Geophysical Journal International

Geophys. J. Int. (2025) **241,** 1243–1261
Advance Access publication 2025 March 17
GJI Seismology

https://doi.org/10.1093/gji/ggaf101

Assimilation of deterministic multicycle earthquake simulations into probabilistic rupture forecasts

Luis A. Vazquez and Thomas H. Jordan

Statewide California Earthquake Center, University of Southern California, Los Angeles, CA 90089–0740*, USA. E-mail: luisalbe@usc.edu*

Accepted 2025 March 10. Received 2025 March 10; in original form 2024 November 26

SUMMARY

A problem of growing importance in earthquake forecasting is how to compare probabilistic forecasting models with deterministic physical simulations and extract physical insights from their differences. Here, we compare the time-independent Uniform California Earthquake Rupture Forecast Version 3 with a long earthquake catalogue simulated by the multicycle Rate-State Quake Simulator (RSQSim). Shaw et al. generated a million-year rupture catalogue for California from RSQSim simulations based on Uniform California Earthquake Rupture Forecast, Version 3 (UCERF3) fault geometries and slip rates and found that the shaking hazard from the synthetic catalogue was in good agreement with the UCERF3 hazard maps. We take this model-to-model comparison to the more granular level of individual faults and ruptures. We map RSQSim ruptures from the Shaw18 catalogue onto equivalent UCERF3 ruptures by maximizing the mapping efficiency and ensuring that every RSQSim realizations is associated with a unique UCERF3 rupture. The full UCERF3 logic tree is used to approximate the prior distributions of individual rupture rates and fault subsection participation rates as independent gamma distributions. We formally test the ontological null hypothesis (ONH) that the empirical RSQSim rupture counts are statistically consistent with the UCERF3 rate distributions, given the sampling uncertainty of the RSQSim catalogue and the epistemic uncertainty of the UCERF3 model. Testing individual rupture rates provides little evidence either for or against the ONH, owing to the predominance of large ruptures with low recurrence rates. However, at the subsection level, the statistically significant discrepancies are much more common than expected under the ONH. We obtain a 25 per cent failure rate at the 5 per cent significance level and a 15 per cent failure rate at 1 per cent level. The false discovery rates estimated by q-value calculations are low, so we can be confident that the same subsections would likely fail if tested against an independent million-year catalogue generated by the same RSQSim model. Bayesian recalibration of the UCERF3 priors using the empirical RSQSim rates yields Gamma posterior distributions that can be derived analytically. The results of testing and recalibration, taken together, quantify how well RSQSim rupture rates agree with, and differ from, the UCERF3 forecast rates. We find that some of the discrepancies can be attributed to the differences in slip rates that drive the models, whereas others are governed by the RSQSim fault dynamics absent from UCERF3.

Key words: Statistical methods; Earthquake prediction; Seismic-event rates; Numerical techniques; Geostatistics.

1 INTRODUCTION

This paper concerns the assimilation of information from a deterministic earthquake rupture simulator (ERS) into a probabilistic earthquake rupture forecast (ERF). The particular problem we investigate is the utilization of a long (∼1 My) catalogue of California seismicity from the Rate-State Quake Simulator (RSQSim; Dieterich & Richards-Dinger 2010; Richards-Dinger & Dieterich 2012; Shaw 2019) to recalibrate the rates of large earthquakes (moment magnitude $M \geq 6.7$) prescribed by the time-independent Uniform California Earthquake Rupture Forecast, Version 3 (UCERF3-TI) (Field *et al.* 2014). The UCERF3 model and RSQSim data set are summarized in Sections 2 and 3.

Our study of the ERS-to-ERF assimilation problem is motivated by the prospects for improving forecasting skill by incorporating more information about fault-system dynamics into probabilistic

ERFs (Tullis *et al.* 2012; Field 2019). Physics-based ERSs such as RSQSim are capable of modelling the dynamical processes of rupture nucleation and quasi-static stress transfer within networks of interacting faults that are as complex as the fault-system ERFs currently used in California (Richards-Dinger & Dieterich 2012) and New Zealand (Shaw *et al.* 2022), so that direct ERS/ERF comparisons are now possible.

