

# Spatiotemporal forecast of extreme events in a chaotic model of slow slip events

Hojjat Kaveh<sup>1</sup>, Jean Philippe Avouac<sup>2</sup> and Andrew M Stuart<sup>3</sup>

<sup>1</sup>Department of Mechanical and Civil Engineering, California Institute of Technology, 1200 East California Boulevard, Pasadena, CA, USA. E-mail: [hkaveh@caltech.edu](mailto:hkaveh@caltech.edu)

<sup>2</sup>Department of Geology and Planetary Science, California Institute of Technology, 1200 East California Boulevard, Pasadena, CA, USA

<sup>3</sup>Department of Computing and Mathematical Science, California Institute of Technology, 1200 East California Boulevard, Pasadena, CA, USA

Accepted 2024 November 11. Received 2024 October 11; in original form 2024 July 15

## SUMMARY

Seismic and aseismic slip events result from episodic slips on faults and are often chaotic due to stress heterogeneity. Their predictability in nature is a widely open question. In this study, we forecast extreme events in a numerical model. The model, which consists of a single fault governed by rate-and-state friction, produces realistic sequences of slow events with a wide range of magnitudes and interevent times. The complex dynamics of this system arise from partial ruptures. As the system self-organizes, the state of the system is confined to a chaotic attractor of a relatively small dimension. We identify the instability regions within this attractor where large events initiate. These regions correspond to the particular stress distributions that are favourable for near complete ruptures of the fault. We show that large events can be forecasted in time and space based on the determination of these instability regions in a low-dimensional space and the knowledge of the current slip rate on the fault.

**Key words:** Seismic cycle; Self-organization; Earthquake interaction, forecasting, and prediction – Numerical modelling.

## 1 INTRODUCTION

Earthquakes and slow slip events (SSEs) result from episodic frictional slip on the faults. Each slip event releases the elastic strain accumulated during an interevent period during which the fault is locked. This principle is often referred to as the elastic rebound theory in reference to Reid (1910). While the elastic rebound theory offers valuable insights into the long-term mean recurrence time of earthquakes and can be used for time-independent earthquake forecasting (Avouac 2015; Marsan & Tan 2020), it falls short of predicting the time or the magnitude of the larger events (Murray & Segall 2002). The difficulty is that earthquakes often exhibit a chaotic behaviour which is manifest in the irregular and rare occurrence of large slip events and various empirical scaling laws, such as the Gutenberg–Richter and the Omori laws (Scholz 1989). The Gutenberg–Richter law (Gutenberg & Richter 1950) states that earthquake magnitudes are distributed exponentially (the number of earthquakes with magnitude larger than  $M$ ,  $N(M)$ , is given by  $\log_{10} N(M) = a - bM$ , where  $b$  is a scaling parameter of the order of one and  $a$  is a constant). The Omori law (Utsu *et al.* 1995) states that the rate of earthquakes during an aftershock sequence decays as  $1/t$  where  $t$  is the time since the main shock. Chaotic behaviour has also been identified in sequences of SSEs in Cascadia (Gualandi *et al.* 2020). These events obey the same scaling laws as regular

earthquakes and produce very similar crack-like and pulse-like ruptures, although with several orders of magnitudes smaller slip rate and stress drop (Michel *et al.* 2019).

The main source of complexities in earthquake sequences is due to stress heterogeneities which can either be of static origin [due to fault geometry (Okubo & Aki 1987), roughness (Sagy *et al.* 2007; Cattania 2019) or heterogeneity of mechanical properties (Kaneko *et al.* 2010)] or dynamical, due stress transfers among faults or within a single fault (Shaw & Rice 2000). As the stress evolves during the earthquake cycle, it generates asperities and barriers that can either facilitate a complete rupture of a fault (a system-size rupture) or impede the propagation of a rupture, resulting in a partial rupture. Partial or complete ruptures of a fault system are therefore observed in nature (Konca *et al.* 2008). Large ruptures, though rare according to the Gutenberg–Richter law, hold greater significance from a seismic hazard perspective.

Advances in the understanding of fault friction (Marone 1998) and in numerical modelling of earthquake sequences (Rice 1993; Lapusta *et al.* 2000; Lapusta & Liu 2009) now make it possible to produce realistic simulations (Barbot *et al.* 2012). When performing those numerical simulations, initial conditions cannot be any arbitrary value, and it is also crucial to recognize that certain initial conditions hold more statistical relevance than others during the evolution of the system. For example, Lapusta & Rice (2003)

and Rubin & Ampuero (2005) advocate for conducting simulations over multiple seismic cycles to mitigate the influence of arbitrary choices in initial conditions. In fact, the space of feasible stress distributions on a fault during earthquake cycles is significantly smaller than the space of arbitrary initial conditions, as the dynamical system self-organizes into a chaotic attractor (Shaw & Rice 2000). When a dynamical system converges to its chaotic attractor, any state outside this attractor is not feasible within the system's evolution. Consequently, the space of feasible states is limited to the attractor itself, resulting in a significantly smaller domain compared to the space of any arbitrary states for the system.

Large events happen rarely in the chaotic evolution of the earthquake cycle so their forecast is extremely challenging. We hypothesize that as for other types of dynamical systems that produce extreme (or rare) events (Blonigan *et al.* 2019; Farazmand & Sapsis 2019), the trajectory of the dynamical system must traverse specific instability regions within the chaotic attractor for large fault ruptures to occur. These instability regions correspond to the optimal distributions of stress (or the states of the frictional interface) that facilitate large ruptures and are also accessible during the evolution of the system because they are part of the chaotic attractor. Despite considerable research on deterministic chaos in earthquake cycle models (Huang & Turcotte 1990; Becker 2000; Anghel *et al.* 2004; Kato 2016; Barbot 2019), certain essential features of the chaotic attractor, particularly modes relevant to instability that are also statistically feasible, have remained elusive in the literature. This is primarily due to the high-dimensional, chaotic and multiscale nature of the problem, as well as the rarity of large events.

The identification of the optimal state of the frictional interface that promotes large events, out of all the statistically feasible distributions is the primary focus of this study. Following the approach of (Farazmand & Sapsis 2017), we use an approximation of the chaotic attractor of the system during the interevent period; this approximation uses Proper Orthogonal Decomposition (POD) to reduce dimension and account for the feasibility constraint. Representing the optimal state of the frictional interface in a low-dimensional space is favourable for the purpose of earthquake forecasting, as the data to constrain the physical parameters and current states of the system are sparse for earthquake cycles. We use the proximity of the current slip rate of the system to the slip rates of optimal solutions to propose an effective forecast method of large events. Our results suggest that this framework can be used to predict events in both space and time when we have access to the slip rate on the fault with certain resolution.

As our case study, we use a quasi-dynamic model with the standard rate-and-state friction with the ageing law (Ruina 1983). We apply this methodology to a 2-D fault within a 3-D medium, using a model setup analogue to a model that has been shown to produce a realistic sequence of SSEs similar to those observed in Cascadia (Dal Zilio *et al.* 2020). We limit the analysis to the case of SSEs as in that case a quasi-dynamic approximation is justified which speeds up the numerical simulations (Rice 1993; Thomas *et al.* 2014). The complexity of events (and in particular the frequency of small events) has been shown to depend on the ratio of the fault length (or width) to the nucleation size (Barbot 2019; Cattania 2019). We benefit from the fact that SSEs have a much larger ratio of nucleation size to the size of the fault compared to regular earthquakes. The range of magnitude of events in our 1000 yr of synthetic data is 5.6–7.4 whereas for a large fault system with typical earthquakes, the range is much bigger. Spatially small-scale processes in regular earthquakes contribute to more complexity of the system. This

might limit the applicability of our method to these events without any further considerations.

## 2 MODEL SET UP

We use a quasi-dynamic model of slip events on a 2-D fault in a 3-D elastic medium, assuming rate-and-state friction with the ageing law for the evolution of the state variable ( $\theta$ ):

$$\frac{\tau}{\bar{\sigma}_n} = \mu^{\text{ref}} + a \ln\left(\frac{V}{V^{\text{ref}}}\right) + b \ln\left(\frac{\theta V^{\text{ref}}}{D_{RS}}\right), \quad (1a)$$

$$\frac{d\theta}{dt} = 1 - \frac{\theta V}{D_{RS}}. \quad (1b)$$

Here,  $V((x, y), t) : \Gamma \times \mathbb{R}^+ \rightarrow \mathbb{R}^+$  is slip rate on the fault,  $\theta((x, y), t) : \Gamma \times \mathbb{R}^+ \rightarrow \mathbb{R}^+$  is the state variable,  $\tau((x, y), t) : \Gamma \times \mathbb{R}^+ \rightarrow \mathbb{R}^+$  is the frictional strength,  $\bar{\sigma}_n$  is the effective normal stress and  $a, b, D_{RS}$  are frictional properties of the surface ( $\Gamma$ ) and are positive.  $\mu^{\text{ref}}$  and  $V^{\text{ref}}$  are reference friction and slip rate, respectively. The sign of  $a - b$  determines the frictional regime of the fault. For  $a - b > 0$ , the fault is velocity strengthening (VS); a jump in the velocity would increase the fault strength. Regions with  $a - b > 0$  suppress the rupture nucleation and acceleration. For  $a - b < 0$  fault is velocity weakening (VW); a jump in the slip rate ( $V$ ) and slipping more than  $D_{RS}$ , decrease the strength and the fault is capable of nucleating earthquakes and accelerating the ruptures.  $a - b$  varies spatially and is plotted in Fig. 1(a).

The stress rate on the fault can also be written as:

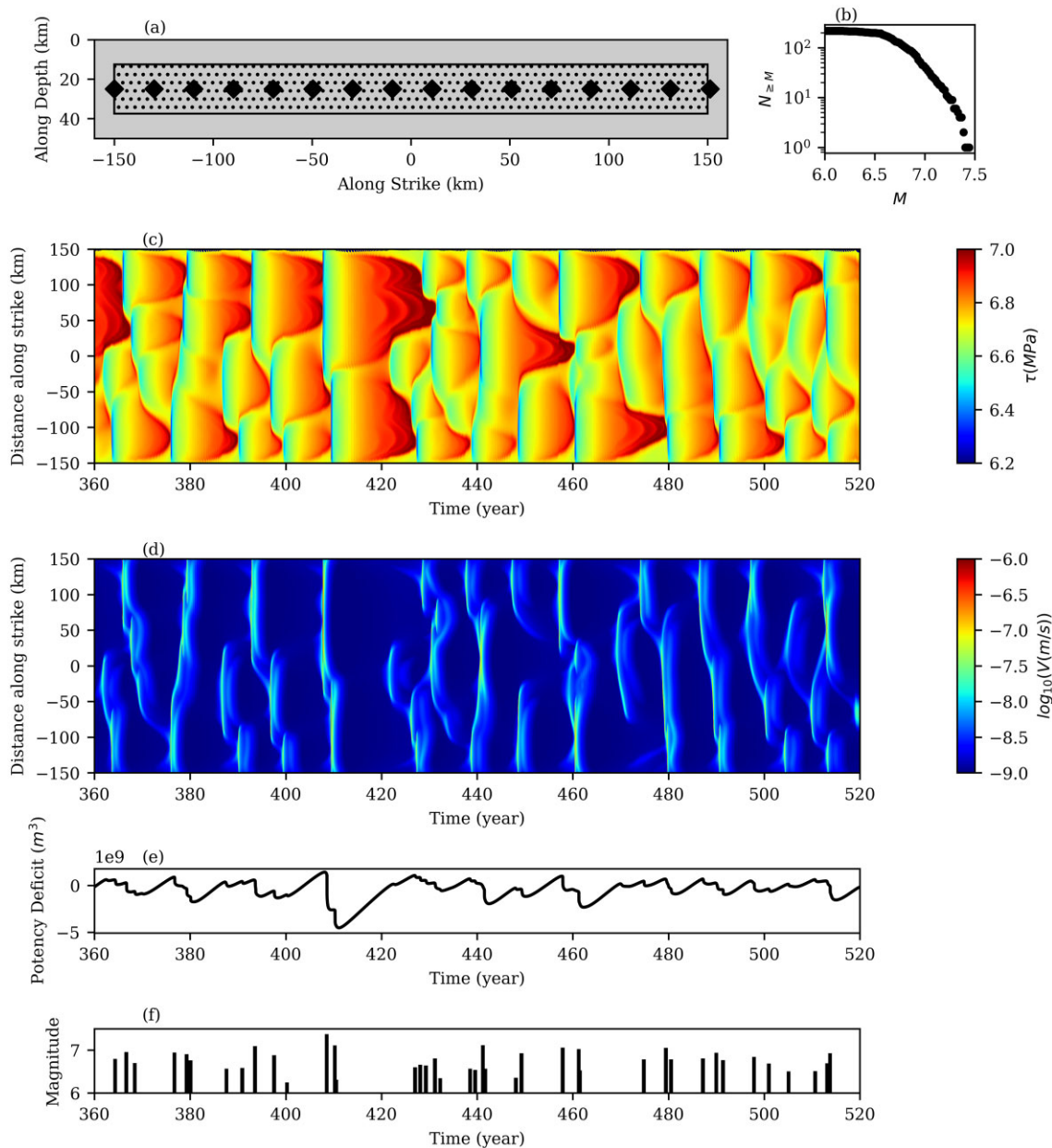
$$\dot{\tau} = \mathcal{L}(V - V_{pl}) - \frac{G}{2c_s} \dot{V}, \quad (2)$$

where  $\mathcal{L}$  is a pseudo-differential operator, and contains elastostatic response (Geubelle & Rice 1995), and  $V_{pl}$  is the plate slip rate.  $G$  and  $c_s$  are shear modulus and shear wave speed, respectively. By taking the time derivative of eq. (1a), and eliminating  $\dot{\tau}$  using eq. (2), we have a dynamical system for  $u = [V, \theta]^T$ . One can also use other pairs of variables such as  $[V, \tau]^T$  to describe this dynamical system.

