Estimating the occurrence of slow slip events and earthquakes with an ensemble Kalman filter

Diab-Montero, Hamed Ali; Li, Meng; van Dinther, Ylona; Vossepoel, Femke C.

DOI
10.1093/gji/ggad154

Publication date
2023

Document Version
Final published version

Published in
Geophysical Journal International

Geophysical Journal International

Geophys. J. Int. (2023) **234,** 1701–1721
Advance Access publication 2023 April 14
GJI Seismology

https://doi.org/10.1093/gji/ggad154

Estimating the occurrence of slow slip events and earthquakes with an ensemble Kalman filter

Hamed Ali Diab-Montero [1], Meng Li,[2] Ylona van Dinther [2] and Femke C. Vossepoel[1]

[1] *Department of Geoscience and Engineering, Delft University of Technology, Mekelweg* 5, 2628 *CD Delft, The Netherlands.*
E-mail: h.a.diabmontero@tudelft.nl , ha.diabmontero@gmail.com
[2] *Department of Earth Sciences, Utrecht University, Princentonlaan* 8*a,* 3584 *CB Utrecht, The Netherlands .*

Accepted 2023 April 13. Received 2022 August 6; in original form 2023 April 4

SUMMARY

Our ability to forecast earthquakes and slow slip events is hampered by limited information on the current state of stress on faults. Ensemble data assimilation methods permit estimating the state by combining physics-based models and observations, while considering their uncertainties. We use an ensemble Kalman filter (EnKF) to estimate shear stresses, slip rates and the state θ acting on a fault point governed by rate-and-state friction embedded in a 1-D elastic medium. We test the effectiveness of data assimilation by conducting perfect model experiments. We assimilate noised shear-stress and velocity synthetic values acquired at a small distance to the fault. The assimilation of uncertain shear stress observations improves in particular the estimates of shear stress on fault segments hosting slow slip events, while assimilating observations of velocity improves their slip-rate estimation. Both types of observations help equally well to better estimate the state θ. For earthquakes, the shear stress observations improve the estimation of shear stress, slip rates and the state θ, whereas the velocity observations improve in particular the slip-rate estimation. Data assimilation significantly improves the estimates of the temporal occurrence of slow slip events and to a large extent also of earthquakes. Rapid and abrupt changes in velocity and shear stress during earthquakes lead to non-Gaussian priors for subsequent assimilation steps, which breaks the assumption of Gaussian priors of the EnKF. In spite of this, the EnKF still provides estimates that are unexpectedly close to the true evolution. In fact, the forecastability for earthquakes for the same alarm duration is very similar to slow slip events, having a very low miss rate with an alarm duration of just 10 per cent of the recurrence interval of the events. These results confirm that data assimilation is a promising approach for the combination of uncertain physics and indirect, noisy observations for the forecasting of both slow slip events and earthquakes.

Key words: Seismic cycle; Inverse theory; Numerical modelling; Probabilistic forecasting; Earthquake interaction, forecasting and prediction; Earthquake dynamics; Data assimilation; Ensemble Kalman filter.

1 INTRODUCTION

Earthquakes are among the deadliest and most damaging natural disasters. They are particularly hazardous because they occur without warning. Seismologists have traditionally focused on two families of methods in an attempt to alert the population. The first family produces Probabilistic seismic hazard assessments (PSHAs) based on historical data of past earthquakes, knowledge about the geology of an area and seismic response models (e.g. Esteva 1967; Cornell 1968; Bommer & Abrahamson 2006; Ordaza & Arroyo 2016). These methods estimate earthquakes' return periods from tens up to tens of thousand of years, but accurately combining very limited, uncertain information is really challenging and often not as successful as needed (Geller 2011). The second group of methods consists of earthquake early warning (EEW) systems, whose success relies on the large improvements in the seismic sensors' sensitivity in the last decades. They detect the less damaging and faster compressional waves (*P* waves) to alert the population seconds before the more destructive shear waves (*S* waves) and surface waves arrive (Allen & Kanamori 2003; Allen & Melgar 2019). Unfortunately, no methods have been proven reliable for the short-term prediction of earthquakes (Holliday *et al.* 2005; Koronovsky *et al.* 2019). Moreover, in between these two families there is a gap in forecasting timescales for earthquakes that goes from minutes to decades. Seismologists look for narrowing this gap by better understanding earthquake sequences, especially earthquake nucleation and earthquake physics

processes. For example, precursors such as foreshocks and slow slip events (SSEs) are pursued due to their systematical temporal and spatial relationship to the subsequent main earthquake. These precursors are believed to trigger the next earthquake or to contribute to the seismogenesis process which makes their study very beneficial for a potential short-term earthquake forecast (Segall & Bradley 2012; Uchida *et al.* 2016; Socquet *et al.* 2017; Pritchard *et al.* 2020).

Significant advances have been made in recent years in the study of seismogenesis. Efforts have concentrated on understanding how fault states control the earthquake nucleation, propagation and arrest (e.g. Ellsworth & Beroza 1995; Rubin & Ampuero 2005; Wibberley *et al.* 2008; Faulkner *et al.* 2010; Galis *et al.* 2017). This led to more realistic, physics-based numerical models (Lapusta *et al.* 2019), which resolve dominant physical processes throughout different phases of the seismic cycle: interseismic, coseismic and post-seismic phases (Lapusta & Liu 2009; Herrendörfer *et al.* 2018; Barbot 2019). Due to the usually simplified modelling assumptions it remains highly challenging to apply such physics-based models to improve short- to medium-term seismic hazard assessment, there are nevertheless progressive efforts to do so (e.g. Barbot *et al.* 2012; Dal Zilio *et al.* 2019). Earthquakes simulators have been developed in an attempt to model earthquake sequences over complex regional fault networks for tens of thousands of years by assuming significant simplifications in solution procedures and physical processes. These models are able to (re-)produce statistical observations such as the Gutenberg–Richter law and the Omori's law. They also shed light on the feasibility of combining physical models with probabilistic seismic hazard assessments (Dieterich & Richards-Dinger 2010; Shaw *et al.* 2018). However, one of the largest difficulties that remains in modelling upcoming sequences of slip on a fault system is that its current conditions (i.e. state of stress and strength) are unknown and cannot be directly measured (Barbot *et al.* 2012; van Dinther *et al.* 2013, 2019).

Observations of past fault slip and *in situ* measurements can be used to constrain physical models. For example, laboratory experiments in controlled, *in situ* conditions allow to constrain a fault's strength, that is its frictional parameters. The improvement in geophysical, geodetic and geological observations and laboratory measurements has offered valuable, albeit noisy and indirect information that may be used to calibrate the stress and velocities in regional, numerical models. For instance, slow slip events have an aseismic energy release (up to months and years) that produces very little surface deformation. This is difficult to record with seismometers and observed only with geodetic measurement systems (Kanamori & Hauksson 1992; Kawasaki *et al.* 1995; Dragert *et al.* 2001; Ide *et al.* 2007; Schwartz & Rokosky 2007). In contrast, the efforts to improve observation of regular earthquakes, which generate seismic waves due to very high slip rates over a short duration, have concentrated on developing more sensitive seismic networks and more complete earthquake catalogues with lower magnitudes of completion. Nonetheless, the sparsity of measurements in space and time, their large uncertainties and the large distances between the observed areas and the modelled ones make the estimation of the current state of stress of faults highly challenging (van Dinther *et al.* 2019; Brodsky *et al.* 2020).

Data assimilation techniques are methods that combine prior knowledge from physics-based simulations with observations to estimate the probability of a state or parameter (Evensen *et al.* 2022; Evensen 2003; van Leeuwen 2010; Bannister 2017). One of the advantages of these techniques is that they help to provide good estimates of variables of interest when having very uncertain initial conditions or parameters. For example, data assimilation is widely used for forecasting the weather (e.g. Evensen 1994; Reichle 2008), ocean currents (e.g. Vossepoel & Behringer 2000; Weaver *et al.* 2003; van Leeuwen 2003), hydrologic processes (e.g. Liu *et al.* 2012), or oil and gas production (e.g. Aanonsen *et al.* 2009; Evensen & Eikrem 2018). Data assimilation can also be used for history-matching of seismic data (e.g. Emerick 2018) or even to make predictions about the SARS-CoV-2 outbreak (e.g. Evensen *et al.* 2021).

In the last decade, different data assimilation approaches that consider observations of different parts of the earthquake process have been developed. These approaches can be classified into three groups: (i) estimation of the seismic wavefield during the coseismic phase of large earthquakes (Maeda *et al.* 2015; Oba *et al.* 2020), (ii) estimation of the slip, slip rates and frictional parameters during the post-seismic phase of an earthquake and during slow slip events (Kano *et al.* 2013; Hori *et al.* 2014; Kano *et al.* 2020) and (iii) the estimation of sequences of fault slip events (van Dinther *et al.* 2019; Hirahara & Nishikiori 2019; Banerjee *et al.* 2022). Most of these approaches have been assessed using perfect model experiments.

In this study, we build on the work of the third group focusing on the use of ensemble-based data assimilation methods to estimate multiple earthquake cycles (van Dinther *et al.* 2019; Hirahara & Nishikiori 2019). The work of van Dinther *et al.* (2019) demonstrates with perfect model experiments that even a single point observation of the shear stress and velocity significantly improves the estimates of fault stress and slip. However, van Dinther *et al.* (2019) mimics a laboratory setup of a slowly accelerating medium, whose fault is described by a simplified friction formulation. Additionally, Hirahara & Nishikiori (2019) focuses on simulating slow slip events instead of earthquakes. Until now the most effective implementation of data assimilation for fast earthquake sequences is yet to be developed, and is unclear whether it is possible to assimilate real measurements to effectively forecast the nucleation and occurrence of future earthquakes.

In this paper, we propose a data-assimilation approach to estimate the occurrence of both slow slip events and earthquake sequences using a numerical model of earthquake sequences, which is based on a rate-and-state friction formulation and adaptive time stepping. We focus here on using data assimilation methods as a means to shorten the time scale of fault slip forecasts. We look for improving our way of forecasting earthquakes by making the most of both physics-based models and observations, that is by leveraging recent progress in earthquake simulations, and benefiting from the abundant data collected from past seismic events.

The outline of the paper is as follows. In Section 2, we summarize the workings of the ensemble-based data assimilation method and we introduce the type of perfect model experiments we perform over slow slip events and earthquake models. In Section 3, we evaluate the estimates of the ensemble Kalman filter (EnKF) for both types of events in two key locations: at the observation location where a single observation of shear stress and velocity has been taken, and at the fault location where the shear stress, the slip rate and the state θ are estimated. We further analyse the evolution of the ensemble distributions during different phases of the seismic cycle, evaluate the forecastability potential from the EnKF and assess the influence of the observations in the estimations of the variables. In Section 4, we discuss the implications of the non-linearity behaviour of the rate-and-state friction law in the assimilation process. Finally, in Section 5 we present our conclusions about the performance of the filter for the estimation of slow slip and earthquake occurrences.

2 METHODOLOGY

2.1 Data assimilation

Data assimilation is an approach that helps to better estimate the evolution of a system by combining knowledge of a dynamic system with observations thereof. We represent variables (or parameters) that we want to estimate as a vector that we call the 'state vector',

$$\mathbf{z}^T = \left(\mathbf{z}_\psi{}^T, \mathbf{z}_\alpha{}^T\right), \tag{1}$$

where $\mathbf{z}$ is the state vector, $\mathbf{z}_\psi$ represent the 'states' of the system and $\mathbf{z}_\alpha$ represent the 'parameters' or physical properties of the system. When applied sequentially, the optimal estimation of the variables is described as a two-step process. The first step is called the 'forecast step', in which we use our knowledge of the dynamics of the system to move it forward from a previous time t_{c-1} to a future time t_c using the following formulation:

$$\mathbf{z}_c = \mathcal{M}_{c:c-1}\left(\mathbf{z}_{c-1}\right) + \boldsymbol{\eta}_c, \tag{2}$$

where $\mathcal{M}_{c:c-1}$ represents the forward model operator from time t_{c-1} to t_c, $\mathbf{z}_{c-1}$ represents the state vector at time t_{c-1} and $\boldsymbol{\eta}_c$ represents the model error which includes missing physics or uncertainty involved in the forward modelling (parameter uncertainties, grid resolution, uncertain initial conditions, etc.).

The second step is the 'analysis step' which is formulated based on Bayes' theorem when data or observations of the system have been made:

$$p(\mathbf{z}|\mathbf{d}) = \frac{p(\mathbf{d}|\mathbf{z})p(\mathbf{z})}{p(\mathbf{d})}, \tag{3}$$

where $p(\mathbf{z})$ represents our prior scientific knowledge of the state vector and its uncertainties, $p(\mathbf{d}|\mathbf{z})$ represents the likelihood of the observations $\mathbf{d}$, given the prior state vector $\mathbf{z}$, $p(\mathbf{d})$ is referred to as the evidence and acts as a normalization factor and $p(\mathbf{z}|\mathbf{d})$ is the posterior estimate of the state of the system given the observed data. We update in this step our knowledge of the system using those observations.

We represent the observation vector $\mathbf{d}$ with:

$$\mathbf{d}_c = \mathcal{H}_c(\mathbf{z}_c) + \boldsymbol{\epsilon}_c, \tag{4}$$

where $\mathbf{d}_c$ is the vector containing the observations taken at time t_c, $\mathcal{H}_c$ is the non-linear observation operator and $\boldsymbol{\epsilon}_c$ are the measurement errors.

One of the benefits of data assimilation is that it helps to estimate states at a location of interest that is not easily accessible to be measured directly. These states receive the name of 'hidden states'. They are estimated using our physical knowledge of the interaction and relationship between the non-observable and the observable locations in the model to transfer the corresponding correction once we update our knowledge of the observable locations. In our case we consider the fault generating the earthquakes as a location of interest whose state variables are hidden states.

2.1.1 The EnKF

For the state estimation in this study, we use a stochastic variant of the EnKF (e.g. Evensen 2003), an ensemble-based data assimilation and a Monte Carlo implementation of the least-squares solution of the Bayesian update problem presented in eq. (3). The EnKF optimally combines the information from the forward numerical model (prior) and its deviation to the observations (likelihood) to produce a posterior estimation of the state vector. The prior, likelihood and posterior probability density functions (pdfs) are approximated by an ensemble of different initial conditions, states or parameters of our model. These distributions are assumed to be Gaussian. The ensemble representation of our state vector, which can contain the state and the parameters, is as follows:

$$\mathbf{z}_n^T = \left(\mathbf{z}_\psi^T, \mathbf{z}_\alpha^T\right)_n, \qquad 1 \le n \le N, \tag{5}$$

where the subscript n refers to a single realization in an ensemble containing N realizations (with $n = 1,..., N$). The prior, is defined as $\mathbf{z}_n^f \sim \mathcal{N}(\overline{\mathbf{z}_n^f}, C_{zz}^f)$. The superscript f stands for forecast and indicates the prior information coming from the forward numerical model which represents our knowledge about the physics of the problem. The covariance C_{zz}^f quantifies the relationship between variables and the uncertainties of the given states. The covariance is defined as follows:

$$\mathbf{C}_{zz}^f = \frac{1}{N-1}\left(\mathbf{z}_n^f - \overline{\mathbf{z}_n^f}\right)\left(\mathbf{z}_n^f - \overline{\mathbf{z}_n^f}\right)^T. \tag{6}$$

We utilize the perturbed-observations scheme, which involves updating each ensemble member with a perturbed observation. It is assumed that the observational errors follow a Gaussian distribution represented by $\boldsymbol{\epsilon}_n \sim \mathcal{N}(0, C_{dd})$. Additionally, we assume we have uncorrelated errors in our states. The perturbed observation vector for each ensemble member is calculated as follows:

$$\mathbf{d}_n = \mathbf{d} + \boldsymbol{\epsilon}_n, \qquad 1 \le n \le N, \tag{7}$$

$$\mathbf{C}_{dd} = \frac{1}{N-1}\sum_{n=1}^{N}\boldsymbol{\epsilon}_n\boldsymbol{\epsilon}_n^T. \tag{8}$$

The EnKF uses the prior distribution $\mathbf{z}_n^f$, the observation vector $\mathbf{d}_n$ and their covariance matrices for estimating the posterior distribution $\mathbf{z}_n^a$ calculated in the analysis step using the following expression:

$$\mathbf{z}_n^a = \mathbf{z}_n^f + \mathbf{K}\left[\mathbf{d}_n - \mathbf{H}\mathbf{z}_n^f\right], \qquad 1 \le n \le N, \tag{9}$$

where the superscript a stands for analysis, $\mathbf{K}$ is the Kalman gain matrix and $\mathbf{H}$ is the linear observation operator matrix. The Kalman gain is interpreted as the relative weight given to the observations information and current state estimate and it is given by:

$$\mathbf{K} = \mathbf{C}_{zz}^f\mathbf{H}^T\left(\mathbf{H}\mathbf{C}_{zz}^f\mathbf{H}^T + \mathbf{C}_{dd}^f\right)^{-1}. \tag{10}$$

A high Kalman gain places a higher weight on the observations and makes the analysis follow them more closely. A low Kalman gain imposes a higher weight on the prediction and an analysis that follows the prediction more closely. Further details can be found in Evensen (2003) and Evensen *et al.* (2022).