Shaw *et al.* (2018) generated a million-year earthquake catalogue from an RSQSim California simulator with UCERF3 fault geometries and geologic slip rates ('Shaw18 catalogue') and directly compared maps of the expected shaking hazard derived from the synthetic catalogue with the published UCERF3 hazard maps. The maps were remarkably similar, suggesting that studies of ERS/ERF similarities and differences could potentially improve the physical framework and predictive performance of fault-system ERFs.

In this study, we take this head-to-head comparison to a more granular level of individual faults and ruptures, formally testing whether the empirical RSQSim rupture counts from the Shaw18 catalogue are statistically consistent with the UCERF3 rate distributions. Basic aspects of the UCERF3 and RSQSim models are described in Sections 2 and 3, respectively. Setting up the comparison requires that we associate each RSQSim rupture realization with a unique, but geometrically much simpler, UCERF3 rupture. We follow Shaw *et al.* (2018) in using an area overlap threshold to map the large ERS event set onto the much smaller ERF event set and we optimize the association by minimizing two types of rupture area differences (Section 4).

The association process generates a set of RSQSim hits for each UCERF3 rupture and each UCERF3 fault subsection, and we use these statistics to test whether the empirical RSQSim rupture counts from the Shaw18 catalogue are statistically consistent with the UCERF3 rate distributions, given the forecasting uncertainty. Forecasting consistency under epistemic uncertainty conforms to what Marzocchi & Jordan (2014) have called the *ontological null hypothesis*, denoted by $\mathbb{O}$. For the problem at hand, $\mathbb{O}$ is a 'collective' null hypothesis comprising many subhypotheses, one for each rupture or fault subsection. These subhypotheses can be collectively tested using standardized frequentist techniques of multihypothesis significance testing, which quantifies significance in terms of p-values (Marzocchi & Jordan 2014, 2018). Our ontological testing procedure, described in Section 5, accounts for the finiteness of the RSQSim sample as well as the epistemic uncertainty in the UCERF3 model.

Rejecting a subhypothesis of $\mathbb{O}$ at a sufficiently small p-value identifies an *ontological error*. Were we testing UCERF3 against actual data, discovery of an ontological error would signal a deficiency in the ability of the model to forecast real-world events. Our study uses no real data, of course; we instead test UCERF3 against synthetic data from a hypothetical 'RSQSim world'. Such a comparison says nothing about UCERF3's real-world performance, but it does provide a statistically formal method for discovering ontological incompatibilities between the deterministic simulator and the probabilistic forecast that can be examined for physical significance.

In Section 6, we apply a Bayesian recalibration scheme to derive the posterior distributions of rupture rates conditioned on the RSQSim data and the UCERF3 prior distribution, and we compare the test results with the recalibrated participation rates. In Section 7, we relate the results of the testing and recalibration to the interpretation of hazard-map discrepancies, and we briefly discuss future applications of the assimilation methods to the challenging problems of time-dependent forecasting.

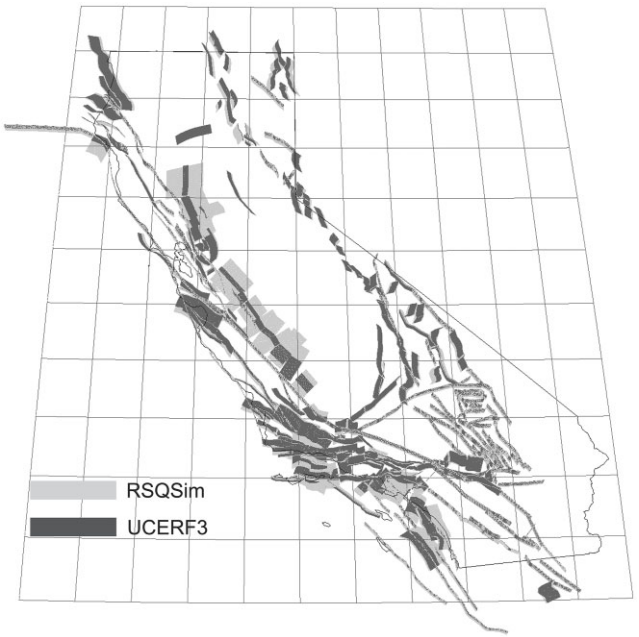

Figure 1. 3-D projection comparing the California fault geometries used in the RSQSim and UCERF3 models. Dark grey areas are UCERF3 faults, which are confined to seismogenic layer of variable thickness and reduced in area by a variable aseismicity factor that accounts for aseismic slip. The average seismogenic thickness for UCERF3 is about 13 km. Light grey areas show the down-dip extensions of the RSQSim faults, which continue the UCERF3 faults to a depth of 18 km.