In practice, we consider a planar thrust fault with 90° dip angle in elastic half-space that consists of a VW patch (dotted area in Fig. 1a), within which ruptures can nucleate and propagate, surrounded by a VS patch where the propagation of seismic ruptures is inhibited (Fig. 1a). The fault is loaded by a surrounding fault that slips at a constant rate.

The model, with the properly selected and piece-wise constant parameters and initial conditions, exhibits a complex sequence of events with a variety of magnitudes distributed with a heavy tail consistent with the Gutenberg–Richter law (Fig. 1b). The shear stress on the locked portion of the fault (Fig. 1c) increases during the interevent period, leading to elastic strain energy build-up. During episodic slip events, the shear stress drops, and elastic strain energy is released and dissipated by frictional sliding and the radiation damping (Fig. 1c). To justify the assumption of ignoring wave propagation effects along the fault, we choose a parameter regime that produces SSEs in which  $V$  is small enough that the wave effects across the faults are negligible. The model parameters are taken from (Dal Zilio *et al.* 2020) to simulate SSEs similar to those in Cascadia. For simplicity we did not include the effect of pore-pressure dilatancy. The frictional and physical properties of our problem are summarized in Table 1 and Fig. 1.

The time-series of the sequence of partial rupture with rare large ruptures is plotted in Figs 1(c) and (d). Since stress is a function of  $\theta$  and  $V$  in the rate-and-state friction, and  $\theta$  is not measurable, we do not have access to stress distribution directly. As a result, in



**Figure 1.** Geometry of the fault (a). The VW patch is the dotted area that is surrounded by the VS patch. The diamonds are the locations of slip rate measurements for the scenario in which we do not have full access to the slip rate on the entire fault. The number of events with a magnitude greater than  $M$ , ( $N_M$ ) is plotted in (b) for 1000 yr of simulation time. Maximum stress along the depth for the VW patch is plotted as a function of distance along strike and time (c). The maximum slip rate for the VW patch along the depth is plotted as a function of distance along strike and time (d). The time-series of the potency deficit and magnitudes are plotted in (e) and (f), respectively.

**Table 1.** Physical properties.

VW region	a	0.004
	b	0.014
VS region	a	0.019
	b	0.014
Characteristic slip weakening distance	$D_{RS}$	0.045 (m)
Reference steady state slip rate	$V^{ref}$	$10^{-6} \frac{m}{s}$
Reference steady-state friction coefficient	$\mu^{ref}$	0.6
Effective normal stress	$\bar{\sigma}_n$	10 (MPa)
Shear modulus	$G$	30 (GPa)
Plate loading velocity	$V_{pl}$	40 (mm yr $^{-1}$ )

this work, we only assume that we have observations of the current slip rate when performing extreme event forecasting. In practice, the current slip rate on the fault can be indirectly constrained by measurements of ground surface displacements which involves an inversion that greatly reduces the spatial resolution of slip rate. Therefore, we will also examine a simplified low-resolution slip rate measurement to mimic the limitations of real observations and assess the performance of our algorithm under such conditions. The slip potency deficit, which is the difference between the slip potency (integral of slip on the fault) and the slip potency if the fault was uniformly slipping at the loading rate, is plotted to show the chaotic behaviour of the system and the rare occurrence of large events. The potency deficit builds up during the interevent period

and drops during the episodic slip events (Fig. 1e). The time-series of the magnitude of events is also plotted in Fig. 1(f). The maximum slip rate on the fault is plotted in Fig. 2 with the dashed line as the threshold that we use for defining an event.

### 3 EXTREME EVENTS FORECASTING METHODS

#### 3.1 Extreme events formulation

The dynamical system that comes from combining eqs (1) and (2) describes the coupled evolution of two functions  $V((x, y), t) : \Gamma \times \mathbb{R}^+ \rightarrow \mathbb{R}^+$  and  $\theta((x, y), t) : \Gamma \times \mathbb{R}^+ \rightarrow \mathbb{R}^+$ . We assume  $u = [V, \theta]^\top$  belongs to an appropriately chosen function space  $\mathcal{U} : (\Gamma \times \mathbb{R}^+) \times (\Gamma \times \mathbb{R}^+) \rightarrow \mathbb{R}^+ \times \mathbb{R}^+$  and characterizes the state of the frictional interface at any given time and position on the fault. In the context of rate-and-state friction, shear stress is a function of the combination of variables  $(V, \theta)$ . Also, the evolution of the system is better rendered in the  $\log_{10} u$  space. Consequently, we use the term ‘pre-event state’ to refer to the spatial distribution of  $w = [\log_{10} V, \log_{10} \theta]^\top$  before a rupture; nonetheless, we formulate the dynamical model in terms of  $u = (V, \theta)$ . To avoid confusion between the term ‘state’ as used to describe the system’s condition before an event and the ‘state variable’ in the friction law, we clarify that ‘pre-event state’ refers to the overall system configuration,  $w = [\log_{10} V, \log_{10} \theta]^\top$ , prior to the event. Meanwhile, the ‘state variable’ ( $\theta$ ) specifically denotes the internal variable in the rate-and-state friction law.

The dynamical system for  $u$  is both multi-scale and chaotic and produces ruptures with a variety of sizes. The governing equation is

$$\frac{\partial u}{\partial t} = \mathcal{N}(u) \quad (3a)$$

$$u(x, y, 0) = u_0(x, y), \quad \forall (x, y) \in \Gamma, \quad (3b)$$

where  $\mathcal{N}$  is a non-linear differential operator<sup>1</sup> that encompasses the quasi-dynamic approximation of the elastodynamics and the friction law (eqs 1 and 2). We denote  $S^t$  as the solution operator for the dynamical system, mapping the current state forward by  $t$  time-units:

$$u(x, y, t) = S^t(u(x, y, 0)); \quad (4)$$

we can break this map into the components  $S_V^t$  and  $S_\theta^t$ :

$$S^t(u(x, y, 0)) = [S_V^t(u(x, y, 0)), S_\theta^t(u(x, y, 0))]^\top. \quad (5)$$

We assume that the dynamical system has a global attractor  $\mathcal{A}$  on which the dynamics are chaotic; we refer to this as the chaotic attractor in what follows.

Inspired by Farazmand & Sapsis (2019), we define event set  $E(V_{\text{thresh}})$  for a prescribed threshold  $V_{\text{thresh}} \in \mathbb{R}^+$  as:

$$E(V_{\text{thresh}}) = \left\{ u \in \mathcal{U} : \sup_{(x,y) \in \Gamma} V(x, y) \geq V_{\text{thresh}} \right\} \quad (6)$$

By setting a proper event threshold ( $V_{\text{thresh}}$ ), the event set includes both partial and full ruptures.

We now seek to determine the optimal feasible distributions of  $\log_{10} u$  (pre-event state) in the interevent period that for a prediction horizon  $T$  lead to large magnitude events. By a ‘feasible pre-event state’, we mean a state that is inside the chaotic attractor of the

system; a combination of  $V$  and  $\theta$  that is likely to be realized during the evolution of the dynamical system. We also want our criteria for optimality of ‘pre-event state’ to be low-dimensional so that it can be captured using observations that are typically sparse in reality. We then use our low-dimensional critical pre-event state and only the current measurable state of the system (slip rate, which can in principle be estimated from geodetic measurements) to forecast the time and location of a possible large event in a time window horizon.

To formulate the question in mathematical terms, we introduce the moment magnitude of fault slip cumulated over the duration of integration  $\Delta t$ .

$$\tilde{M}(u(x, y, t); \Delta t) = \frac{2}{3} \log_{10} \left( G \int_0^{\Delta t} \int_{\Gamma} S_V^{t'}(u(x, y, t)) dx dy dt' \right) - 6, \quad (7)$$

where  $G$  is the elastic shear modulus.  $\tilde{M}$  measures the seismic moment on the fault in the  $\log_{10}$  scale during  $\Delta t$  time-units (Scholz 1989).  $\tilde{M}$  is slightly different from the definition of the moment magnitude ( $M$ ) for one event because in  $\tilde{M}$ , we take  $\Delta t$  to be a constant rather than being the actual duration of a particular event. In practice, we set it to be larger than the longest duration of events in our model. While we make use of  $\tilde{M}$  in our problem setup and benefit from its continuity over  $u$ , we will report the performance of the forecast of extreme events with a regular definition of moment magnitude ( $M$ ).

We next define a cost function:

$$F(u; \Delta t, T) = \sup_{t \in [0, T]} \tilde{M}(S^t(u); \Delta t), \quad (8)$$

where function  $F : \mathcal{U} \rightarrow \mathbb{R}$  takes  $u$  as input and, for a prescribed prediction horizon ( $T$ ) and event duration ( $\Delta t$ ), finds the largest moment magnitude generated by the initial condition  $u$ . The optimal (most dangerous) feasible pre-event state conditions are determined by finding the local maxima ( $U^*$ ) of  $F(u; \Delta t, T)$  over  $u \in \mathcal{A} \setminus E(V_{\text{thresh}})$  through an optimization process:

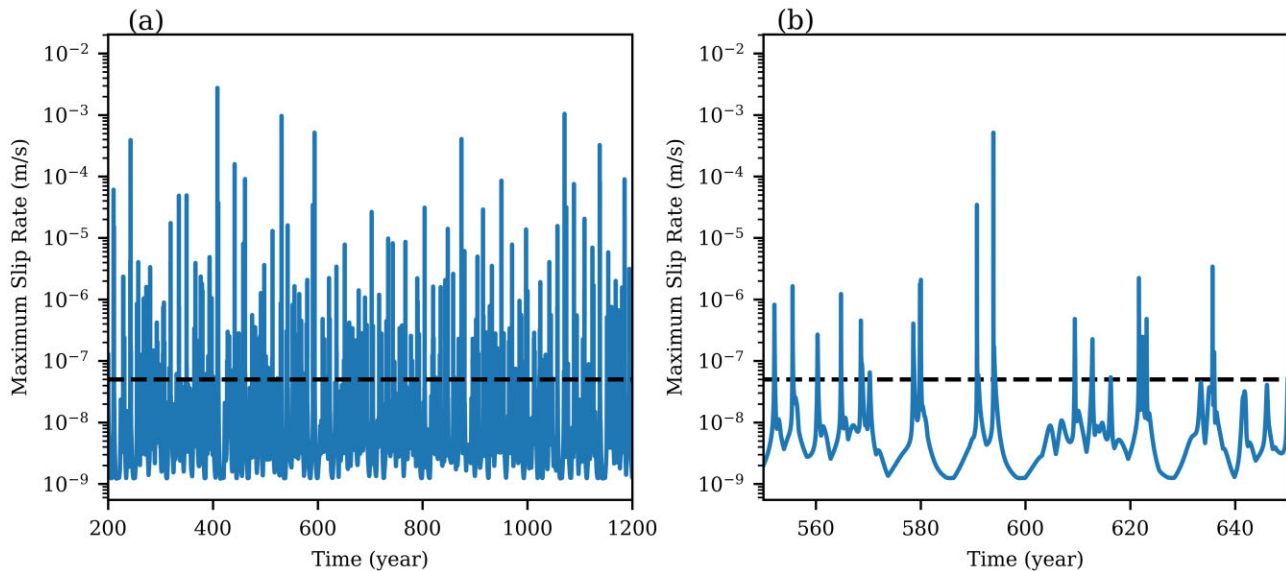
$$U^* = \{u^* | u^* \in \mathcal{A} \setminus E(V_{\text{thresh}}), u^* \text{ is a local maximizer of } F(u; \Delta t, T), F(u^*; \Delta t, T) > F_e^*\}, \quad (9)$$

where  $F_e^*$  is some threshold for the magnitude to define a ‘large’ event. eq. (9) encompasses the main question of this work; that is finding optimal and statistically feasible pre-event state on the fault during the interevent period that makes large events in a short time window. In eq. (9),  $u^* \in \mathcal{A} \setminus E(V_{\text{thresh}})$  ensures that  $u^*$  is inside the chaotic attractor (statistical feasibility constraint) and also in the interevent period; any state ( $u^*$ ) outside  $\mathcal{A}$  is inaccessible during the system’s evolution because of the self-organization. After solving the optimization problem (eq. 9), we use the ‘similarity’ of the current states of the system to solutions of eq. (9), as an indicator of an upcoming large event. We use the current slip rate as our only knowledge of the current state of the system as  $\theta$  is not measurable. Solutions to eq. (9) are instability regions inside the chaotic attractor that generate large ruptures within the time span of  $[0, T]$ .

