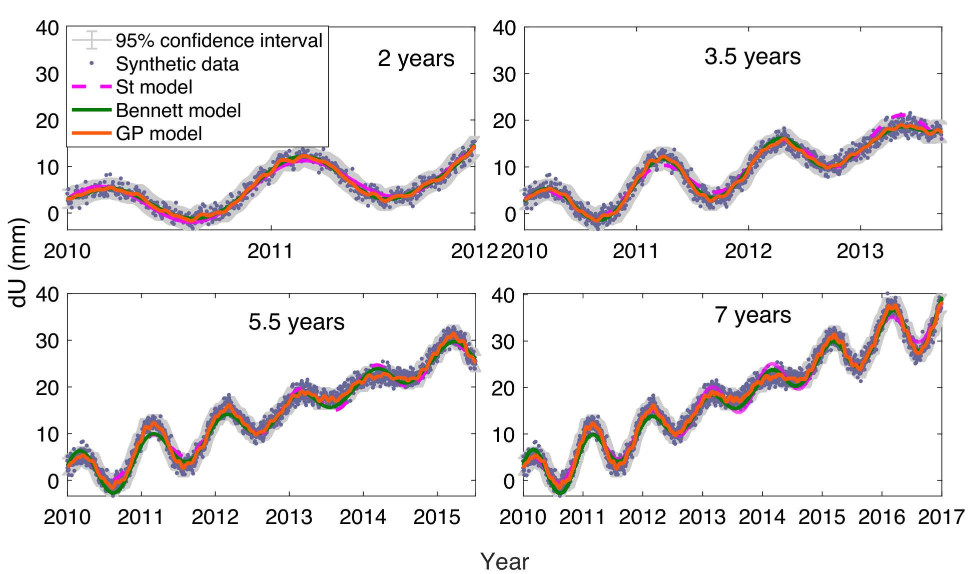
Fig. 1 Composition of the synthetic time series

Results of modeling the quasi-periodic signals

In order to validate the performance of the GP model, we perform the comparisons with the two other methods: (i) Standard model (St model) with the constant periodic amplitude; (ii) Bennett model considering the time-varying seasonal signal. The combination of the Generalized Gauss Markov (GGM) and white noise model is used to parameterize the noise (Koulali and Clarke 2021). The parameters can be estimated by maximum likelihood estimation (MLE) (Bos et al. 2013). Figure 2 shows the fitting results of the three models for the synthetic time series with time spans of 2, 3.5, 5.5, and 7 years. The GP model shows a better fitting effect for the different time spans than the other two models. With the increase of the time span, the fitting effect of the St model becomes worse and worse, while the GP and Bennett model are more stable.

Figure 3 shows the model residuals corresponding to the synthetic time series with a time spans of 7 years. The residuals from the GP model show random fluctuation, while the Standard model result shows obvious periodicity remained in the residual series. This may be due to inadequate modeling of seasonal signals in the Standard model. Although the residual effect of the Bennett model is better than that of the Standard model, it still shows clearer periodic characteristics than the GP model, suggesting that the covariance

Fig. 2 Modeling comparison for the synthetic time series with different time spans based on Standard model, Bennett model, and GP model



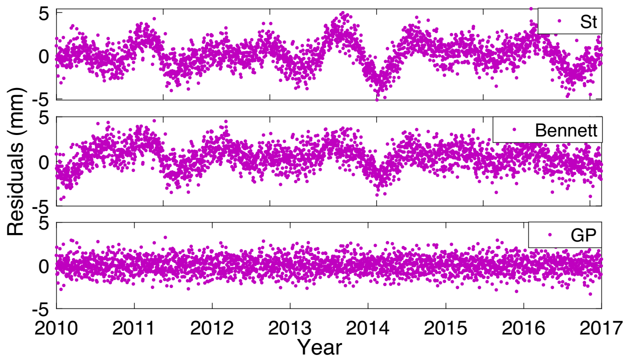


Fig. 3 Model residuals for synthetic time series based on various models

kernel function of GP model is more flexible and can capture the quasi-periodic signals in the time series.

In order to demonstrate the impact of time-varying seasonal signals on estimation of linear motion velocity, we compare the estimated linear velocity from three models and the simulated real value ($5mm/yr$); the results are as shown in Table 1. The Model_ RMSE (Root-Mean-Square Error) is employed as the evaluation index of model fitting results, calculated with the observation and model values, $Model\_\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(observed_t - predicted_t)^2}$. It can be found that whether short or long time span, the velocity estimation effect of the GP model is better than the Standard model and Bennett model. With the increase of time span, the reliability of the velocity estimation of the three models has improved to different degrees; however, the uncertainty of the GP model is much smaller, and its model fitting RMSE is more stable for short- or long-term time spans. This may be related to ignoring the seasonal variation in the Standard model and inadequately modeling quasi-periodic signals in the Bennett model because any not properly modeled and removed seasonal signal or residual periodicity may move the stochastic part to much more correlated noise, causing the uncertainties to be artificially overestimated. It should be noted that the differences of the estimated velocity and its uncertainty from the three models are more significant for the short time series. For the time series with the time span of 2 years, compared with the simulated true value (5 mm/yr), the error of the estimated velocity from the Standard model, the Bennett model, and the GP model is 0.5, 0.4, and 0.3 mm/yr, respectively. The estimation accuracy of the GP model is improved by 42.1% and 29.5% compared with the other two models. Previous research has shown that when the observation time of the GPS continuous observation station time series is less than 2.5 years, the influence of the seasonal term (periodic term) on the velocity field estimation will be enlarged, which will reduce the reliability of the results (Bevis and Brown 2014; Blewitt and Lavallée 2002). Nevertheless, GP model shows a great advantage of modeling time series spanning short periods.