2.2 Forward modelling

We use a numerical model setup inspired by a large-scale biaxial friction apparatus consisting of two sandstone block specimens in a direct-shear configuration (Fig. 1, Fukuyama *et al.* 2014; Spiers *et al.* 2017). We assume a symmetric setup and model only the lower half-space. To limit computational resources we discretize a 1-D line across the block boarded by a single, 0-D fault point, which provides a reasonable approximation when evaluating temporal estimates (Li *et al.* 2022). Discretization on a fully staggered grid is solved using finite differences by adopting the C++ library for non-linear coupled problems 'Garnet' (Pranger 2020). The numerical method and conservation and constitutive equations for this 1-D model are

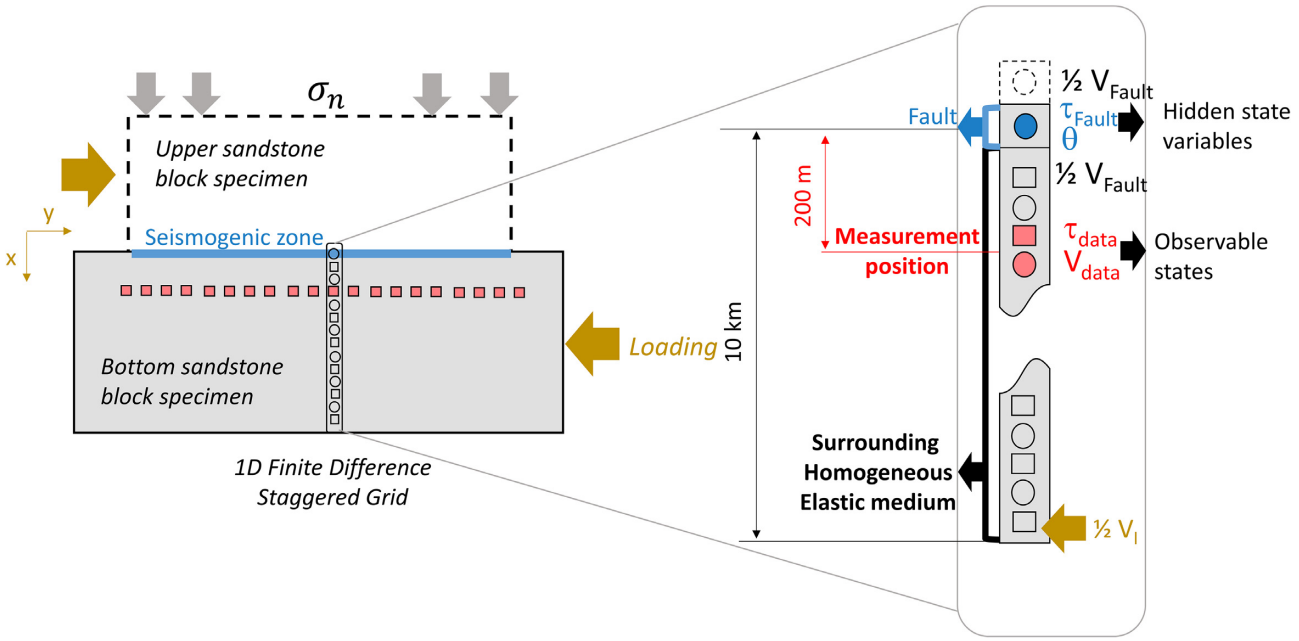

Figure 1. Schematic representation of the 1-D model used to represent a horizontal straight fault. The shear stress (τ) is estimated in the nodes represented by circles and the velocity (v_y) in the ones represented by squares The fault (blue) corresponds to the uppermost node of the grid and the location where shear stress, slip rate and state θ are estimated by the EnKF. Shear-stress and velocity observations (red) at 200 m away from the fault in the surrounding homogeneous elastic medium are assimilated into a 10-km model. The spacing between nodes is 20 m.

adopted from and described in detail in Li *et al.* (2022). Below we summarize the equations relevant for our forward model setup.

2.2.1 Fault model

The fault's temporal evolution is modelled as an initial value problem. The slip along the fault is assumed to be governed by a rate-and-state friction formulation, which was proposed based on laboratory friction experiments by Dieterich (1978, 1979) and Ruina (1983) (eq. 11). We use a regularization near zero slip rate according to Rice (1993) and Ben-Zion & Rice (1997), such that the friction formulation that defines the relation between shear stress τ_{fault} and normal stress σ_n on the fault is given by

$$\tau_{\text{fault}} = a\sigma_n \text{arcsinh}\left\{\frac{V}{2V_0} exp\left[\frac{\mu_0}{a} + \frac{b}{a}\left(\ln\frac{\theta V_0}{L}\right)\right]\right\} + \eta V, \quad (11)$$

where τ_{fault} is the shear stress at the fault, μ_0 is the reference friction coefficient at slip rate V_0, V is the slip rate, L is the characteristic slip distance, a is the empirical parameter representing the direct effect and b is the parameter representing the evolution effect. The 'state' θ has a unit of time and is a scalar that increases during the interseismic phase (stick), when the asperities are in contact and locked in the fault, and decreases during the coseismic phase (slip). The state θ is governed by the evolution equation (Ruina 1983) given by

$$\dot{\theta} = 1 - \frac{V\theta}{L}. \quad (12)$$

The fault is 'velocity-weakening' and potentially frictionally unstable when $a - b < 0$, and 'velocity-strengthening' and generally frictionally stable when $a - b > 0$. Finally, the parameter η used in eq. (11) refers to the 'radiation damping term' used in the quasi-dynamic approximation of inertia (Rice 1993), which is used in earthquake cycle simulations to reduce the computational costs (Cochard & Madariaga 1994; Ben-Zion & Rice 1995; Liu & Rice 2007; Crupi & Bizzarri 2013). However, this is known to introduce qualitative and quantitative differences compared to fully dynamic modelling results (Thomas *et al.* 2014). The damping viscosity $\eta = G/(2c_s)$ is equal to half the shear impedance of the elastic material surrounding the fault.

2.2.2 Medium model

Following the simplification made in Li *et al.* (2022), we directly write out the physical equations in 1-D. In this scenario, all variables are invariant along dip and strike thus only the shear stress component τ_{xy} and the velocity component v_y need to be solved. The fault reduces to a 0-D point at $x = 0$ in a computational domain $\mathbf{\Omega}(x) = [0, H]$, where H is the distance from the fault interface until the bottom of the model. We choose the fault point to be velocity-weakening to be able to nucleate sequences of earthquakes. The model needs to satisfy the momentum balance equation and Hooke's elasticity:

$$\begin{aligned} \dot{\tau}_{xy} &= G\frac{\partial v_y}{\partial x}, \\ \frac{\partial \tau_{xy}}{\partial x} &= 0. \end{aligned} \quad (13)$$

where G is the shear modulus.

We assume boundary conditions corresponding to the values of the velocity in both extremes of the model

$$\begin{aligned} v_y(x = 0) &= \frac{1}{2}V, \\ v_y(x = H) &= \frac{1}{2}V_l. \end{aligned} \quad (14)$$

Our assumption about the symmetry of the setup allows us to consider the velocity in the fault interface as half of the total slip-rate V, and the one at the bottom of the model as half the loading rate V_l.

The initial conditions are chosen to allow the fault to creep at the imposed slip velocity V_l in a steady state at $t = 0$, namely

$$V(t=0) = V_l\,,$$
$$\theta(t=0) = \frac{L}{V_l}\,, \tag{15}$$
$$\tau_{\text{fault}}(t=0) = a\sigma_n \text{arcsinh}\left\{\frac{V_l}{2V_0}\exp\left[\frac{\mu_0}{a} + \frac{b}{a}\ln\left(\frac{V_0}{V_l}\right)\right]\right\} + \eta V_l\,. \tag{16}$$

2.3 Perfect model experiments

In the data-assimilation experiments, we use an ensemble of 50 members. Each member is initialized with a different value of initial shear stress at the fault node following a Gaussian distribution with a standard deviation of 2.5 MPa. The synthetic true value of initial shear stress is set in 20 MPa and the mean of the ensemble has a bias of 3 MPa with respect to that true value (Fig. 2). The total simulation time is 1500 yr from which we sample observations at 2.5-yr intervals for slow slip events and at 5-yr intervals for earthquakes. The workflow is illustrated in Fig. 3. To compare the results with and without data assimilation, we run the 50 members up to 200 yr where the first observations are available. Data are assimilated then, and subsequently whenever sampled synthetic observations are available.

In an operational data assimilation application, the observations that are assimilated are real observations. In our case, we assimilate synthetic observations, such that we can evaluate how our fault estimates compare to the known truth. We also assume our model is perfect, that is that our parameters shown in Table 1 are correct. The synthetic observations are samples of the simulated shear stress and the velocity as if they were measured at a single location at a short distance away from the fault, with noise added to represent measurement error. The selection of observation error amplitudes was based on widely accepted instrument sensitivity values (van Dinther *et al.* 2019) and the maximum observed variations in stress and velocity during the coseismic phase. For slow slip events, observation error values were set at 0.25 MPa for shear stress measurements and 0.25 for the logarithm of velocity observations. For earthquakes, the observation errors were assigned as 0.75 MPa for shear stress measurements, and 0.75 for the logarithm of velocity observations.

We follow the workflow presented in Fig. 3 for the perfect model experiments. In our case, the state vector $\mathbf{z_n}$ includes the shear stress of the fault, the shear stress in the medium, the slip rate, the velocity in the medium and the state θ:

$$\mathbf{z}_n^T = \left(\tau_{\text{Fault}}, \boldsymbol{\tau}_{xy}^T, \ln(\text{V}), \ln(\mathbf{v}_\text{y})^T, \ln(\theta)\right)_n\,, \qquad 1 \le n \le N. \tag{17}$$

After the analysis step the updated shear stress and state θ are used to calculate the slip rate V. This is done by using an implicit solver that finds the corresponding slip rate that satisfies eq. (11). The velocity values in the medium v_y are calculated for the next step solving the system of eqs (12)–(14), where the estimated posterior values become part of the velocity history of the model.

3 RESULTS AND ANALYSIS

The results of the EnKF estimation are analysed at the observation location at 200 m from the fault to analyse how observations affect the assimilation and at the unknown target location, the fault.

3.1 State estimation in the homogeneous elastic medium

The time evolution of the shear stress with respect to the grid τ_{xy} at the observation location shows that the estimates of the EnKF (represented in green) for both the slow slip events and the earthquakes are in sync with the truth (Figs 4a and b, respectively). The EnKF analysis resolves well both the interseismic and coseismic phases, despite the large errors of the observations. For example, we observe large errors around time 834 yr for the slow slip events and around time 980 yr for the earthquakes. The EnKF is especially effective for the slow slip events (Fig. 4a), where the ensemble mean captures the coseismic phase reasonably well. The shear-stress estimates for the earthquakes during the coseismic phase are comparatively less precise (Fig. 4b). The very large and fast stress drop during earthquakes is inherently difficult for any probabilistic or averaging method to precisely reconstruct. Nonetheless, the timing of earthquakes is generally anticipated by a drop in the mean shear stress about 2–3 yr prior to the earthquake occurrence and an increase in spread of the ensemble with an approximate standard deviation of 2 MPa.

As depicted in Figs 4(c) and (d), the uncertainty of the time evolution of the velocity at the observation location differs considerably between slow slip events and earthquakes. For both the slow slip and the earthquake events, the analysis of the EnKF captures the true velocity in the interseismic phase very well. In the coseismic phase it is more difficult to provide an accurate estimate due to fast changes in velocity. For the slow slip events (Fig. 4c) the EnKF captures the evolution of the velocity very well and, remarkably, it even accurately captures the peak velocities in terms of magnitude and timing. However, for the earthquakes the magnitude of the estimated peak velocity still has a large uncertainty (Fig. 4d). This is a result of averaging over 50 ensemble members, which each only show 7–9 orders of magnitude higher velocities for seconds. However, the ensemble mean correctly traces the increases in velocity up to the loading rate. This is important to realize for estimating the timing of nucleation, since after the increase in slip rates is identified, its maximum is capped by the activation of inertia and thus already known to be in the order of meters per second.

3.2 State estimation at the fault

The EnKF estimates the shear stress and slip velocities relatively accurately at the locations where we have observations (Fig. 4). Nevertheless, good estimates are not guaranteed for locations where measuring devices do not have easy access, such as the fault (Fig. 5). This is especially difficult if the state variables experience rapid changes. Therefore, we evaluate the performance of the EnKF in the same experiments from Fig. 4 but at the location where the estimation is most challenging, that is at the fault. At the same time, this is the location where correct estimates are critical and most needed, as movements in the medium respond to displacements at the fault and the state variables are unobserved there.

Prior to data assimilation ensemble members are not aligned, such that their mean can not separate between interseismic and coseismic phases and almost becomes a constant value (Fig. 5). After 200 yr there is a significant decrease in the spread of the ensemble for all hidden states as a result of the assimilation of the first observations. As shear stress is increased in many ensemble members, the occurrence of the first slow slip event is synchronized even after assimilation of a single observation (Fig. 5a). The first earthquake is preceded by assimilation of a second observation, which introduces uncertainties in the shear stress estimates, but does

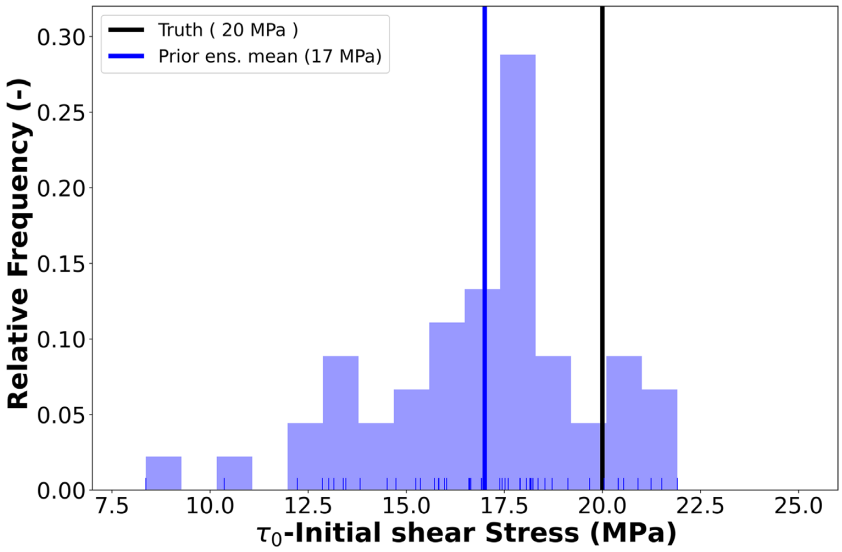

Figure 2. Histogram illustrating the initial distribution of shear stress at the fault used to initialize the ensemble. We define a synthetic truth with an initial shear stress at the fault of 20 MPa. The vertical blue line is the mean of the Gaussian distribution, the black vertical line the truth, the marks in the horizontal axis represent the 50 different initial stress values used for the ensemble.