Set  $\mathcal{A} \setminus E(V_{\text{thresh}})$  is a complicated set in the high-dimensional function space  $\mathcal{U}$ . Even if we can solve this optimization problem in this large space, it would be impractical to represent pre-event state in this high-dimensional space because the sparse data generally available in reality can only yield a low-dimensional model of the slip rate distribution on a fault. As a result, we approximate this set with a simpler set, characterized in a low-dimensional space using the POD method. This approach is developed in the next part.

<sup>1</sup>Technically a pseudo-differential operator.





**Figure 2.** Time-series of the maximum slip rate for a period of 1000 yr (a) and 100 yr (b) with threshold velocity denoted by a dashed line.

### 3.2 Model reduction and forecast scheme

Many high-dimensional chaotic dynamical systems can be approximated by a low-dimensional system (Brandstätter *et al.* 1983; Li *et al.* 2023; Rowley & Dawson 2017; Taira *et al.* 2017). Although the underlying dynamics of earthquakes and Slow Slip cycles are often chaotic (Huang & Turcotte 1990; Anghel *et al.* 2004; Kato 2016; Barbot 2019; Becker 2000), in certain examples, it has been observed that the chaotic attractors are low dimensional (Gualandi *et al.* 2020, 2023) which mathematically implies that we can approximate the evolution of the sequence of events using parameters in a finite-dimensional space instead of an infinite-dimensional function space. We use this property to reduce the dimensionality and approximate the chaotic attractor during the interevent period.

We approximate and reduce the dimensionality of the chaotic attractor of the system during the inter-event period using the POD technique (explained in Appendix A). The POD approach is widely adopted in the study of turbulent fluid flow (Taira *et al.* 2017); it is a linear model reduction method that uses singular value decomposition on a data set of snapshot time-series of the field, with the time-average removed. This process identifies spatial modes that are ranked according to their statistical significance in the data set. Since the evolution of the system is better realized in the  $w = \log_{10} u$  space, we apply the POD on the  $w$  rather than  $u$ . We denote by  $\bar{w}$  the average of the snapshots of the field ( $w$ ) (with variable time stepping) during the interevent period. POD technique inputs snapshots of  $w - \bar{w}$  during the interevent period and gives orthonormal basis functions  $\phi_i : \Gamma \times \Gamma \rightarrow \mathbb{R} \times \mathbb{R}$  and their associated variance  $\lambda_i$  for  $i \geq 1$  where  $\lambda_1 > \lambda_2 > \dots$  which quantifies the statistical importance of each mode in the data set. The subtraction of the mean is crucial because it ensures that the covariance matrix in the POD algorithm accurately reflects the variability and relationships within the data set, rather than being influenced by the absolute positions of the data points. Then we can describe  $w$ , and consequently  $u$ , using a new coordinate system with the basis functions defined by  $\phi_i$ 's. Since the basis functions are ordered by the variance they capture in the data, the truncation and approximation of the field  $w - \bar{w}$ , with the first  $N_m$  POD modes captures a maximal statistical relevance

(in the variance sense) of data between all possible  $N_m$ -dimensional linear subspaces of  $\log_{10} \mathcal{U}$ .

We approximate  $w : w \in \log_{10} (\mathcal{A} \setminus E(V_{\text{thresh}}))$  as perturbations around  $\bar{w} = [\bar{w}^V, \bar{w}^\theta]$  along those basis functions. Since we want to approximate only the interevent period we should exclude the event period ( $E(V_{\text{thresh}})$ ) from the data set of snapshots that are used to find POD modes ( $\phi_i$ 's). Following Blonigan *et al.* (2019), we constrain the perturbations along those eigenvectors to lie within a hyperellipse with a radius along each eigenvector proportional to the standard deviation of the data captured by each mode. In other words, we allow more perturbation along those directions that capture more statistical relevance in the data. The approximation of the chaotic attractor during the interevent period can be written as:

$$\log_{10} (\mathcal{A} \setminus E(V_{\text{thresh}})) \approx \left\{ \bar{w} + \sum_{i=1}^{N_m} a_i \phi_i \mid \sum_{i=1}^{N_m} \frac{a_i^2}{\lambda_i} \leq r_0^2 \right\}, \quad (10)$$

where  $\phi_i$ 's ( $i \geq 1$ ) are the orthonormal basis functions ordered by the data variance they capture ( $\lambda_i$ ) in the centred data set of time snapshots of  $w - \bar{w}$  excluding the event period  $E(V_{\text{thresh}})$ . Here  $a_i$  is the amplitude of perturbation along  $\phi_i$  and  $N_m$  is the number of basis functions we keep in our model reduction. The maximum perturbation along each basis function ( $\phi_i$ ) is constrained by the corresponding variance  $\lambda_i$ . One can play with the amplitude of the allowed perturbation which is represented by  $r_0$ .

Then eq. (9), which is an optimization problem in a high-dimensional function space  $\mathcal{U}$ , constrained on a complicated set  $\mathcal{A} \setminus E(V_{\text{thresh}})$ , can be approximated as an optimization problem in a low-dimensional ( $\mathbb{R}^{N_m}$ ) space constrained within a hyperellipse. To solve the constrained optimization problem, we use optimal sampling in the framework of Bayesian optimization as it is useful when the objective function is costly to evaluate (Blanchard & Sapsis 2021). The optimization method is described in Appendix B. Alternative approaches, such as adjoint-based optimization methods, can also be used to enhance the efficiency of solving the optimization problem (Stiernström *et al.* 2024; Blonigan *et al.* 2019). During the optimization process, we collect all optimal pre-event states ( $w^* = [(\log V)^*, (\log \theta)^*]^\top$ ) in a set  $W^*$  that satisfies the feasibility constraint ( $w^* \in \log_{10} (\mathcal{A} \setminus E(V_{\text{thresh}}))$ ) and has the value of

$F(10^{w^*}; \Delta t, T)$  above the threshold  $F_e^*$ :

$$W^* := \left\{ w^* = \bar{w} + \sum_{i=1}^{N_m} a_i \phi_i \mid \sum_{i=1}^{N_m} \frac{a_i^2}{\lambda_i} \leq r_0^2, F(10^{w^*}; \Delta t, T) > F_e^* \right\}, \quad (11)$$

$W^*$  corresponds to the set of all of the pre-event states leading to extreme events. To perform the spatial forecast, we need to record the evolution of each  $w^*$  for up to time  $T$ .

We use the proximity of the current state of the system to optimal states as an indicator of an upcoming large event. The current state of the system ( $w$ ) is not measurable because  $\theta$  is not measurable. Slip rate is the measurable component in  $w$  and we use it as a proxy of the current state of the system. Then, following Blonigan *et al.* (2019), we use the maximum cosine similarity between the  $\log_{10}$  of the current slip rate ( $\log V(t)$ ) and all of the optimal slip rates ( $\log V_i^*$ 's) in the set  $W^*$  as an indicator that signals an upcoming large event.

$$I(t) = \max_i \frac{\langle \log V(t) - \bar{w}^V, \log V_i^* - \bar{w}^V \rangle_{L^2}}{\| \log V(t) - \bar{w}^V \|_2 \| \log V_i^* - \bar{w}^V \|_2} \quad (12)$$

where  $\langle \cdot, \cdot \rangle_{L^2}$  is the  $L^2$  inner product,  $\bar{w}^V$  is the snapshot-average slip rate during interevent periods in the data set,  $\log V_i^*$  is the velocity component of the  $i$ th optimal pre-event state ( $w_i^*$ ) and  $\| \cdot \|_2$  is the  $L^2$  norm. Note that  $I(t)$  is only a function of the current slip rate on the fault.

## 4 RESULTS

### 4.1 Extreme event forecast

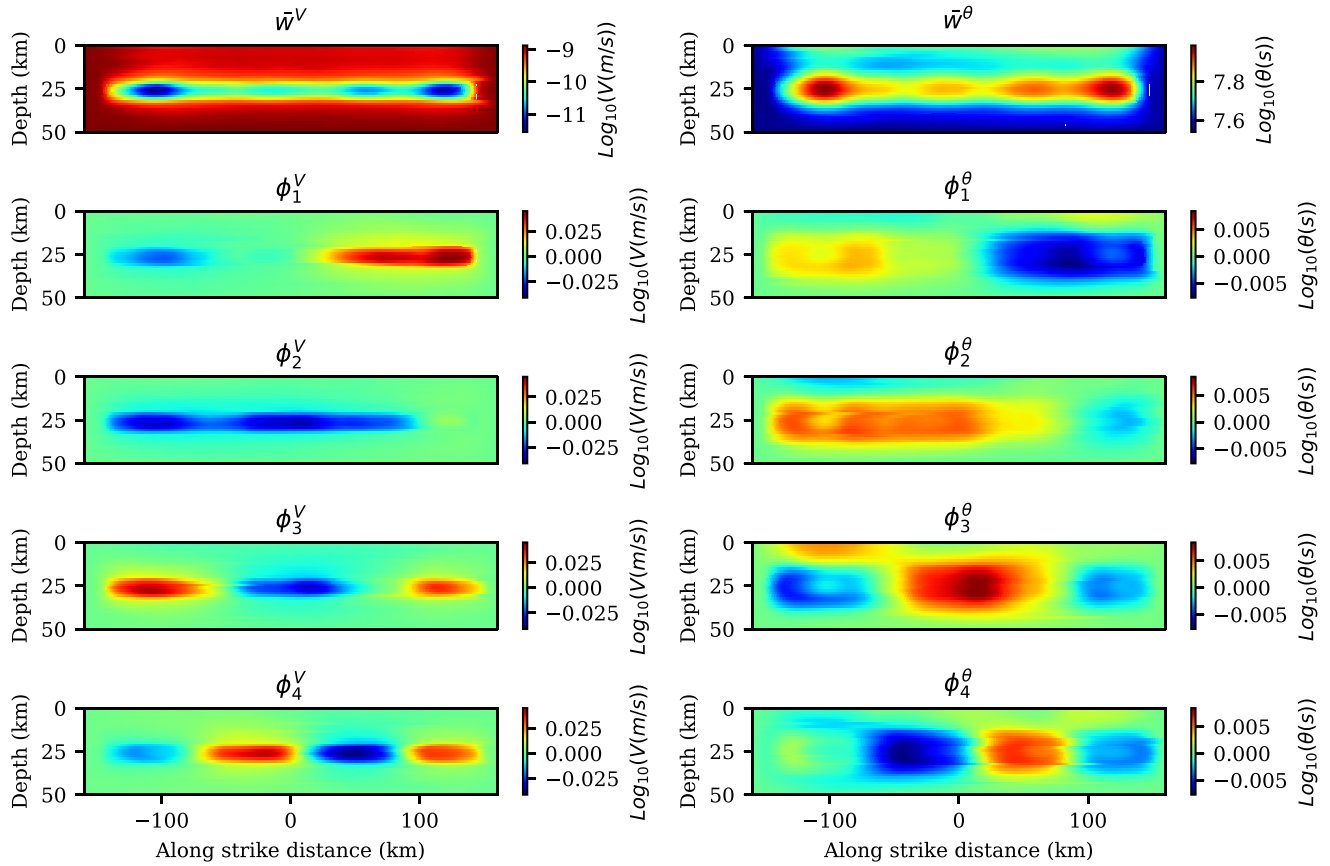
We use a simulation run for a total duration of 2200 yr. We exclude the initial 200 yr to eliminate the transient behaviour, letting the system converge to its chaotic attractor. To define the event set (eq. 6), we set the event threshold  $V_{\text{thresh}} = 5 \times 10^{-8}$  (m s<sup>-1</sup>). The event threshold is chosen such that we get reasonable scaling properties and also, don't lose many events. The time-series of the maximum slip velocity on the fault is plotted in Fig. 2 in which  $V_{\text{thresh}}$  is denoted by a dashed line. We use data from  $t = 200$  to  $t = 1200$  yr to perform the model reduction and find basis functions  $\phi_i$ 's and their corresponding variances  $\lambda_i$ 's. We approximate  $\mathcal{A} \setminus E(V_{\text{thresh}})$  using eq. (10) with a number of modes  $N_m = 13$  which capture more than 85 per cent variance of the data (based on the discussion in Appendix A). The mean of the field ( $\bar{w} = [\bar{w}^V, \bar{w}^\theta]^\top$ ) together with the first four eigenfunctions  $\phi_i = [\phi_i^V, \phi_i^\theta]^\top$  for interevent periods for the time range  $t \in [200, 1200]$  (yr) are plotted in Fig. 3 with  $\bar{w}$  as the empirical mean of the interevent states of the system  $w$ ,  $\phi_i^V$  as the  $i$ th eigenfunction of the  $\log_{10} V$  and  $\phi_i^\theta$  as the  $i$ th eigenfunction of the  $\log_{10} \theta$ . Using  $\phi_i$ 's and  $\lambda_i$ 's, we solve the optimization problem which has  $T$  (prediction horizon),  $\Delta t$  (event duration) and  $r_0$  (amount of perturbation around  $\bar{w}$ ) as hyperparameters. We set the prediction horizon to  $T = 0.5$  (yr) and  $\Delta t = 0.25$  (yr) as the maximum duration of events in the time window of  $t \in [200, 1200]$  yr. With the increase of  $T$ , because of the effect of chaos, the predictability decreases and we would expect the performance of the algorithm to decrease.

