




Article

Landslide Forecast by Time Series Modeling and Analysis of High-Dimensional and Non-Stationary Ground Motion Data

Guoqi Qian ^{*,†} , Antoinette Tordesillas  and Hangfei Zheng [†] 

School of Mathematics and Statistics, University of Melbourne, Parkville 3010, Australia;
atordes@unimelb.edu.au (A.T.); hangfei.zheng@unimelb.edu.au (H.Z.)

* Correspondence: qguoqi@unimelb.edu.au; Tel.: +61-3-8344-4899

† These authors contributed equally to this work.

Abstract: High-dimensional, non-stationary vector time-series data are often seen in ground motion monitoring of geo-hazard events, e.g., landslides. For timely and reliable forecasts from such data, we developed a new statistical approach based on two advanced econometric methods, i.e., error-correction cointegration (ECC) and vector autoregression (VAR), and a newly developed dimension reduction technique named empirical dynamic quantiles (EDQ). Our ECC–VAR–EDQ method was born by analyzing a big landslide dataset, comprising interferometric synthetic-aperture radar (InSAR) measurements of ground displacement that were observed at 5090 time states and 1803 locations on a slope. The aim was to develop an early warning system for reliably forecasting any impending slope failure whenever a precursory slope deformation is on the horizon. Specifically, we first reduced the spatial dimension of the observed landslide data by representing them as a small set of EDQ series with negligible loss of information. We then used the ECC–VAR model to optimally fit these EDQ series, from which forecasts of future ground motion can be efficiently computed. Moreover, our method is able to assess the future landslide risk by computing the relevant probability of ground motion to exceed a red-alert threshold level at each future time state and location. Applying the ECC–VAR–EDQ method to the motivating landslide data gives a prediction of the incoming slope failure more than 8 days in advance.

Keywords: empirical dynamic quantiles; error-correction cointegration; landslide forecast; vector autoregression time series



Citation: Qian, G.; Tordesillas, A.; Zheng, H. Landslide Forecast by Time Series Modeling and Analysis of High-Dimensional and Non-Stationary Ground Motion Data. *Forecasting* **2021**, *3*, 850–867. <https://doi.org/10.3390/forecast3040051>

Academic Editor: Sonia Leva

Received: 7 September 2021

Accepted: 8 November 2021

Published: 12 November 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Geo-hazard events, including earthquakes and landslides, can present significant damage to the environment and society; cf. [1,2]. Developing the capability to reliably predict the imminent occurrence of such a disaster is crucial for enhancing society's preparedness and mitigating the disaster's impact. However, behaviors of the geological processes underpinning such geo-hazard events are mostly complicated in both space and time, in that the multivariate time series data collected from monitoring these processes by modern techniques, such as interferometric synthetic-aperture radar (InSAR), are often high-dimensional and non-stationary. To date, there exist few statistical methods capable of analyzing high-dimensional non-stationary InSAR time series for timely and reliable forecasts of geo-hazard events [3]. It was therefore the goal of this research to develop an effective statistical method for modeling high-dimensional, non-stationary vector time series and accordingly for timely forecasting with precision. The capability of the method was tested on a real landslide dataset detailed in [4]. Briefly, the landslide dataset comprises 1803 time series of InSAR data on ground displacement, observed every 6 min for 5090 times and at 1803 locations.

Analyzing multivariate time series is often based on a vector auto-regression (VAR) model framework; see Chapter 2 in [5]. A prerequisite for the parameter estimation involved in the VAR-based analysis to be statistically consistent is that the underlying

multivariate time series be stationary. This condition is, however, not satisfied for our landslide data which can be shown to be unit-root non-stationary with significant statistical evidence. Therefore, we cannot use the VAR model to consistently fit the landslide data and estimate the involved parameters directly. Instead, we resort to searching for certain linear transformation of the data consisting of lagged time-difference operations, by which the non-stationary landslide time series can be converted to be stationary before applying the stationary VAR methodology. The technique underpinning such linear transformation is the so-called error-correction cointegration (ECC) method; cf. Chapter 5 in [5].

Using ECC and VAR to analyze non-stationary vector time series can be computationally infeasible if the vector dimensionality is too high, e.g., $k = 1803$ for the landslide data, because the number of unknown parameters involving statistical inference will be of order $O(k^2)$. Recently, reference [6] proposed a dimension reduction technique, called empirical dynamic quantiles (EDQ), to represent a high-dimensional vector time series by a low-dimensional subset of the former with negligible loss of information. The referenced subset is made of the EDQ series with pre-specified quantile levels. For example, the 1803 landslide time series may be represented by 11 EDQ series at levels $0.0(\min), 0.1, \dots, 0.9, 1.0(\max)$. Details are provided later in Section 4. VAR model fitting can then be computationally feasibly applied to analyze the small number of EDQ series just determined. Moreover, the results from analyzing the EDQ series are readily extendable to the original high-dimensional time series by numerical interpolation, because the EDQ technique is able to calculate a quantile level value for each scalar time series in the original data.

In the light of all the discussion so far, we propose an ECC–VAR method to analyze high-dimensional, non-stationary VAR time series through the corresponding EDQ series found beforehand. The involved unknown parameters will be estimated using general least squares (GLS) or maximum pseudo likelihood (MPL) method based on the derived EDQ series. Once the model fit is obtained, we compute the forecasts together with their 80% forecast intervals at any future time and location, and get other statistical inference results (e.g., goodness-of-fit statistic R^2 and cointegration test outcome etc.) to demonstrate the efficacy of our method. Moreover, we can compare the forecasts with a threshold value determined by the domain knowledge—cf. [7]—so that the probability of the process under observation reaching or exceeding the threshold value can be computed for each future time. This probability, conventionally named the *red alert warning probability*, is easy to understand and widely adopted for predicting the risk of an impending geo-hazard event.

This paper is organized as follows. In Section 2, we describe the landslide data motivating our work in this paper. Next, the ECC–VAR model and its statistical inference are developed in Section 3. The dimension reduction technique EDQ is described in Section 4. The ECC–VAR–EDQ methodology is then demonstrated through analysis of the landslide InSAR data via estimation and forecasting in Section 5. The forecast intervals and probabilities of red alert warning on the landslide data are derived and presented in Section 6 before concluding the paper in Section 7.

2. Motivational Data on Ground Motion in Landslide

Landslide, a type of common geo-hazard event, is caused by significant down-slope movement of soil and/or rocks under the direct influence of gravity. Slope movement occurs when forces acting down-slope (mainly due to gravity) exceed the strength of the earth materials composing the slope. A widely used technique to determine the instability level of a slope is InSAR, by which the precursory movement of the slope surface prior to collapse is constantly monitored. Slope stability radar (SSR) is a mobile InSAR that can remotely scan a slope surface and detect land movement with sub-millimeter precision at high space-time resolutions. Details about using SSR to monitor and detect the slope instability are given in [8–10].

The landslide data to be analyzed in this paper come from a use of SSR to monitor the instability of a slope surface over an undisclosed area stretching 200 m wide and 40 m high. The data comprise time series of the cumulative surface displacement along the reference

line-of-sight from the stationary, ground-based monitoring station to each of 1803 observed locations (a minute area of about 4 m^2 for each location) on the slope. The displacement observations were updated every 6 min over a 3-week period, 10:07 31 May to 23:55 21 June, for 5090 updates in total; cf. [4]. Hence, we have a vector time series with dimensions 1803 and length 5090. The displacement observations and their locations are shown in Figure 1.

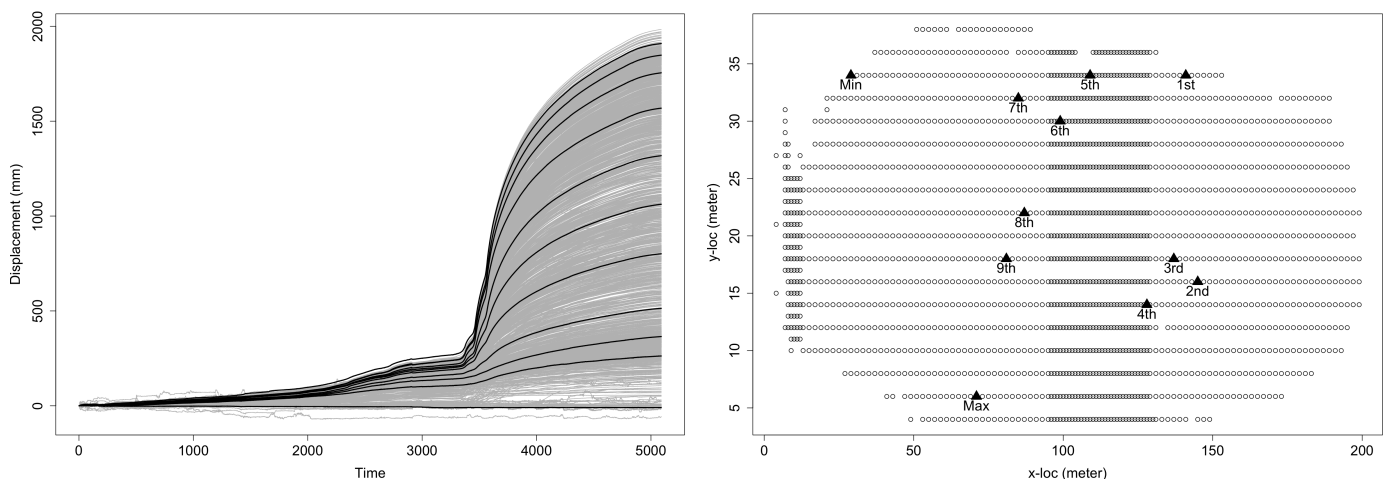


Figure 1. (Left) Slope surface displacement from 31 May to 21 June with 1083×5090 observations in total. (Right) The 1803 small areas (locations). The 11 locations highlighted in black correspond to the 11 EDQ series at quantile levels 0.0 (min), 0.1, \dots , 0.9 and 1.0 (maximum).