Results of recovering and predicting the missing data

Obtaining a reliable and continuous time series is of great significance for analyzing geophysical events; however, a lot of data are usually missing in geodetic time series. To validate the recovery performance of the missing data from the GP model, we simulate the synthetic time series with the different data-deleted proportions of 10, 20%, 30%, and 40% in the interior and the end of the sequence, respectively. The recovery and prediction results of the missing data are shown in Figs. 4 and 5. The GP model can better reflect the variety of characteristics of the deleted data even when 40% of the data are missed. The recovery effect is less good with the increased proportions of missing data; however, the model RMSE at the missing-data epochs varies very little, with the maximum differences of 0.8 and 0.6 mm for recovery and prediction (see Fig. 6), respectively. The GP model shows few differences in recovering the interior and the end for the same percentage of data missing. Thus, the quantity and distribution of the training samples significantly impact the recovery and prediction of the missing data.

Table 1 Linear rate estimation results and model RMSE for the synthetic time series with the different time spans based on the Standard model, Bennett model, and GP model

Time spans	2 years		3.5 years		5.5 years		7 years	
Modeling results	Velocity (mm/yr)	Model RMSE (mm)	Velocity (mm/yr)	Model RMSE (mm)	Velocity (mm/yr)	Model RMSE (mm)	Velocity (mm/yr)	Model RMSE (mm)
Standard model	5.5 ± 0.5	1.2	5.4 ± 0.7	1.5	5.0 ± 0.8	1.6	5.0 ± 0.7	1.6
Bennett model	4.6 ± 1.6	1.0	5.1 ± 0.1	1.1	5.3 ± 0.4	1.4	5.2 ± 0.3	1.4
GP model	5.3 ± 0.3	0.9	5.0 ± 0.2	1.0	5.1 ± 0.1	0.9	5.0 ± 0.1	1.0

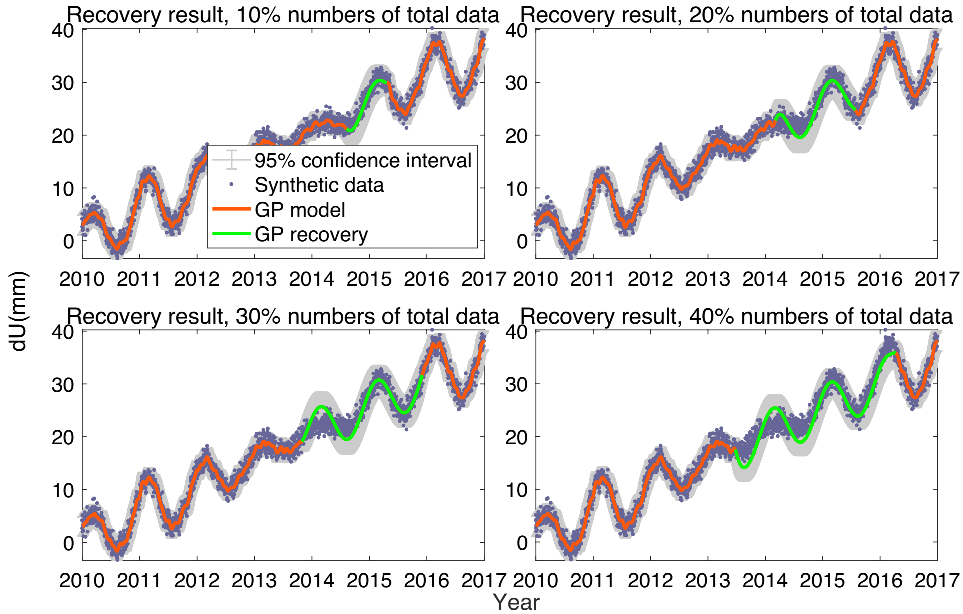


Fig. 4 Recovery results of the different missing data proportions (10, 20, 30, and 40%) based on GP model

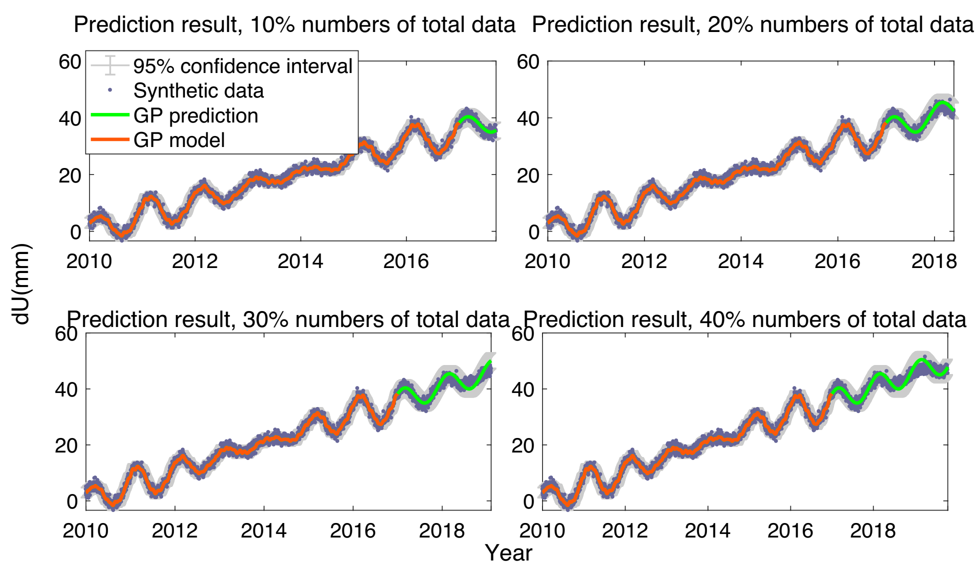


Fig. 5 Prediction results for the different time length (10, 20, 30, and 40%) based on GP model