The value of  $r_0$  in the eq. (10) controls the size of the hyperellipse which is the constraint of the optimization problem. We perform the optimization for different values of  $r_0$  (in Appendix B). For perturbations constrained within a small hyperellipse (small  $r_0$ ), the algorithm does not find any optimal pre-event state that leads

to a large event. This makes sense because, for small  $r_0$ ,  $w$  is close to the  $\bar{w}$  which is the snapshot-average state of  $w$  during interevent periods. For very large  $r_0$ , the approximation of  $\mathcal{A} \setminus E(V_{\text{thresh}})$  with a hyperellipse is less valid because we let the perturbation have amplitudes much larger than the standard deviation of each component along each eigenfunction. So, one should find an intermediate  $r_0$  whose values of the cost function at the local maxima are larger but close to the maximum magnitude observed in the data set. Here, we report results for  $r_0 = 3$  which means that we don't let the pre-event state go outside the total 3 standard deviation range from  $\bar{w}$  in  $\mathbb{R}^{N_m}$ . Unlike Blonigan *et al.* (2019) that, for a fluid flow problem, found a unique solution for their similar optimization problem, we see convergence to multiple local maxima ( $w^* = [(\log V)^*, (\log \theta)^*]^\top$ ) for different algorithm initiations. As a result, to make our algorithm robust, we solve the optimization problem multiple times with random initiations.

The snapshot-average stress during the interevent period for the VW patch, and the pre-stress corresponding to one of the optimal solutions is plotted in Figs 4(a) and (b). We have also plotted the dimensionless quantity  $\log_{10}(V\theta/D_{RS})$  in Fig. 4(c). The term  $V\theta/D_{RS}$  indicates whether the fault is above or below steady state in the rate-and-state friction law. When  $V\theta/D_{RS} > 1$ , the fault is above steady state, signalling the nucleation phase, while  $V\theta/D_{RS} < 1$  means the fault is below steady state, in a healing phase (Rubin & Ampuero 2005). The cumulative slip distribution corresponding to the event with magnitude 7.5 led by the optimal pre-event state is plotted in Fig. 4(d). We have plotted the slip rate ( $V$ ), and the state variable ( $\theta$ ) corresponding to this particular optimal solution, together with the convergence of the optimization algorithm, in Appendix B. We record the rupture extent of optimal solutions (a total of 12 local maxima) that have  $F_e^* > 7.4$  to use for spatial prediction. These optimal pre-event state distributions are relatively complex with heterogeneities both along the strike and along the dip directions. Because we have only approximated the chaotic attractor by a hyperellipse, the solutions of the optimization problem are unlikely to be exactly observed in the simulation of the dynamical system evolution. However, when initiating from sufficiently close points within the chaotic attractor, the trajectories remain close together during the early stages of their evolution. We rely on this principle to forecast the time and location of large slip events. It is interesting to note that with the defined event threshold, we do not see any full-system size rupture in the forward simulation. However, if we start from homogeneous initial conditions, we see periodic fault-size ruptures. This solution is probably unstable or stable with a small basin of attraction because a relatively small perturbation from the homogeneous initial condition leads to the convergence of the system to its chaotic attractor.

The indicator  $I(t)$  (eq. 12), can effectively forecast large events for the data set from  $t = 1200$  to  $t = 2200$  yr with a prediction horizon of  $T = 0.5$  (yr). To illustrate,  $I(t)$  is plotted alongside  $F$  in Fig. 5(a). A high value of  $F$  shows an upcoming large event in the time interval  $[0, T]$  and we observe that when  $F$  rises, the indicator signals a large event by rising to large values. We define a threshold  $I_e$  above which we signal an upcoming large event. We also define  $F_e$  as the threshold for extreme events; whenever  $F$  is larger than  $F_e$  we say that an extreme event is going to happen in the next  $T$  year(s). The values of  $F_e$  and  $I_e$  are determined such that the proportion of the true positive and true negative forecasts of extreme events are maximized. By recording the values of  $I(t)$  and  $F(t)$ , we can empirically find the conditional probability  $P(F|I)$  (Fig. 5b). Values of  $F_e$  and  $I_e$  are denoted by the white vertical and horizontal dashed lines in Fig. 5(b). The probability in this context is with respect to the



**Figure 3.** Snapshot-average of the  $\log_{10}$  of slip rate ( $\bar{w}^V$ ) and state variable ( $\bar{w}^\theta$ ) during the interevent periods, and first four eigenfunctions for  $\log_{10}$  of slip rate ( $\phi_i^V$  for  $1 \leq i \leq 4$ ) and state variable ( $\phi_i^\theta$  for  $1 \leq i \leq 4$ ) that are ordered by the variance they capture in the data sets. The data set contains interevent snapshots of  $\log_{10}$  of slip rate and state variable during the interevent periods from the year 200 to 1200.

invariant measure of the chaotic attractor. Different quadrants of this plot show four conditions including true negative, false negative, true positive, and false positive from bottom left counter-clockwise to top left. While most of the high values of  $P(F|I)$  lie inside the true negative and true positive regions, it is essential to acknowledge that the probabilities of false negative and false positive are not zero. We also plot the empirical probability of observing an event greater than  $F_e$  given the knowledge of  $I$ , ( $P[F > F_e | I]$ ). This value which is denoted by  $P_{ee}$  is plotted in Fig. 5(c).  $P_{ee}$  consistently rises to values close to one, which is another way to show that the indicator  $I$  can be used as a predictor of large events. We plot the forecast of rupture extent in Fig. 5(d) which shows the effectiveness of both spatial and temporal forecasts of large events. Since we have recorded the rupture extent of optimal solutions (elements in set  $W^*$ ), as soon as the current state of the system gets close to the  $i$ th optimal solution and the indicator signals an upcoming event ( $I(t) > I_e$ ), we propose the recorded rupture extent of the  $i$ th optimal solution as the spatial forecast. Fig. 5(e) shows the temporal forecast of events with the magnitude of events plotted in blue. Whenever the indicator has a value greater than  $I_e$ , we forecast (red region) that an event larger than  $F_e = 6.9$  (black dashed line) will happen. Red shows the temporal prediction of events larger than  $F_e$ . The magnitude in Fig. 5(e) is calculated according to the regular definition of the magnitude of an event (i.e. by integrating the slip velocity above the threshold over the exact duration of the event). In Movie S1, an animation of this prediction is available.

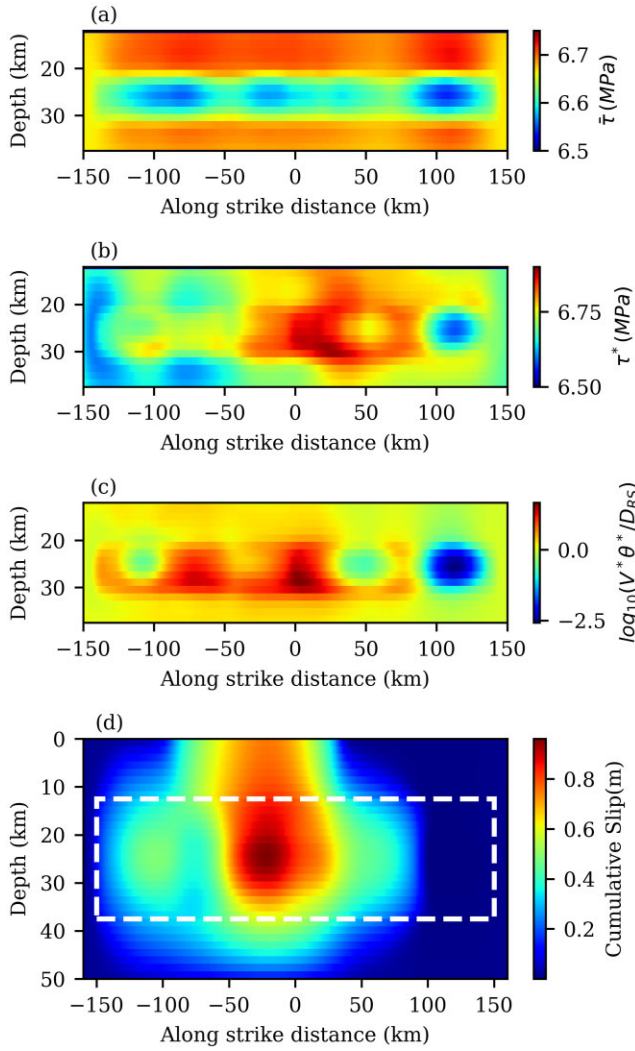
## 4.2 Forecast with partial observation of slip rate

So far, we have assumed that we have full access to the slip rate on the fault. Here, we relax this assumption and use slip rate measurements only at a few points on the fault (diamonds in Fig. 1a). We denote  $\hat{V} : \mathbb{R}^{N_p} \times \mathbb{R}^+ \rightarrow \mathbb{R}^+$  as the time-series of partial slip rate observation, where  $N_p$  is the number of points of slip rate measurements and we take it to be 16 in this case study. We assume that these points are at the centre of the fault along the depth and have equal distances along the strike. We redefine the indicator  $I(t)$  for this special case as follows:

$$I(t) = \max_i \frac{\langle \log \hat{V}(t) - \hat{w}^V, \log \hat{V}_i^* - \hat{w}^V \rangle_{\mathbb{R}^{N_p}}}{\| \log \hat{V}(t) - \hat{w}^V \|_2 \| \log \hat{V}_i^* - \hat{w}^V \|_2}, \quad (13)$$

where  $\hat{V}_i^*$  is the slip rate at the measurement points (diamonds in Fig. 1a) of the  $i$ th optimal solution in the set  $W^*$ .  $\hat{w}^V$  is the average slip rate at the measurement points during the interevent period.  $\langle \cdot, \cdot \rangle_{\mathbb{R}^{N_p}}$  is the inner product in  $\mathbb{R}^{N_p}$ . Fig. 6 shows the forecast performance in the limited slip rate measurement scenario. The general consistent increase in  $I(\cdot)$  when the function  $F(\cdot)$  is visible in Fig. 6(a). Figs 6(b) and (c) show statistically the performance of the predictor. While most of the probability mass of  $P(F|I)$  belongs to true positive and true negative we should appreciate that there is more probability mass in the false positive quadrant compared to the scenario in which we have full access to the slip rate. This can be observed better in Figs 6(c), (d) and (e). Although as  $I$  increases,





**Figure 4.** Snapshot-average of the shear stress on the VW patch of the fault during the interevent period (a). One of the local optimal pre-stress distributions that leads to an event with a magnitude of 7.5 (b). The dimensionless quantity  $\log_{10}(V^*\theta^*/D_{RS})$  for the optimal pre-event  $V^*$  and  $\theta^*$  is plotted in (c). The corresponding cumulative slip of the event that happens right after starting from optimal pre-event state (d). To increase the readability (a, b, c) are plotted only for the VW patch. The VW patch in (d) is denoted by the dashed white line.

$P_{ee}$  increases consistently,  $P_{ee}$  is almost 0.9 when  $I$  is the maximum which suggests that there is a 10 per cent chance of a false positive signal when  $I$  takes its maximum value. This false positive can also be observed in Figs 6(d) and (e) around the year 1610. While it is important to appreciate the limitations, the overall performance is satisfying. To reduce this limitation, one can use filtering methods to invert and approximate slip rates at a few more points on the fault to improve the performance.

### 4.3 Impact of low resolution observation on prediction accuracy

In this part, we illustrate a limitation of our method as we lose more and more information with loosing the resolution of the data. Real-world slip inversion on the fault has inherent low-pass filter because the process of finding slip on the fault from surface

displacements involves filtering techniques that inevitably introduce this type of limitation. These techniques are necessary due to the measurement limitations, which cannot capture high-frequency variations accurately, leading to a smoother and potentially less precise representation of the actual slip rates. We apply a Gaussian kernel to the synthetic slip rate data, mimicking the characteristics of realistic data sets. This approach enables us to systematically assess the impact of reduced resolution in the observed slip rate on the performance of extreme event prediction. By varying the standard deviation of the Gaussian kernel, we evaluate how different resolutions affect the algorithm's accuracy. The standard deviation is expressed in a dimensionless form relative to the width of the VW zone.