Significant displacement of more than 0.5 m and up to 1.5 m, showing accelerating slope deformation, can be found in the original data to have developed in the western side of the slope since early June 15, with velocities reaching 75 mm/h at around $t = 3576$ (13:52 15 June) near the head of the landslide; cf. [4]. The actual landslide occurred at around $t = 3568$ (13:02 15 June), preceded by a precursory slope deformation at around $t = 1600$ (5:42 7 June). It would be a remarkable achievement if one is able to forecast the occurrence of landslide shortly after a precursory slope deformation appears (here $t = 1600$ is 196+ hours earlier than $t = 3568$). However, this cannot be achieved before an effective statistical model is developed to characterize the dynamic system underpinning the observed data, and neither can it before a computationally feasible procedure is constructed to perform forecast. We show in Sections 5 and 6 that our ECC–VAR–EDQ methodology is able to overcome these challenges.

Once landslide forecasts are obtained, the subsequent challenge is to interpret and inform the results to the public in easy to understand language. In addition to using the displacement and velocity forecasts, we use the red alert probability concept introduced in Section 1 to interpret and inform the results.

3. The Error-Correction Cointegration (ECC) Approach for VAR Time Series

In this section, as introduced in [5], we outline an error-correction cointegration (ECC) approach for analyzing unit-root non-stationary vector auto-regression (VAR, of order p) time series data. Provided that the approach is effective in fitting the training sample data, we can use the fitted ECC–VAR(p) model to forecast. We present a rigorous introduction of the ECC–VAR(p) model first. We then derive the method and procedure for parameter estimation and other statistical inference tasks for the ECC–VAR(p) model. Finally, we show how to use the model for forecasting. Readers who are interested in only the applications of this model just need to take Equations (1), (4), (5), (6) and (10), and the definition of EDQ, and then jump to Section 5 and forward.

3.1. ECC-VAR(p) Model

Let $\{z_t, t = 1, \dots, T\}$ be a $k \times 1$ vector time series with $z_t = \{z_{1t}, z_{2t}, \dots, z_{kt}\}^\top$. For each given t , z_t can be treated as being collected from k locations. The vector time series z_t is said to follow a VAR(p) model if

$$z_t = \Phi_0 + \sum_{i=1}^p \Phi_i z_{t-i} + a_t, \quad t = 1, \dots, T \quad (1)$$

where Φ_0 is a $k \times 1$ unknown vector; Φ_i s are $k \times k$ unknown matrices for $i > 0$ with $\Phi_p \neq 0$; and $\{a_t\}$ is a sequence of independent and identically distributed random vectors with mean zero and covariance matrix Σ_a . Denote B as the back shift operator such that $Bz_t = z_{t-1}$ and $B^i z_t = z_{t-i}$. Then the VAR(p) model (1) can be re-written as

$$\Phi(B)z_t = \Phi_0 + a_t \quad \text{where } \Phi(B) = I_k - \sum_{i=1}^p \Phi_i B^i \text{ with } \Phi_p \neq 0. \quad (2)$$

The unknown parameters involved in (2) can be estimated by a general least squares (GLS) method when regarding (2) as similar to a multivariate linear regression equation. However, such parameter estimators are valid and possess those large-sample asymptotic properties in multivariate linear regression models only when the time series data under analysis are stationary and invertible. It is well established that a VAR(p) model characterizes stationary and invertible vector time series if and only if all solutions, with respect to λ , of the determinant equation

$$\left| I_k - \sum_{i=1}^p \Phi_i \lambda^i \right| = |\Phi(\lambda)| = 0 \quad (3)$$

are greater than 1 in modulus.

An important type of non-stationary vector time series is unit-root non-stationarity, of which some solutions in (3) equal 1 and the rest are greater than 1 in modulus; i.e., $|\Phi(\lambda)| \neq 0$ for $|\lambda| \leq 1$ except $|\Phi(1)| = 0$. A unit-root, non-stationary time series can be converted to a stationary one by taking difference operations for certain number of times. However, there is a danger of over-differencing, and an overly differenced time series may not be invertible, so that the GLS estimator for VAR(p) is not valid. A sequential hypothesis testing approach can be used to test whether or not there is a unit-root non-stationarity against stationarity, and how many times the difference operations should be taken to achieve stationarity. Details can be found in, e.g., [11], for cases of univariate time series.

For vector time series, we deal with the unit-root non-stationarity by cointegration and error-correction. The basic idea behind cointegration is to find a linear combination between two order- d integrated, i.e., $I(d)$, processes that yield a process with a lower order of integration; cf. [12]. Specifically, the cointegration idea can be applied to an order-1, unit-root, non-stationary vector time series, i.e., an $I(1)$ process, through a difference operation. That is, $\Delta z_t = z_t - z_{t-1}$ will be stationary if $\{z_t\}$ is order-1, unit-root non-stationary. However, Δz_t is not necessarily invertible but admits an error-correction form, as explained by [5]:

$$\Delta z_t = \Pi z_{t-1} + \sum_{i=1}^{p-1} \Phi_i^* \Delta z_{t-i} + c(t) + a_t \quad (4)$$

where $\Pi = \Phi_1 + \dots + \Phi_p - I_k$; $\Phi_j^* = -(\Phi_{j+1} + \dots + \Phi_p)$ with $j = 1, \dots, p-1$; and $c(t)$, as a deterministic vector trend function of t , generalizes Φ_0 in (1). Model (4) is called the error-correction cointegration (ECC) form for z_t ; cf. [5].

Since Δz_t is stationary, it follows that Πz_{t-1} must be stationary, although z_{t-1} is non-stationary. This suggests there exist k linear combinations of unit-root, $k \times 1$ vector, non-stationary time series z_t that each becomes stationary. One needs to get more details about these k linear combinations in order to further analyze z_t . This can be achieved

by looking into the rank of Π , denoted as $r = \text{rank}(\Pi)$, which consists of three cases: $r = 0$, $0 < r < k$ and $r = k$. If $r = 0$, $\Pi = 0$, implying Δz_t is a stationary VAR($p - 1$) time series. If $r = k$, it means the k linear combinations are linearly independent of each other. If $0 < r < k$, the k linear combinations are determined by r linearly independent combinations. In the latter two cases, we can write $\Pi = \alpha\beta^\top$ with the two $k \times r$ matrices α and β , satisfying $\text{rank}(\alpha) = \text{rank}(\beta) = r$. Then (4) becomes

$$\Delta z_t = \alpha\beta^\top z_{t-1} + \sum_{i=1}^{p-1} \Phi_i^* \Delta z_{t-i} + c(t) + a(t) \quad (5)$$

where $\beta = (\beta_1, \dots, \beta_r)$ is called the *cointegration matrix* with β_1, \dots, β_r the *cointegration vectors*, and $\alpha = (\alpha_1, \dots, \alpha_r)$ is referred to as the *loading or adjustment matrix*. Clearly, β and α are not unique, although the spaces spanned by them are uniquely defined. This can be fixed by restricting the first r rows of β to be I_r .

References [13,14] used ideas of canonical correlation analysis (CCA) to develop a cointegration test-estimation procedure to determine the cointegration rank r and estimate the cointegration vectors β_1, \dots, β_r based on model (5) and the observed vector time series $\{z_t\}$. A test statistic derived for the cointegration test is essentially a likelihood ratio statistic in the form of either a *trace statistic* or a *maximum eigenvalue statistic*. Once r is determined from the cointegration test, the cointegration matrix $\beta = (\beta_1, \dots, \beta_r)$ can be consistently estimated using the eigenvectors corresponding to the eigenvalues involved in the trace or maximum eigenvalue statistic. Details can be found in the aforementioned references. In practice, one can use R function `ca.jo()` in package `urca` to perform the cointegration test-estimation of r and β . The estimators of r and β are denoted as \hat{r} and $\hat{\beta} = (\hat{\beta}_1, \dots, \hat{\beta}_{\hat{r}})$, respectively. Writing $y_{t-1} = \hat{\beta}^\top z_{t-1}$; and by the ECC form (5), the observed vector time series z_t can be asymptotically characterized by the following ECC(\hat{r})-VAR(p) model:

$$\Delta z_t = \alpha y_{t-1} + \sum_{i=1}^{p-1} \Phi_i^* \Delta z_{t-i} + c(t) + a_t \quad (6)$$

which is a stationary process. Then the unknown parameters α , Φ_i^* s, etc., can be consistently estimated by the general least squares (GLS) method. See [14] for details.

3.2. Making Statistical Inferences from the ECC-VAR Model

For simplicity of presentation, assume $c(t) = c_0$, an unknown constant vector indicating z_t 's deterministic drift. The ECC(\hat{r})-VAR(p) model (6) now becomes

$$\Delta z_t = \alpha y_{t-1} + \sum_{i=1}^{p-1} \Phi_i^* \Delta z_{t-i} + c_0 + a_t$$

which can be expressed in a multivariate linear regression form

$$\Delta z_t^\top = x_t^\top \Gamma + a_t^\top, \quad \text{or in matrix form} \quad Z = X\Gamma + A \quad (7)$$

where $x_t^\top = (1, y_{t-1}^\top, \Delta z_{t-1}^\top, \dots, \Delta z_{t-p+1}^\top)$ is a $[k(p-1) + \hat{r} + 1] \times 1$ column vector of observations, and $\Gamma = [c_0, \alpha, \Phi_1^*, \dots, \Phi_{p-1}^*]^\top$ is a $[k(p-1) + \hat{r} + 1] \times k$ unknown matrix parameter to be estimated. Furthermore, $Z = [\Delta z_{p+1}, \dots, \Delta z_T]^\top$ is a $(T-p) \times k$ response observation matrix, $X = [x_{p+1}, \dots, x_T]^\top$ is the $(T-p) \times [k(p-1) + \hat{r} + 1]$ design matrix and $A = [a_{p+1}, \dots, a_T]^\top$ is a $(T-p) \times k$ error or residual matrix.