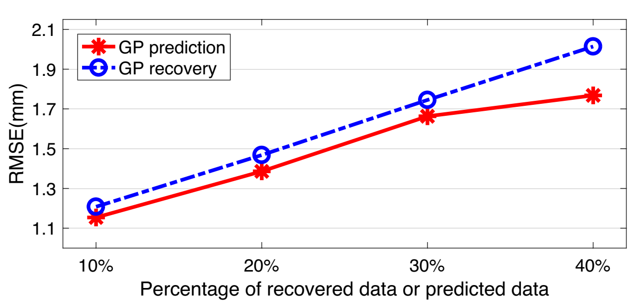


Fig. 6 RMSE variation of recovery and prediction based on GP model with the different missing data percentage

Applications in real geodetic time series

To testify our method, in the following sections, we take North China for research region. We adopt a long-time series for BJFS site from the International GNSS Service (IGS) observation net from 2000 to 2022. In the meantime, we also model the GRACE time series with the same time span in this region for comparisons.

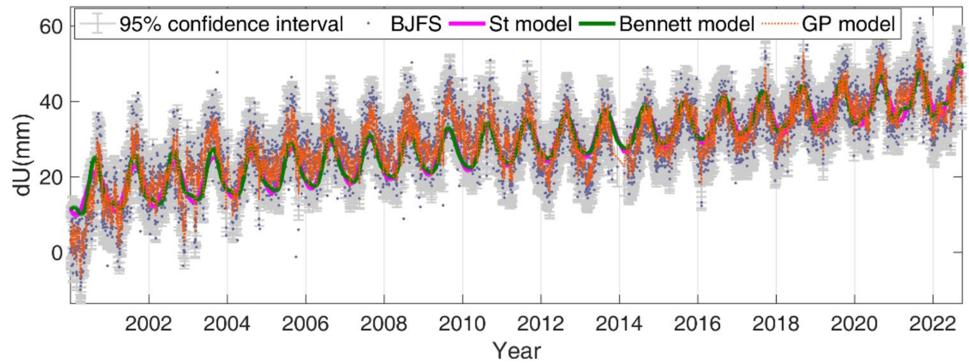


Fig. 7 Modeling comparison of the vertical time series for BJFS site based on three models

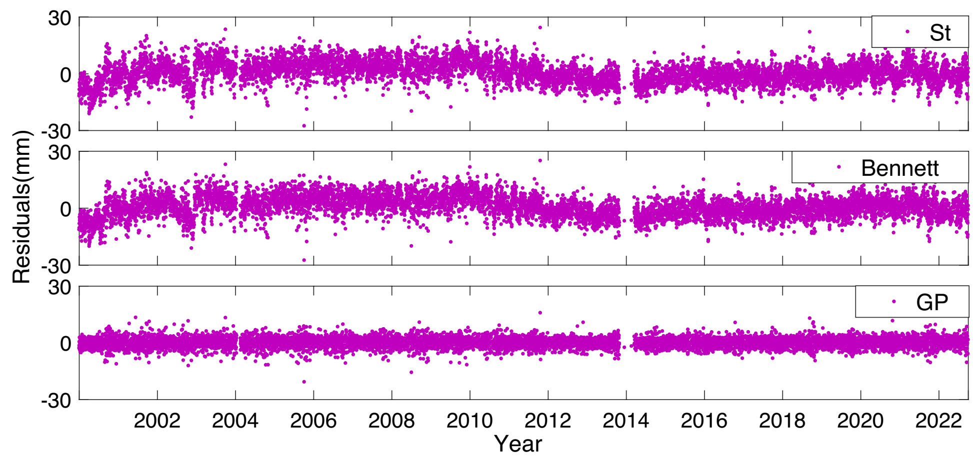


Fig. 8 Model residuals of the vertical time series for BJFS site based on three models (Standard model, Bennett model, and GP model)

GNSS position time series

With the rapid development of GNSS satellite geodesy technology, it has been widely used in monitoring crustal deformation and revealing the dynamics mechanism. GNSS observation time series contain all kinds of geophysical phenomena, including surface and crustal movement, such as the linear variation trend caused by tectonic movement among plates or blocks and the nontectonic deformation caused by the seasonal changes from the atmosphere and hydrosphere loading. It is as yet a challenge to separate the multiple source signals. Most GNSS observation sites show discontinuities and quasi-periodic position signals in the time series, which are usually retained in the time series when estimating velocity fields (Bennett 2008). We selected the GNSS site BJFS, located in North China, from the International GNSS Service (IGS) network for demonstration. The daily coordinate solution for 24-year time span (from October 1999 to July 2022) is obtained by GIPSY/OASIS software (Webb and Zumberge 1993). Figure 7 shows the original coordinate time series of the vertical component and the modeling results. The original time series of this site shows a significant seasonal variation characteristic; however, the GP model can effectively reproduce the complex quasi-periodic seasonal signals much better than others. This may be attributed to the consequence of the kernel function because GP covariance function absorbs not only the quasi-seasonal periods but also other short-term systematics.