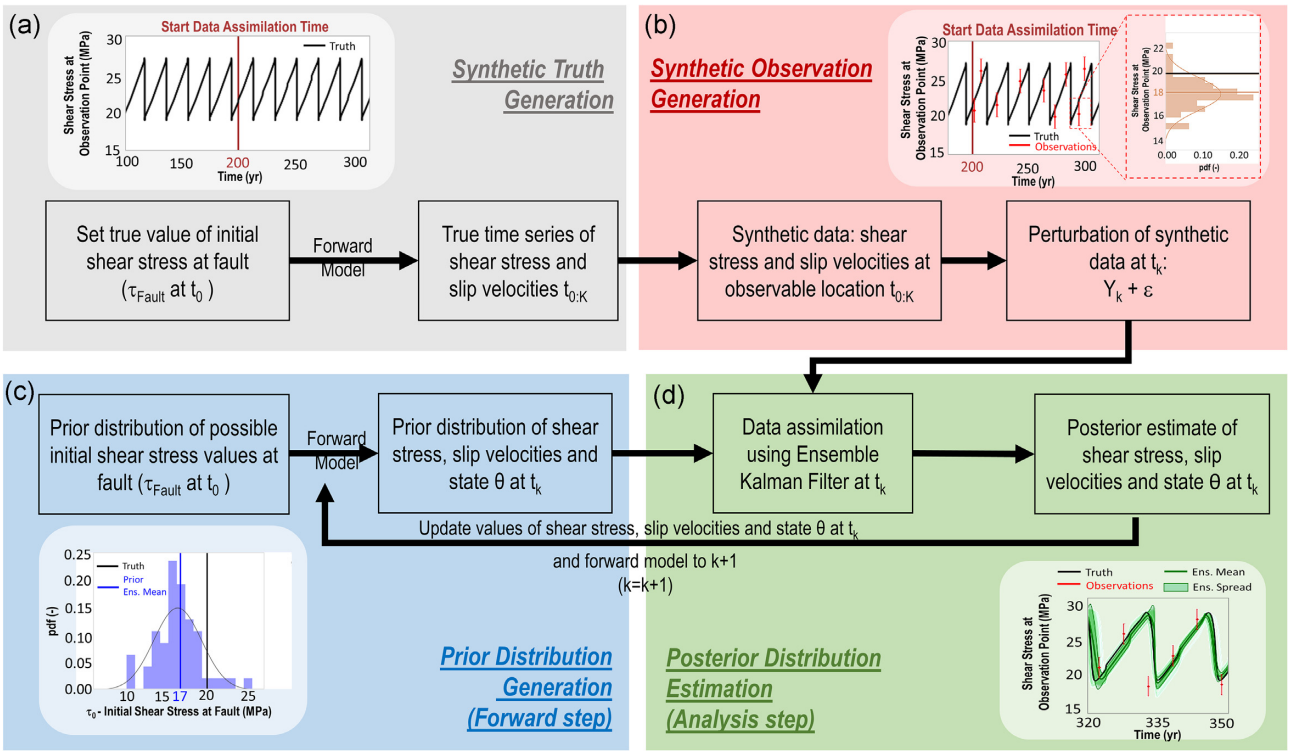

Figure 3. Schematic representation of a perfect model experiment and sequential data assimilation. The flowchart describes four steps of the synthetic experiment. (a) First, in grey the steps for generating the synthetic truth to be estimated. (b) Second, in red the creation of the synthetic observations to be assimilated. (c) Third, in blue the creation of the ensemble of realisations that form the prior distribution. (d) Fourth, in green the data assimilation step where a data-assimilation method (in this case the EnKF) is used to estimate a posterior distribution of the state using the synthetic observations and the prior ensemble distribution. The last two steps are done sequentially until the last time step K where observations are available. In our case, we assimilate shear stress and slip velocity observations measured in the medium.

allow the ensemble to accurately predict the timing of the stress drop (Fig. 5b). Assimilation of a rapid slip rate is also apparent from the slip rate jump at $t = 210$ yr (Fig. 5d), which is immediately corrected by the forward model using shear stress to estimate the next slip rate. Assimilation during the next slip sequences shows that the estimates of the EnKF are very well synchronized with the truth in our case of state estimation with known and constant parameters. Generally, for a large part of the interseismic period, slip rates and states from all ensemble members show that a slip event is not to be expected (Figs 5c–f). Interestingly, for some sequences all three

Table 1. Material and rate-and-state friction parameters for the 1-D model setup for both slow slip events and earthquakes.

Parameter	Symbol	Slow slip events	Earthquake events
Shear modulus	G	32 GPa	32 GPa
Density	ρ	2670 kg m^{-3}	2670 kg m^{-3}
Initial mean stress	σ_n	40 MPa	40 MPa
Static friction coefficient	μ_0	0.6	0.6
Reference slip rate	V_0	1 μm s^{-1}	1 μm s^{-1}
Characteristic slip distance	L	0.24 m	0.18 m
RSF direct effect	a	0.0060	0.0060
RSF evolution effect	b	0.0158	0.0160
Depth model	H	10 km	10 km
Loading slip rate	V_l	10 nm s^{-1}	10 nm s^{-1}
Grid spacing	Δx	20 m	20 m

hidden states are tracked with extreme accuracy, for example around the earthquake at about 230 yr (Figs 5b, d and f).

The peak shear stress have a larger spread in the posterior of the earthquake model when compared to the slow slip events when approaching the coseismic phase (Figs 5a and b). The ensemble spread around the coseismic phase of the earthquakes has a large spread around the peak slip rate compared to the slow slip events (Figs 5c and d). The ensemble mean does not reach the peak slip rates nor the peak shear stresses due to the spread of the ensemble in the interseismic and coseismic phases which is related to the short duration of the coseismic phase. However, individual ensemble members give a good indication of peak slip rates and shear stresses. The ensemble provides good state θ estimates, whose timing is synchronized with the truth after the data assimilation starts for both fault-slip events (Figs 5e and f). The ensemble prior PDF of the three hidden states show a change in spread around 2–3 yr before and after the coseismic phase. This distinct increase in the spread of the ensemble prior to the event may be an indication for an upcoming earthquake.

3.3 Non-Gaussianity

In the EnKF, it is assumed that the prior distributions of the state vector's components and observation errors are Gaussian distributed. Additionally, the EnKF produces a low-rank representation of the prior covariance matrices by using a finite ensemble of members, which makes the EnKF very sensitive to outliers. This is particularly true for cases where the ensemble size is small. We verify that the ensemble size used for our experiments is large enough by doing an analysis of the eigenvalues of the outer and inner product covariance matrices of our ensemble. Our analysis indicates that an ensemble of 30 members would already be enough to resolve the significant principal components of the covariance matrix, which means that our current ensemble size of 50 members is adequate. For a more extended explanation of this verification see Appendix A.

In this section, we evaluate how well preserved the assumption of Gaussian distributions is in the different phases of the cycle of slow slip events and earthquakes (Figs 6 and 7). For both types of events we provide the histograms of the ensembles during the interseismic phase (Figs 6b and 7b) and shortly before the start of the coseismic phase (Figs 6c and 7c). For the slow slip events the ensemble distributions are very close to a Gaussian. The preservation of the Gaussian assumption allows the EnKF to have estimations as shown in Fig. 6(a). The range of the differences between the ensemble mean and the truth corresponds to the expected range of uncertainty. The maximum errors are around 0.4–0.8 MPa, which are between 25 per cent of the average stress drop of the slow slip events. In terms of the uncertainties on the timing of slip occurrence, it is about 1 yr, which is 20 per cent of the recurrence interval of 5 yr of the slow slip events.

In contrast, for earthquake sequences, the Gaussian distribution in the prior is only well preserved during the interseismic phase (Figs 7a and b). This is because the members that experience the sudden large change of shear stress in the coseismic phase deviate strongly from the ensemble mean (Fig. 7c). Some of the ensemble members are finishing the interseismic phase of the previous cycle and entering the coseismic phase, while others are starting the next cycle. This results in a skewed distribution of the prior shear stress, which translates into a bimodal distribution for the prior (Fig. 7c). This effect is also seen in Fig. 7(d) in the evolution of the absolute errors and the standard deviation of the ensemble. Similar to the slow slip events the errors of the filter are in the same range of the values of the standard deviation of the ensembles and in the range of expected uncertainties. The maximum errors are around 50 per cent of the stress drop of 10 MPa. The errors in terms of the earthquake occurrence are about 5 yr which is 25 per cent of the recurrence interval of 20 yr of the earthquake sequences. This similarity in performance suggests that the non-Gaussianity of prior estimates may not significantly limit the effectiveness of the EnKF in our case of state estimation with known and constant parameters.

3.4 Sensitivity of analysis to observation type

We observe that the EnKF gives reasonable estimates of the shear stress, slip rate and state θ at the fault despite the rapid changes in the stresses and velocities in the coseismic phase. In this section, we assess how the estimation of these rapid changes is realized by the EnKF and what is the influence of the different types of observations on this. We do this by analysing the elements of the Kalman gain matrix at every analysis step of the data assimilation. As explained in Section 2.1, a high Kalman gain places a higher weight on the observations information when estimating the posterior distribution of the ensemble. Every element of the matrix maps the 'innovation', that is the difference at the observation location between the observed values and the estimated values from the forward model, to an update of a variable in the state vector. From this we can assess, for example, the relative influence of the shear-stress observations at the fault on the estimated slip rate. Fig. 8 illustrates the Kalman gain matrix elements from all analysis steps of the slow slip events and the earthquakes that relate the shear-stress observations and the velocity observations to the estimates of shear stress, slip rate and state θ at the fault. By comparing the time when the observations are assimilated relative to the slip occurrence, we evaluate how these Kalman gain elements change after every occurrence of a slow slip event or earthquake.

In both type of events, the Kalman-gain values that relate the shear-stress estimates to the shear-stress observations are higher than those that relate to the velocity observations, suggesting that the shear-stress estimates are more strongly influenced by the shear-stress observations than by the velocity observations. This is especially notable for the earthquakes, where the Kalman gain matrix elements that relate to the shear-stress observations are between four and eight times higher than the elements that relate to the velocity observations. In contrast, the Kalman gain matrix elements of the slip rate estimates that relate to velocity observations for the slow slip events are two to four times higher than the elements that relate

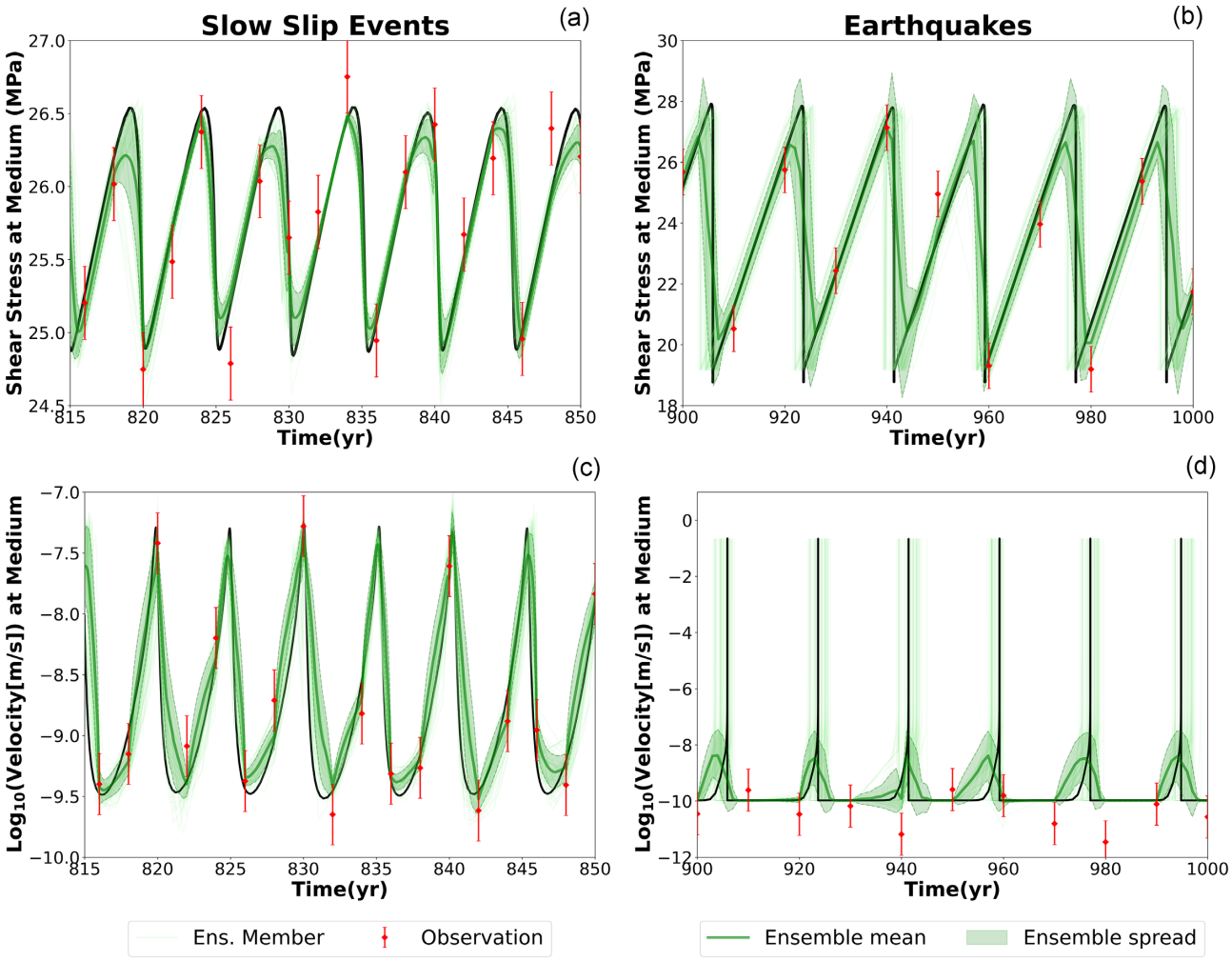

Figure 4. Estimated evolution of (a, b) shear stress and (c, d) slip rate at the observation location for (a, c) slow slip events and (b, d) earthquakes. The solid black line represents the true evolution, the red markers are the observations with error bars indicating the standard deviation of the observational error, the green solid line is the ensemble mean and the light green lines represent the ensemble members. The ensemble spread is shown as the light green hatched area, which is one standard deviation below and above the ensemble mean for a normal distribution and corresponds to the ensemble members between the percentiles P16 and P84 of the ensemble distribution.

to the shear-stress observations. This suggests that the slip-rate estimates are influenced the most by the velocity observations, and the shear-stress estimates are influenced the most by the shear-stress observations.

We observe a difference in the relative influence of the slip-rate observations on the slip-rate estimates between the slow slip events and the earthquakes. For slow slip events (Fig. 8c), the slip-rate estimates are mostly influenced by the velocity observations, but for earthquakes, both types of observations appear to have equal influence in the period before the slip occurrence. However, after the slip occurrence, the elements that relate the velocity observations to the slip-rate estimates are higher than the ones that relate to the shear-stress observations.