In practice, fitting (7) can simply be carried out by fitting k multiple linear regression models separately by the GLS method. Using matrix differentiation, it is easy to show that the GLS estimator of Γ is

$$\hat{\Gamma} = (X^\top X)^{-1}(X^\top Z) \quad \text{or} \quad \text{vec}(\hat{\Gamma}) = (I_k \otimes [(X^\top X)^{-1}X^\top])\text{vec}(Z) \quad (8)$$

where \otimes is the Kronecker product. It can also be shown that

$$\text{cov}(\text{vec}(\mathbf{Z})) = \mathbf{\Sigma}_a \otimes \mathbf{I}_{T-p} \quad \text{and} \quad \widehat{\text{cov}}(\text{vec}(\hat{\mathbf{\Gamma}})) = \hat{\mathbf{\Sigma}}_a \otimes (\mathbf{X}^\top \mathbf{X})^{-1}$$

with $\hat{\mathbf{\Sigma}}_a = [(T-p) - (k(p-1) + \hat{r} + 1)]^{-1} \mathbf{Z}^\top (\mathbf{I} - \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top) \mathbf{Z}$. The confidence interval for any element of $\mathbf{\Gamma}$ or function of $\mathbf{\Gamma}$ can be accordingly computed from these results. Moreover, the coefficient of determination of model (7) can be calculated as

$$R^2 = 1 - \frac{SS_{res}}{SS_{total}} = 1 - \frac{\text{tr}([\mathbf{Z} - \mathbf{X}\hat{\mathbf{\Gamma}}][\mathbf{Z} - \mathbf{X}\hat{\mathbf{\Gamma}}]^\top)}{\text{tr}(\mathbf{Z}^\top (\mathbf{I}_{T-p} - (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top) \mathbf{Z})} \quad (9)$$

which can be used to assess the goodness of fit of the model.

Computations of $\hat{\mathbf{\Gamma}}$ and the associated statistical inference results can be carried out using the `cajorls()` function in R package `urca`. The AR order p can be best estimated by a standard model selection criterion such as AIC or BIC, which are commonly used in time series analysis. Details are not pursued here.

3.3. Forecasting Based on the Fitted ECC–VAR Model

The fitted ECC–VAR model (6) can be used to forecast values of $\Delta \mathbf{z}_t$ and \mathbf{z}_t at future time $t = T + h$, $h = 1, 2, \dots$. Namely,

$$\widehat{\Delta \mathbf{z}_{T+h}} = \hat{\mathbf{\Pi}} \widehat{\mathbf{z}_{T+h-1}} + \sum_{i=1}^{p-1} \hat{\mathbf{\Phi}}_i^* \widehat{\Delta \mathbf{z}_{T+h-i}} + c(T+h) \quad \text{and} \quad \widehat{\mathbf{z}_{T+h}} = \widehat{\Delta \mathbf{z}_{T+h}} + \widehat{\mathbf{z}_{T+h-1}} \quad (10)$$

where $\hat{\mathbf{\Pi}} = \hat{\mathbf{\alpha}} \hat{\mathbf{\beta}}^\top$, $\hat{\mathbf{\Phi}}_i^*$, and $c(T+h)$ can be calculated based on the results of cointegration test-estimation and $\hat{\mathbf{\Gamma}}$. Furthermore, $\widehat{\Delta \mathbf{z}_{T+h-i}} = \Delta \mathbf{z}_{T+h-i}$ and $\widehat{\mathbf{z}_{T+h-i}} = \mathbf{z}_{T+h-i}$ if $T+h-i \leq T$. Variance of $\widehat{\Delta \mathbf{z}_{T+h}}$ and $\widehat{\mathbf{z}_{T+h}}$ can be estimated based on $\widehat{\text{cov}}(\text{vec}(\hat{\mathbf{\Gamma}}))$ and the multivariate δ -method or empirical plug-in method, from which the prediction intervals for $\Delta \mathbf{z}_{T+h}$ or \mathbf{z}_{T+h} can be computed.

Statistical estimation, testing and forecast based on the ECC–VAR model, can be carried out without significant computing complications if the dimension k of the data $\{\mathbf{z}_t\}$ is small. It will become computationally infeasible if k is large, when the number of unknown parameters under estimation is of size $O(k^2)$. The landslide time series data considered in this paper has dimensionality of 1803; therefore, some dimension-reduction technique was needed for a feasible computational analysis.

Peña et al. [6] introduced a concept of EDQ for dimension reduction in high-dimensional vector time series and developed a technique to find a small number of EDQ series which are used to represent the original high-dimensional vector time series data. In the light of the EDQ concept and technique, we obtained various statistical inferences about the ECC(r)-VAR(p) model based on only a small number of EDQ series found from the original data with pre-specified quantile levels. Since the EDQ series are representative of the original data, we could use the statistical inference results about these EDQ series to gather statistical inferences on all individual time series in the data by interpolation and approximation. Details are provided later in the paper after the EDQ dimension-reduction method is described next.

4. The EDQ Technique for Vector Time Series Dimension Reduction

For a $k \times 1$ vector time series $\{\mathbf{z}_t = (z_{1t}, \dots, z_{kt})^\top, t = 1, \dots, T\}$, which is equivalently a set of k scalar time series where k can be large, we want to choose a small subset of those k series to describe $\{\mathbf{z}_t = (z_{1t}, \dots, z_{kt})^\top, t = 1, \dots, T\}$ without compromising its overall temporal dynamics, so that information drawn from this small subset is representative of $\{\mathbf{z}_t = (z_{1t}, \dots, z_{kt})^\top, t = 1, \dots, T\}$ and also efficient. This can be achieved by using

the empirical dynamic quantiles (EDQ) technique introduced by [6]. Given a probability value $p \in [0, 1]$, the level- p EDQ series of $\{z_t = (z_{1t}, \dots, z_{kt})^\top, t = 1, \dots, T\}$ is defined as

$$\mathbf{q}^{(p)} \equiv \{q_t^{(p)}, t = 1, \dots, T\} := \arg \min_{\mathbf{q} = \{q_1, \dots, q_T\} \in \mathcal{C}_k} \left[\sum_{t=1}^T \left(p \sum_{z_{it} \geq q_t} |z_{it} - q_t| + (1-p) \sum_{z_{it} \leq q_t} |z_{it} - q_t| \right) \right]$$

where $\mathcal{C}_k = \{\{z_{it}, t = 1, \dots, T\}, i = 1, \dots, k\}$ is the set of k scalar time series in $\{z_t = (z_{1t}, \dots, z_{kt})^\top, t = 1, \dots, T\}$. Note that the level p empirical quantile of $\{z_{it}, i = 1, \dots, k\}$ for fixed t is

$$q_t^{*(p)} = \arg \min_{q \in \mathbb{R}} \left[p \sum_{z_{it} \geq q} |z_{it} - q| + (1-p) \sum_{z_{it} \leq q} |z_{it} - q| \right],$$

and $q_t^{*(p)}$ is not necessarily equal to $q_t^{(p)}$ for any t . Specifically, $\{q_t^{(p)}, t = 1, \dots, T\}$ is one of the k scalar time series in $\{z_t = (z_{1t}, \dots, z_{kt})^\top, t = 1, \dots, T\}$, and $\{q_t^{*(p)}, t = 1, \dots, T\}$ is not necessarily one. Hence, an EDQ series keeps the temporal dynamics in $\{z_t = (z_{1t}, \dots, z_{kt})^\top, t = 1, \dots, T\}$, and $\{q_t^{*(p)}, t = 1, \dots, T\}$ does not. Let $Q_t^{(p)}$ be the level p population quantile corresponding to $q_t^{*(p)}$. It is shown in [6] that

$$\lim_{k \rightarrow \infty} \Pr \left(\sum_{t=1}^T |q_t^{(p)} - Q_t^{(p)}| < \varepsilon \right) = 1 \quad \text{for any } \varepsilon > 0 \text{ and given } T.$$

In practice, we use the EDQ technique to find a quantile level for each of the k time series in $\{z_t = (z_{1t}, \dots, z_{kt})^\top, t = 1, \dots, T\}$ or equivalently \mathcal{C}_k . Then we choose a small number, say, $m = 11$, of them to construct m EDQ series $\mathbf{q}^{(\mathbb{P}_m)} = \{(q_t^{(p_1)}, \dots, q_t^{(p_m)})^\top, t = 1, \dots, T\}$, with the associated quantile levels $0 \leq p_1 < p_2 < \dots < p_m \leq 1$. Now the aforementioned statistical inference made on $\{z_t, t = 1, \dots, T\}$ can be made to the m EDQ series $\mathbf{q}^{(\mathbb{P}_m)}$ in the same way, which will return the same type of estimation and forecast results. In addition, the results obtained from the m EDQ series can be used to draw estimations and forecasts for any other scalar time series in $\{z_t, t = 1, \dots, T\}$ or \mathcal{C}_k . For example, for each such scalar time series, we can find its quantile level, say, p' . This determines two of the m EDQ series of levels, say, p_j and $p_{j'}$, which are the two closest to p' . Then forecasts for the level p' quantile series can be obtained through interpolating the forecasts for the level p_j and $p_{j'}$ EDQ series.

5. Applying the ECC-VAR-EDQ Method to Analyze the InSAR Landslide Data

The InSAR landslide data shown in Figure 1 contain 1803 time series observed at 1803 min areas (each of size $\sim 4 \text{ m}^2$) in a slope of 200 m long and 40 m high, and each time series consists of 5090 cumulative displacement observations, recorded once every 6 min from 31 May to 21 June. As found in Section 2, the actual landslide occurred at around $t = 3568$ (13:02, 15 June), preceded by a precursory slope deformation at around $t = 1600$ (4:41, 7 June); and $t = 3820$ (14:39, 16 June) is well after the landslide. Since an important objective of this analysis is to assess the forecasting capability of the ECC-VAR-EDQ method, we will use the data from $t = 1$ to $t = 1600$ as the training sample for parameter estimation and model fitting, then use the fitted model to forecast at $t = 1601$ to $t = 3820$. Since 1803 is a very high number of dimensions, we reduced the dimensions using the EDQ technique, by which and the training sample 11 EDQ series at levels $\mathbb{P}(11) = (\text{Min}, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, \text{Max})$, with $\text{Min} = 1/1803$ and $\text{Max} = 1802/1803$, were determined together with their locations. These 11 EDQ series (extended to include data at all 5090 time stamps) and their locations in the slope are displayed in Figure 1 in black, and Table 1.

Table 1. Positions (pixels), in 1 to 1803, of the selected 11 EDQ time series.