Table 2 Linear rate estimation and model RMSE for GNSS time series

Model	Velocity (mm/yr)	Model RMSE (mm)
Standard model	1.2 ± 0.1	5.8
Bennett model	1.2 ± 0.1	5.7
GP model	1.1 ± 0.0	2.7

Fitting residuals from three models (St model, Bennett model, and GP model) are shown in Fig. 8. The Standard model and Bennett model fail to fit the time series, and there are clear systematics left in the residuals, whereas the GP model captures the quasi-periodic signal so well that the residual shows a good random distribution.

Table 2 shows the linear rate estimation value and model RMSE of three models. Although the estimated velocity values from different models are closer, the uncertainty and RMSE differences among three models are very significant.

GP model shows a higher reliability of the velocity. Its accuracy is improved by 82 and 80%, and the model fitting RMSE is reduced by 54 and 52%, respectively. The small differences of the estimated velocity may be mainly attributed to two aspects: One is that the time span is long enough (24 years), which results in less effect on the estimation of long-term motion velocity, and another is that the motion velocity magnitude itself is indeed small in the region.

GRACE gravity time series

GRACE gravity satellite data can help analyze the balance of water quality and water exchange among the atmosphere, land, ocean, and ice sheet and estimate the global or regional changes in land water storage. Due to the massive exploitation of groundwater, there has been a serious problem of water supply and demand in North China, which affects the human living environment and leads to serious vertical motion, seawater intrusion, ecological environment degradation, and perennial rivers drying up or becoming seasonal rivers. Therefore, obtaining the long-term change trend of land water storage is of great practical significance. GRACE time series also contain a significant natural inter-annual change in addition to the prominent seasonal cycles, resulting in a great deviation when estimating the long-term linear trend of quality change or loading deformation. We here apply the GP model to the equivalent water height (EWH) time series in North China (from April 2002 to March 2022) from satellite gravity observations to estimate the long-term linear change trend of land water storage and recover the data blank for 11 months between GRACE and GRACE-FO. We first use the first-order term calculated by Chambers to replace the first-order term in the spherical harmonic coefficient and then employ the data provided by SLR to replace the second-order term. Finally, the decorrelation method of the Duan sliding window removes the north–south stripe error, and the Gaussian filter suppresses the high-order noise with a smoothing radius of 300 km. The calculated EWH time series its spectral analyses are shown in Fig. 9. Obviously, the EWH series contains the complex quasi-periodic signals and the trend variability.

Figure 10 shows the modeling comparison and model residuals for the EWH time series based on the three different models. Although the noise characteristics and sampling rate of GRACE are different from GNSS observations, the GP model can still capture the time-varying signals very well. The residuals do not show significant periodicity compared to the Standard and Bennett models.

We obtain a linear rate of -7.9 ± 0.8 mm/yr and model RMSE of 4.8 mm based on the GP model, which is significantly better than the other two models with the linear rate of -6.3 ± 1.7, -6.3 ± 1.5 mm/yr and model RMSE of 36.0, 33.8 mm (Table 3). Its accuracy is improved by 53% and 47%, and the model fitting RMSE is reduced by 87 and 86%, respectively. Obviously, the quasi-periodic characteristics in the GRACE EWH time series are well reproduced by the GP model, so it has a much better fitting effect, lower RMSE, and higher reliability of linear rate estimation. In the meantime, the missing data for about 11 months between GRACE and GRACE-FO were recovered effectively. It can be clearly seen from Fig. 10 that the shallow groundwater level in North China has risen for two consecutive years (2021–2022) after continuous decline and deficit for about 20 years. Thus, we conclude that the shallow groundwater is reaching a balance of production and replenishment, which can mainly be attributed to the implementation of China's South to North water diversion project.

Interpretation of the quasi-periodic signals in GNSS and GRACE time series

The change of surface hydrological loading may be an important factor causing regional seasonal surface deformation. To demonstrate the seasonal and inter-annual variation extracted by our model, we first obtain the residuals by removing the long-term linear changes from the original GNSS and GRACE time series based on the GP model and then calculate the vertical mass load deformation derived

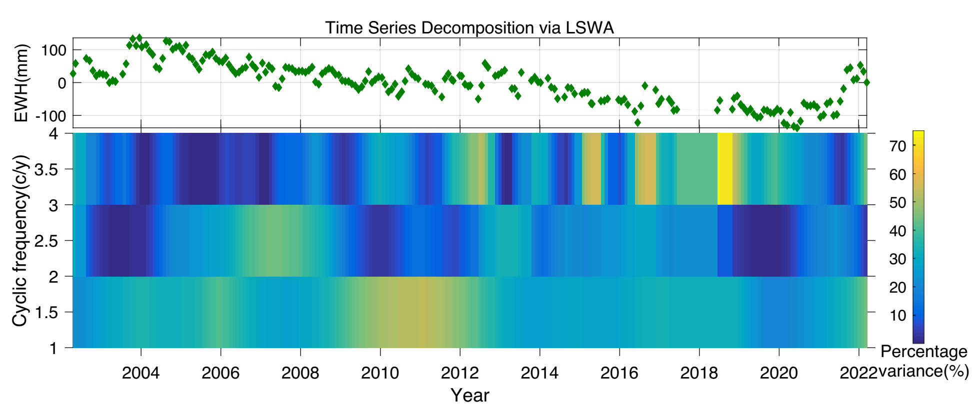


Fig. 9 EWH time series from GRACE data in North China (top) and its spectral analyses (bottom)