2 UCERF3

UCERF3-TI has been the California component of the National Seismic Hazard Model (NSHM) for the past decade (Field *et al.* 2014; Petersen *et al.* 2020), providing authoritative estimates of the magnitude, location and time-averaged rate of potentially damaging earthquakes across the state. UCERF3-TI is based on the statistical assumption that ruptures occur randomly in time according to independent and stationary Poisson processes; that is, the rupture rates are constant and the occurrence times of different ruptures are statistically independent.

UCERF3-TI models two types of earthquake sources: ruptures of one or more faults within a fault network (fault-system sources) and sources from a spatial grid of distributed seismic activity (gridded 'off-fault' seismicity). In this study, we focus on the fault-system sources and ignore the gridded seismicity, which is only a minor contributor to California earthquake hazard at the high return periods ($\geq$ 475 yr) considered in most applications of probabilistic seismic hazard analysis (PSHA; Field *et al.* 2014; Shaw *et al.* 2018).

The fault-system sources constitute a finite set of possible supraseismogenic ruptures, which we index by an integer i. Each has a fixed fault geometry and a constant, but uncertain, rupture rate given by the random variable λ_i, usually expressed as number of ruptures per year. We restrict our calculations to the rupture set derived from a single fault model, FM3.1, which is one of two supported by UCERF3 (Field *et al.* 2014). The simplification is acceptable because the hazard differences between the two fault models are known to be small (Field *et al.* 2023).

FM3.1 comprises 2606 rectangular subsections with along-strike lengths about half of their down-dip widths (Fig. 1). By definition, supraseismogenic ruptures comprise two or more contiguous subsections. The average magnitude of the two-subsection ruptures is

about 6.2. Larger ruptures are built by adding neighbouring subsections according to a set of plausibility criteria, which include Coulomb compatibility and a 5-km limit on fault jumps (Milner *et al.* 2013). In total, FM3.1 defines a set of 252 945 supraseismogenic ruptures with non-zero rates. Owing to the multiplicity of multifault ruptures, most of these sources produce earthquakes that are large but occur at low rates, many vanishingly small. The fraction of distinct UCERF3 ruptures with $M \geq 6.7$ is 96.9 per cent of the total (245 042), but the model assigns mean rates greater than 10^{-6} per year to only 4.7 per cent; that is only 11 473 of the $M \geq 6.7$ ruptures have mean recurrence intervals of 1 My or less.

UCERF3 characterizes the aleatory variability of the supraseismogenic ruptures by the rate variables λ_i, and it represents the epistemic uncertainty of λ_i by a logic tree of weighted model alternatives (branches). The eight-level logic tree for UCERF3-TI/FM3.1 generates 720 end nodes (leaves), $\{\lambda_i^{(l)}\}$. Weights $\{w_l\}$ are assigned to the branches by expert judgement and normalized to a unit sum; they are assumed to be the same for all ruptures (Field *et al.* 2014; Fig. 3). The discrete, empirical distribution of epistemic uncertainty can thus be represented as a finite, weighted set of hazard-curve alternatives,

$$\mathcal{E}(\lambda_i) = \left\{\lambda_i^{(l)}, w_l \ : \ l = 1, \cdots, 720\right\}, \tag{1}$$

where $\lambda_i^{(l)} \varepsilon (0, \infty)$, $w_l \varepsilon [0, 1]$ and $\sum_l w_l = 1$. To emphasize that the branch weighting process is inherently subjective, Marzocchi & Jordan (2014) called $\mathcal{E}$ the *experts' ensemble*. The expectation of the aleatory variable λ_i under $\mathcal{E}$ defines the mean UCERF3 rupture rate,

$$\bar{\lambda}_i = \sum_{l=1}^{720} w_l \, \lambda_i^{(l)}. \tag{2}$$

In PSHA, these mean rates are fed into the ground motion models to compute maps of the expected shaking hazard (Petersen *et al.* 2015).