Quantile Level	Selected Pixel	Quantile Level	Selected Pixel
Min	202	0.6	827
0.1	1432	0.7	672
0.2	1454	0.8	685
0.3	1392	0.9	630
0.4	1307	Max	534
0.5	995		

It was confirmed that some of the 11 EDQ series are non-stationary, which is attributed to either unit root or inside unit-circle roots of the determinant Equation (3). The type of non-stationarity attributed to the inside unit-circle roots is also referred to as the explosive non-stationarity; cf. [15]. To simplify the analysis, we took logarithm transformation on the 11 EDQ time series to remove the explosive non-stationarity as much as possible. That is, define

$$z_{it} = \log(x_{it} - \min_{i,t}(x_{it}) + 0.5), \quad i = 1, \dots, 11; \quad t = 1, \dots, 5090 \quad (11)$$

where x_{it} is the original cumulative displacement observation at time t and area (pixel) i corresponding to the 11 quantile levels $\mathbb{P}(11)$. The 11 EDQ series defined in (11) are displayed in Figure 2. We used as the training sample the 11 EDQ series $\{z_t = (z_{1,t}, \dots, z_{11,t})^\top, t = 1, \dots, 1600\}$ to perform statistical analysis by the ECC-VAR-EDQ method in the sequel.

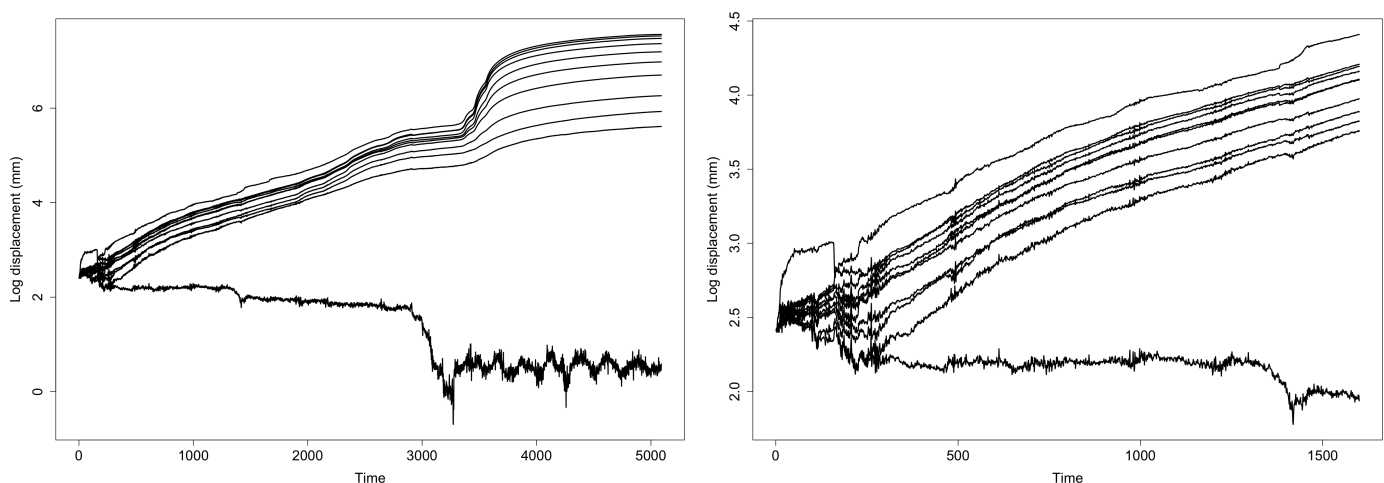


Figure 2. (Left) The 11 log-displacement EDQ series from $t = 1$ to $t = 5090$. (Right) The 11 log-displacement EDQ series from the training sample $t = 1$ to $t = 1600$.

5.1. Unit Root Test and Cointegration Test for the EDQ Series

First, we ran the augmented Dickey–Fuller (ADF) unit root test for the 11 EDQ series $\{z_t = (z_{1,t}, \dots, z_{11,t})^\top, t = 1, \dots, 1600\}$ and their first-order differences $\{\Delta z_t\}$ to see whether they each are stationary or not. The level 1%, 5% and 10% critical values of the ADF test statistic, obtainable from the R package *urca*, are -3.43 , -2.86 and -2.57 , respectively. By the ADF test, the null hypothesis of unit root non-stationarity should be rejected at a significance level if the absolute value of the computed ADF statistic is greater than the absolute value of the corresponding critical value. The ADF unit test results for the 22 times series in $\{z_t\}$ and $\{\Delta z_t\}$ are provided in Table 2, from which we see statistical evidence, at the 1% or 5% significance level, of stationarity for 16 series, and there is no significant evidence to reject the unit-root non-stationarity for the level-(Min, 0.6, 0.7, 0.8, 0.9, Max) EDQ series in $\{z_t\}$.

Table 2. Unit root test results for the level/first-order difference of the 11 selected EDQ time series (log transformed). Superscript *** indicates rejecting the null hypothesis at 1% level; ** indicates rejecting the null hypothesis at 5% level. All these tests were conducted by including a trend term and up to 13 lags.

Quantile Level p	Pixel ID	Augment Dickey–Fuller(ADF)
Min	202	−0.9334/−26.8762 ***
0.1	1432	−2.9837 **/−21.3401 ***
0.2	1454	−4.6048 ***/−21.5254 ***
0.3	1392	−3.1837 **/−17.0134 ***
0.4	1307	−3.3355 **/−17.7037 ***
0.5	995	−2.8695 **/−16.1240 ***
0.6	827	−1.7187/−14.2224 ***
0.7	672	−1.4623/−13.3679 ***
0.8	685	−1.3172/−13.9375 ***
0.9	630	−1.5063/−11.7859 ***
Max	534	−1.3808/−17.9545 ***

Next, we used `ca.jo()` to perform a sequence of cointegration tests based on the trace statistic for the vector time series $\{z_t\}$ determined by the 11 EDQ series. The null hypothesis of $r = \text{rank}(\Pi) \leq r_0$ was to be rejected at significance level α if the trace statistic was greater than the level α critical value. The cointegration test results for $r_0 = 0$ to 10 are provided in Table 3.

Table 3. Null hypotheses, trace statistic and level-(10%, 5%, 1%) critical values of cointegration tests.

Hypothesis	Statistic	10%	5%	1%
$r \leq 10$	0.42	6.50	8.18	11.65
$r \leq 9$	5.82	15.66	17.95	23.52
$r \leq 8$	16.77	28.71	31.52	37.22
$r \leq 7$	35.19	45.23	48.28	55.43
$r \leq 6$	57.56	66.49	70.60	78.87
$r \leq 5$	120.14	85.18	90.39	104.20
$r \leq 4$	253.29	118.99	124.25	136.06
$r \leq 3$	404.85	151.38	157.11	168.92
$r \leq 2$	628.63	186.54	192.84	204.79
$r \leq 1$	998.73	226.34	232.49	246.27
$r \leq 0$	1869.85	269.53	277.39	292.65

Table 3 suggests the best estimate of the cointegration rank is $\hat{r} = 6$. Using `ca.jo()`, the CCA based estimate of the cointegration vectors β , with its top 6 rows constituting a 6×6 identity matrix, is found to be

$$\hat{\beta} = \begin{pmatrix} 1 & 0 & \cdots & \cdots & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & \cdots & 0 & 1 \\ 177.6742 & -36.5351 & -25.6465 & -21.2280 & -8.0301 & -15.0115 \\ -397.0240 & 62.0006 & 44.8457 & 36.3807 & 13.7202 & 26.0822 \\ -538.0373 & 103.0722 & 63.2254 & 51.6327 & 12.8647 & 42.5018 \\ 635.1576 & -109.8483 & -71.2795 & -57.6532 & -16.9223 & -46.6249 \\ 125.5005 & -20.1797 & -12.3700 & -10.3835 & -2.6405 & -8.1719 \end{pmatrix}$$

5.2. Estimating and Fitting the ECC(\hat{r})-VAR(p) Model for the EDQ Series

Given $\hat{r} = 6$ and $c(t) = c_0$, we still need an estimate of p before we can fit the ECC(\hat{r})-VAR(p) model (6) by fitting the multivariate linear regression model (7) using the least squares method. Instead of using a model selection criterion, we used the coefficient of determination R^2 given by (9) to assess the adequacy of the model. Accordingly, $p = 2$ was found to be adequate with $R^2 = 0.99$ based on the associated multivariate linear regression analysis. Then we used the R function `cajorls()` to fit an ECC(6)-VAR(2) model for the 11 EDQ series $\{z_t\}$ shown in the right panel in Figure 2, giving the following results:

$$\widehat{\Delta z}_t = \hat{\alpha} y_{t-1} + \hat{\Phi}_1^* \Delta z_{t-1} + \hat{c}_0, \quad \text{where} \quad (12)$$

$$\hat{\alpha} = \begin{pmatrix} -0.0029 & -0.1037 & 0.1410 & 0.0772 & -0.2227 & -0.0291 \\ 0.0008 & -0.0694 & 0.0814 & 0.0581 & -0.1458 & 0.0298 \\ 0.0000 & 0.0059 & -0.0739 & 0.0075 & 0.0859 & 0.0586 \\ -0.0002 & -0.0099 & -0.0140 & -0.0060 & 0.0364 & 0.0368 \\ 0.0000 & -0.0031 & -0.0222 & 0.0067 & 0.0306 & 0.0219 \\ 0.0009 & -0.0379 & 0.0697 & 0.0285 & -0.1006 & -0.0049 \\ 0.0011 & -0.0106 & 0.0151 & 0.0138 & -0.0206 & 0.0077 \\ 0.0011 & -0.0218 & 0.0448 & 0.0181 & -0.0612 & -0.0026 \\ -0.0001 & -0.0082 & -0.0102 & 0.0156 & 0.0192 & 0.0076 \\ 0.0000 & 0.0112 & -0.0283 & -0.0055 & 0.0552 & 0.0048 \\ -0.0009 & 0.0079 & -0.0069 & -0.0081 & 0.0206 & -0.0174 \end{pmatrix},$$