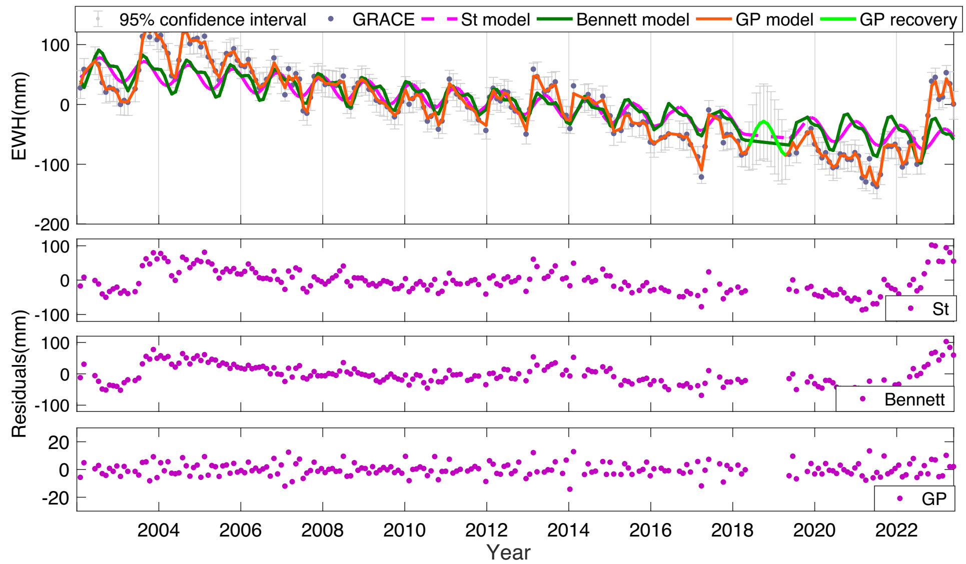


Fig. 10 Modeling comparison (top) and model residuals (bottom) of the EWH time series from GRACE data in North China

Table 3 Linear rate estimation results and model RMSE for GRACE time series

Model	Velocity (mm/yr)	Model RMSE (mm)
Standard model	−6.3 ± 1.7	36.0
Bennett model	−6.3 ± 1.5	33.8
GP model	−7.9 ± 0.8	4.8

from GRACE time series for comparison. Due to the lack of spherical harmonic coefficients of about 11 months between GRACE and GRACE-FO, the deformation value during the corresponding time period cannot be effectively inverted. Therefore, we use the GP model to recover the missing data of GRACE deformation, as shown in Fig. 11. It can be clearly seen that the GNSS vertical displacement time series fitted by the GP model and the surface vertical deformation time series retrieved by GRACE have a good synchronization fluctuation, indicating that the variation of GNSS station coordinate time series in North China is mainly due to the mass loading deformation.

It is worth noting that there is a certain disagreement in the amplitude between the both, which may be mainly attributed to the thermoelastic displacements, temperature variations, discrepancy of spatial resolution, atmospheric, oceanic tides, and other effects except for continental water quality changes. For example, Horwath et al. (2010) find orbit mismodeling, such as solar radiation pressure or Earth albedo, to be the most likely source for inducing large-scale residual patterns. Tregoning et al. (2009) suppose that local processes or site-specific analysis errors dominate their GNSS height estimates as the main error sources. Not considering the atmospheric tides can also cause certain semiannual and annual signals in time series (Tesmer et al. 2011). The annual amplitude in vertical direction caused by temperature variation can reach ~ 2 mm (Wei et al. 2015). Further, due to the fact that short-wavelength loads dominate the signal in a small scale. GNSS is more likely influenced by local effects, such as local site instability, compaction, and decompaction associated with aquifer drawdown and recharge. GRACE,

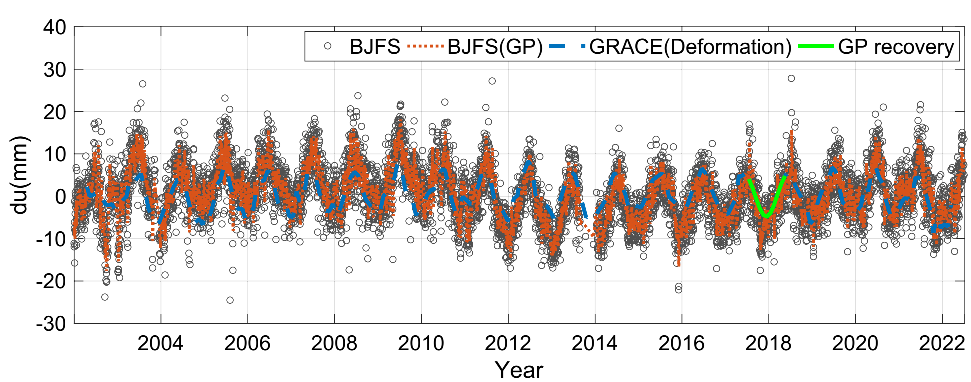


Fig. 11 Modeling comparison between the vertical time series for BJFS site and the deformation time series from GRACE data in North China

on the other hand, produces a broader but lower-amplitude vertical deformation field driven by long-wavelength average of mass load variations. Our GNSS solutions here show a higher amplitude than the GRACE.

Figure 12 compares the modeled EWH change time series with the surface water quality change from the Global Land Data Assimilation System (GLDAS) with the trend removed and the precipitation product from North China's Tropical Rainfall Measuring Mission (TRMM). There is a significantly high correlation among the three, with the peak value of EWH well corresponding to the peak value of precipitation, indicating that the seasonal and inter-annual change is mainly derived from the hydrological loading and water reserve change. In fact, EWH time series from GRACE is terrestrial water storage (TWS) change, including surface and ground water, while GLDAS only represents surface water. The seasonal fluctuation and annual change may be synchronous, but both trends have significant discrepancies. In order to more clearly show the annual fluctuations and the time-varying period signals at the seasonal timescale, we here removed the trend at long-term scale from the modeled EWH and GLDAS time series.