For the state θ both types of observations seem to influence equally the estimates for the slow slip events, while for earthquakes, the Kalman gain elements that relate to the shear-stress observations are two times higher than the elements that relate to the velocity observations.

We observe that all Kalman gain-matrix elements for the slow slip events tend to be higher after the slip occurrence. For the earthquakes the higher values occur shortly before and after the event. The Kalman gain is a function of the observational error and the prior covariance of the state vector as shown in eq. (10). In our perfect model experiments shown here, we assume that the uncertainties of the synthetic observations are constant in time. Therefore, the Kalman gain values, as indicated by the boxes in Fig. 8, become a function of the ensemble spread. Since the ensemble spread is high shortly before and after the earthquake occurrence the higher Kalman gain values will result effectively in a stronger influence of the observation information on the analysis during these periods. For the slow slip events a large ensemble spread occurs mainly after the coseismic phase because of the smoother transition of the phases. We observe higher Kalman gain values after the seismic event. Interestingly, the slip rate (blue line) changes more notably than the other variables before the earthquake event as illustrated by Figs 8(b) and (d). This suggests that slip rate is more sensitive to data assimilation before the earthquake event than other variables.

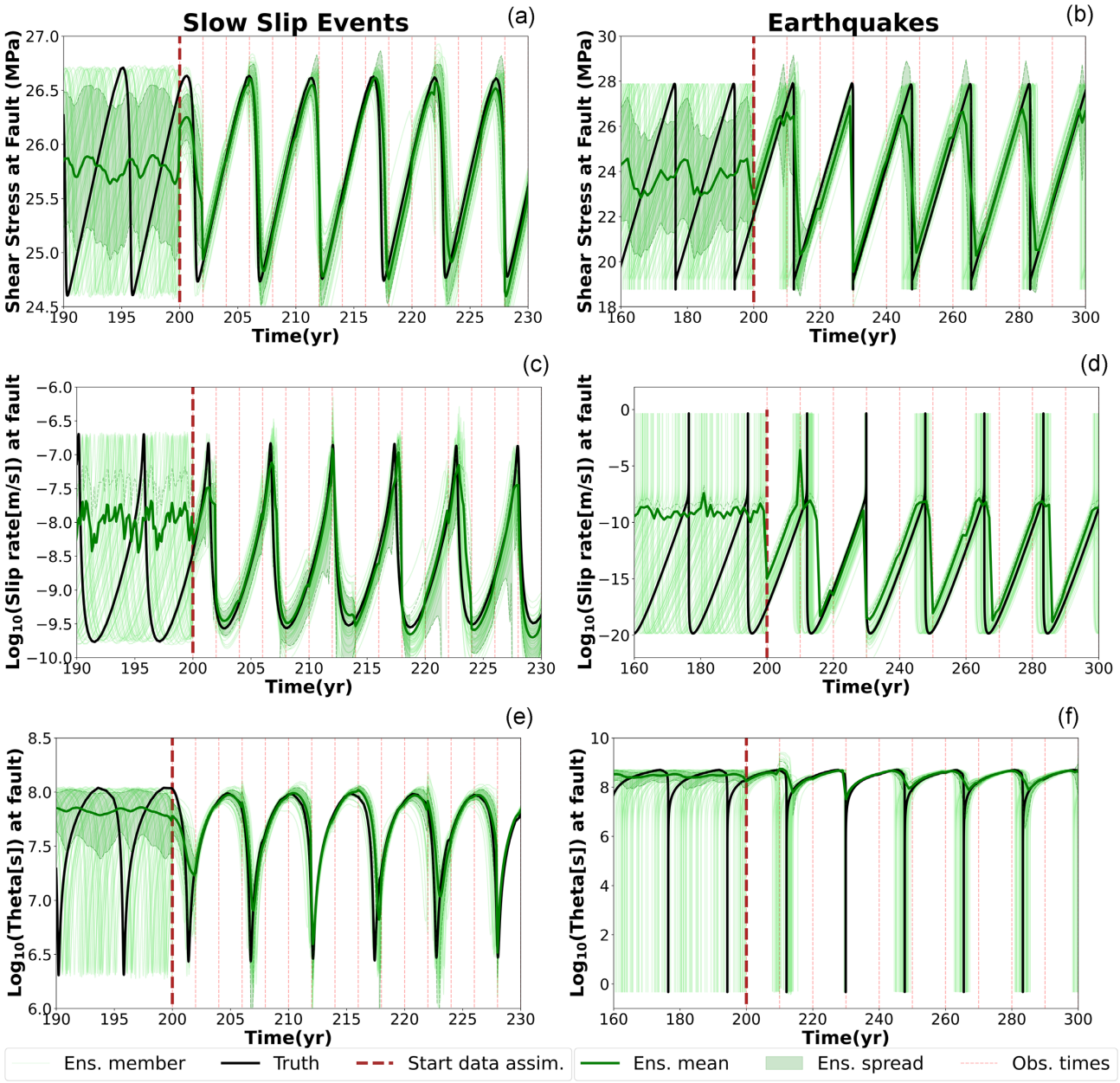

Figure 5. Estimated evolution of (a, b) shear stress, (c, d) slip rate and (e, f) state θ at the fault for (a, c, e) slow slip events and (b, d, f) earthquakes. The brown line is the start of the data assimilation. The black solid line is the true evolution of the variables. The red dashed lines indicate the observation and assimilation times. The colouring of the lines is as in Fig. 4.

3.5 Forecastability

We evaluate the forecastability of the EnKF for slow slip events and earthquakes by comparing time-series of estimated relative frequency of fault-slip occurrence to the synthetic true occurrences (Fig 9). Relative frequencies indicate the percentage of ensemble members that estimate an earthquake within five different rolling window lengths (1, 2, 3, 4 and 5 yr) and a lag of 1 yr. The lag in this scenario means that we shift our window's interval each time by 1 yr. We observe that the relative frequencies of occurrence for the rolling windows of 1-yr length are typically rather small, that is smaller than 0.2, although for both fault-slip types exceptions up to 0.5 exist. However, for longer windows from 2 up to 5 yr it is possible to reach relative frequencies from 0.75 up to 1.0. Interestingly, particularly for rolling windows of 5 yr, the peak of relative frequencies in most cases occurs at the same time as the true slip events, with maximum deviations of less than 1 yr. Moreover, the maximum relative frequencies are rather comparable for both fault-slip types, indicating that the forecastability of earthquakes is also comparable to those of slow slip events.

We also evaluate the forecastability of the data-assimilation framework by using a Molchan or error diagram to analyse what is known just prior to the earthquake, instead of forecasting a slip event once it has happened, that is hind-casting (Fig 10). The diagram is constructed considering 72 earthquakes and 259 slow slip events in a time span of 1300 yr. We assume that an alarm is ringed once a percentage of the ensemble (10, 20 and 30 per cent) has reached its peak shear stress. We calculate the period between the moment that the alarm rings and the actual slip occurrence and divide this time

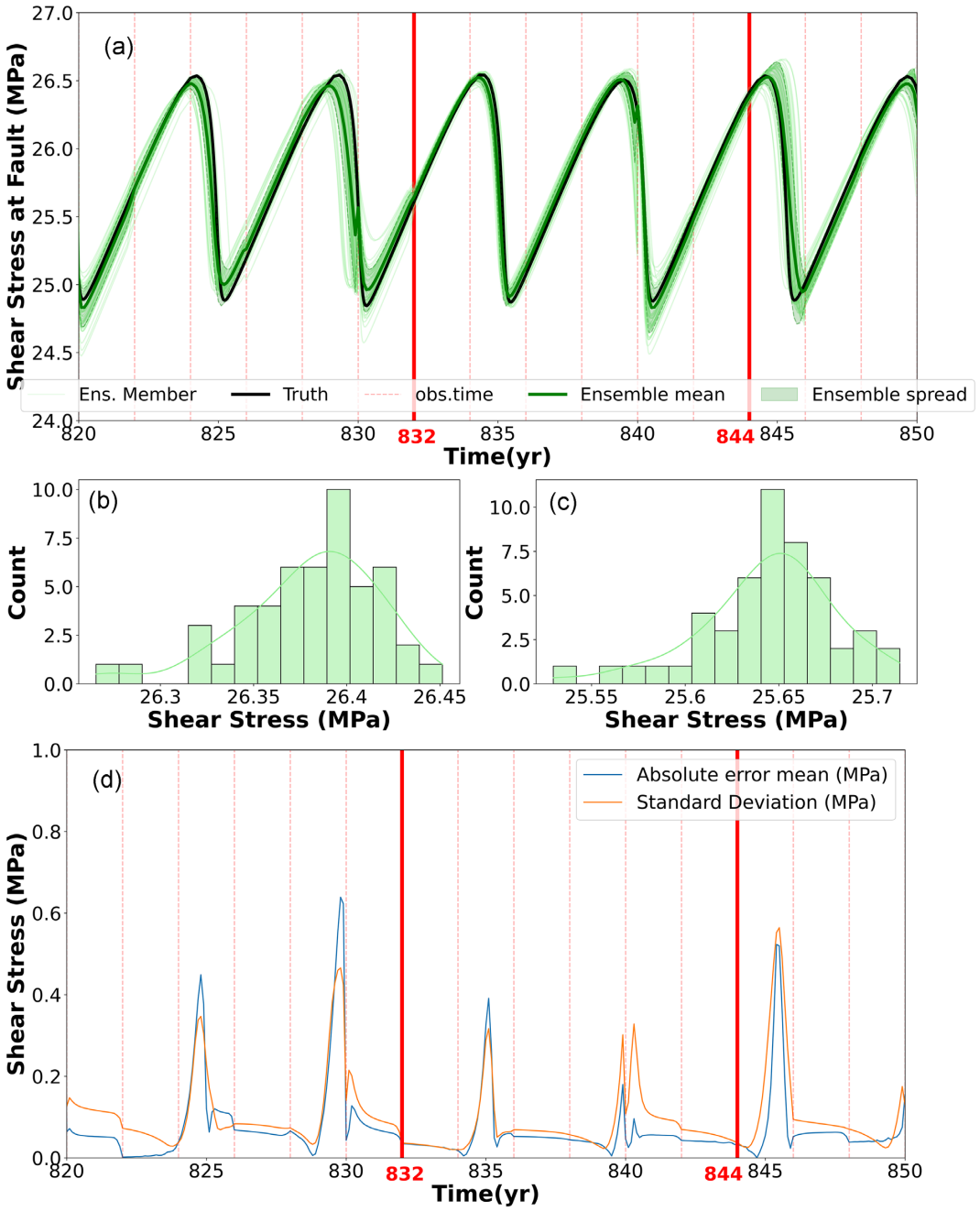

Figure 6. (a) Shear stress estimation of slow slip events with prior distributions during the (b) interseismic and (c) just prior to the coseismic phase. (d) Evolution of shear stress errors with respect to the ensemble spread. The vertical red solid lines indicate the time of the ensemble distributions shown in (b) and (c). The red dashed lines indicate the observation and assimilation times The blue line represents the absolute value of the difference between the truth and the ensemble mean estimate. The orange line represents the standard deviation of the ensemble distribution.

by the recurrence interval of the event to obtain the alarm duration. We estimate the failure rate as the fraction of events that are not forecast from the total of events that occurred.

In the Molchan diagram, a forecast located at (0,0) corresponds to having a zero prediction failure, while ringing the alarm of an upcoming slip event for a very short period. Theoretically, if we ring the alarm all the time we will not miss any slip events (i.e. 1,0) and if we never ring the alarm we will miss all events (0,1). An optimal forecast would follow a curve that comes as close as possible to the origin and to both axes. The results show that the forecastability of the EnKF for both slow slip events and earthquakes is very similar when ringing an alarm at times when 10 per cent of the ensemble members have reached their peak stress. Almost 90 per cent of the slow slip events and earthquakes are forecast when ringing the alarm just for 10 per cent of the recurrence interval duration. That means that the slow slip events were forecast as early as half a year before the slip occurrence and around 2 yr before for the earthquakes. We also evaluate how the forecastability changes when considering a different proportion of ensemble members (20 and 30 per cent). We observe that the EnKF shows lower failure rates for earthquakes for the same alarm duration than for slow slip events. This may be connected to the larger spread of the ensemble

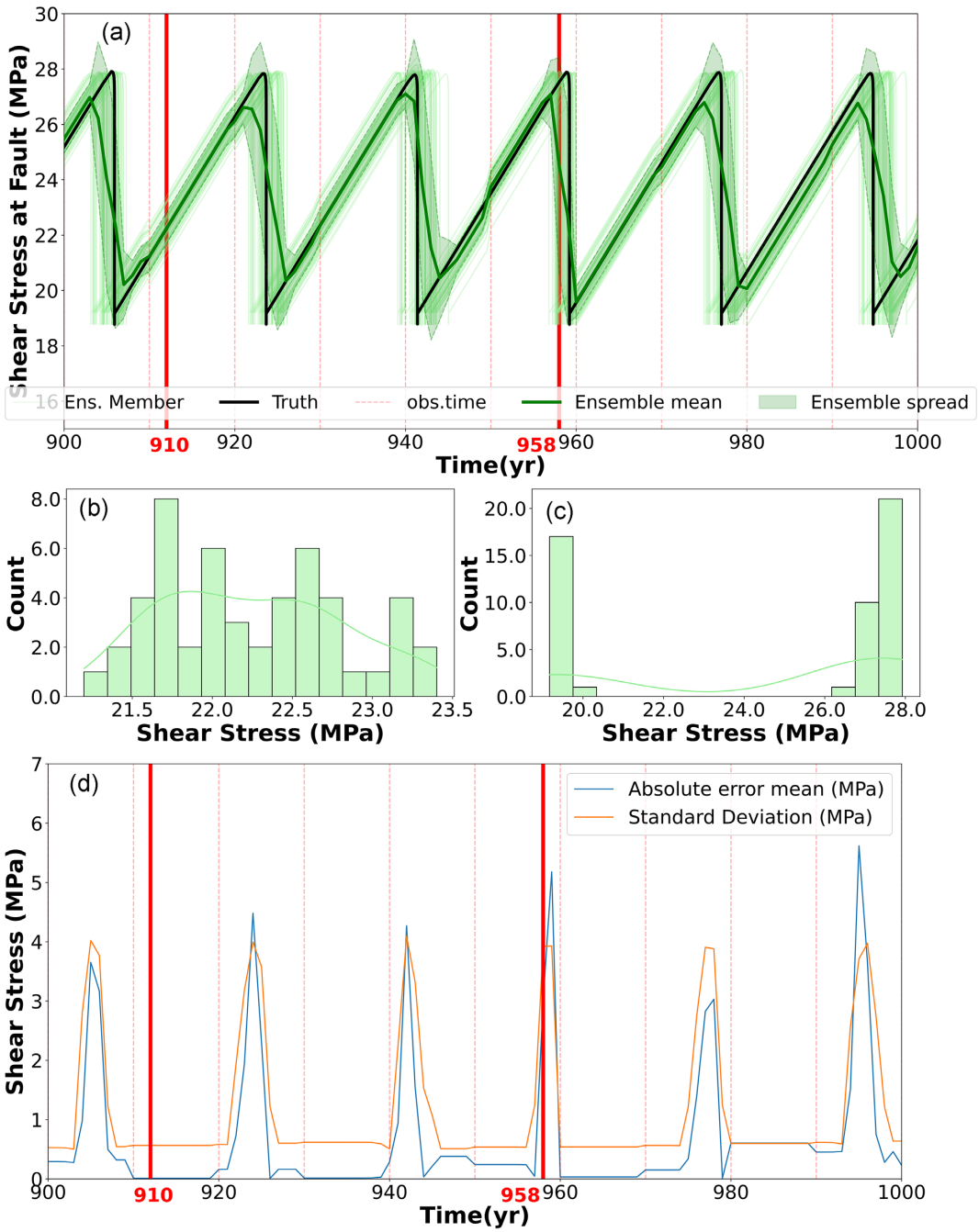

Figure 7. As in Fig. 6 but for earthquakes instead of slow slip events.

experienced when estimating earthquakes that may lead to earlier alarms when compared to the slow slip events that tend to stick closer to the truth.