We assume that the slip rate is corrupted by a Gaussian kernel which is defined mathematically as:

$$G(x, y) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{x^2 + y^2}{2\sigma^2}\right), \quad (14)$$

where  $\sigma$  is the standard deviation of the Gaussian kernel, controlling the extent of the smoothing effect. By convolving this kernel with the original slip rate data  $V(x, y)$ , we obtain the low resolution slip rate  $V'(x, y)$ :

$$V'(x, y) = \int_{\Gamma} V(x', y') \cdot G(x - x', y - y') dx' dy'. \quad (15)$$

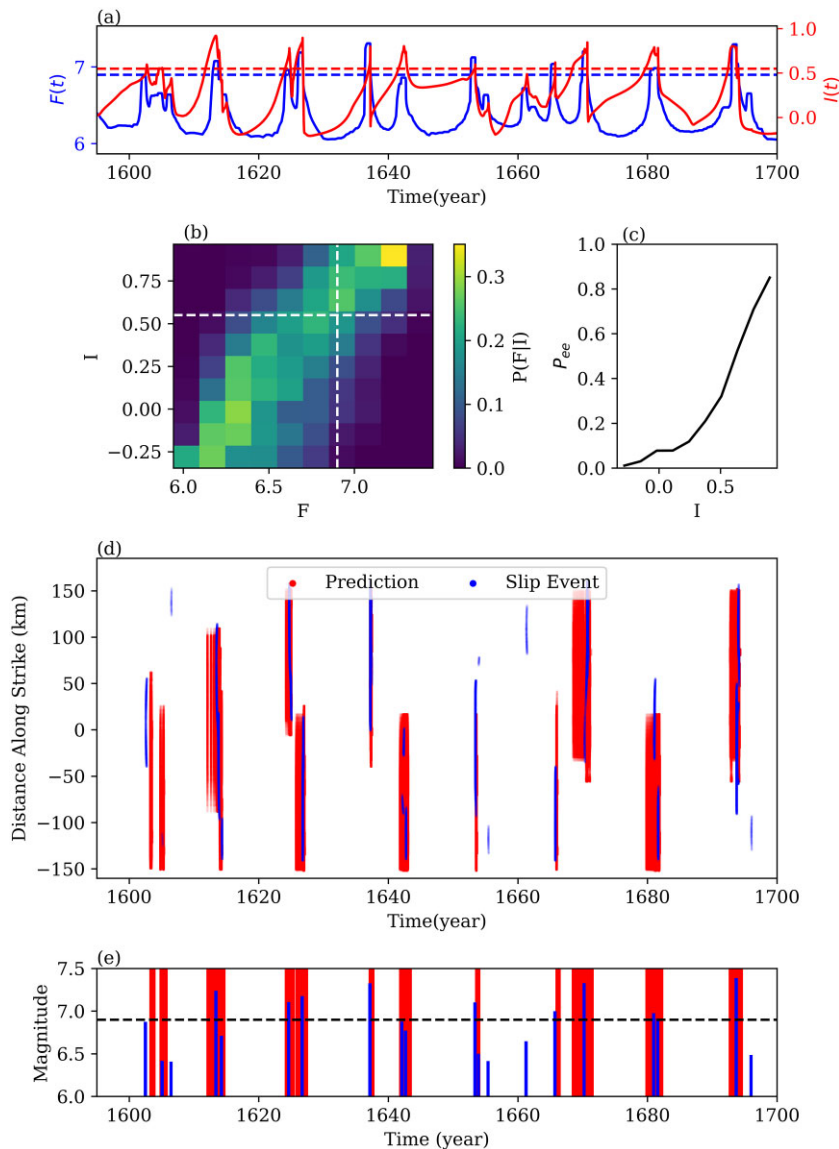
To visually demonstrate the effect of the kernel on the data, we plotted one snapshot of slip rate without applying the low-pass filter in Fig. 7(a) and then applied the low-pass filter with different standard deviation on that snapshot of the velocity and plot them in Figs 7(b)–(d). The conditional probability  $P(F|I)$  for a 1000-yr-long data that are corrupted by these Gaussian kernels are plotted in Figs 7(e)–(g). As the resolution decreases the probability mass in the upper left (false positive) and lower right (false negative) increases. Figs 7(f) and (g) show that with a standard deviation greater than  $0.5W_{VW}$ , we have a large probability of a false signal. This is a limitation of our work and potentially considering more POD modes, using data assimilation techniques to more accurately invert for slip on the fault, and considering the history of the time-series are some of the methods that can be used to improve the performance when the slip rate on the fault is not well constrained.

## 5 DISCUSSION

Our results demonstrate the possibility of predicting the time, size and spatial extent of extreme events in a simplified dynamical model of earthquake sequences based on the instantaneous observation of fault slip rate. By constraining the pre-event state on a fault to the only feasible ones and solving an optimization problem, we found the optimal pre-event state in a low-dimensional space. Optimal pre-event state refers to configurations of slip rate and state variable heterogeneity on the fault triggering large events within small time windows. Identifying the optimal pre-event state distributions that are also statistically accessible during the earthquake cycle is pivotal.

States of the system self-organize into a chaotic attractor which occupies only a small fraction of all possible distributions on the fault. The identification of the optimal pre-event state within this reduced set is crucial for two reasons. First, it helps establish a low-dimensional representation of optimal pre-event state; the significance of reduced-order proxy of critical pre-event state is even more important for earthquakes than SSEs, primarily due to the scarcity of observational data obtained from palaeoseismic records.





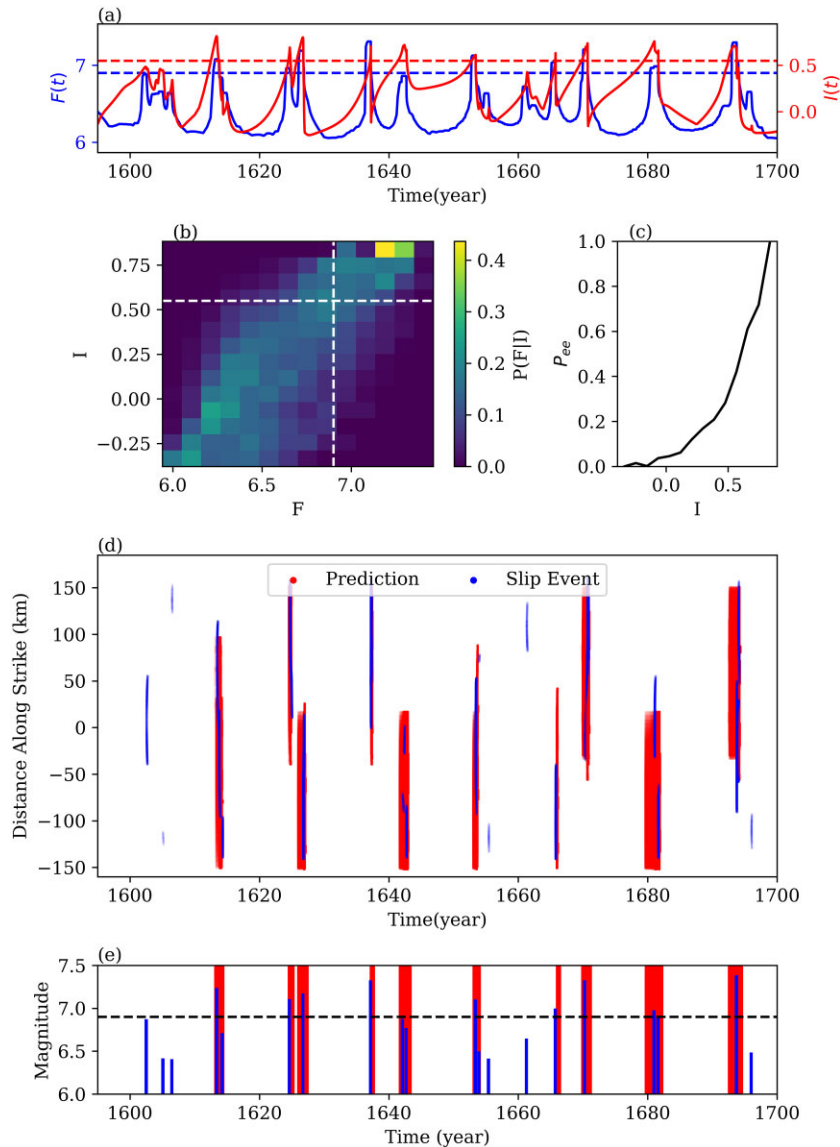
**Figure 5.** Spatiotemporal prediction of events. The time-series of the functions  $F$  and  $I$  show that  $I$  rises when there is an upcoming large event ( $F$  is large), and it goes down when there is no upcoming large event. The blue and red dashed lines correspond to  $F_e$  and  $I_e$  (a). The empirical conditional probability  $P(F|I)$ . The vertical and horizontal dashed lines are  $F_e$ , and  $I_e$ , respectively (b). The empirical probability of having an event with the value  $\tilde{M}$  greater than  $F_e$  in the next 0.5 (year) as a function of the value of the indicator  $I$  (c). The spatiotemporal prediction of events is plotted by red where blue is the actual events in the data set (d). Prediction of the magnitudes with the blue bars as the magnitude of events in the data set. The horizontal axis for the blue bars denotes the time when an event starts. Red regions denote the times of high probability of large events [above magnitude 6.9 (dashed line)] based on our indicator (e). The statistical plots (b, c) are calculated based on 1000 yr of data in the test set (data from the year 1200 to 2200).

Secondly, everything outside this set remains unseen during the earthquake cycle's evolution. If that was not the case, the space of hypothetical stress distribution possibly leading to large events would be intractable.

In Section 4.2, we studied a scenario in which the slip rate is known at only a few points on the fault. The results are almost as good as when we have full access to the slip rate on the fault because the slip evolution at neighbouring points on the fault is strongly correlated due to elastic coupling. This result most likely benefits from large nucleation length for SSEs which is generally not true for earthquakes. The nucleation length for a 1-D fault for mode III is given by  $h_{ra} = \frac{2GD_{RS}b}{\pi\bar{\sigma}(b-a)^2}$  (Rubin & Ampuero 2005), where  $G$  is shear modulus,  $\bar{\sigma}$  is the effective normal stress and  $a$ ,  $b$ ,  $D_{RS}$  are

frictional parameters. For a 2-D fault, the nucleation size is given by  $h = (\pi^2/4)h_{ra}$  (Chen & Lapusta 2009), and is 29.7 (km) in our model, whereas the width of the VW zone is  $W_{VW} = 25$  (km).

Slip rate data of a fault is determined through the inversion of surface displacement, which results in low spatial resolution. We therefore studied the performance of extreme event prediction when the synthetic slip rate is corrupted by a low pass filter. Our results Fig. 7 indicate that predictability is compromised when the standard deviation of the low-pass filter kernel gets larger and larger. This finding highlights a limitation in the application of our study in its current form when the slip rate on the fault is not well resolved. Addressing this limitation will be a focus of our future work. Potential approaches include incorporating additional components into



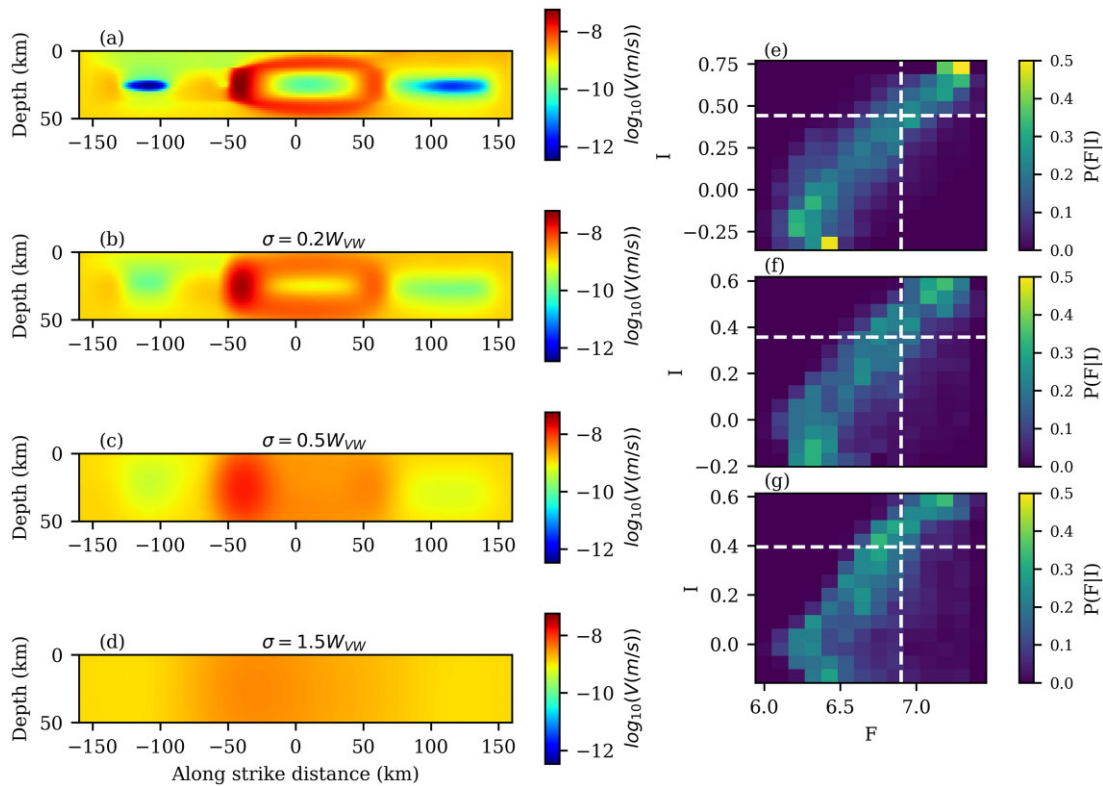
**Figure 6.** Spatio-temporal prediction of events same as in Fig. 5 but using slip rate only at 16 points on the fault (denoted in Fig. 1a by diamonds).

the extreme event criteria and solving a data assimilation problem, such as using the Ensemble Kalman filter, to more accurately invert for slip rates on the fault.