$$\hat{\Phi}_1^* = \begin{pmatrix} -0.409 & -0.022 & 0.067 & -0.227 & -0.626 & -0.378 & 0.666 & 0.156 & -0.323 & 0.783 & 0.304 \\ 0.001 & -0.632 & -0.056 & -0.171 & -0.370 & 0.174 & 0.711 & -0.303 & -0.165 & 0.567 & 0.244 \\ -0.002 & 0.037 & -0.458 & -0.091 & -0.174 & 0.051 & 0.079 & -0.076 & -0.020 & 0.272 & 0.133 \\ -0.001 & -0.000 & -0.062 & -0.457 & -0.106 & 0.169 & 0.127 & -0.138 & -0.064 & 0.065 & 0.176 \\ -0.001 & 0.027 & 0.051 & -0.152 & -0.612 & -0.014 & 0.084 & -0.031 & 0.002 & 0.281 & 0.129 \\ 0.001 & -0.143 & -0.050 & -0.129 & -0.332 & -0.353 & 0.437 & -0.154 & -0.113 & 0.523 & 0.274 \\ 0.001 & -0.087 & 0.026 & -0.163 & -0.160 & 0.007 & -0.169 & 0.010 & -0.157 & 0.343 & 0.269 \\ 0.001 & -0.134 & -0.069 & -0.144 & -0.213 & 0.035 & 0.346 & -0.466 & -0.145 & 0.497 & 0.304 \\ -0.000 & -0.079 & 0.069 & -0.154 & -0.232 & 0.123 & 0.102 & 0.043 & -0.601 & 0.296 & 0.329 \\ -0.001 & -0.003 & 0.019 & -0.214 & -0.074 & -0.038 & 0.065 & 0.106 & -0.167 & 0.124 & 0.013 \\ 0.001 & -0.237 & -0.082 & 0.066 & -0.357 & 0.249 & 0.077 & 0.041 & 0.051 & 0.037 & 0.160 \end{pmatrix}$$

and $\hat{c}_0 = (0.008, -0.0068, 0.0074, -0.003, 0.0016, -0.011, -0.0085, -0.0103, -0.0014, -0.0003, 0.0063)^\top$.

5.3. Landslide Displacement Forecasting

Based on the fitted ECC(6)-VAR(2) model for the 11 EDQ series, we could obtain the forecasts of Δz_t and z_t for $t = 1601$ until $t = 3820$, i.e., those time states in the test sample. Equation (10) was used for all calculations. The forecasts of the displacement and velocity were then calculated by the formulas $\widehat{x_{i,T+h}} = e^{\widehat{z_{i,T+h}}}$ and $\widehat{\Delta x_{i,T+h}} = \widehat{x_{i,T+h}} - \widehat{x_{i,T+h-1}}$, respectively. These forecasts are plotted in Figure 3.

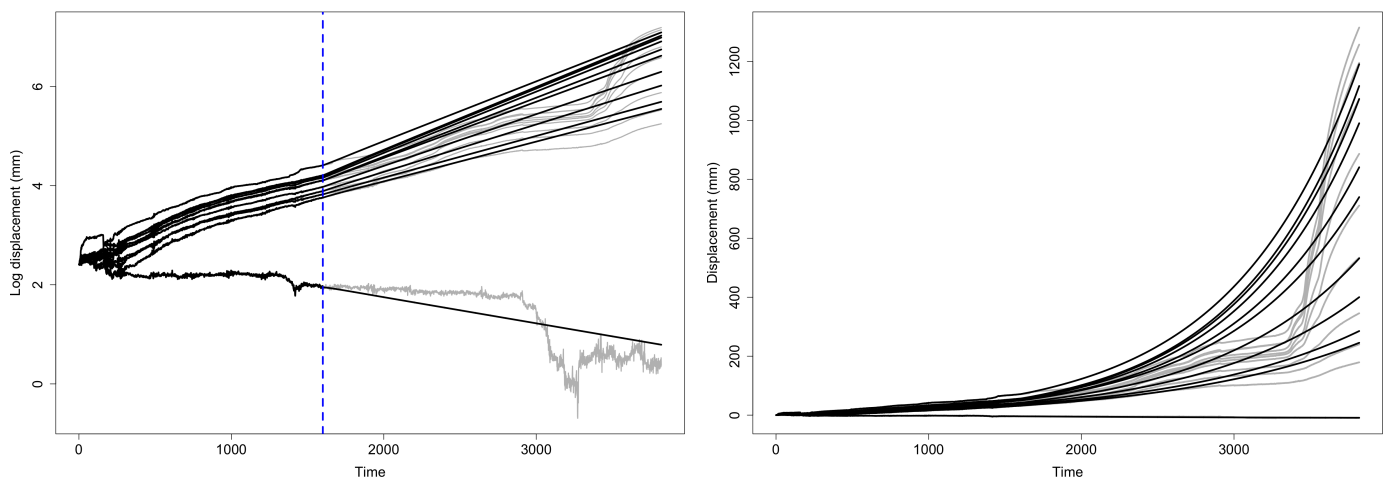


Figure 3. (Left) Log-displacements of the 11 selected EDQ series from $t = 1$ to $t = 3820$. (Right) Displacements of the 11 selected EDQ series from $t = 1$ to $t = 3820$. The black lines represent the forecasts and the gray lines represent the actual observations.

6. Probabilistic Landslide Prediction via the ECC-VAR-EDQ Method

6.1. Forecast Intervals for Displacement and Velocity

In order to see the variations in the forecasts $\widehat{x}_{i,T+h}$ and $\widehat{\Delta x}_{i,T+h}$, we resorted to find the corresponding forecast intervals. This required finding the variances of $\widehat{x}_{i,T+h}$ and $\widehat{\Delta x}_{i,T+h}$ first, which can be achieved by applying the δ -method or an empirical sampling plug-in method based on the forecasts $\widehat{\Delta z}_{i,T+h}$ and $\widehat{z}_{i,T+h}$ and their variance and covariance matrices. In fact, each component in these variance and covariance matrices could be estimated from the predicted values of Δz_t and z_t , with $t = T + 1, \dots, T + h$, by fitting the multivariate linear regression model (7). For simplicity of presentation, we use $\text{Var}(\widehat{\Delta z}_{it})$ and $\text{Var}(\widehat{z}_{it})$ to denote the estimated variances of $\widehat{\Delta z}_{it}$ and \widehat{z}_{it} , respectively. By asymptotic normal approximation for GLS estimators, an approximately 80% forecast interval for Δz_{it} is given as

$$\left[\widehat{\Delta z}_{it} - 1.28 \sqrt{\text{Var}(\widehat{\Delta z}_{it})}, \widehat{\Delta z}_{it} + 1.28 \sqrt{\text{Var}(\widehat{\Delta z}_{it})} \right], \quad t = T + 1, \dots, T + h, \dots \quad (13)$$

where 1.28 is the value of 0.9 quantile of standard normal distribution $\Phi^{-1}(0.9) \approx 1.28$. Furthermore, an approximately 80% forecast interval for z_{it} is given as

$$\left[\widehat{z}_{it} - 1.28 \sqrt{\text{Var}(\widehat{z}_{it})}, \widehat{z}_{it} + 1.28 \sqrt{\text{Var}(\widehat{z}_{it})} \right], \quad t = T + 1, \dots, T + h, \dots \quad (14)$$

Results from (13) and (14) can be used to find 80% forecast intervals for $x_{it} = e^{z_{it}} + \min_{i,t}(x_{it}) - 0.5$ by either the δ -method or transformation. As an example, let us find an approximately 80% forecast interval for $x_{i,T+1}$. By the δ -method,

$$E(\widehat{x}_{i,T+1}) \approx e^{\widehat{z}_{i,T}} \left(e^{E(\widehat{\Delta z}_{i,T+1})} \right) + \min_{i,t}(x_{it}) - 0.5 \approx e^{\widehat{z}_{i,T+1}} + \min_{i,t}(x_{it}) - 0.5 = \widehat{x}_{i,T+1} \quad (15)$$

$$\text{Var}(\widehat{x}_{i,T+1}) \approx e^{2\widehat{z}_{i,T} + 2E(\widehat{\Delta z}_{i,T+1})} \text{Var}(\widehat{\Delta z}_{i,T+1}) \approx e^{2\widehat{z}_{i,T+1}} \text{Var}(\widehat{\Delta z}_{i,T+1}) \quad (16)$$

Therefore, using (15) and (16), an approximately 80% forecast interval for the cumulative displacement $x_{i,T+1}$ can be computed from

$$\left[\widehat{x}_{i,T+1} - 1.28 \sqrt{\text{Var}(\widehat{x}_{i,T+1})}, \widehat{x}_{i,T+1} + 1.28 \sqrt{\text{Var}(\widehat{x}_{i,T+1})} \right].$$

Similarly to (15) and (16), we can get

$$E(\widehat{x_{i,T+h}}) \approx \widehat{x_{i,T+h}}, \quad \text{and} \quad \text{Var}(\widehat{x_{i,T+h}}) \approx e^{2\widehat{z_{i,T+h}}} \text{Var}(\widehat{\Delta z_{i,T+h}}),$$

from which we can compute an approximately 80% forecast interval for $x_{i,T+h}$. However, the approximation error involved in estimating $\text{Var}(\widehat{x_{i,T+h}})$ may be excessive so as to render the interval inaccurate. We propose to compute an approximately 80% forecast interval for $x_{i,T+h}$ based on a monotonic transformation of that for the log-displacement $z_{i,T+h}$. Namely, the referenced 80% forecast interval for $x_{i,T+h}$ is

$$\left[e^{\widehat{z_{i,T+h}} - 1.28\sqrt{\text{Var}(\widehat{z_{i,T+h}})}} + \min_{i,t}(x_{it}) - 0.5, e^{\widehat{z_{i,T+h}} + 1.28\sqrt{\text{Var}(\widehat{z_{i,T+h}})}} + \min_{i,t}(x_{it}) - 0.5 \right]. \quad (17)$$