4 DISCUSSION

The results show that the EnKF provides good estimates, and also the lowest deviations from the truth, during the interseismic phase of both the slow slip events and the earthquakes. The state estimates are most uncertain during and around the coseismic phase of the earthquakes. This is more pronounced for the slip-rate's and state variable θ's estimates, which experience variations of many orders of magnitude between interseismic and coseismic values. This is not surprising, as the estimate is constructed by averaging 50 ensemble members that have largely varying slip rate and state variables (e.g. Fig. 4). This aspect of the data assimilation may be different when using a different assimilation method [see, for example, chapter 19 in Evensen *et al.* (2022)].

4.1 Impact of non-linearity and non-Gaussianity

We imposed a rate-and-state friction formulation (eq. 11) on the fault to generate earthquake sequences. This dynamical friction model is non-linear in the sense that small changes in one variable (e.g. stress) trigger disproportionally large and non-linear changes in another variable (e.g. slip rate) during the coseismic phase. This type of non-linearity can be visualized by analysing the hidden state variables in a phase diagram (Fig. 11), while its impact on

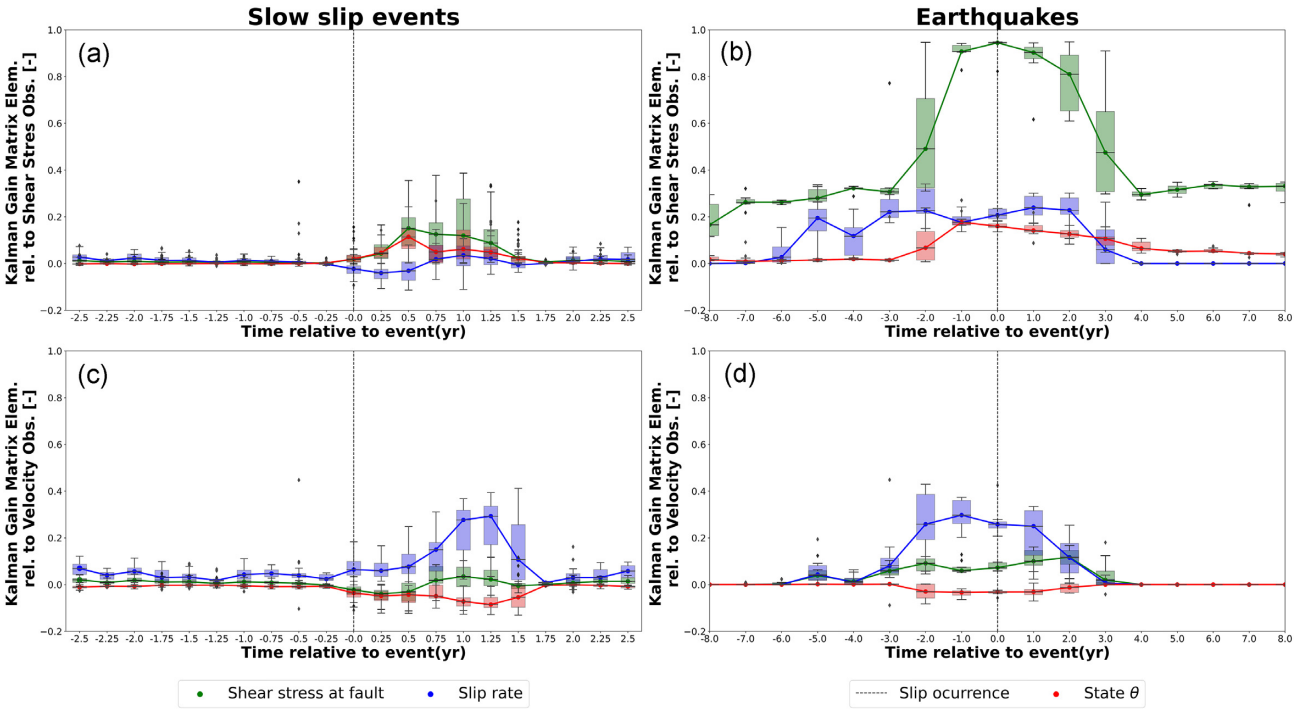

Figure 8. Kalman gain matrix elements versus the relative time to the slip occurrence for slow slip events (a, c), and earthquakes (b, d). The top row (a, b) shows the Kalman gain matrix elements with respect to the shear stress observations, and the bottom row (c, d) the Kalman gain matrix elements with respect to the velocity observations. The green boxplots are the gain values for the shear stress, the blue boxplots for the slip rate and the red boxplots for the state θ at the fault. The solid lines connect the median values for all the boxplots for a given variable.

data assimilation is analysed by visualizing the trajectories of the ensemble members during the analysis step in both phases of the seismic cycle (Fig. 12).

The phase diagrams of the slow slip events have a smoother transition between the interseismic and coseismic phases, presenting almost no sharp corners in the variables' relationships (Figs 11a and c). As shown in Fig. 6 the ensemble distributions are Gaussian, and in these phase diagrams there is a smaller lag in time between the truth and the ensemble mean observed by the closeness of the members to the truth. Besides, the truth is better captured by the ensemble in this smoother trajectory. In contrast, the earthquakes phase diagrams reveal very pronounced corners and abrupt transitions, which lead to a split of the ensemble members' distribution into two groups, corresponding to time snapshots prior and after the coseismic phase (Fig. 11d). This bi-modality of the ensemble is also seen in Fig. 7(c) and although in the interseismic phase the ensemble members surround the truth they have a more significant lag in time when compared to the slow slip events since the transitions are sharper and restrictive. This lag of time can be visualized as the distance between the members that stay further away for the truth and that will stay in the interseismic phase of a previous earthquake once the coseismic phase of the truth has already occurred.

As mentioned before, the movements in the medium are a response to the displacements at the fault. This means that sharp transitions in state variables at the fault should translate into sharp transitions in the medium. We further analyse these relationships and effects between variables by making a cross-plot of the estimated hidden states and the observed variables during a single analysis step just prior to a coseismic period (Fig. 12). This shows how the non-linearities identified in the phase diagrams translate into a disruption of the estimates of the EnKF (Fig. 12). For the slow slip events we see that the ensemble members stay close to each other and the updates from the EnKF are not as large as in the earthquakes estimates. We also observe a deviation of the ensemble members from the true cycle of the slow slip events in Figs 11(a) and (c). This effect can be explained if we consider the 1-D earthquake model as a 1-D spring-slider model (e.g. Burridge & Knopoff 1967). If we use this model to simulate slow slip events, the ratio between the stiffness of the spring and the so-called critical stiffness of the 1-D spring-slider model will determine if the events will have a decaying, increasing, or no effect in their trajectories. This ratio is slightly less than 1 for the slow slip events shown in Fig. 11 which will result in cycles which have a slightly growing orbit in the phase diagram. In the earthquake models, this effect does not occur because of the added effects of inertia and radiation damping that force the ensemble members to stay in the same cycle.

The cross-plot from the earthquake models shows that a few members are already experiencing a higher velocity, because they are accelerating towards an earthquake. These accelerating members behave as outliers in the sense that when corrected, they lie outside the overall trend observed for slip rates as illustrated by the prior ensemble members and truth trajectory. Nonetheless, despite the clear presence of bi-modality in the prior (e.g. Fig. 11d), the estimated shear stresses at the fault are projected back to end up close to the true shear stress (black star). Having a good shear-stress estimate is more important than having a good slip-rate estimate, because the forward model re-calculates slip rates using the estimated shear stresses (Section 2.3). This means that the slip-rate estimate of the data assimilation will be overwritten in the next propagation time step.

Interestingly, the comparison of the analysis for slow slip events and earthquakes shows that the update of the ensemble members in the earthquake experiments bring the ensemble members' shear stress and velocity closer to the truth than in the case of the slow

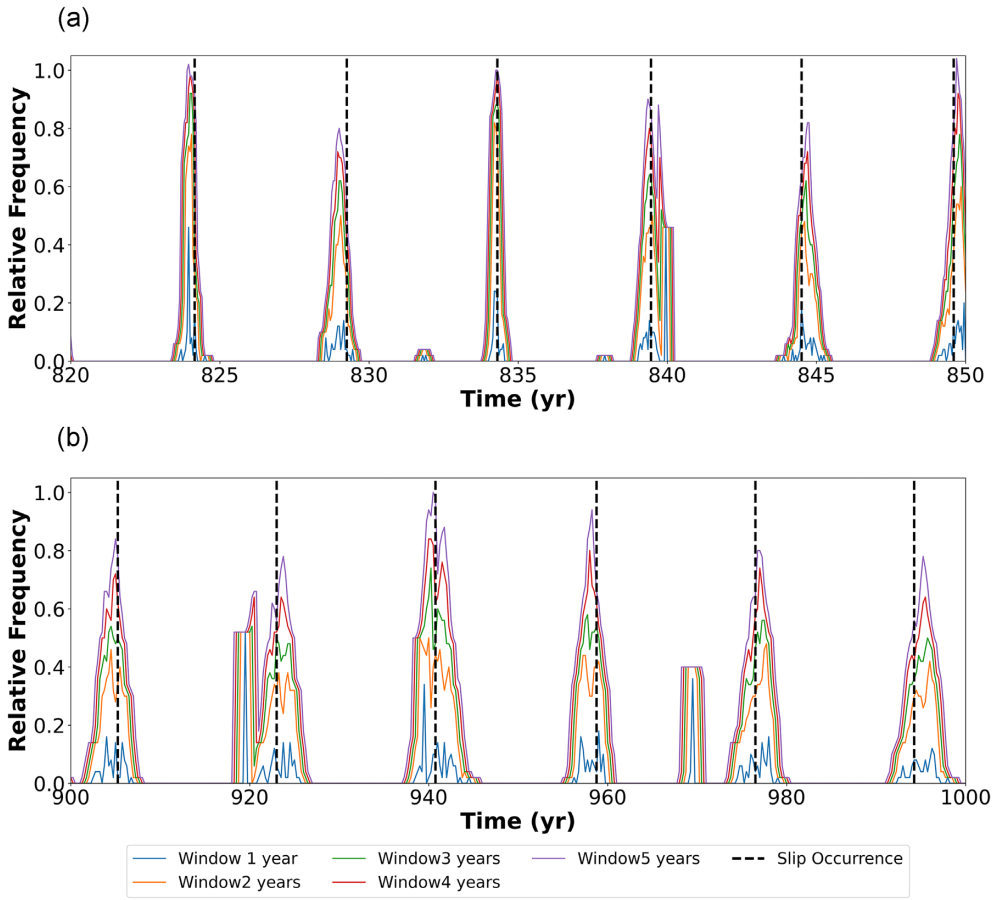

Figure 9. Time-series of the probability of fault-slip occurrence for (a) slow slip and (b) earthquakes. These estimates are calculated using a rolling time window using different window lengths. Probabilities are given as the relative frequency of the ensemble members that forecast the earthquakes. The light grey line indicates the time of the slip occurrence. The solid blue, orange, green, red and purple lines are the estimations for a window of 1, 2, 3, 4 and 5 yr, respectively.

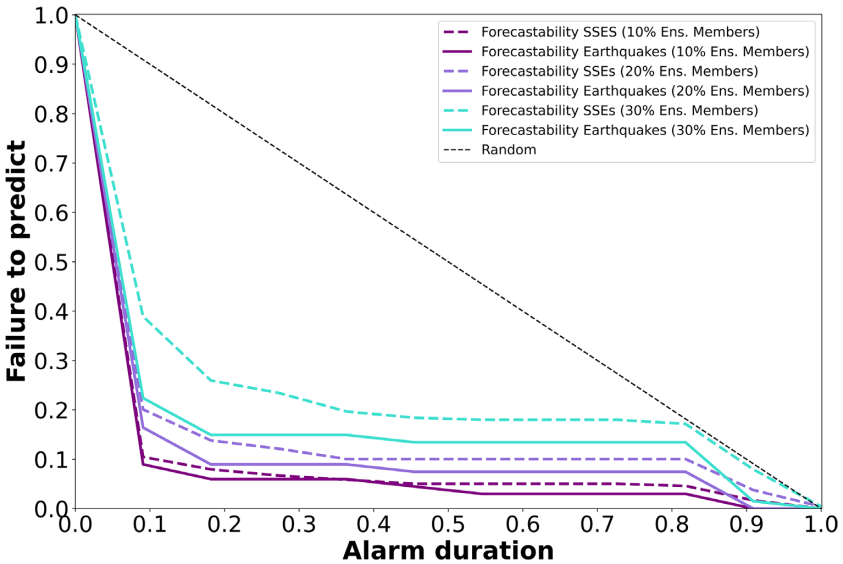

Figure 10. Molchan diagram comparing the forecasting potential of the EnKF for slow slip events and earthquakes. The horizontal axis is the alarm duration that was calculated as a percentage of the recurrence interval of the seismic event. The vertical axis is the failure to predict events and was calculated considering the percentage of slip occurrences, which were correctly forecast during the alarm duration. The dashed curves correspond to slow slip events and the solid line curves are the estimates for earthquakes.

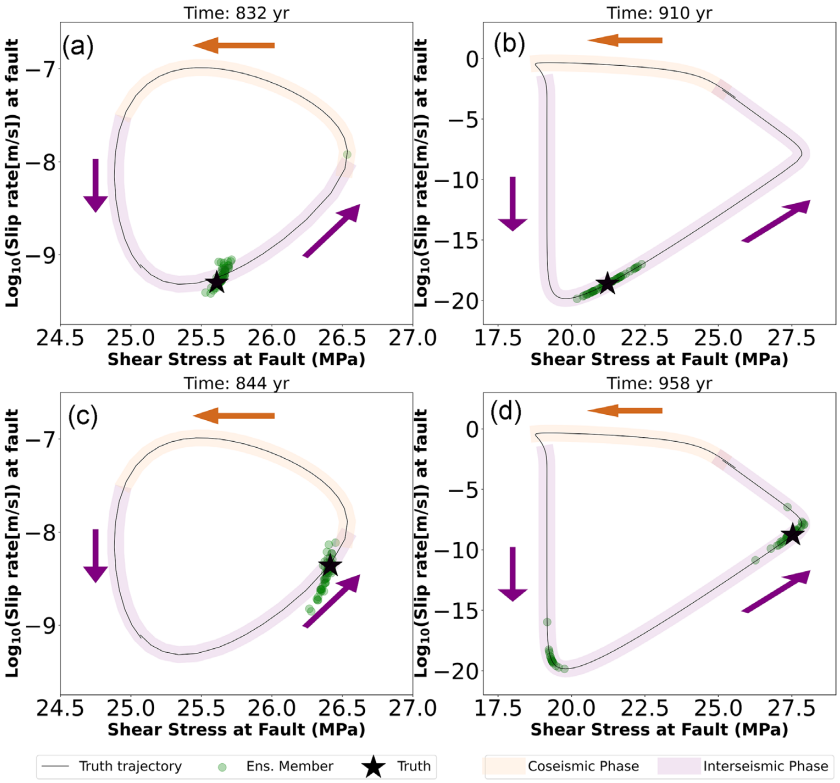

Figure 11. Phase diagrams showing the evolution of slip rate versus shear stress for a complete cycle (black line) of (a, c) slow slip events and (b, d) earthquakes. The top row shows the ensemble members (green circles) and the truth (black star) during the interseismic phase: (a) at time 832 yr, and (b) at time 910 yr, as shown in Figs 6(a) and 7(b), respectively. The bottom row shows the ensemble members and truth before the coseismic phase: (c) at time 844 yr and (d) at time 958 yr as shown in Figs 6(a) and 7(b), respectively. The purple hatched area represents the interseismic phase and the light-orange hatched area the coseismic phase.

slip events. Despite a more angular trajectory of the true shear stress in the earthquake experiment, the average slope represented in the Kalman gain (grey lines) suggests an effective update of the shear stresses at the fault. These findings suggest that the data assimilation is almost equally effective in estimating the occurrence of earthquakes as it is in estimating the occurrence of slow slip events. This interpretation is supported by the evaluation of the fit of the shear stress evolution of the ensemble members to the truth (Fig. 4), and by the quantification of the shear stress error (Figs 6 and 7), and forecastability (Fig. 10).