For earthquakes, the ratio of the nucleation size to fault dimensions is much smaller than in SSEs. Rupture dynamics considerations indicate that the initial shear stress must be sufficiently high and well-correlated across the entire fault for a system-spanning earthquake to occur. Therefore, having information at least at the scale of the fault dimension is essential to predict *whether* a big rupture will happen. However, to predict when the event will nucleate, it might be necessary to resolve the system at the scale of the nucleation length, as constraints on the slip rate at this scale are crucial. A key question remains: for earthquakes, is resolving the system at the nucleation length scale necessary for time predictability, and is resolution at the fault dimension sufficient to predict the extent of rupture? Investigating the role of observational resolution in the predictability of both the timing and extent of future seismic events remains a significant challenge, which we aim to address in future works.

## 6 CONCLUSION

Our study suggests that the chaotic nature of earthquake sequences is not an insurmountable obstacle to time-dependent earthquake forecasting. However, we acknowledge that we considered a favourable model setup designed to produce SSEs. It would be now interesting to test this approach in the case of a model setup producing regular earthquakes (i.e. with slip rates of  $1 \text{ cm s}^{-1}$  to  $1 \text{ m s}^{-1}$  to be comparable to real earthquakes) with larger ratios of fault dimensions to nucleation size and with a larger range of earthquake magnitudes (Cattania 2019; Barbot 2021; Lambert & Lapusta 2021). This is doable although computationally challenging. The amplitude of the stress heterogeneity would be more substantial for regular earthquakes, where dynamical wave-mediated stresses allow for rupture propagation over lower stress conditions than for aseismic slip, particularly in models with stronger dynamical weakening or with persistent heterogeneity such as normal stress perturbations (Noda *et al.* 2009; Lambert *et al.* 2021).



**Figure 7.** Impact of the lowering the observed slip rate resolution on prediction. One snapshot of the slip rate is plotted in (a). To visualize the effect of reduction of resolution, the low-pass filter applied to the snapshot in (a) is plotted in (b, c, d) with different standard deviations. The conditional probability of  $P(F|I)$  when the slip rate is corrupted with a Gaussian low-pass filter with different standard deviations ( $\sigma = 0.2W_{VW}$ ,  $0.5W_{VW}$ ,  $1.5W_{VW}$ ) are plotted in (e, f, g) respectively.

It is expected that earthquake sequences would then show more complexity due to the cascading effects which are responsible for foreshocks and aftershocks in natural earthquake sequences, and which are not present in our simulations. In that regard, Blonigan *et al.* (2019) reported that the performance of their prediction of rare events diminishes with the increase in Reynolds number in their turbulent flow case. It is possible that we have the same limitation as the ratio of the nucleation size to the dimensions of the fault decreases.

## ACKNOWLEDGMENTS

The authors HK and J-PA express their sincere gratitude to the National Science Foundation (NSF) for their financial support of this research project, through the Industry-University Collaborative Research Center Geomechanics and Mitigation of Geohazards (award #1822214). Author AMS is grateful to DoD for support as a Vannevar Bush Faculty Fellow. Additionally, the authors are grateful for the valuable input and discussion provided by Nadia Lapusta, Brendan Meade, Themis Sapsis, Adriano Gualandi, Alba Rodriguez, Camilla Cattania, Elif Oral, Mateo Acosta, Kelian Dascher-Cousineau, Zachary R Ross, Jan Dirk Jansen, Kyungjae Im. The authors also extend their appreciation to Eric Dunham and the other anonymous reviewer for their constructive feedback and insightful comments.

## SUPPORTING INFORMATION

Supplementary data are available at *GJIRAS* online.

Supplemental Videos: Movie S1 and Movie S2

Please note: Oxford University Press is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the paper.

## DATA AVAILABILITY

We used a model of a 2-D thrust fault in a 3-D medium governed by rate-and-state friction with ageing law for the evolution of state variable ( $\theta$ ). The model parameters are summarized in Table 1. To simulate the forward model, we use the QDYN software,<sup>2</sup> which is an open-source code to simulate earthquake cycles (Luo *et al.* 2017). We use the POD technique to reduce the dimensionality of the problem. This method is reviewed in Appendix A. To solve the optimization problem we use the Bayesian optimization method (Brochu *et al.* 2010; Blanchard & Sapsis 2021) that is reviewed in Appendix B. We used the open source code available on GitHub<sup>3</sup> for solving the optimization problem.

## REFERENCES

Anghel, M., Ben-Zion, Y. & Rico-Martinez, R., 2004. Dynamical system analysis and forecasting of deformation produced by an earthquake

<sup>2</sup><https://github.com/ydluo/qdyn>.

<sup>3</sup><https://github.com/ablancha/gpsearch>.



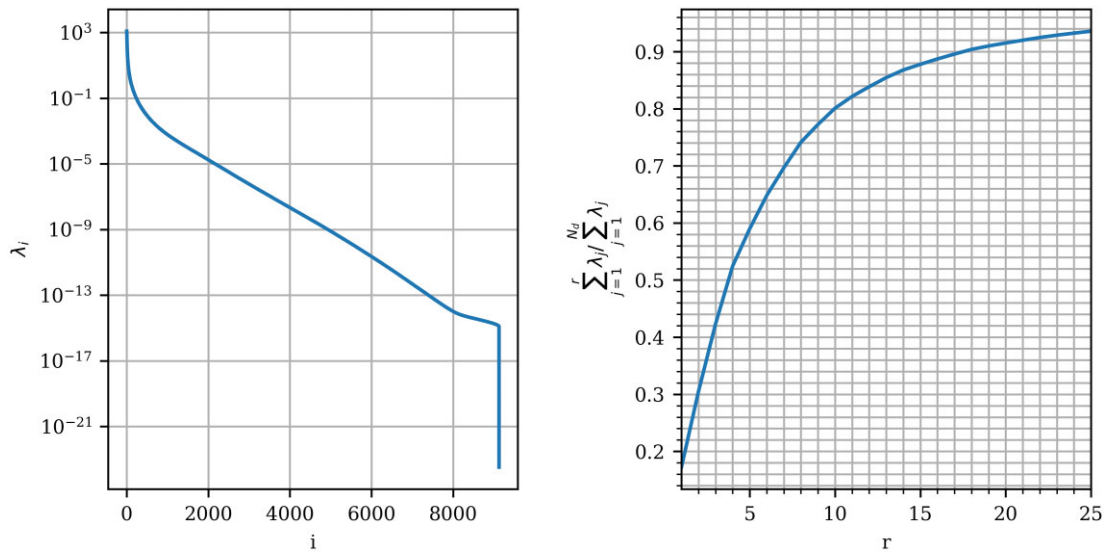
- fault, in *Computational Earthquake Science Part I, Pageoph Topical Volumes*, pp. 2023–2051, eds, Donnellan, A., Mora, P., Rico-Martinez, R., Matsu'ura, M. & Yin, X.-C., Birkhäuser.
- Avouac, J.-P., 2015. From geodetic imaging of seismic and aseismic fault slip to dynamic modeling of the seismic cycle, *Annu. Rev. Earth planet. Sci.*, **43**(1), 233–271.
- Barbot, S., 2019. Slow-slip, slow earthquakes, period-two cycles, full and partial ruptures, and deterministic chaos in a single asperity fault, *Tectonophysics*, **768**, 228171, doi:10.1016/j.tecto.2019.228171
- Barbot, S., 2021. A spectral boundary–integral method for quasi–dynamic ruptures of multiple parallel faults, *Bull. seism. Soc. Am.*, **111**(3), 1614–1630.
- Barbot, S., Lapusta, N. & Avouac, J.-P., 2012. Under the hood of the earthquake machine: toward predictive modeling of the seismic cycle, *Science*, **336**(6082), 707–710.
- Becker, T.W., 2000. Deterministic chaos in two state-variable friction sliders and the effect of elastic interactions, in *Geocomplexity and the Physics of Earthquakes*, Vol. **120**, pp. 5–26, eds Rundle, J. B., Turcotte, D. L. & Klein, W., American Geophysical Union.
- Blanchard, A. & Sapsis, T., 2021. Bayesian optimization with output-weighted optimal sampling, *J. Comput. Phys.*, **425**, doi:10.1016/j.jcp.2020.109901.
- Blonigan, P.J., Farazmand, M. & Sapsis, T.P., 2019. Are extreme dissipation events predictable in turbulent fluid flows?, *Phys. Rev. Fluids*, **4**(4), doi:10.1103/PhysRevFluids.4.044606.
- Brandstätter, A., Swift, J., Swinney, H.L., Wolf, A., Farmer, J.D., Jen, E. & Crutchfield, P.J., 1983. Low-dimensional chaos in a hydrodynamic system, *Phys. Rev. Lett.*, **51**(16), 1442–1445.
- Brochu, E., Cora, V.M. & de Freitas, N., 2010. A tutorial on bayesian optimization of expensive cost functions, with application to active User Modeling and Hierarchical Reinforcement Learning, doi:10.48550/arXiv.1012.2599
- Cattania, C., 2019. Complex earthquake sequences on simple faults, *Geophys. Res. Lett.*, **46**(17–18), 10 384–10 393.
- Chen, T. & Lapusta, N., 2009. Scaling of small repeating earthquakes explained by interaction of seismic and aseismic slip in a rate and state fault model, *J. geophys. Res.*, **114**(B1), doi:10.1029/2008JB005749
- Dal Zilio, L., Lapusta, N. & Avouac, J., 2020. Unraveling scaling properties of slow–slip events, *Geophys. Res. Lett.*, **47**(10), doi:10.1029/2020GL087477.
- Farazmand, M. & Sapsis, T.P., 2017. A variational approach to probing extreme events in turbulent dynamical systems, *Sci. Adv.*, **3**(9), e1701533, doi:10.1126/sciadv.1701533
- Farazmand, M. & Sapsis, T.P., 2019. Extreme events: mechanisms and prediction, *Appl. Mech. Rev.*, **71**(5), 050801, doi:10.1115/1.4042065
- Geubelle, P.H. & Rice, J.R., 1995. A spectral method for three-dimensional elastodynamic fracture problems, *J. Mech. Phys. Solids*, **43**(11), 1791–1824.
- Gualandi, A., Avouac, J.-P., Michel, S. & Faranda, D., 2020. The predictable chaos of slow earthquakes, *Sci. Adv.*, **6**(27), eaaz5548, doi:10.1126/sciadv.aaz5548
- Gualandi, A., Faranda, D., Marone, C., Cocco, M. & Mengaldo, G., 2023. Deterministic and stochastic chaos characterize laboratory earthquakes, *Earth planet. Sci. Lett.*, **604**, 117995, doi:10.1016/j.epsl.2023.117995
- Gutenberg, B. & Richter, C.F., 1950. Seismicity of the Earth and associated phenomena, *MAUSAM*, **1**(2), 174–176.
- Huang, J. & Turcotte, D.L., 1990. Are earthquakes an example of deterministic chaos?, *Geophys. Res. Lett.*, **17**(3), 223–226.
- Kaneko, Y., Avouac, J.-P. & Lapusta, N., 2010. Towards inferring earthquake patterns from geodetic observations of interseismic coupling, *Nat. Geosci.*, **3**(5), 363–369.
- Kato, N., 2016. Earthquake cycles in a model of interacting fault patches: complex behavior at transition from seismic to aseismic slip, *Bull. seism. Soc. Am.*, **106**(4), 1772–1787.
- Konca, A.O. et al., 2008. Partial rupture of a locked patch of the Sumatra megathrust during the 2007 earthquake sequence, *Nature*, **456**(7222), 631–635.
- Lambert, V. & Lapusta, N., 2021. Resolving simulated sequences of earthquakes and fault interactions: implications for physics-based seismic hazard assessment, *J. geophys. Res.*, **126**(10), e2021JB022193, doi:10.1029/2021JB022193
- Lambert, V., Lapusta, N. & Faulkner, D., 2021. Scale dependence of earthquake rupture prestress in models with enhanced weakening: implications for event statistics and inferences of fault stress, *J. geophys. Res.*, **126**(10), e2021JB021886, doi:10.1029/2021JB021886
- Lapusta, N. & Liu, Y., 2009. Three-dimensional boundary integral modeling of spontaneous earthquake sequences and aseismic slip, *J. geophys. Res.*, **114**(B9), doi:10.1029/2008JB005934
- Lapusta, N. & Rice, J.R., 2003. Nucleation and early seismic propagation of small and large events in a crustal earthquake model, *J. geophys. Res.*, **108**(B4), doi:10.1029/2001JB000793
- Lapusta, N., Rice, J.R., Ben-Zion, Y. & Zheng, G., 2000. Elastodynamic analysis for slow tectonic loading with spontaneous rupture episodes on faults with rate- and state-dependent friction, *J. geophys. Res.*, **105**(B10), 23765–23789.
- Li, M., Jain, S. & Haller, G., 2023. Model reduction for constrained mechanical systems via spectral submanifolds, *Nonlin. Dyn.*, **111**(10), 8881–8911.
- Luo, Y., Ampuero, J.P., Galvez, P., Van Den Ende, M. & Idini, B., 2017. QDYN: a Quasi-DYNamic earthquake simulator (v1.1), doi:10.5281/zenodo.322459
- Marone, C., 1998. Laboratory-derived friction laws and their application to seismic faulting, *Annu. Rev. Earth planet. Sci.*, **26**(1), 643–696.
- Marsan, D. & Tan, Y.J., 2020. Maximum earthquake size and seismicity rate from an ETAS model with slip budget, *Bull. seism. Soc. Am.*, **110**(2), 874–885.
- Michel, S., Gualandi, A. & Avouac, J.-P., 2019. Similar scaling laws for earthquakes and Cascadia slow-slip events, *Nature*, **574**(7779), 522–526.
- Murray, J. & Segall, P., 2002. Testing time-predictable earthquake recurrence by direct measurement of strain accumulation and release, *Nature*, **419**(6904), 287–291.
- Noda, H., Dunham, E.M. & Rice, J.R., 2009. Earthquake ruptures with thermal weakening and the operation of major faults at low overall stress levels, *J. geophys. Res.*, **114**(B7), doi:10.1029/2008JB006143
- Okubo, P.G. & Aki, K., 1987. Fractal geometry in the san andreas fault system, *J. geophys. Res.*, **92**(B1), 345–355.
- Reid, H., 1910. The Mechanics of the Earthquake, The California Earthquake of April 18, 1906, Report of the State Investigation Commission, Vol. 2, Caknegie Institution of Washington.
- Rice, J.R., 1993. Spatio-temporal complexity of slip on a fault, *J. geophys. Res.*, **98**(B6), 9885–9907.
- Rowley, C.W. & Dawson, S.T., 2017. Model reduction for flow analysis and control, *Annu. Rev. Fluid Mech.*, **49**(1), 387–417.
- Rubin, A.M. & Ampuero, J.-P., 2005. Earthquake nucleation on (aging) rate and state faults, *J. geophys. Res.*, **110**(B11), doi:10.1029/2005JB003686
- Ruina, A., 1983. Slip instability and state variable friction laws, *J. geophys. Res.*, **88**(B12), 10 359–10 370.
- Sagy, A., Brodsky, E.E. & Axen, G.J., 2007. Evolution of fault-surface roughness with slip, *Geology*, **35**(3), 283–286.
- Scholz, C.H., 1989. Mechanics of faulting, *Annu. Rev. Earth planet. Sci.*, **17**, 309–334.
- Shaw, B.E. & Rice, J.R., 2000. Existence of continuum complexity in the elastodynamics of repeated fault ruptures, *J. geophys. Res.*, **105**(B10), 23 791–23 810.
- Stiernström, V., Almquist, M. & Dunham, E.M., 2024. Adjoint-based inversion for stress and frictional parameters in earthquake modeling, doi:10.48550/arXiv.2310.12279
- Taira, K. et al., 2017. Modal analysis of fluid flows: an overview, *AIAA J.*, **55**(12), 4013–4041.
- Thomas, M.Y., Lapusta, N., Noda, H. & Avouac, J.-P., 2014. Quasi-dynamic versus fully dynamic simulations of earthquakes and aseismic slip with and without enhanced coseismic weakening, *J. geophys. Res.*, **119**(3), 1986–2004.
- Utsu, T., Ogata, Y., S, R. & Matsu'ura, 1995. The centenary of the Omori formula for a decay law of aftershock activity, *J. Phys. Earth*, **43**(1), 1–33.