Hereafter, we refer to the specific model UCERF3-TI/FM3.1 simply as UCERF3.

3 RSQSim

RSQSim is a boundary element, event-driven ERS developed by Dieterich & Richards-Dinger (2010), Richards-Dinger & Dieterich (2012) and Shaw (2019). The simulator captures three essential features of earthquake physics: pre-seismic nucleation, represented in terms of rate and state dependent friction; coseismic stress transfer, implemented as quasi-static Coulomb interactions within a homogeneous half-space; and steady stress accumulation, modelled by the hybrid tectonic loading scheme of Shaw (2019). Shaw's three-layer loading model improves the depth distribution of hypocentres, avoiding the seismicity artifacts observed in simulations based on the simpler backslip approximation used in the original version of the simulator. The rupture process includes radiation damping and dynamic overshoot.

The simulator is coded as an efficient three-state, event-stepping algorithm that has been efficiently parallelized on supercomputers. The computational efficiency of RSQSim allowed Shaw *et al.* (2018) to generate a 996 956-yr ('million-year') catalogue, here referred to as the Shaw18 catalogue, and numerically document the sensitivity of the catalogue to changes in the main model parameters.

From a dynamical perspective, RSQSim is a rather crude multicycle ERS that ignores important physical phenomena, such as elastic waves, viscoelasticity and off-fault seismicity. Despite these limitations, its ability to simulate very long earthquake catalogues makes it a promising tool for studying complex fault interactions (Tullis *et al.* 2012; Field 2019), and its utility in seismic hazard analysis has been demonstrated in a series of applications. RSQSim California simulations were used to guide the rupture plausibility filters of UCERF3 (Field *et al.* 2014) and the 2023 revision of NSHM earthquake rupture forecast (Milner *et al.* 2022; Field *et al.* 2023). RSQSim simulations incorporating New Zealand fault models played a similar role in the 2022 update of the New Zealand NSHM (Gerstenberger *et al.* 2022; Shaw *et al.* 2022), and they have more recently been used to study the eastern Taiwan fault system (Chia-Cheng & Hung-Yu 2024).

As previously noted, the RSQSim simulator, when stocked with the UCERF3 faults and slip rates, produces earthquake rates and hazard maps that are similar to UCERF3 (Shaw *et al.* 2018). RSQSim catalogues have recently been used to drive CyberShake ground motion simulations in the Los Angeles region, producing the first realistic seismic hazard model derived entirely from deterministic, physics-based simulations (Milner *et al.* 2021). Here, we show that ERS catalogues can provide important ontological consistency checks on probabilistic, empirically based ERFs that may prove useful in assessing the compatibility of the ERFs with salient aspects of earthquake physics.

The RSQSim California fault model used by Shaw *et al.* (2018) was derived from the UCERF3 fault model FM3.1 but differs in the scale of the discretization and the down-dip extent of faulting. RSQSim discretizes FM3.1 into 265 464 triangular patches with an average area of 1.35 km^2, almost two orders of magnitude smaller than the average UCERF3 subsection area ($\sim$90 km^2). The ruptures produced by RSQSim consequently show much more spatial complexity than the ruptures represented by UCERF3.

UCERF3 restricts fault rupturing to a variable-depth seismogenic layer with a thickness estimated from the local seismicity, whereas RSQSim allows rupturing to extend downwards to a uniform depth cut-off at 18 km, which is about 5 km greater than the average UCERF3 seismogenic thickness of 13 km. Consequently, the depth extent of large RSQSim ruptures is systematically greater than the maximum depth of comparable UCERF3 ruptures (Fig. 1).

The magnitude–frequency distribution (MFD) of the Shaw18 catalog is quite different from the UCERF3 MFD (Fig. 2). The simulation produced 203 072 ruptures with $M \geq 6.7$, giving an average rate of 0.203 events yr^{-1}, which is 60 per cent higher than the UCERF3 rate of 0.128 events yr^{-1}. The cumulative moment rate for RSQSIM is 1.1×10^{20} Nm yr^{-1}, about 50 per cent higher than the UCERF3 value of 7.2×10^{19} Nm yr^{-1}. RSQSim slip rates were chosen to match the UCERF3 geologic rates, so that the increase in cumulative moment rate is primarily due to the increase in the depth of faulting. In contrast, the increase in cumulative rupture rates reflects the shape of the MFD. The RSQSim distribution bulges above the UCERF3 distribution in the magnitude range 6.5–7.7 and falls below it above magnitude 7.9. The largest magnitude in the Shaw18 catalogue is 8.1, whereas UCERF3 forecasts events with $M \geq 8.3$ at an average rate greater than 10^{-4} yr^{-1}, that is, at mean return periods less than 10 000 yr.