Finally, the above analyses and the relatively minor difference in performance between slow slip events and earthquakes also suggest that the non-Gaussianity of the prior and the strong non-linearity of the forward model do not significantly hamper the effectiveness of the EnKF.

4.2 Limitations

A major simplification in this study is that we use a 1-D model that is inherently periodic as we consider our frictional properties to be constant in time and there are no spatial heterogeneities. One of the advantages of working with such a simple setup is that we can understand the impact of the rate-and-state friction law in a simple case. Another advantage is that this type of perfect model experiments with such simplifications also helps to evaluate how the assumed settings affect the assimilation results. This is very useful when moving to a more realistic application since it helps to evaluate the impact of factors such as the amplitudes of the observation errors, observation intervals, the observed variables, and measurement location. However, we expect that to fully exploit the capability of the physics-based model to estimate the shear stresses at the fault level through data assimilation, we should use higher dimensional models.

These limitations are tackled by the use of higher dimensional models in 2-D and 3-D that are already available in Garnet (Li *et al.* 2022). The use of 3-D models especially help to evaluate the benefits of the assimilation of observations in the estimation of the spatial distribution of the shear stress at the fault. Our future work is oriented into using ensemble data assimilation methods for assimilating data observed in a metre-scale biaxial shear stress experiments (Spiers *et al.* 2017). These experiments have the advantage of offering data over a wide range of loading rates in which in every single step the parameters are relatively constant. This allows to evaluate the effect of the uncertainty in the frictional parameters separately from the complexity of the current stress field experienced by the fault.

The study presented here, is also a first step towards the application of data assimilation with real data. The synthetic experiments presented here, assumed that the shear stress and velocities close to the fault are directly observed. In reality, one of the likely difficulties is that we cannot observe the stress and velocity directly in a real case. This means that we cannot use a linear observation operator. Including more complex observation operators may introduce an additional source of non-linearity that complicate the estimation of earthquake occurrences. An alternative approach to overcome this difficulty would be to estimate the shear stress and the slip rate along

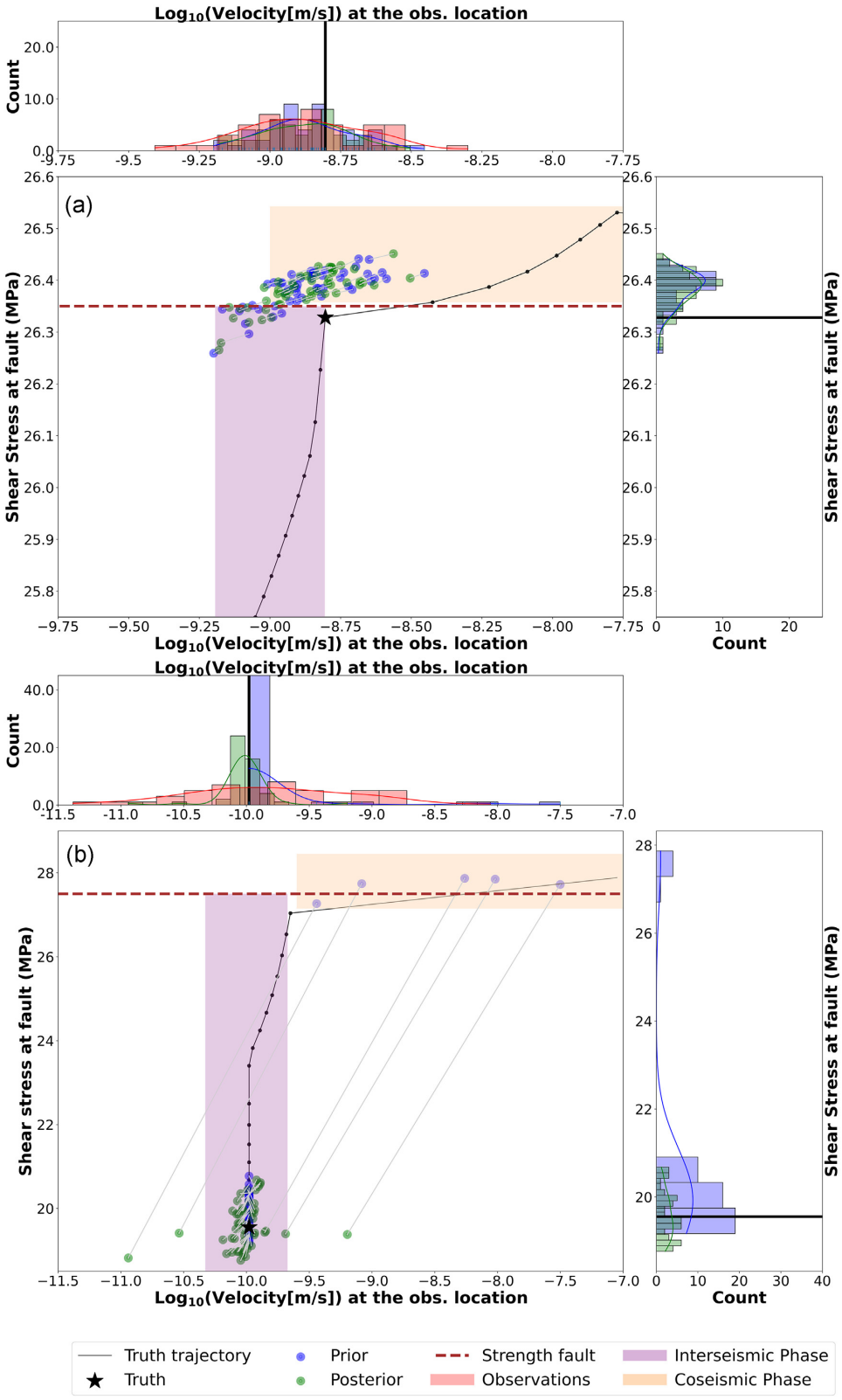

Figure 12. Ensemble results explaining how the update of the analysis step occurs when getting close to the coseismic phase at (a) 844 yr for the slow slip events and (b) 958 yr for the earthquakes. The horizontal axes show observed velocity and the vertical axis show the unobserved shear stress at the fault. The ensemble members are represented with circles. The blue circles are the prior values before the analysis step. The green circles are the posterior values after the analysis step. The light grey lines connecting the circles are the updates whose slope is the Kalman gain component corresponding to each ensemble member. The histograms on top and on the right of the cross-plot correspond to the marginal distributions of the prior (blue) and the posterior (green). The red distribution shown in the upper histogram are the perturbed observations of velocity measured at the medium. The black lines show the trajectory that the truth follows during the earthquake cycle for the current event that is being estimated. The kink in this trajectory indicates the sudden transition from the interseismic phase towards the coseismic phase. The black star is the current shear stress at the fault and velocity in the medium for the truth.

the plate interface based on crustal deformation through kinematic inversion and assimilate this information as a direct observation. However, it is important to consider that the kinematic inversion can introduce additional errors which may be hard to trace given the complexity of the system.

Another important simplification of our state estimation approach for a real application is that we assume that our parameters are correct, known and constant. As studied by Banerjee *et al.* (2022); Hirahara & Nishikiori (2019) having a bias in models with constant frictional parameters affects the accuracy of the shear stress and velocity estimates even in the case of seismic events of long duration such as slow slip events. As we have shown in this study the non-linearities in seismic events of short duration, such as earthquakes, already challenges the accuracy of the estimates. Therefore, an approach with state and parameter estimation or that includes model error could be an alternative for a real-data application where the recurrence intervals are more variable and parameters are uncertain. The estimation of parameters helps the reanalysis of past earthquakes and also gives additional degrees of freedoms to the ensemble of shear stresses, slip rates and state θ as defined by the rate-and-state friction law. However, it is important to remark the inclusion of uncertainties in the parameters may lead to different nucleation lengths across models and as a consequence lead to different time-step lengths in the model. This would make the data assimilation more computationally expensive and would reduce the comparability between members.

Comparisons of the effectiveness of the EnKF for non-periodic cases have been made by van Dinther *et al.* (2019) who benchmarked the forecastability of event alarms using ensemble data assimilation of non-periodic events compared to periodic recurrence models. Their results suggest that data assimilation in models of earthquakes with slow slip rates outperforms the periodic recurrence models especially for moderately large events. Their study highlighted the advantages of using ensemble data assimilation for including physics-based information into probabilistic hazard assessment. We further tested the state estimation of a non-periodic truth using an ensemble of periodic models (Appendix B). For a non-periodic case, we see that the accuracy of the EnKF estimates is less than in the state estimation of a periodic truth. The results suggest that when the earthquake recurrence interval of the truth is shorter than the earthquake recurrence interval of the prior, the EnKF gives a relatively accurate estimate of the shear stresses and velocities.

Finally, other data assimilation methods than the EnKF can be implemented like the particle filter which has no assumptions of the Gaussianity of the state variable distributions. The use of the particle filter or other data assimilation methods for non-linear models can be explored to have better estimates around the coseismic phase [see, for example, chapter 19 in Evensen *et al.* (2022)].

4.3 Implications

There is an increased interest in dynamic source inversion as a way of combining information from physics-based models and observations. Multiple advances have been done in kinematic inversion and dynamic inversion of past earthquakes. However, these techniques are limited to past earthquakes. Data assimilation has the potential for combining both sources of information and not only provide useful reanalysis of past events but also forecast future earthquakes. The results from this study suggest that ensemble data assimilation effectively estimates both slow slip events and earthquakes. In particular, the forecastability results show that with a very short alarm duration many future slow slip events and earthquakes occurrences can be estimated with our setup. With this proof of concept we hope to catch the attention of seismologists for the use of data assimilation to advance the field of earthquake forecasting.

5 CONCLUSIONS

In this study, an EnKF on a 1-D model representing a 0-D fault point loaded by a displaced, elastic medium, assimilates synthetic, noisy and indirect observations of shear stress and velocity. Assuming that the parameters of our physical model are perfect and using an ensemble with 50 members we estimate the shear stress, slip rate and state variable θ at the fault. We further evaluate the forecastability of the filter to estimate future occurrences of both slow slip events and earthquakes. Our results suggest that the EnKF is a useful and promising method for quantifying the uncertainty of the current state of stress, slip rates, and strength of faults.

We conclude that the estimates of the EnKF are most accurate during the interseismic phase of both the slow slip- and earthquake cycle. As an example, the absolute errors in the shear stress estimates are around 3–5 per cent of the stress drop during the interseismic phase while the standard deviation is slightly higher between about 4 and 7 per cent of the stress drop. In contrast, the largest estimation errors are found during and around the coseismic phase of the earthquakes where the shear stress errors, for example, can reach about 20–25 per cent of the stress drop shortly before and after the coseismic phase. The fast changes in the state of stress and velocity in this phase result in a sudden change in the distributions of the estimated variables. The distribution of variables becomes broader and in the case of the earthquakes it becomes bimodal which can introduce biases in the mean estimates of the variables.

The EnKF effectively estimates the occurrence of earthquakes that last only seconds, while observations are available over decadal time scales. An analysis of the influence of the observations during the analysis step shows that the assimilation of shear stress observations is very useful for the system to better estimate the shear stress on the fault for slow slip events and earthquakes. The velocity observations are most influential on estimates of slip rate for slow slip events. For earthquakes, both types of observations are relevant for the estimation of the slip rate. For the estimates of the state θ both type of observations are equally important for slow slip events but for earthquakes the shear stress are the most influential. Finally, a comparison of the evolution of the influence of the observations in time indicates that observations taken after the slip occurrence are specially important for the estimates of the filter. This is more evident in earthquakes than in slow slip events and shows the importance of the assimilation of observations recorded from previous earthquakes for better estimating the next ones.

An additional analysis of the forecastability of the EnKF for slow slip events and earthquakes shows that for both types of events there is very low forecasting failure rate of about 10 per cent when ringing very short alarms of just 10 per cent of the recurrence interval of the events. That means that most slow slip events could be forecasted half a year before their occurrence and around 2 yr before the earthquakes with an occurrence interval of approximately 20 yr. Our results suggest that data assimilation has the capacity to improve estimates of fault-slip occurrence for both slow slip events and earthquakes, and the potential to eventually advance the field of earthquake forecasting.

ACKNOWLEDGMENTS

This publication is part of the 'InFocus: An Integrated Approach to Estimating Fault Slip Occurrence' project (grant number: DEEP.NL.2018.037) funded by NWO's (Dutch Research Council) DeepNL programme, which aims to improve the fundamental understanding of the dynamics of the deep subsurface under the influence of human interventions. Additionally, the authors thank Casper Pranger who developed the code library *Garnet*, and also Lars Nerger and the Alfred Wegener Institute for Polar and Marine Research(AWI) team for the code library Parallel Data Assimilation Framework (PDAF). HDM thanks Simone Spada and Andreas Stordal, who helped to improve this paper with fruitful discussions. We thank the reviewers, Masayuki Kano and Takane Hori, for giving valuable suggestions for improvement.

Author contributions following the CReDiT taxonomy: Conceptualization: HDM, YvD, FV; Data Curation: HDM; Formal Analysis: HDM, FV; Funding Acquisition: YvD, FV; Investigation: HDM, YD, FV; Methodology: HDM, ML, YvD, FV; Project Administration: YvD, FV; Resources: YvD, FV; Supervision: YvD, FV; Validation: HDM, ML, FV; Visualization: HDM, YvD, FV; Writing-Original Draft: HDM, YvD, FV; Writing-Review and Editing: HDM, ML, YvD, FV; FV was daily supervisor of this work.

DATA AVAILABILITY

The forward model for (a)seismic slip sequences utilizing the numerical modelling package Garnet is made accessible via repository https://bitbucket.org/cpranger/garnet/src/master/. The specific code version used for this paper is available at https://bitbucket.org/cpranger/garnet/src/hamed-lisa/experiments/rate_and_state_1d/. The data produced and analysed in this study (Diab-Montero et al. 2023) is available via 4TU.ResearchData (http://doi.org/10.4121/20260932).

REFERENCES

Aanonsen, S.I., Naevdal, G., Oliver, D.S., C. R.A. & Valles, B., 2009. Ensemble Kalman filter in reservoir engineering—a review, *SPE J.,* **14,** 393–412.

Allen, R.M. & Kanamori, H., 2003. The potential for earthquake early warning in southern California, *Science,* **300,** 786–789.

Allen, R.M. & Melgar, D., 2019. Earthquake early warning: advances, scientific challenges, and societal needs, *Ann. Rev. Earth planet. Sci.,* **47,** 361–388.

Banerjee, A., van Dinther, Y. & Vossepoel, F.C., 2022. On parameter bias in earthquake sequence models using data assimilation, *Nonlin. Process. Geophys.,* **30**(2), 101–115.

Bannister, R.N., 2017. A review of operational methods of variational and ensemble-variational data assimilation, *Quart. J. R. Meteorol. Soc.,* **143,** 607–633.

Barbot, S., 2019. Slow-slip, slow earthquakes, period-two cycles, full and partial ruptures, and deterministic chaos in a single asperity fault, *Tectonophysics,* **768,** doi:10.1016/j.tecto.2019.228171.