**APPENDIX A: PROPER ORTHOGONAL DECOMPOSITION (POD): METHOD AND RESULT**

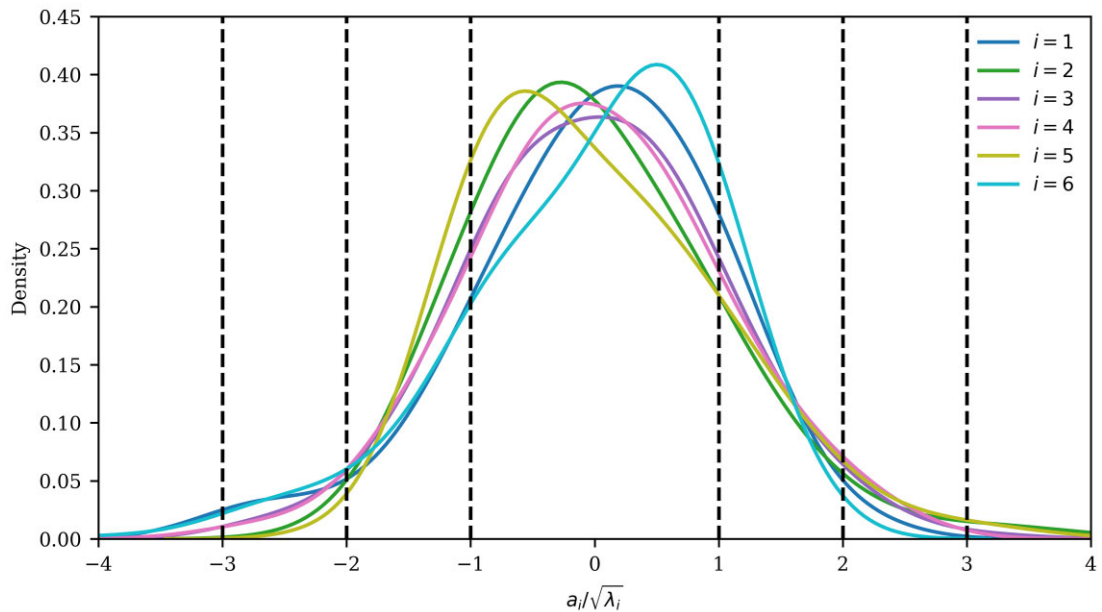
In this section, we review how to reduce the dimension of the data set consisting of slip rate and state variable using the POD method. We use this method to find critical pre-event state in a low-dimensional space instead of the high-dimensional function space. Another reason to use this method is because eq. (9) is an optimization problem constrained on the chaotic attractor of the system with the event period excluded. To solve the constraint optimization problem (eq. 9), one method (Farazmand & Sapsis 2017) is to exclude the extreme events from the chaotic attractor and approximate the remaining using the POD technique. Here, we exclude the event period from the data set to only approximate the interevent period. The method of approximating the chaotic attractor using POD modes is used in

different fields. As an example, the work in (Blonigan *et al.* 2019) used 50 POD modes to approximate the chaotic attractor of a turbulent channel flow. One behavioral difference between our model of the earthquake cycle and the turbulent channel flow example is that the time stepping in our problem is adaptive due to the system’s multi-scale behavior; there are more sample data when the dynamical system is stiff. However, since we are removing the event period from the data, we only include the slow part of the system in our data set.

In the following paragraphs, we describe the POD analysis on our data set of simulations. The data set comprises snapshots within the time span from the year 200 to 1200 excluding the event set ( $E(V_{\text{thresh}})$ ). We use the time snapshots of discretized states of the system ( $\theta$  and  $V$ ) which belong to a high but finite-dimensional space. After discretization,  $V : \mathbb{R}^{M_x \times N_y} \times \mathbb{R}^+ \rightarrow \mathbb{R}^+$



**Figure A1.** Convergence of the eigenvalues (left) and the ratio of a truncated sum of eigenvalues to the total sum of eigenvalues (right).



**Figure A2.** The distribution of  $a_i(t)/\sqrt{\lambda_i}$  in the data set of the interevent periods. The vertical lines correspond to  $a_i/\sqrt{\lambda_i} = \pm 1, \pm 2, \pm 3$  and are plotted to give insight for selecting proper  $r_0$  in eq. (10).

and  $\theta : \mathbb{R}^{N_x \times N_y} \times \mathbb{R}^+ \rightarrow \mathbb{R}^+$ .  $N_x = 256$  and  $N_y = 32$  are the numbers of grid points along the strike and depth respectively.

Since the evolution of the system is better realized in  $\log_{10}$  space, we apply the POD on the  $\log_{10}$  of the data set. We define vectors  $w_1(t_k)$  and  $w_2(t_k)$  both in  $\mathbb{R}^{N_x N_z}$  for time  $t_k$  as the vectorized form of the logarithm of  $V$  and  $\theta$  at time  $t_k$ .

$$w_1(t_k) = \log_{10} \begin{pmatrix} V_{1,1} \\ V_{1,2} \\ \vdots \\ V_{1,N_x} \\ V_{2,1} \\ V_{2,2} \\ \vdots \\ V_{2,N_x} \\ \vdots \\ V_{N_z,1} \\ V_{N_z,2} \\ \vdots \\ V_{N_z,N_x} \end{pmatrix}_{t=t_k} \quad (\text{A1})$$

$$w_2(t_k) = \log_{10} \begin{pmatrix} \theta_{1,1} \\ \theta_{1,2} \\ \vdots \\ \theta_{1,N_x} \\ \theta_{2,1} \\ \theta_{2,2} \\ \vdots \\ \theta_{2,N_x} \\ \vdots \\ \theta_{N_z,1} \\ \theta_{N_z,2} \\ \vdots \\ \theta_{N_z,N_x} \end{pmatrix}_{t=t_k} \quad (\text{A2})$$

where for example, by  $[V_{i,j}]_{t_k}$ , we mean slip rate at  $i$ th element along strike and  $j$ th element along the depth at  $k$ th snapshots in the data set. Then, we stack pairs of  $w_1$  and  $w_2$  to make a vector  $w$ :

$$w(t_k) = \begin{bmatrix} w_1(t_k) \\ w_2(t_k) \end{bmatrix} \in \mathbb{R}^{2N_x N_z}. \quad (\text{A3})$$

We define  $\bar{w} = [\bar{w}^V, \bar{w}^\theta]^\top$  as the average of  $w(t_i)$  for all snapshots in the dataset, where the dataset consists of snapshots of the field captured with variable time stepping during the interevent period.

$$\bar{w} = \frac{1}{N_d} \sum_{i=1}^{N_d} w(t_i) \quad (\text{A4})$$

where  $N_d$  is the total number of snapshots in the data set.  $\bar{w}^V$  and  $\bar{w}^\theta$  are plotted in Fig. 3. We define  $p(t_k) = w(t_k) - \bar{w}$  and then we define a matrix  $P \in \mathbb{R}^{2N_x N_z \times N_d}$  with the following entries:

$$P = [p(t_1) \ p(t_2) \ \cdots \ p(t_{N_d})] \in \mathbb{R}^{2N_x N_z \times N_d}. \quad (\text{A5})$$

Then, we define the covariance matrix  $R$  as the following:

$$R = \frac{1}{(N_d - 1)} P P^T \in \mathbb{R}^{2N_x N_z \times 2N_x N_z} \quad (\text{A6})$$

Now, we can find the eigenvectors of matrix  $R$ :

$$R \phi_j = \lambda_j \phi_j \quad \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{2N_x N_z} \geq 0. \quad (\text{A7})$$

Eigenvalues show how well each eigenvector captures the original data in  $L^2$  sense. Eigen-vectors of matrix  $R$  can be found using Singular Value Decomposition (SVD) of matrix  $P$ :

$$P = \Phi \Sigma \Psi^T \quad (\text{A8})$$

where in general  $\Phi \in \mathbb{R}^{2N_x N_z \times 2N_x N_z}$  and  $\Psi \in \mathbb{R}^{N_d \times N_d}$  are orthogonal ( $\Phi \Phi^T = I_{2N_x N_z \times 2N_x N_z}$  and  $\Psi \Psi^T = I_{N_d \times N_d}$ ) and determine, through columns, the left and right singular vectors of  $P$ ; and diagonal matrix  $\Sigma \in \mathbb{R}^{2N_x N_z \times N_d}$  has singular values on its diagonal (Taira *et al.* 2017). We can write:

$$R = \frac{1}{(N_d - 1)} P P^T = \frac{1}{(N_d - 1)} \Phi \Sigma \Psi^T \Psi \Sigma^T \Phi^T$$

$$R \Phi = \frac{1}{(N_d - 1)} \Phi \Sigma \Sigma^T \quad (\text{A9})$$

because of the special form of  $\Sigma$  that will be discussed shortly, the columns of  $\Phi$  (denoted here by  $\phi_i$  and are plotted in Fig. 3 for  $i \leq 4$ ) are eigenvectors of matrix  $R$  that are ordered by the variance they capture in data. Note that  $\phi_i \in \mathbb{R}^{2N_x N_z}$  and we can separate it into eigenvectors of the slip rate ( $\phi_i^V$ ) and the state variable  $\phi_i^\theta$ :

$$\phi_i = \begin{bmatrix} \phi_i^V \\ \phi_i^\theta \end{bmatrix} \quad (\text{A10})$$

Assuming the number of time snapshots is much smaller than the dimension of the problem  $N_d \ll 2N_x N_z$ ,  $\Sigma$  has the following form:

$$\Sigma = \begin{bmatrix} \sigma_1 & 0 & 0 & 0 \\ 0 & \sigma_2 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \sigma_{N_d} \\ 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 \end{bmatrix}_{2N_x N_z \times N_d} \quad (\text{A11})$$

Then, using eqs (A7), (A9), and (A11),  $\frac{1}{(N_d - 1)} \sigma_j^2 = \lambda_j$ .  $\lambda_j$  corresponds to the variance of the data along  $\phi_j$ . If  $\lambda_j$  goes to zero very fast, it suggests that we can explain the data set in a low-dimensional subspace consisting of a finite number of eigenfunctions. The ratio  $\sum_{j=1}^r \lambda_j / \sum_{j=1}^{N_d} \lambda_j$  shows the proportion of the variance of the data that are captured in the first  $r$  eigenfunctions. Based on Fig. A1, the first 13 modes of the data capture almost 85 per cent of the data.