A large number of studies have shown that in addition to the colored noise, there are also CME with unknown sources of space–time correlation between different GNSS reference stations (Xu et al. 2022). The CME may be the main source of GNSS time series error, affecting satellite geodetic time series analysis and speed estimation and signal extraction (Bian et al. 2021). So far, there are few studies on the generation mechanism, cycle and other characteristics, and influencing factors of common mode error.

Discussion

Satellite geodetic time series have been widely used to interpret geophysical phenomena. However, there are usually large discontinuities and complex quasi-periodic position signals that remain in the time series that geometric or geophysical models cannot well explain. GP model proposed in this paper can be used as an alternative to the deterministic method. The method has been successfully applied to model reconstruction and signal extraction of GNSS station position time series and GRACE gravity time series, demonstrating a great advantage in modeling quasi-periodic signals, estimating the long-term motion velocity, and recovering the missing data.

The original purpose of the research is to model the complex quasi-periodic signals in the geodetic series and optimize the trend parameters estimation and precision evaluation. This is the topic of the paper; therefore, the special attention is paid to the modeling of quasi-periodic signals and parameters estimation in the paper. The advantage of GP model is that it is a nonparametric modeling method, which can flexibly model the complex nonlinear signals without obtaining the signal form in advance. The mean function depicts long-term trend change, while the quasi-periodic and the other spurious signals in time series are reproduced by the covariance kernel function. Meantime, model parameters are estimated by Maximum Likelihood Estimation and the accuracy is evaluated based on the error propagation law. The new approach demonstrates better performance and effectiveness with a higher reliability of linear rate estimation, a lower uncertainty, and model fitting RMSE with respect to the conventional models. It can be used as an alternative to the deterministic method to extract long-term motion change in geodetic time series with missing data and time-varying seasonal signals, thus impacting maintenance of geodetic reference frames, geodynamics, geophysics, and crustal deformation analysis.

However, some aspects of the GP model warrant further research. Just as a multivariate normal distribution, the Gaussian process is fully determined by a mean function and a covariance function. The mean function has a vital impact on obtaining the long-term changes of the time series, and the covariance kernel function is the core of the Gaussian process, which determines the effect of reproducing the seasonal variability. Different mean vectors and covariance kernel matrices may significantly impact modeling the time series and estimating the hyperparameters. Therefore, the custom mean and kernels need to be specified according to the need of the practical application.

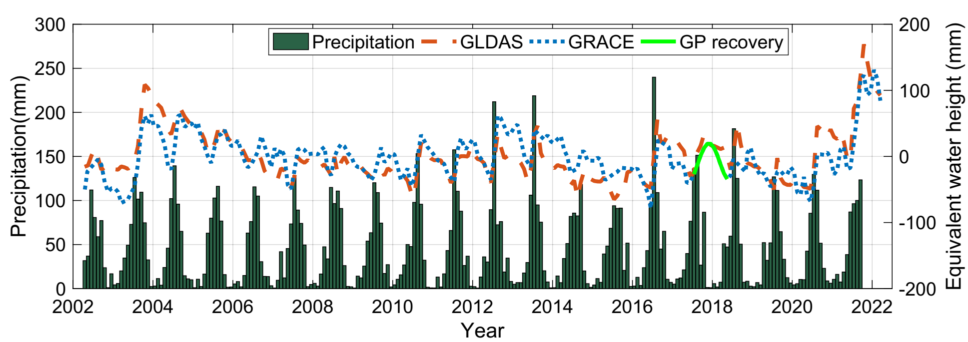


Fig. 12 Comparison between the modeled EWH time series from GRACE, GLDAS, and precipitation

The approach is only based on training and learning of individual components. In fact, three components of the GNSS time series are mutually correlated. Furthermore, all stations are spatially coherent for a regional observation net. If this correlated information can be fully used, it will be great in analyzing signals composition and filling missing data. Due to this spatio-temporal nature, machine learning with large datasets training in both time and space domains are explored; hence, better performance can be expected. It will be our further research work. The GP model proposed is not restricted to GNSS and GRACE observation time series but can also apply to any geodetic time series type. The potential research direction is exploring the possible application of the Gaussian model in the extraction of tectonic signals, for example, separating the common mode errors or quasi-periodic signals from the satellite geodetic time series, improving the signal to noise ratio of observation data, including other aspects of geodetic time series modeling such as transient deformation signals and slow slip of faults (Xu et al. 2019, 2015). Hines and Hetland (2018) have used Gaussian process regression to detect transient strain resulting from slow slip events in the Pacific Northwest.

In addition, in the processing of modeling, we also found the model can be used to restore the missing data or predict the forward series. We try to test the prediction effectiveness only by the simulated experiments (the real value is as known). The feasibility of GP model in restoring and predicting the missing data is preliminarily proved, thus enhancing and extending the applications of the model in GNSS time series analysis for geodesy and geodynamics. Future research may extend the adoption of the various models to restore and predict GNSS time series for three coordinate components and other types of geodetic time series. This will be another topic; various methods have been proposed to reconstruct the missing data such as Singular Spectrum Analysis (SSA), Principal Component Analysis (PCA), Wavelet Decomposition (WD), Kalman Filter (KF).Least Squares (LS) fitting, Boosting Tree (BT), Gradient Boosting Decision Tree (GBDT), Long Short-Term Memory(LSTM), Support Vector Machine (SVM). The effectiveness of reconstructing the missing data has a strong dependence with the percentages of missing data and noise levels of time series. It may enable the inclusion of more and different impact factors or features for individual sites or data types and generate more complete knowledge about using GP approaches for geodetic data analytics and applications.