Barbot, S., Lapusta, N. & Avouac, J.-P., 2012. Under the hood of the earthquake machine: toward predictive modeling of the seismic cycles, *Science,* **336,** 707–710.

Ben-Zion, Y. & Rice, J.R., 1995. Slip patterns and earthquake populations along different classes of faults in elastic solids, *J. geophys. Res.,* **100**(B7), 12 959–12 983.

Ben-Zion, Y. & Rice, J.R., 1997. Dynamic simulations of slip on a smooth fault in an elastic solid, *J. geophys. Res.,* **102**(B8), 17 771–17 784.

Bommer, J.J. & Abrahamson, N.A., 2006. Why do modern probabilistic seismic-hazard analyses often lead to increased hazard estimates?, *Bull. seism. Soc. Am.,* **96,** 1967–1977.

Brodsky, E.E. *et al.*, 2020. The state of stress on the fault before, during, and after a major earthquake, *Ann. Rev. Earth planet. Sci.,* **48**(1), 49–74.

Burridge, R. & Knopoff, L., 1967. Model and theoretical seismicity, *Bull. seism. Soc. Am.,* **57**(3), 341–371.

Cochard, A. & Madariaga, R., 1994. Dynamic faulting under rate-dependent friction, *Pure appl. Geophys.,* **142**(3), 419–445.

Cornell, C.A., 1968. Engineering seismic risk analysis, *Bull. seism. Soc. Am.,* **63,** 1583–1606.

Crupi, P. & Bizzarri, A., 2013. The role of radiation damping in the modeling of repeated earthquake events, *Ann. Geophys.,* **56**(1), doi:10.4401/ag-6200.

Dal Zilio, L., van Dinther, Y., Gerya, T. & Avouac, J.-P., 2019. Bimodal seismicity in the Himalaya controlled by fault friction and geometry, *Nat. Commun.,* **10**(1), 1–11.

Diab-Montero, H.A.,Li, M.,van Dinthe, Y.& Vossepoel, F.C.,2023.*Data underlying the publication: Estimating the Occurrence of Slow Slip Events and Earthquakes with an Ensemble Kalman Filter. Version 1*. 4TU.ResearchData. Dataset. doi:10.4121/20260932.v1.

Dieterich, J.H., 1978. Time-dependent friction and the mechanics of stick-slip, *Pure appl. Geophys.,* **116,** 790–806.

Dieterich, J.H., 1979. Modeling of rock friction: 1. Experimental results and constitutive equations, *J. geophys. Res.,* **84**(B5), 2161–2168.

Dieterich, J.H. & Richards-Dinger, K.B., 2010. Earthquake recurrence in simulated fault systems, in *Seismogenesis and Earthquake Forecasting: The Frank Evison Volume II,* pp. 233–250, Springer.

Dragert, H., Wang, K. & James, T.S., 2001. A silent slip event on the deeper Cascadia subduction interface, *Science,* **292**(5521), 1525–1528.

Ellsworth, W.L. & Beroza, G.C., 1995. Seismic evidence for an earthquake nucleation phase, *Science,* **268,** 851–855.

Emerick, A.A., 2018. Deterministic ensemble smoother with multiple data assimilation as an alternative for history matching seismic data, *Comput. Geosci.,* **22,** 1175–1186 .

Esteva, L., 1967. Criteria for the construction of spectra for seismic design, in *Proceedings of the 3rd Panamerican Symposium of Structures,* Caracas, Venezuela.

Evensen, G. *et al.*, 2021. An international initiative of predicting the SARS-CoV-2 pandemic using ensemble data assimilation, *Foundat. Data Sci.,* **3**(3), 413–477.

Evensen, G., 1994. Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics, *Comput. Geosci.,* **99,** 10 143–10 162.

Evensen, G., 2003. The ensemble Kalman filter: theoretical formulation and practical implementation, *Ocean Dyn.,* **53,** 343–367.

Evensen, G. & Eikrem, K.S., 2018. Strategies for conditioning reservoir models on rate data using ensemble smoothers, *Comput. Geosci.,* **22,** 1251–1270.

Evensen, G., Vossepoel, F.C. & van Leeuwen, P.J., 2022. *Data Assimilation Fundamentals: A Unified Formulation of the State and Parameter Estimation Problem,* 1st edn, Springer.

Faulkner, D., Jackson, C., Lunn, R., Schlische, R., Shipton, Z., Wibberley, C. & Withjack, M., 2010. A review of recent developments concerning the structure, mechanics and fluid flow properties of fault zones, *J. Struct. Geol.,* **32,** 1557–1575.

Fukuyama, E. *et al.*, 2014. Large-scale biaxial friction experiments using a NIED large-scale shaking table, *Rep. Nat'l. Res. Inst. Earth Sci. Disast. Prevent.,* **81,** 15–35.

Galis, M., Ampuero, J.P., Mai, P. & Cappa, F., 2017. Induced seismicity provides insight into why earthquake ruptures stop, *Sci. Adv.,* **3**(12), doi:10.1126/sciadv.aap7528.

Geller, R., 2011. Shake-up time for Japanese seismology, *Nature,* **472,** 407–409.

Herrendörfer, R., Gerya, T. & van Dinther, Y., 2018. An invariant rate- and state-dependent friction formulation for visco-elasto-plastic earthquake cycle simulations, *J. geophys. Res.,* **123**(6), 5018–5051.

Hirahara, K. & Nishikiori, K., 2019. Estimation of frictional properties and slip evolution on a long-term slow slip event fault with the ensemble Kalman filter: numerical experiments, *J. geophys. Int.,* **219,** 2074–2096.

Holliday, J.R., Nanjo, K.Z., Tiampo, K.F., Rundle, J.B. & Turcotte, D.L., 2005. Earthquake forecasting and its verification, *Nonlin. Process. Geophys.,* **12**(6), 965–977.

Hori, T., Miyazaki, S., Hyodo, M., Nakata, R. & Kaneda, Y., 2014. Earthquake forecasting system based on sequential data assimilation of slip on the plate boundary, **62,** 179–189.

Ide, S., Beroza, G.C., Shelly, D.R. & Uchide, T., 2007. A scaling law for slow earthquakes, *Nature,* **447**(7140), 76–79.

Kanamori, H. & Hauksson, E., 1992. A slow earthquake in the Santa Maria Basin, California, *Bull. seism. Soc. Am.,* **82**(5), 2087–2096.

Kano, M., Miyazaki, S., Ishikawa, Y. & Hirahara, K., 2020. Adjoint-based direct data assimilation of GNSS time series for optimizing frictional parameters and predicting postseismic deformation following the 2003 Tokachi-Oki earthquake, *J. geophys. Res.* **, 72,** doi:10.1186/s40623-020-01293-0.

Kano, M., Miyazaki, S., Ito, K. & Hirahara, K., 2013. An adjoint data assimilation method for optimizing frictional parameters on the afterslip area, *Earth, Planets Space,* **65,** 1575–1580.

Kawasaki, I., Asai, Y., Tamura, Y., Sagiya, T., Mikami, N., Okada, Y., Sakata, M. & Kasahara, M., 1995. The 1992 Sanriku-oki, Japan, ultra-slow earthquake, *J. Phys. Earth,* **43**(2), 105–116.

Koronovsky, N.V., Zakharoy, V.S. & Naimark, A.A., 2019. Short-term earthquake prediction: reality, research promise, or a phantom projection, *Moscow Univ. Geol. Bull.,* **74**(4), 333–341.

Kuehn, N.M., Hainzl, S. & Scherbaum, F., 2008. Non-Poissonian earthquake occurrence in coupled stress release models and its effect on seismic hazard, *J. geophys. Int.,* **174**(2), 649–658.

Lapusta, N. *et al.*, 2019. Modeling earthquake source processes: from tectonics to dynamic rupture, Report to National Science Foundation, Available at: . http://www.seismolab.caltech.edu/pdf/MESP_White_Paper_Main_Text_8_March_2019.pdf

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.

Li, M., Pranger, C. & van Dinther, Y., 2022. Characteristics of earthquake cycles: a cross-dimensional comparison of 0D to 3D numerical models, *J. geophys. Res.,* **127**(8),.

Liu, Y. *et al.*, 2012. Advancing data assimilation in operational hydrologic forecasting: progresses, challenges, and emerging opportunities, *Hydrol. Earth Syst. Sci.,* **16,** 3863–3887.

Liu, Y. & Rice, J.R., 2007. Spontaneous and triggered aseismic deformation transients in a subduction fault model, *J. geophys. Res.,* **112**(B9), 17 771–17 784.

Maeda, T., Obara, K., Shinohara, M., Kanazawa, T. & Uehirao, K., 2015. Successive estimation of a tsunami wavefield without earthquake source data: a data assimilation approach toward real-time tsunami forecasting, *Geophys. Res. Lett.,* **42**(19), 7923–7932.

Oba, A., Furumura, T. & Maeda, T., 2020. Data assimilation-based early forecasting of long-period ground motions for large earthquakes along the Nankai Trough, *J. geophys. Res.,* **125**(6), doi:10.1029/2019JB019047.

Ordaza, M. & Arroyo, D., 2016. On uncertainties in probabilistic seismic hazard analysis, *Earthq. Spectra,* **32,** 1405–1418.

Pranger, C., 2020. Unstable physical processes operating on self-governing fault system, improved modeling methodology, *PhD thesis*, ETH Zurich.

Pritchard, M.E. *et al.*, 2020. New opportunities to study earthquake precursors, *Seismol. Res. Lett.,* **91**(5), 2444–2447.

Reichle, R.H., 2008. Data assimilation methods in the earth sciences, *Adv. Water Resour.,* **31**(11), 1411–1418.

Rice, J.R., 1993. Spatio-temporal complexity of slip on a fault, *J. geophys. Res.,* **98 (B6),** 9885–9907.

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.

Schwartz, S.Y. & Rokosky, J.M., 2007. Slow slip events and seismic tremor at circum-Pacific subduction zones, *Rev. Geophys.,* **45**(3), doi:10.1029/2006RG000208.

Segall, P. & Bradley, A.M., 2012. Slow-slip evolves into megathrust earthquakes in 2D numerical simulations, *Geophys. Res. Lett.,* **39**(18), doi:10.1029/2012GL052811.

Shaw, B.E., Milner, K.R., Field, E.H., Richards-Dinger, K., Gilchrist, J.J., Dieterich, J.H. & Jordan, T.H., 2018. A physics-based earthquake simulator replicates seismic hazard statistics across California, *Sci. Adv.,* **4**(8), doi:10.1126/sciadv.aau0688.

Socquet, A. *et al.*, 2017. An 8 month slow slip event triggers progressive nucleation of the 2014 Chile megathrust, *Geophys. Res. Lett.,* **44**(9), 4046–4053.

Spiers, C., Hangx, S. & Niemeijer, A., 2017. New approaches in experimental research on rock and fault behaviour in the Groningen gas field, *Netherl. J. Geosci.,* **96,** s55–s69.

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.

Uchida, N., Iinuma, T., Nadeau, R.M., Bürgmann, R. & Hino, R., 2016. Periodic slow slip triggers megathrust zone earthquakes in northeastern Japan, *Science,* **351**(6272), 488–492.

van Dinther, Y., Gerya, T., Dalguer, L., Mai, P., Morra, G. & Giardini, D., 2013. The seismic cycle at subduction thrusts: Insights from seismo-thermo-mechanical models: seismo-thermo-mechanical modeling, *J. geophys. Res.,* **118**(12), 6183–6202.

van Dinther, Y., Künsch, H.R. & Fichtner, A., 2019. Ensemble data assimilation for earthquake sequences: probabilistic estimation and forecasting of fault stresses, *J. geophys. Int.,* **217**(3), 1453–1478.

van Leeuwen, P.J., 2003. Nonlinear ensemble data assimilation for the ocean. In recent developments in data assimilation for atmosphere and ocean, in *Proceedings of the ECMWF Seminar on Recent Developments in Data Assimilation for Atmosphere and Ocean,* 8–12 September, Shinfield Park, Reading, pp. 265–286.

van Leeuwen, P.J., 2010. Nonlinear data assimilation in geosciences: an extremely efficient particle filter, *Quart. J. R. Meteorol. Soc.,* **136,** 1991–1999.

Vossepoel, F.C. & Behringer, D.W., 2000. Impact of sea level assimilation on salinity variability in the western equatorial Pacific, *J. Phys. Oceanogr.,* **30,** 1706–1721.

Weaver, A.T., Vialard, J. & Anderson, D.L.T., 2003. Three- and four-dimensional variational assimilation in a general circulation model of the tropical Pacific Ocean. Part 1: formulation, internal diagnostics and consistency checks, *Mon. Wea. Rev.,* **131,** 1360–1378.

Wibberley, C.A., Yielding, G. & Di Toro, G., 2008. Recent advances in the understanding of fault zone internal structure: a review, *Geol. Soc. Spec. Publ.,* **299,** 5–33.

APPENDIX A: DIMENSIONAL ANALYSIS TO EVALUATE ENSEMBLE DEGENERACY

We perform a singular value decomposition (SVD) of the covariance matrix of our ensemble to verify whether the ensemble size is appropriate to sample and explain the variability of our system. The covariance matrix is by definition symmetric, positive semi-definite, Hermitian and all its eigenvalues are non-negative. One of the properties of square real symmetric matrices with non-negative eigenvalues is that the eigenvalues and the singular values coincide. In this section, we compare the singular values of the inner product and the outer product of the ensemble-anomaly matrix of the state vector $\boldsymbol{z_n}$ to check whether the covariance matrix of $\boldsymbol{z_n}$ is well-sampled. The reasoning behind this test is that eigenvalues of the matrices resulting from the outer product and the inner product of a column vector should coincide. If we have enough ensemble members for constructing the covariance matrix, the difference between the eigenvalues (and singular values) between both matrices

should be very small. Moreover, the ensemble size should be larger than the number of significant components in the decomposition to assure that our ensemble has enough variability to represent the numbers of 'modes' in our data. First, we demonstrate our statement that the eigenvalues for the matrices resulting from the outer and inner products of our state vector should coincide. Let $\boldsymbol{Z}$ be the ensemble-anomaly matrix of $\boldsymbol{z_n}$ given by

$$\mathbf{Z} = \frac{1}{\sqrt{N-1}}(\mathbf{z}_n - \overline{\mathbf{z}_n}). \tag{A1}$$

Let $\boldsymbol{C_{zz,outer}}$ be the outer product matrix of the ensemble-anomaly matrix, factorized as:

$$\mathbf{C_{zz,outer}} = \mathbf{ZZ}^T, \tag{A2}$$

where $\boldsymbol{Z}^T$ is the transpose of $\boldsymbol{Z}$ and $\boldsymbol{C}_{zz,\text{outer}}$ corresponds to the covariance matrix estimated by eq. (6). Let $\boldsymbol{C_{zz,inner}}$ be the inner product matrix of the ensemble,

$$\mathbf{C_{zz,inner}} = \mathbf{Z}^T\mathbf{Z} = \mathbf{LDL}^T, \tag{A3}$$

where the right hand side is the eigenvalue decomposition of $\boldsymbol{C_{zz,inner}}$ (thus, $\boldsymbol{LL}^T = \boldsymbol{I}$, and $\mathbf{D}$ is a diagonal matrix).