Results from (13) and (14) can also be used to find 80% forecast intervals for velocity $v_{i,T+h} = \Delta x_{i,T+h}$ at location i and time $t = T + 1, \dots, T + h, \dots$. To achieve this, we first applied the δ -method to find an estimated variance of $\widehat{v_{i,T+h}} = \widehat{\Delta x_{i,T+h}}$, which is

$$\begin{aligned} \text{Var}(\widehat{v_{i,T+h}}) &= \text{Var}(\widehat{x_{i,T+h}}) + \text{Var}(\widehat{x_{i,T+h-1}}) - 2\text{Cov}(\widehat{x_{i,T+h}}, \widehat{x_{i,T+h-1}}) \\ &\approx e^{\widehat{z_{i,T+h}}} \left[e^{\widehat{z_{i,T+h}}} - e^{\widehat{z_{i,T+h-1}}} \right] \text{Var}(\widehat{\Delta z_{i,T+h}}) \\ &\quad + e^{\widehat{z_{i,T+h-1}}} \left[e^{\widehat{z_{i,T+h-1}}} - e^{\widehat{z_{i,T+h}}} \right] \text{Var}(\widehat{\Delta z_{i,T+h-1}}) \\ &\quad - e^{\widehat{z_{i,T+h}} + \widehat{z_{i,T+h-1}}} \text{Cov}(\widehat{\Delta z_{i,T+h}}, \widehat{\Delta z_{i,T+h-1}}). \end{aligned} \quad (18)$$

Then, a normal approximation based 80% forecast interval for $\widehat{v_{i,T+h}}$ is

$$\left[\widehat{v_{i,T+h}} - 1.28\sqrt{\text{Var}(\widehat{v_{i,T+h}})}, \widehat{v_{i,T+h}} + 1.28\sqrt{\text{Var}(\widehat{v_{i,T+h}})} \right]. \quad (19)$$

Alternatively, an approximately 80% forecast interval for $v_{i,T+h} = x_{i,T+h} - x_{i,T+h-1}$ can be obtained by transforming that for $x_{i,T+h}$ and $x_{i,T+h-1}$ as follows:

$$\left[e^{\widehat{z_{i,T+h}} - 1.28\sqrt{\text{Var}(\widehat{z_{i,T+h}})}} - e^{\widehat{z_{i,T+h-1}} + 1.28\sqrt{\text{Var}(\widehat{z_{i,T+h-1}})}}, e^{\widehat{z_{i,T+h}} + 1.28\sqrt{\text{Var}(\widehat{z_{i,T+h}})}} - e^{\widehat{z_{i,T+h-1}} - 1.28\sqrt{\text{Var}(\widehat{z_{i,T+h-1}})}} \right]. \quad (20)$$

Results of the forecasts and approximately 80% forecast intervals for the 11 EDQ series of cumulative displacement x_{it} and velocity v_{it} for t from 1601 to 3820 are displayed in the first two columns of Figures 4–6. A horizontal line indicating the red-alert velocity of 10mm/hr is added to each velocity plot there.

6.2. Probability of Future Risk of Landslide

Forecasts of the displacement and velocity and their respective 80% forecast intervals plotted in Figures 4–6 show that they correctly capture the overall trends of ground movement in terms of displacement and velocity. However, they also show the need for improving the forecasting accuracy, and it is difficult to interpret these results for the general public. On the other hand, interpreting the risk of an incoming landslide in terms of a probability is much more accessible to general public. Hence, we propose to measure this risk by a probability of the forecast ground motion velocity exceeding a red-alert level. The risk may be alternatively evaluated by comparing the forecast displacement with a red-alert displacement level, but it will not be pursued here due to lack of consensus about setting the red-alert displacement level.

In field of slope failure research, it is commonly accepted that a ground motion velocity of 10 mm per hour or 1 mm per 6 min is an alarming number indicating slope failure. Since the cumulative displacement in our InSAR data is updated roughly every 6 min, the referenced velocity v_{it} is actually the change of displacement per 6 min. Therefore, we

will compute the probability of v_{it} exceeding 1 at location i and $t = T + 1, \dots, T + h, \dots$. Calculating this probability requires at least an approximate probability distribution for v_{it} be available. By the asymptotic properties of the GLS estimation in multivariate linear regression and the δ -method, it can be shown that

$$v_{it} \stackrel{asym.}{\sim} N(\widehat{v_{it}}, \text{Var}(\widehat{v_{it}})), \quad \text{equivalently} \quad \frac{v_{it} - \widehat{v_{it}}}{\sqrt{\text{Var}(\widehat{v_{it}})}} \stackrel{asym.}{\sim} N(0, 1)$$

where $\text{Var}(\widehat{v_{it}})$ is given in Equation (18). Now the probability of the velocity v_{it} exceeding the red-alert threshold level is

$$\Pr(v_{it} \geq 1) \approx 1 - \Phi\left(\frac{1 - \widehat{v_{it}}}{\sqrt{\text{Var}(\widehat{v_{it}})}}\right), \quad i = 1, \dots, k; t = T + 1, \dots, T + h, \dots \quad (21)$$

Calculated values of $\Pr(v_{it} \geq 1)$ for the 11 EDQ series at t from 1601 to 3820 are displayed in the right panels of Figures 4–6.

The probability plots in Figures 4–6 provide very accurate probability forecasting on the impending landslide. For example, $\Pr(v_{11,t} \geq 1) > 0.5$ for the level Max EDQ series when $t \geq 3513$, 5.5 h ahead of the actual landslide occurred around $t = 3568$; and $\Pr(v_{10,t} \geq 1) > 0.5$ for the level-0.9 EDQ series when $t \geq 3575$, within 42 min after the actual landslide. Recalling that these forecast and probability results are computed using the observations at t from 1 to 1600, about 196.8 h before the actual landslide. Therefore, we can conclude that our ECC–VAR–EDQ method is able to accurately predict an impending landslide, with greater than 50% red-alert probability, more than 196 h in advance.

6.3. Landslide Prediction for All Locations

In addition to getting the probabilistic forecasting results for the 11 EDQ series, we used the ECC–VAR–EDQ method to compute the displacement forecasts x_{it} , the velocity forecasts v_{it} and the forecast probabilities $\Pr(v_{it} \geq 1)$ for all 1803 locations and at time t from 1601 to 3820. The key method used is interpolation. For example, for a location with EDQ level 0.83, all the forecasts at this location were calculated as the weighted average of those at level 0.8 and level 0.9 EDQ locations. Namely, for $t = 1601$ to 3820

$$\begin{aligned} \widehat{x_{(0.83),t}} &= 0.7 \cdot \widehat{x_{(0.8),t}} + 0.3 \cdot \widehat{x_{(0.9),t}} \\ \widehat{v_{(0.83),t}} &= 0.7 \cdot \widehat{v_{(0.8),t}} + 0.3 \cdot \widehat{v_{(0.9),t}} \\ \Pr(v_{(0.83),t} \geq 1) &= 0.7 \cdot \Pr(v_{(0.8),t} \geq 1) + 0.3 \cdot \Pr(v_{(0.9),t} \geq 1) \end{aligned}$$

An mp4 video mine_r6.mp4 is provided online which displays the aforementioned forecasts of displacement and the red-alert probability for all 1803 locations and at time t from 1601 to 3820. Two screenshots of this video at time $t = 3513$ and 3568 are shown in Figures 7 and 8, respectively. The video and the screenshot show that the displacement forecasts capture the spatiotemporal dynamics and trends of the actual slope surface movement at all locations very well, and the red-alert probability forecasts provide timely predictions of the landslides at all locations.

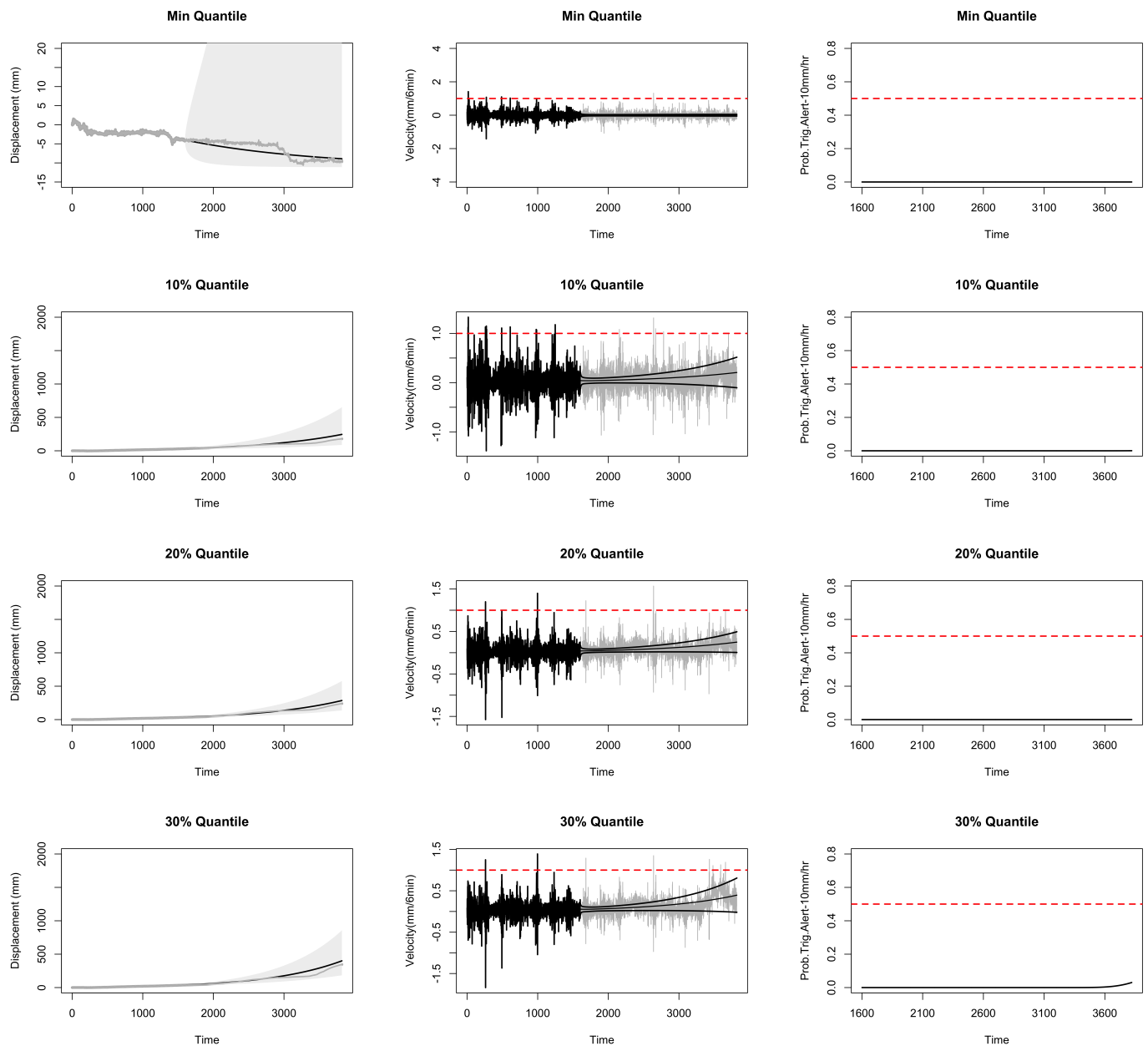


Figure 4. Slope surface displacement (**Left**), velocity (**Middle**) and probability of triggering a red alert (**Right**) in the Min, 10%, 20% and 30% EDQ series. The forecast displacement and velocity are plotted in black with their estimated 80% forecast intervals, and the corresponding observed data are plotted in gray lines. The red dashed line in each velocity plot is the velocity alert threshold, which is 1 mm/6 min (10 mm/hr), and indicates 50% probability of triggering a red alert in the probability plots.