4 RUPTURE ASSOCIATION

Rupture association is the process of selecting a unique member of the UCERF3 rupture set to represent each RSQSim rupture realization. The first step is to map the RSQSim ruptures onto subsets of

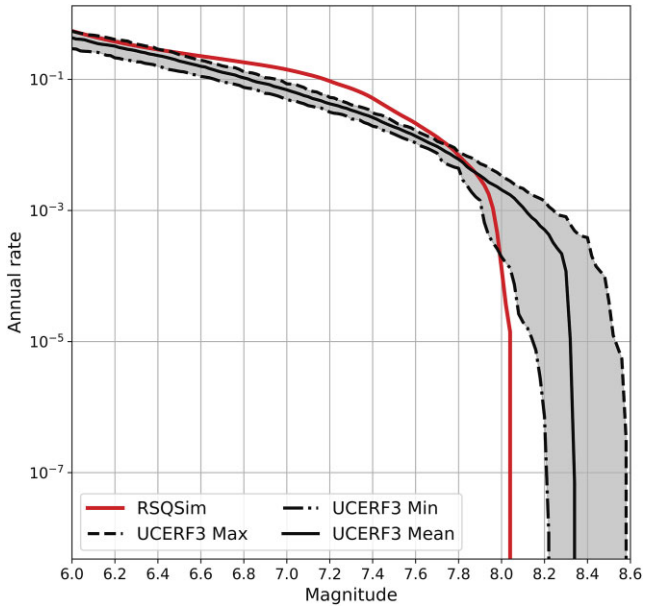

Figure 2. Cumulative magnitude–frequency distributions for UCERF3-TI/FM3.1 (solid line within the dashed region) and the million-year RSQSim catalogue (solid line outside the dashed region). Dashed and dash–dotted lines are respectively the maximum and minimum annual rates from the experts' distribution given by 720 branches of the UCERF3-TI/FM3.1 logic tree.

the FM3.1 subsections, and the second is to associate these subsets with UCERF3 ruptures. We follow Shaw *et al.* (2018) in using an area overlap threshold to perform the mapping, and we optimize the association by minimizing the subsection differences.

4.1 UCERF3 rupture set

We let $\{s : 1, 2, \cdots, N_S\}$ be the index set comprising all fault subsections defined by FM3.1 ($N_S = 2606$), and we identify a UCERF3 rupture by the indices of its participating subsections; for example, the i^{th} rupture is represented by $\{s \varepsilon U_i\}$, where U_i is the index set of the participating subsections. The number of subsections comprised by a rupture (the size of U_i) ranges from 2 to 211. The complete collection of UCERF3 supraseismogenic ruptures $\{U_i\}$ consists of 252 945 index sets.

In UCERF3, the rupture area of a participating subsection is the total fault area of that subsection, denoted a_s^{F}, reduced by a variable aseismicity factor d_s, which accounts for the observed or inferred aseismic slip (Field *et al.* 2014). The reduced fault area, $a_s^{\text{U}} := (1 - d_s) a_s^{\text{F}}$, is the same for all UCERF3 ruptures that involve subsection s. The area of the i^{th} UCERF3 rupture is the sum of these reduced areas,

$$A_i^{\text{U}} := \sum_{s \varepsilon U_i} a_s^{\text{U}} = \left[1 - D_i\right] A_i^{\text{F}}. \tag{3}$$

Here $A_i^{\text{F}} := \sum_{s \varepsilon U_i} a_s^{\text{F}}$ is the total (unreduced) fault area and

$$D_i = 1 - \frac{A_i^{\text{U}}}{A_i^{\text{F}}} = \frac{\sum_{s \varepsilon U_i} a_s^{\text{F}} d_s}{\sum_{s \varepsilon U_i} a_s^{\text{F}}} \tag{4}$$

is the aseismicity factor for the i^{th} rupture.