Second, we build an eigenvalue decomposition for $\boldsymbol{C_{zz,outer}}$. For this, we define a matrix $\boldsymbol{Z}$ as a diagonal matrix whose elements are given by the square root of the inverse of the elements of $\boldsymbol{D}$,

$$\boldsymbol{D^{-1}} = \boldsymbol{SS}^T \tag{A4}$$

, Thus $\boldsymbol{SDS}^T = \boldsymbol{I}$. Additionally we prove that,

$$(\mathbf{ZLS})^T(\mathbf{ZLS}) = \mathbf{I}, \tag{A5}$$

by multiplying both sides of the eq. (A3) by $(\boldsymbol{LS})$ and $(\boldsymbol{LS})^T$. This results in

$$(\mathbf{LS})^T\mathbf{Z}^T\mathbf{Z}(\mathbf{LS}) = (\mathbf{LS})^T\mathbf{LDL}^T(\mathbf{LS}) = \mathbf{SDS}^T = \mathbf{I}, \tag{A6}$$

Finally, we can rewrite the eigenvalue decomposition of $\boldsymbol{C}_{zz,\text{outer}}$ as

$$\mathbf{C_{zz,outer}} = (\mathbf{ZLS})\,\mathbf{D}\,(\mathbf{ZLS})^T, \tag{A7}$$

Fig. A1 shows the comparison between the SVD analysis for four different ensemble sizes, namely: 10, 20, 30 and 50 members. For the smaller ensemble sizes (10, 20 and 30 members) the scree plot of the inner product matrix do not reach the plateau of the outer product indicating that it is not well sampled. For an ensemble size of 50 members, the two curves overlap around 40 components where the singular values of both scree plots become very small and flatten. The ensemble size of 50 members is larger than the crossing of the two plots which means that this ensemble size resolves well the variability of the system.

APPENDIX B: ANALYSIS OF NON-PERIODIC EVENTS

In our perfect model experiments we assumed that the parameters were known and constant. When simulations are adequately resolved and fault properties are homogeneous, this assumption produces a fully periodic behaviour of fault slip sequences. In this appendix, we further study the performance of the EnKF for estimating non-periodic earthquake sequences. Li *et al.* (2022) derive an equation that theoretically predicts the recurrence interval T of an earthquake sequence for 1-D, 2-D and 3-D models. For a 1-D model, the equation is as follows:

$$T = \frac{\Delta\tau_{\text{dyn}}}{\dot{\tau}_{h*}} = \frac{\sigma_n(b-a)H}{2GV_l}\ln\frac{V_{\text{dyn}}}{V_l}, \tag{B1}$$

where $\Delta\tau_{\text{dyn}}$ is the stress drop, $\dot{\tau}_{h*}$ is the stress rate at a fault location, and the dynamic slip velocity V_{dyn} is approximated as 1 m s^{-1} for simplicity. We perturb the loading rate V_l through time to generate a non-periodic truth and keep the other variables in eq. (B1) constant. We assume a multiplicative noise β to follow a Gaussian distribution $\beta \sim \mathcal{N}(1, 0.1)$ and apply it to the loading rate V_l in the boundary conditions shown in eq. (14). The new boundary conditions are:

$$\begin{aligned} v_y(x=0) &= \frac{1}{2}V, \\ v_y(x=H) &= \beta\frac{1}{2}V_l. \end{aligned} \tag{B2}$$

Fig. B1 shows the variability of the recurrence interval for a synthetic truth from a non-periodic model generated by the perturbation of the loading rate. The events in the periodic model shown in the main text have a recurrence interval of 17.8 yr. The non-periodic events generated using the multiplicative noise on the loading rate have a mean recurrence interval of about 18 yr with a standard deviation of approximately 5 yr (Fig. B1). The recurrence interval thus has a coefficient of variation C_v of 0.28, which suggests this sequence is quasi-periodic (Kuehn *et al.* 2008).

We perform perfect model experiments where we assimilate synthetic observations of shear stress and velocities from the non-periodic model and estimate the shear stress, slip rate and state θ at the fault as shown in Fig. B2. The ensemble size is 50. The ensemble members can only produce periodic sequences with a recurrence interval of 17.8 yr, but each of them has a different initial shear stress value at the fault at $t = 0$.

The accuracy of the EnKF is lower for the non-periodic events when compared to the periodic events shown in Fig. 5. We observe that the EnKF gives relatively accurate estimates of the shear stresses, velocities and state θ for earthquakes with a shorter recurrence interval than 17.8 yr. However, for earthquakes with a longer recurrence interval than 17.8 yr, there are large errors during the interseismic phase. For some of the earthquakes with longer recurrence intervals, the coseismic phase is well captured by the EnKF estimates when the following earthquake cycle corresponds to a shorter recurrence interval earthquake. The RMSE for the shear stress, slip rate and state θ is larger for the non-periodic case than for the periodic case, with the difference between the slip-rate RMSE for these two cases being relatively small compared to the difference in RMSE of the other two variables (Fig. B3).

Given that each ensemble member in our forward model was unable to adapt their loading rates and the duration of the recurrence interval, we find it encouraging that shorter interseismic or unexpected coseismic periods can still be estimated regularly. The consistent early prediction for long intervals (e.g. for 260–380 yr) also suggests that including a variable assimilation time step should improve our non-periodic results. Ones we know an earthquake occurred one should always account for that wealth of new information and observations. Furthermore, van Dinther *et al.* (2019)'s results on quasi-periodic sequences for a model with a heterogeneous fault zone demonstrated significant out-performance of data assimilation with respect to periodic conceptual models in terms of forecastability.

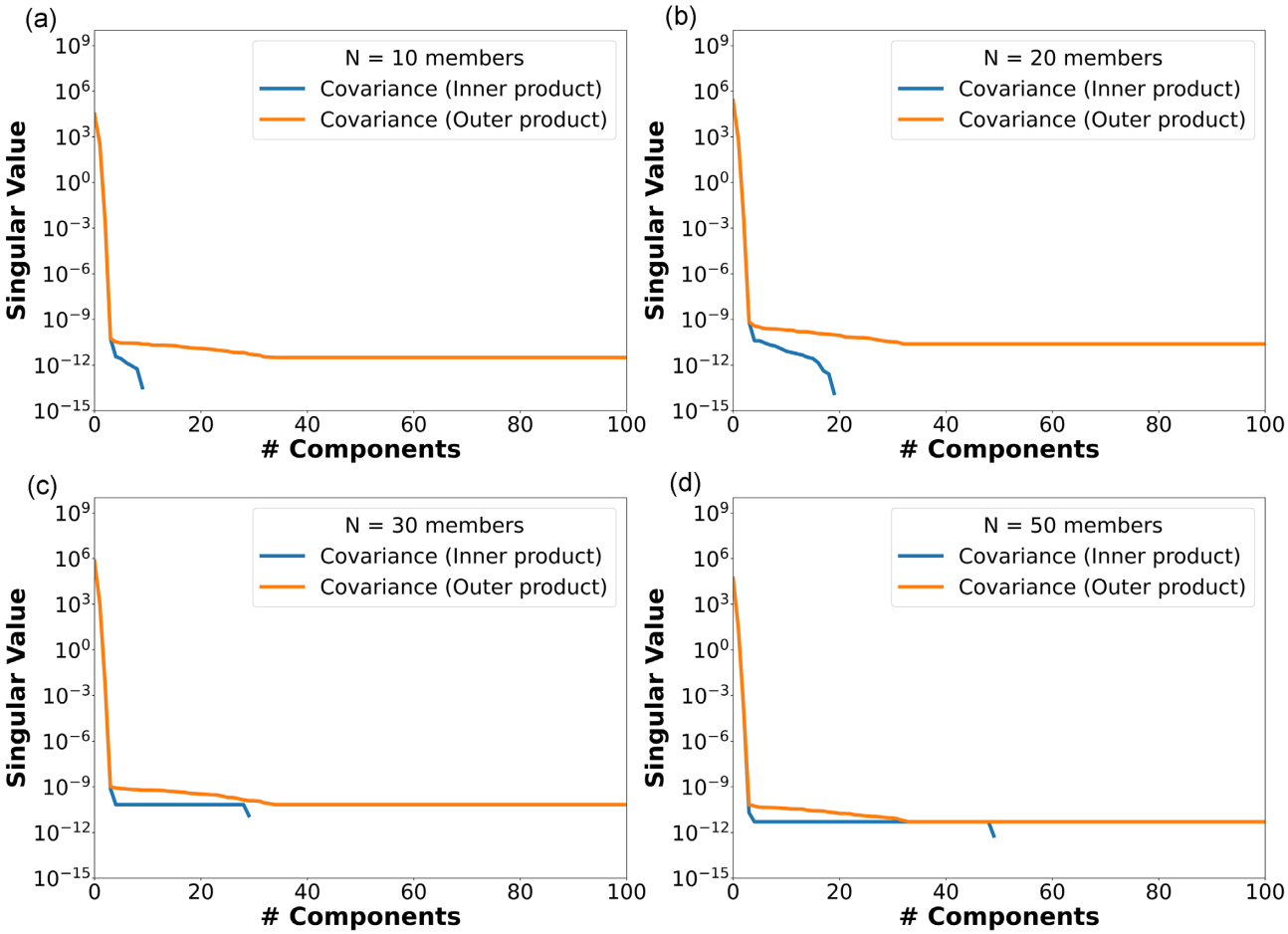

Figure A1. Scree plot for comparing the singular values and number of significant components for the inner and outer product covariance matrices obtained for an ensemble size of (a) 10 members, (b) 20 members, (c) 30 members and (d) 50 members. The solid orange line is the scree plot for the outer product of the ensemble-anomaly matrix, and the solid blue line for the inner product.

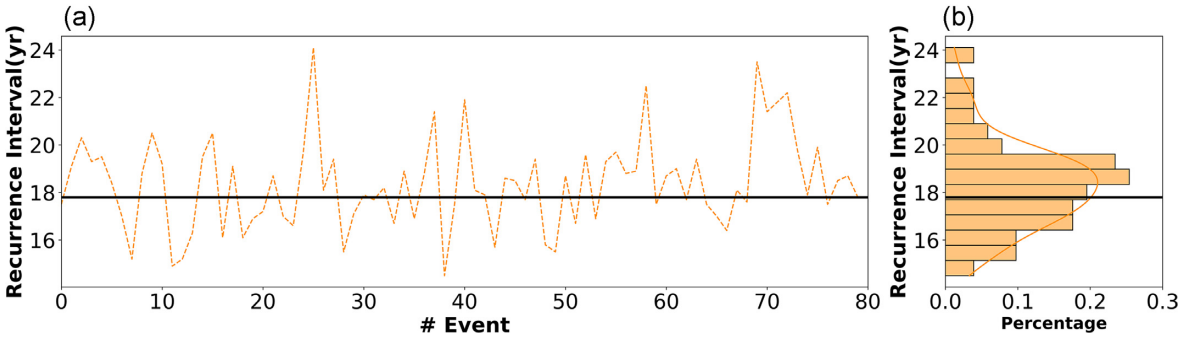

Figure B1. (a) Variability of the recurrence interval of the synthetic truth generated using a perturbation of the loading rate for a sequence of 80 earthquakes and (b) a histogram showing the distribution of the recurrence interval for the earthquakes from the sequence. The orange dashed line shows the variability of the recurrence interval for the earthquakes in the non-periodic model. The solid black line shows the recurrence interval of 17.8 yr of the periodic model.

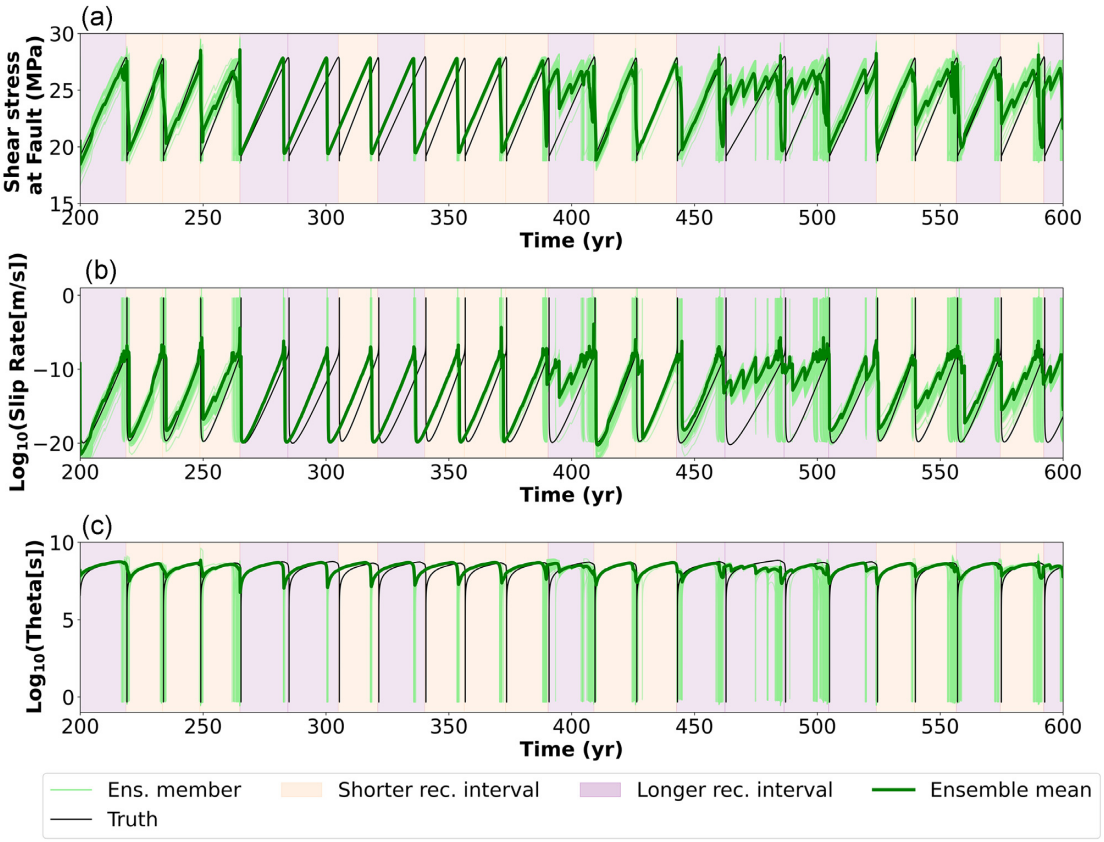

Figure B2. Estimated evolution of (a) shear stress, (b) slip rate and (c) state θ at the fault for the non-periodic earthquake sequences. The black solid line is the true evolution of the variables. The estimates of the EnKF are shown in green. The colour of the hatched zones differentiates between the events with a recurrence interval shorter than 17.8 yr (light orange) and longer than 17.8 yr (light purple).

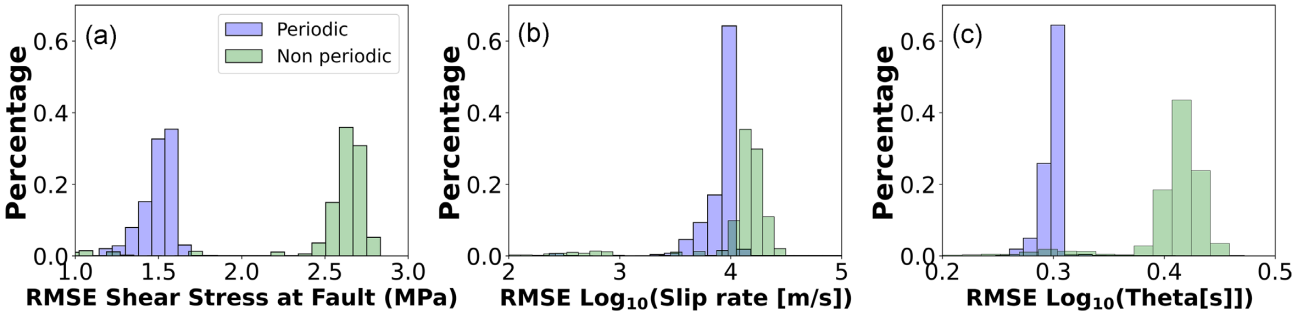

Figure B3. Comparison of the RMSE for the shear stress (a), slip rate (b) and state θ (c) for the periodic (blue) and non-periodic events (green).