Conclusion

Due to the influence of the complex environment and observation error, the amplitude of seasonal signals is time-varying from year to year, which brings certain difficulties and challenges in modeling and recovering the satellite geodetic time series. Here, we propose an excellent method for modeling the quasi-periodic signals based on the Gaussian process for machine learning. The experiment results based on the synthetic and real time series show that the fitting effect of the GP model is significantly better than the traditional Standard model with the periodic signals by the constant amplitude and the existing Bennett model with time-varying signals. The accuracy of parameter estimation is improved by more than 80%, and the model fitting RMSE is reduced by more than 52%. The residual of the GP model shows random distribution, while the traditional method still leaves the clear periodic systematics not fully modeled, which means the noise on the velocity estimation for GP model observes little influence. The GP constitutes an excellent approach for modeling complex seasonal signals and noise, especially in mitigating biases associated with complex quasi-periodic signals when estimating secular velocities or other signals. It also shows a great advantage in the recovery or prediction of large missing data in geodetic time series.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10291-024-01616-8.

Acknowledgements The research is supported by the National Natural Science Foundation of China (41774041). We are grateful to International GNSS Service (IGS) for providing GNSS data.

Author's contribution KX proposed this study and wrote the manuscript. SH conducted the numerical computation and experiments. SJ modified the manuscript. JL and JW wrote the program code. WZ reviewed and modified the manuscript. YZ and AR analyzed and validated the method. KL and YL revised the manuscript. KX and All authors were involved in discussions throughout the development.

Data availability The software package used in this paper can be downloaded from the website (https://github.com/SBH08180815/GP_Time_Series_Tool). GNSS data used in this study were obtained from Nevada Geodetic Laboratory (http://geodesy.unr.edu). The data for GRACE are available at http://icgem.gfz-potsdam.de/series. And the data for GLDAS and Precipitation are available at https://disc.gsfc.nasa.gov/.

Declarations

Conflict of interest The authors declare no conflict of interest.

References

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

Bennett RA (2008) Instantaneous deformation from continuous GPS: contributions from quasi-periodic loads. Geophys J Int 174:1052–1064

Bevis M, Brown A (2014) Trajectory models and reference frames for crustal motion geodesy. J Geodesy 88:283–311

Bian Y, Yue J, Ferreira VG, Cong K, Cai D (2021) Common mode component and its potential effect on gps-inferred crustal deformations in Greenland. Pure Appl Geophys 178:1805–1823

Blewitt G, Lavallée D (2002) Effect of annual signals on geodetic velocity. J Geophys Res: Solid Earth 107:ETG 9-1-ETG 9-11

Bogusz J (2015) Geodetic aspects of GPS permanent station nonlinearity studies. Acta Geodyn Et Geomater 12(4):180. https://doi.org/10.13168/AGG.2015.0033

Bogusz J, Figurski M (2014) Annual signals observed in regional GPS networks. Acta Geodyn Et Geomater 11:125–131

Bogusz J, Klos A (2016) On the significance of periodic signals in noise analysis of GPS station coordinates time series. GPS Soluti 20:655–664

Bos MS, Fernandes RMS, Williams SDP, Bastos L (2013) Fast Error Analysis of Continuous GNSS Observations with Missing Data. J Geodesy 87:351–360

Chen Q, van Dam T, Sneeuw N, Collilieux X, Weigelt M, Rebischung P (2013) Singular spectrum analysis for modeling seasonal signals from GPS time series. J Geodyn 72:25–35

Didova O, Gunter B, Riva R, Klees R, Roese-Koerner L (2016) An approach for estimating time-variable rates from geodetic time series. J Geodesy 90:1207–1221

Dong D, Fang P, Bock Y, Webb F, Prawirodirdjo L, Kedar S, Jamason P (2006) Spatiotemporal filtering using principal component analysis and Karhunen-Loeve expansion approaches for regional GPS network analysis. J Geophys Res: Solid Earth 111(B03405):1–16

Ghaderpour E, Ghaderpour S (2020) Least-squares spectral and wavelet analyses of V455 andromedae time series: the life after the super-outburst. Publ Astron Soc Pac 132:114504

Ghaderpour E, Pagiatakis SD (2019) LSWAVE: a MATLAB software for the least-squares wavelet and cross-wavelet analyses. GPS Solut 23:50

Ghaderpour E, Vujadinovic T (2020) Change detection within remotely sensed satellite image time series via spectral analysis. Remote Sens 12:4001

Hines TT, Hetland EA (2018) Revealing transient strain in geodetic data with Gaussian process regression. Geophys J Int 212:2116–2130

Horwath M, Rülke A, Fritsche M, Dietrich R (2010) Mass variation signals in GRACE products and in crustal deformations from GPS: a comparison. Springer, Berlin Heidelberg

Hyvärinen A, Oja E (1997) A fast fixed-point algorithm for independent component analysis. Neural Comput 9:1483–1492

Jiang W, Deng L, Zhao L, Zhou X, Liu H (2014) Effects on noise properties of GPS time series caused by higher-order ionospheric corrections. Adv Space Res 53:1035–1046