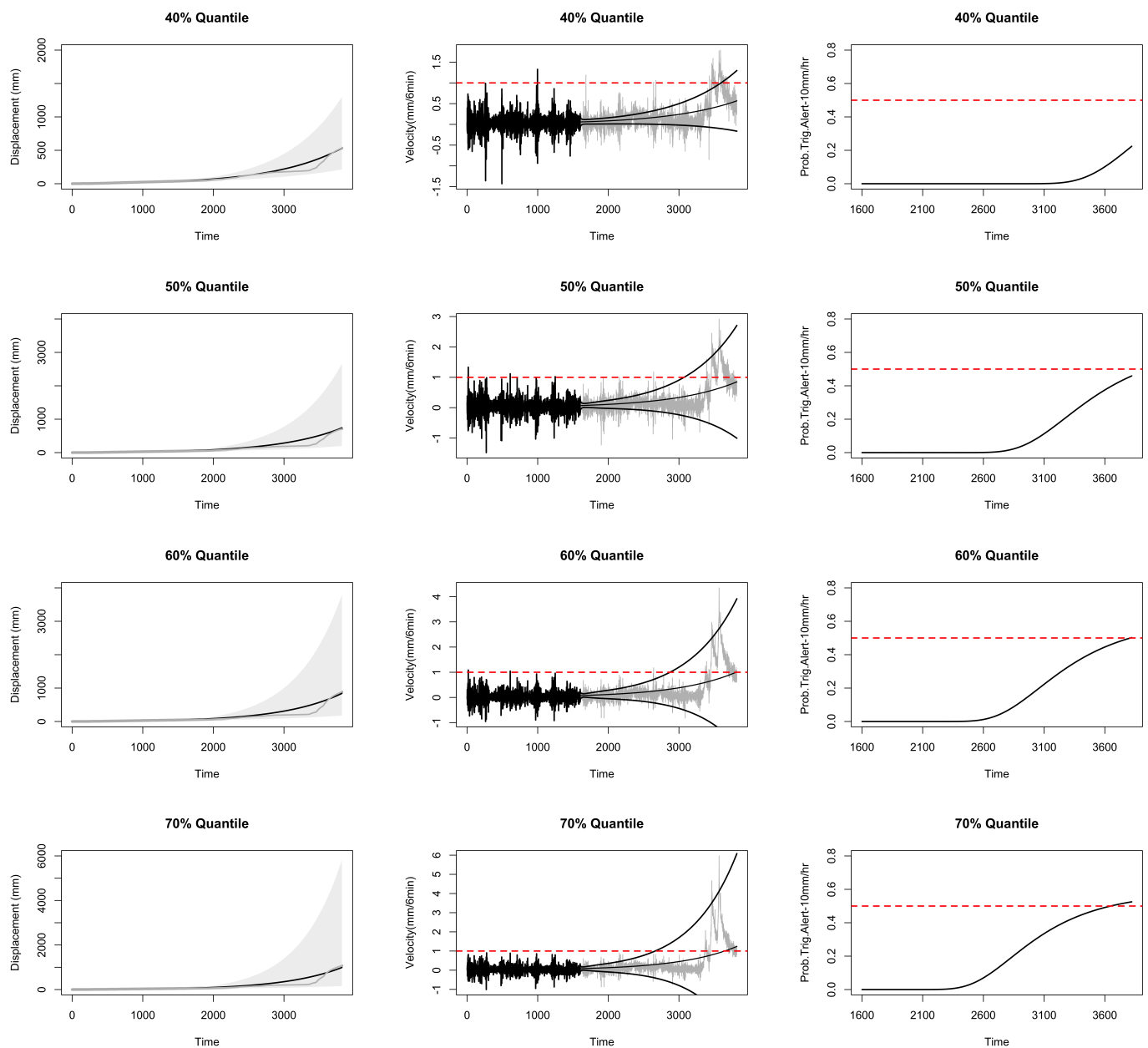


Figure 5. Slope surface displacement (Left), velocity (Middle) and probability of triggering red alert (Right) for the 40%, 50%, 60% and 70% EDQ series. The forecast displacement and velocity are plotted in black with their estimated 80% forecast intervals, and the corresponding observed data are plotted in gray lines. The red dashed line in each velocity plot is the velocity alert threshold, which is 1 mm/6 min (10 mm/hr), and indicates 50% probability of triggering a red alert in the probability plots.

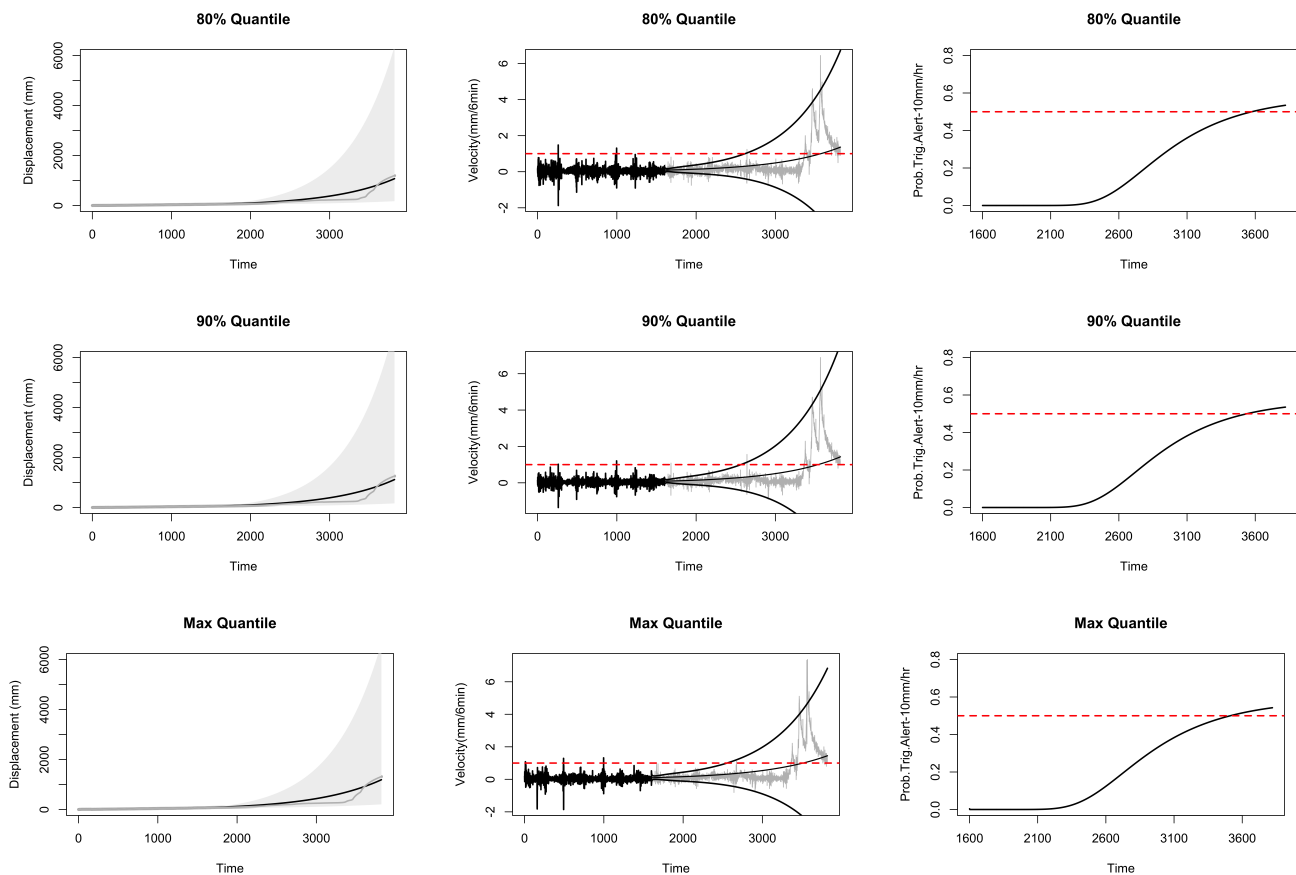


Figure 6. Slope surface displacement (**Left**), velocity (**Middle**) & probability of triggering red alert (**Right**) for the 80%, 90% & Max EDQ series. The forecast displacement and velocity are plotted in black with their estimated 80% forecast intervals, and the corresponding observed data are plotted in gray lines. The red dashed line in each velocity plot is the velocity alert threshold, which is 1 mm/6 min (10 mm/hr), and indicates 50% probability of triggering a red alert in the probability plots.