Using this explanation, we can approximate the interevent period ( $\mathcal{A} \setminus E(V_{\text{thresh}})$ ) by:

$$\log_{10} (\mathcal{A} \setminus E(V_{\text{thresh}})) \approx \left\{ w = \bar{w} + \sum_{i=1}^{N_m} a_i \phi_i \mid \sum_{i=1}^{N_m} \frac{a_i^2}{\lambda_i} \leq r_0^2 \right\}. \quad (\text{A12})$$

where  $N_m$  is the number of modes (eigenfunctions) that are considered in the truncation. One can play with  $r_0$  to enlarge the set. For very large  $r_0$  the approximation is not valid anymore. The value of  $r_0$  determines how much we let perturbation around the average of the data set  $\bar{w}$ . As an example, taking  $N_m = 1$  and  $r_0 = 1$  would let perturbation around  $\bar{w}$  along  $\phi_1$  with an amplitude equal to the standard deviation of the data set along that eigenvector ( $\sqrt{\lambda_1}$ ).

Using the orthonormality of  $\phi_i$ 's, we can find the projection of any  $w(t)$  onto  $\phi_i$  using the following inner product:

$$a_i(t) = \langle w(t) - \bar{w}, \phi_i \rangle \quad (\text{A13})$$



where  $a_i(t)$  is the projection of  $w(t) - \bar{w}$  onto eigenvector  $\phi_i$  and  $\langle, \rangle$  denotes the inner product. We can find  $a_i(t_k)$  for all of the time snapshots in the data set and plot the distribution of  $a_i/\sqrt{\lambda_i}$  (Fig. A2). We see that the distribution is close to the standard normal distribution. Looking at this figure gives us intuition about choosing a value for  $r_0$ . For example, selecting  $r_0$  to be large ( $> 4$ ), would lead to exploring low-probability regions. The dashed lines in the figure, correspond to  $a_i/\sqrt{\lambda_i} = 1, 2, 3$ .

Using the approximation in eq. (A12), we reduce the dimensionality of the system from  $\mathbb{R}^{2N_x N_z}$  to  $\mathbb{R}^{N_m}$  and approximate a complicated set ( $\mathcal{A} \setminus E(V_{\text{thresh}})$ ) by a hyperellipse which is a straightforward constraint for our optimization problem. With the mentioned approximation, and denoting  $w^* = \bar{w} + \sum_{i=1}^{N_m} a_i^* \phi_i$ , we write an optimization problem in the low dimensional  $\mathbb{R}^{N_m}$  space which is an equivalent approximate of eq. (9):

$$A^* = \{a^* \mid \sum_{i=1}^{N_m} \frac{a_i^{*2}}{\lambda_i} \leq r_0^2, w^* \text{ is a local maximizer of } F(10^{w^*}; \Delta t, T), F(10^{w^*}; \Delta t, T) > F_e^*\} \quad (\text{A14})$$

where  $a^* \in \mathbb{R}^{N_m}$  whose  $i$ th element is  $a_i^*$ . eq. (A14) ensures that the optimal solutions are not too far from the mean states ( $\bar{w}$ ).

To show the applicability of the POD model reduction outside the application of this paper, we also applied the method to a data set including all snapshots within the period of 200 years to 1200 years (without removing the event period). The result of this model reduction is available in Movie S2. This video shows that we can capture all phases of earthquake cycles using a few POD modes.

## APPENDIX B: OPTIMIZATION

Here we revisit optimal sampling in the framework of Bayesian optimization as discussed in (Brochu *et al.* 2010) and is improved in (Blanchard & Sapsis 2021) for finding the precursors of extreme events. The optimization algorithm works by exploring the input space ( $a = [a_1, \dots, a_{N_m}] \in \mathbb{R}^{N_m}$ ) using a Gaussian surrogate model. Suppose that we want to solve the constrained optimization problem of eq. (9) with the approximation in eq. (10). Without loss of generality, we study the minimization of the minus sign of the cost function ( $G = -F$ ) instead of maximizing it. The cost function can be evaluated using a forward simulation of a given initial condition. Here we assume that the observation is contaminated by a small Gaussian noise with variance  $\sigma_e^2 = 10^{-4}$ .

$$z = G(a; T, \Delta t) + \epsilon \quad \epsilon \sim \mathcal{N}(0, \sigma_e^2) \quad (\text{B1})$$

where  $\epsilon$  is the observational noise, and  $T$  and  $\Delta t$  are hyperparameters of the cost function  $G$  that are determined before the optimization process. The iterative approach starts from some randomly sampled  $N_{\text{init}}$  points  $\{a_k \in \mathbb{R}^{N_m}\}_{k=1}^{N_{\text{init}}}$  that each of them corresponds to a point in the set defined in (10). Using the forward model of eq. (B1) we find the input-output pair  $\mathcal{D}_0 = \{a_k, z_k\}_{k=1}^{N_{\text{init}}}$ .  $a_k \in \mathbb{R}^{N_m}$  is the vector of POD coefficients with  $N_m$  as the number of POD modes we have decided to consider, and  $z_k$  comes from eq. (B1). Using a Gaussian surrogate model, the expected value and variance of the process, condition on the input/output at each step  $i$  ( $\mathcal{D}_i$ ) is given by the following equation:

$$\begin{aligned} \mu(a) &= m_0 + k(a, A_i) K_i^{-1} (z_i - m_0) \\ \sigma^2(a) &= k(a, a) - k(a, A_i) K_i^{-1} k(A_i, a) \end{aligned} \quad (\text{B2})$$

where  $K_i = k(A_i, A_i) + \sigma_e^2 I$ ,  $A_i = \{a_k\}_{k=1}^{N_{\text{init}}+i}$ , and  $z_i = \{z_k\}_{k=1}^{N_{\text{init}}+i}$ . We consider the Radial Basis Function (RBF) with Automatic Relevance Determination (ARD):

$$k(a, a') = \sigma_f^2 \exp(-(a - a')^T \Theta^{-1} (a - a')/2) \quad (\text{B3})$$

where  $\Theta$  is a diagonal matrix containing the length scale for each dimension. At each iteration, we construct a surrogate model (eq. B2). Then, the next point in the input space is found by minimizing an acquisition function ( $g: \mathbb{R}^{N_m} \rightarrow \mathbb{R}$ ). We use the **Lower Confidence Bound** (LCB) acquisition function which is defined as the following:

$$g_{LCB}(a) = \mu(a) - \kappa \sigma(a) \quad (\text{B4})$$

where  $\kappa$  is a positive number that balances exploration and exploitation. For small  $\kappa$ , we do not consider uncertainties of the surrogate model and trust the mean of the conditional Gaussian process. For large  $\kappa$ , minimizing eq. (B4) is equivalent to finding a point that has the largest uncertainty. We use  $\kappa = 1$  in this study. The algorithm is extracted from Ref (Blanchard & Sapsis 2021) and is summarized in Algorithm 1. We start the algorithm by randomly sampling 10 initial points inside the hyper-ellipse (eq. 10) and then augmenting the input-output pairs by minimizing the acquisition function until the size of the input-output points reaches 200. To show the effectiveness of the algorithm in finding optimal solutions, we define the function  $c$  as the following:

$$c(i) = - \min_{1 \leq j \leq i} \min_a \mu(a \mid \mathcal{D}_j) \quad (\text{B5})$$

To find  $c(i)$ , we need to find the minimum of the Gaussian process in each iteration  $i$  and report the minimum over all  $1 \leq j \leq i$ . The algorithm does not guarantee finding all of the local maxima. As a result, the algorithm is repeated for 30 trials with different randomly chosen initial points. The behaviour of  $c(i)$  for different values of  $r_0$  is plotted in Fig. B1 (a). The solid line is the median of  $c(i)$  for different trials as a function of iteration and the shaded band shows half of the median absolute deviation. One of the optimal solutions is plotted in Fig. B1 (b,c). During the optimization process, we augment the set  $W^*$  if the condition in eq. (11) is satisfied.

---

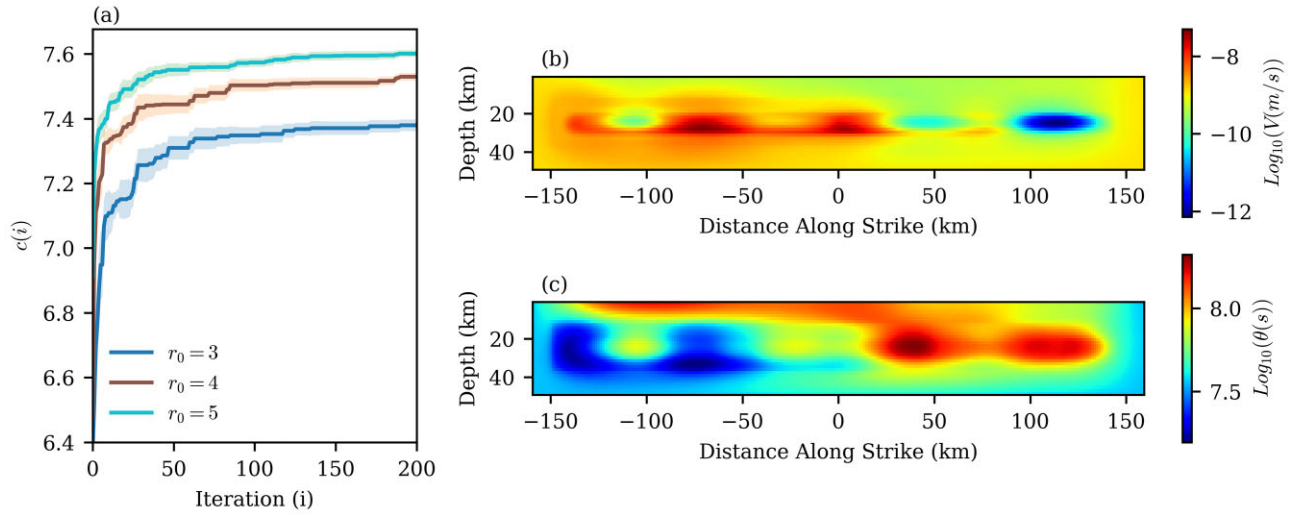
### Algorithm 1 Bayesian Optimization

---

- 1: **Input:** Number of initial points  $n_{\text{init}}$  and number of iterations  $n_{\text{iter}}$
- 2: **Initialize:** Surrogate model on initial dataset  $\mathcal{D}_0 = \{a^{(k)}, z^{(k)}\}_{k=1}^{n_{\text{init}}}$
- 3: **for**  $n=0$  to  $n_{\text{iter}}$  **do**
- 4:   Select best next point  $a_{n+1}$  by minimizing acquisition function constrained inside the hyperellipse (eq. (10)):

$$a^{(n+1)} = \arg \min_{\sum_{i=1}^{N_m} \frac{a_i^2}{\lambda_i} \leq r_0^2} g_{LCB}(a; \bar{G}, \mathcal{D}_n)$$

- 5:   Evaluate objective function  $G$  at  $a^{(n+1)}$  and record  $z^{(n+1)}$
  - 6:   If  $z^{(n+1)} < -F_e^*$  augment the set  $W^*$  (eq. (11))
  - 7:   Augment dataset  $\mathcal{D}_{n+1} = \mathcal{D}_n \cup \{a_{n+1}, z_{n+1}\}$
  - 8:   Update surrogate model
  - 9: **end for**
-



**Figure B1.** Convergence of the optimization for different values of  $r_0$  (a).  $\log_{10}(V)$  and  $\log_{10}\theta$  of one of the optimal solutions with  $r_0 = 3$  which leads to a magnitude 7.5. The optimal solution is highly heterogeneous and shows the effect of favorable stress heterogeneity in generating big events (b,c). The stress calculated from this optimal solution is plotted in Fig. 4.