Klos A, Bos MS, Bogusz J (2018) Detecting time-varying seasonal signal in GPS position time series with different noise levels. GPS Solut 22(1). https://doi.org/10.1007/s10291-017-0686-6

Klos A, Bos MS, Fernandes RMS, Bogusz J (2019) Noise-dependent adaption of the wiener filter for the GPS position time series. Math Geosci 51:53–73

Kondrashov D, Ghil M (2006) Spatio-temporal filling of missing points in geophysical data sets. Nonlin Process Geophys 13:151–159

Koulali A, Clarke PJ (2021) Modelling quasi-periodic signals in geodetic time series using Gaussian processes. Geophys J Int 226:1705–1714

Kreemer C, Blewitt G (2021) Robust estimation of spatially varying common-mode components in GPS time series. J Geodesy. https://doi.org/10.1007/s00190-020-01466-5

Liu N, Dai W, Santerre R, Kuang C (2017) A MATLAB-based Kriged Kalman Filter software for interpolating missing data in GNSS coordinate time series. GPS Solut 22:25

Mao A (1999) Noise in GPS coordinate times series. J Geophys Res. https://doi.org/10.1029/1998JB900033

Matthias S (2008) Gaussian Processes for Machine Learning. Int J Neural Syst 14(02):69–106. https://doi.org/10.1142/S0129065704001899

Ren AK, Xu KK, Shao ZH, Liu XQ, Wang XY (2023) Effect of the 2011 Tohoku-Oki earthquake on continuous GNSS station motions. GPS Solut 27:50

Shen Y, Li W, Xu G, Li B (2014) Spatiotemporal filtering of regional GNSS network's position time series with missing data using principle component analysis. J Geodesy 88:1–12

Tesmer V, Steigenberger P, Dam TV, Mayer-Guerr T (2011) Vertical deformations from homogeneously processed GRACE and global GPS long-term series. J Geodesy 85:291–310

Tregoning P, Watson C, Ramillien G, Mcqueen H, Zhang J (2009) Detecting hydrologic deformation using GRACE and GPS. Geophys Res Lett. https://doi.org/10.1029/2009GL038718

Webb FH, Zumberge JF (1993) An introduction to the GIPSY-OASIS-II. JPL Publ. D-11088. In

Wei N, Shi C, Liu JN (2015) Annual variations of 3-D surface displacements observed by GPS and GRACE data:a comparison and explanation. Chin J Geophys 58:3080–3088

Williams Simon DP (2004) Error analysis of continuous GPS position time series. J Geophys Res Solid Earth. https://doi.org/10.1029/2003JB002741

Williams S (2003) The effect of coloured noise on the uncertainties of rates estimated from geodetic time series. J Geodesy 76:483–494

Wu H, Li K, Shi W, Clarke KC, Zhang J, Li H (2015) A wavelet-based hybrid approach to remove the flicker noise and the white noise from GPS coordinate time series. GPS Solut 19:511–523

Xu C, Yue D (2015) Monte Carlo SSA to detect time-variable seasonal oscillations from GPS-derived site position time series. Tectonophysics 665:118–126

Xu KK, Gan WJ, Wu JC (2019) Pre-seismic deformation detected from regional GNSS observation network: a case study of the 2013 Lushan, eastern Tibetan Plateau (China), M_s 70 earthquake. J Asian Earth Sci. https://doi.org/10.1016/j.jseaes.2019.05.004

Xu KK, He R, Li KZ, Ren AK, Shao ZH (2022) Secular crustal deformation characteristics prior to the 2011 Tohoku-Oki earthquake detected from GNSS array, 2003–2011. Adv Space Res 69:1116–1129

Xu KK, Wu JC, Wu WW (2015) Detection of transient aseismic slip signals from GNSS spatial-temporal data. Chin J Geophys 58:2330–2338

Zhang N, Xiong J, Zhong J, Leatham K (2018) Gaussian process regression method for classification for high-dimensional data with limited samples. 358–363. https://doi.org/10.1109/ICIST.2018.8426077

Keke Xu is a professor at Henan Polytechnic University. He received his Ph.D. degree at the School of Geodesy and Geomatics, Tongji University. His research interests are satellite geodesy and application in geoscience.

Wei Zheng is a professor at Henan Polytechnic University. He received his Ph.D. degree from Huazhong University of Science and Technology. His research interests are satellite gravity and altimetry.

Shaobin Hu received his BS. degree from Henan Polytechnic University in 2023. His research focuses on satellite geodesy.

Jian Wang received his BS. degree from Henan Polytechnic University in 2023. His research focuses on GNSS data processing.

Shuanggen Jin is a professor at Henan Polytechnic University and a researcher at Shanghai Astronomical Observatory. He received his Ph.D. degree from the University of Chinese Academy of Sciences. His research interests are satellite navigation and geodesy.

Yongzhen Zhu received his BS. degree from Minjiang University in 2022, and is an MS. student at Henan Polytechnic University. His research focuses on the satellite gravity inversion.

Jun Li received his BS. degree from Henan Polytechnic University in 2023. His research focuses on satellite geodesy.

Kezhao Li is a professor at Henan Polytechnic University. He received his Ph.D. degree from Northwestern Polytechnical University. His research interests are GNSS navigation and positioning.

Ankang Ren received his MS. degree from Henan Polytechnic University in 2022. His research focuses on the GNSS time series analysis.

Yifu Liu received his BS. degree from Henan Polytechnic University in 2022, and is an MS. student at Henan Polytechnic University. His research focuses on the geodetic time series analysis.