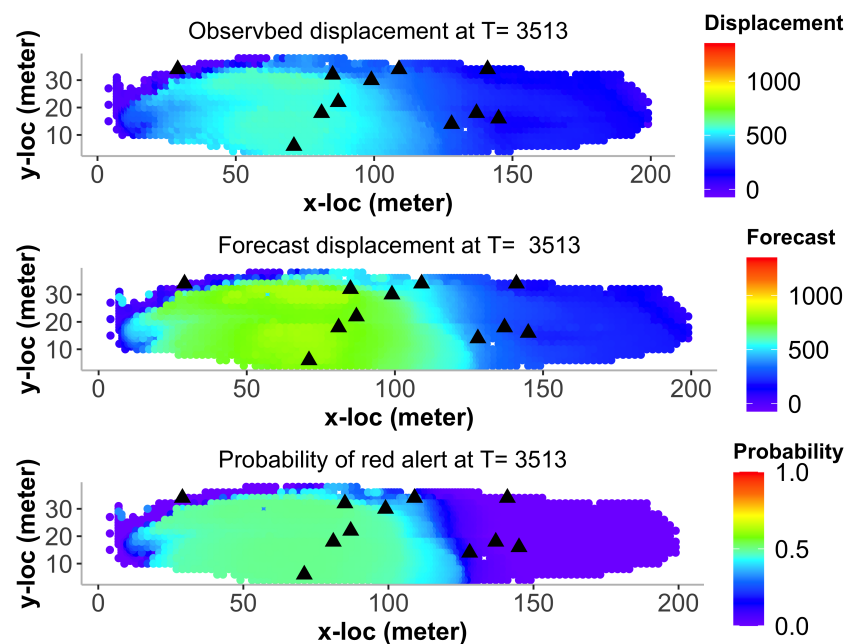


Figure 7. Screenshot of mine_r6.mp4 at time $t = 3513$. The **top** (middle) panel displays the actual (forecast) displacement, and the **bottom** panel displays the probability forecasts at all locations. The \blacktriangle s indicate the 11 EDQ locations. Results were obtained based on the data at $t = 1, \dots, 1600$.

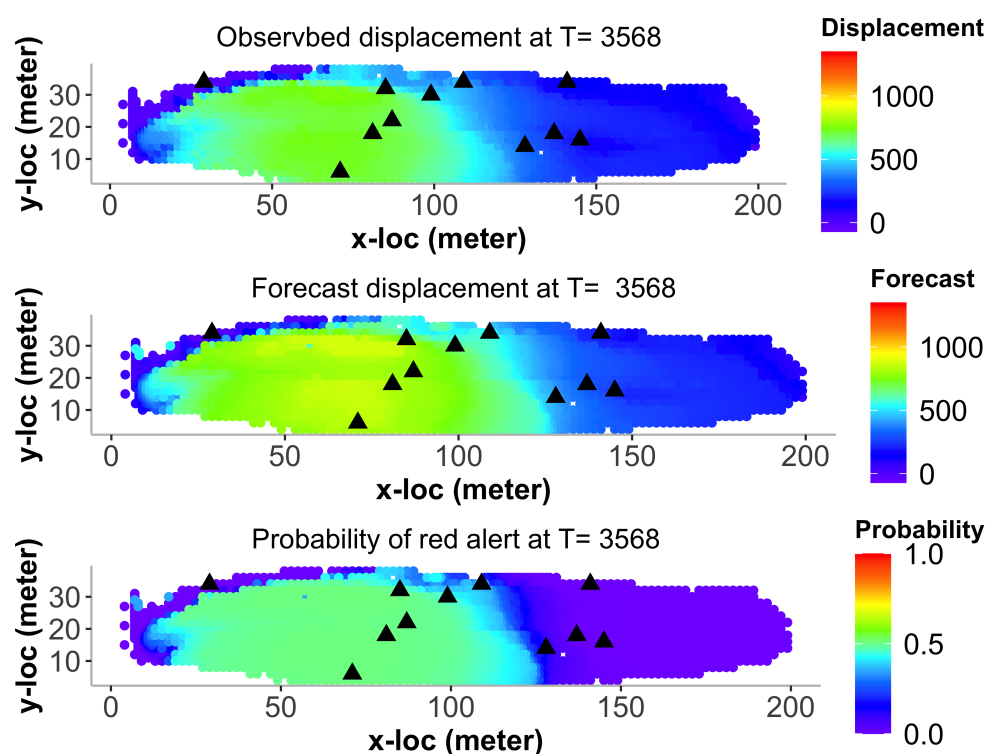


Figure 8. Screenshot of mine_r6.mp4 at time $t = 3568$. The **top (middle)** panel displays the actual (forecast) displacement, and the **bottom** panel displays the probability forecasts at all locations. The \blacktriangle s indicate the 11 EDQ locations. Results were obtained based on the data at $t = 1, \dots, 1600$.

7. Conclusions

We have developed an ECC–VAR–EDQ method to characterize and analyze high-dimensional, unit-root non-stationary vector time series. The model underpinning the new method has a vector autoregression form with an error-correction cointegration representation. We showed how to use the new method to make statistical inferences, including parameter estimation, goodness of model fit test, forecasts and forecast intervals, for high-dimensional non-stationary vector time series data. An advantage of our method is that we can use a small number of EDQ series to reduce the dimension and apply an ECC–VAR model to these EDQ series. Thus, the procedure is computationally feasible, with negligible loss of spatial-temporal information of the data.

We inferred an ECC(6)–VAR(2)–EDQ(11) model to analyze the InSAR landslide data observed at 1803 locations and across 5090 time states. The statistical inference results were shown to be remarkably informative. For example, the goodness of fit statistic $R^2 = 0.99$ was close to 1. The results also estimated 80% forecast intervals that can be used to calculate the probability of a possible red-alert event in the future based on the up-to-date observations. Once this probability is higher than 50%, we can raise an early warning alert about a probable incoming landslide. For the motivational landslide case study in this paper, the results by our method can predict the incoming landslide more than 8 days in advance, which means we will have enough time to prepare an effective mitigation plan and reduce the damage caused by landslides.

Although our ECC–VAR–EDQ method has a capability for timely forecasting an incoming landslide, there is still room for improvement. For example, the ECC approach is only able to deal with the unit-root non-stationary vector time-series, but the real data can be non-stationary of other types. To enable using ECC, we took a logarithm transform to our original data, but we still found there are some sharp increases (diffusion) in those high level quantile series, suggesting the log-transformation is successful but not ideal. Hence in future work, we will investigate improving ECC to handle other types of non-stationarity in high-dimensional VAR time series.

Finally, note that most of the current landslide monitoring and forecast research are not configured for computational scalability and as such has focused on only uni-variate time series or low-dimensional vector time series; cf. [3]. Thus, their methods are not able to handle spatial dependence and temporal dynamics jointly for InSAR landslide data. Our ECC–VAR–EDQ method is able to effectively handle this spatial dependence and temporal dynamics simultaneously in a computationally feasible way.

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

Funding: This research was partly funded by the U.S. DoD High Performance Computing Modernization Program (HPCMP) and RDECOM International Technology Center-Pacific (ITC-PAC) contract number: FA5209-18-C-0002.

Acknowledgments: The authors thank the Assistant Editor and the four anonymous reviewers for providing valuable comments and suggestions, leading to an improvement of the presentation of the paper.

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

References

1. Dick, G.J.; Eberhardt, E.; Cabrejo-Liévano, A.G.; Stead, D.; Rose, N.D. Development of an early-warning time-of-failure analysis methodology for open-pit mine slopes utilizing ground-based slope stability radar monitoring data. *Can. Geotech. J.* **2014**, *52*, 515–529.
2. Glade, T.; Crozier, M.J. Landslide Hazard and Risk-Concluding Comment and Perspectives. In *Landslide Hazard and Risk*; John Wiley & Sons, Ltd.: Hoboken, NJ, USA, 2012; Chapter 26, pp. 765–774. doi:10.1002/9780470012659.ch26.
3. Intrieri, E.; Carlá, T.; Gigli, G. Forecasting the time of failure of landslides at slope-scale: A literature review. *Earth-Sci. Rev.* **2019**, *193*, 333–349.
4. Wang, H.; Qian, G.; Tordesillas, A. Modeling big spatio-temporal geo-hazards data for forecasting by error-correction cointegration and dimension-reduction. *Spat. Stat.* **2020**, *36*, 100432. doi:10.1016/j.spasta.2020.100432.
5. Tsay, R.S. *Multivariate Time Series Analysis: With R and Financial Applications*; John Wiley & Sons: Hoboken, NJ, USA, 2014.
6. Peña, D.; Tsay, R.S.; Zamar, R. Empirical dynamic quantiles for visualization of high-dimensional time series. *Technometrics* **2019**, *61*, 429–444.
7. Wessels, S.D.N. Monitoring and Management of a Large Open Pit Failure. Ph.D. Thesis, University of the Witwatersrand, Johannesburg, South Africa, 2009.
8. Harries, N.; Noon, D.; Rowley, K. Case studies of slope stability radar used in open cut mines. In *Stability of Rock Slopes in Open Pit Mining and Civil Engineering Situations*; SAIMM Johannesburg: South Africa, 2006; pp. 335–342.
9. Casagli, N.; Catani, F.; Del Ventisette, C.; Luzi, G. Monitoring, prediction, and early warning using ground-based radar interferometry. *Landslides* **2010**, *7*, 291–301.
10. Stacey, T. *Stability of Rock Slopes in Open Pit Mining and Civil Engineering Situations*; The South African Institute of Mining and Metallurgy: Johannesburg, South Africa, 2006.
11. Kwiatkowski, D.; Phillips, P.C.; Schmidt, P.; Shin, Y. Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *J. Econom.* **1992**, *54*, 159–178.
12. Pfaff, B. *Analysis of Integrated and Cointegrated Time Series with R*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2008.
13. Johansen, S. Statistical analysis of cointegration vectors. *J. Econ. Dyn. Control* **1988**, *12*, 231–254.
14. Johansen, S.; Juselius, K. Maximum likelihood estimation and inference on cointegration—With applications to the demand for money. *Oxf. B Econ. Stat.* **1990**, *52*, 169–210.
15. Phillips, P.C.; Wu, Y.; Yu, J. Explosive behavior in the 1990s Nasdaq: When did exuberance escalate asset values? *Int. Econ. Rev.* **2011**, *52*, 201–226.