The mean UCERF3 magnitude is computed from the reduced area using the UCERF3 ensemble of magnitude–area relationships (Shaw 2013; Field *et al.* 2014). We distinguish this reduced magnitude M_i^{U} from the average unreduced (full-fault) magnitude M_i^{F}. The two magnitude-area distributions, M_i^{F} versus A_i^{F} (Fig. 3a) and M_i^{U} versus A_i^{U} (Fig. 3b), are very similar; both can be approximated by a bilinear scaling relation of the form $M \sim C_n^{\text{U}} \log A$, where the slope transitions from $C_1^{\text{U}} = 1.1$ to $C_2^{\text{U}} = 1.0$ at $M = 7.7$. The two magnitudes are related by

$$M_i^{\text{U}} = M_i^{\text{F}} + \log\left[1 - D_i\right]. \tag{5}$$

Scatter plots of A_i^{U} versus A_i^{F} and M_i^{U} versus M_i^{F} (Figs 3c and d) show the area and magnitude reductions that were applied to the UCERF3 ruptures.

4.2 RSQSim realizations

There are 547 563 ruptures in the Shaw18 catalogue. The j^{th} realization is defined by the set of triangular elements on the native RSQSim grid that were active (i.e. achieved simulator state 2) during that event. The median of the RSQSim magnitude–area distribution (Fig. 4) can be approximated by a trilinear scaling relation where the slope decreases from $C_1^{\text{R}} = 0.9$ at lower magnitudes to $C_3^{\text{R}} = 0.7$ at higher magnitudes, close to the value of 2/3 expected from stress-drop models. A transitional value, $C_2^{\text{R}} = 0.8$, holds in the range $7.2 \leq M < 7.7$. This trilinear relation is in fair agreement with Shaw's (2013, 2023) compilations of the magnitude–area observations.

By extending the FM3.1 subsections to the maximum depth of the RSQSim rupture (18 km), we can identify each of the active triangular elements with a unique subsection. Summing over these elements gives the area of the j^{th} realization mapped onto subsection s, which we denote by a_{sj}^{R}. For fully ruptured subsections, a_{sj}^{R} is usually greater than a_s^{U}, owing to the rupture area reductions by the UCERF3 aseismicity factors and the down-dip extensions of the RSQSim ruptures below the seismogenic zone. In extreme instances, d_s can be as large as 0.9, and the area ratio can exceed a factor of eight.

The RSQSim realizations are resolved on 1-km^2 triangular patches, much smaller in scale than the UCERF3 subsections, and they can be very complex, featuring discontinuous rupture areas, strong depth variations, jumps across fault gaps and coseismic triggered events isolated from the main rupture (Richards-Dinger & Dieterich 2012). Owing to the small-scale heterogeneity, RSQSim ruptures commonly involve subsections for which the rupture area is much smaller than the prescribed UCERF3 rupture area; that is, $a_{sj}^{\text{R}} \ll a_s^{\text{U}}$. Retaining these minor participants can inflate the number of subsections and thus bias the UCERF3 representation to high rupture areas. We reduce this bias by including a subsection as a participant in the RSQSim realization if and only if the RSQSim/UCERF3 area ratio is greater than a fixed overlap threshold $\eta > 0$. The RSQSim realizations are thus represented by a threshold-dependent collection of subsection index sets $\{R_j(\eta) : j = 1, 2, \cdots, 547, 563\}$, where $s \varepsilon R_j(\eta)$ only if $a_{sj}^{\text{R}} / a_s^{\text{U}} \geq \eta$.

The area of the RSQSim rupture mapped onto the UCERF3 subsection set, $\tilde{A}_j^{\text{R}}(\eta) := \sum_{s \varepsilon R_j} a_{sj}^{\text{R}}(\eta)$, is less than or equal to the true rupture area A_j^{R}, which reduces the rupture magnitude from M_j^{R} to $\tilde{M}_j^{\text{R}}(\eta)$. Because the latter approaches the former as η goes to zero, the magnitude is better preserved when η is smaller. Shaw *et al.* (2018) adopted $\eta = 0.2$ as the rupture threshold and showed that the quality of the mapping was fairly insensitive to threshold values between 0.1 and 0.4.

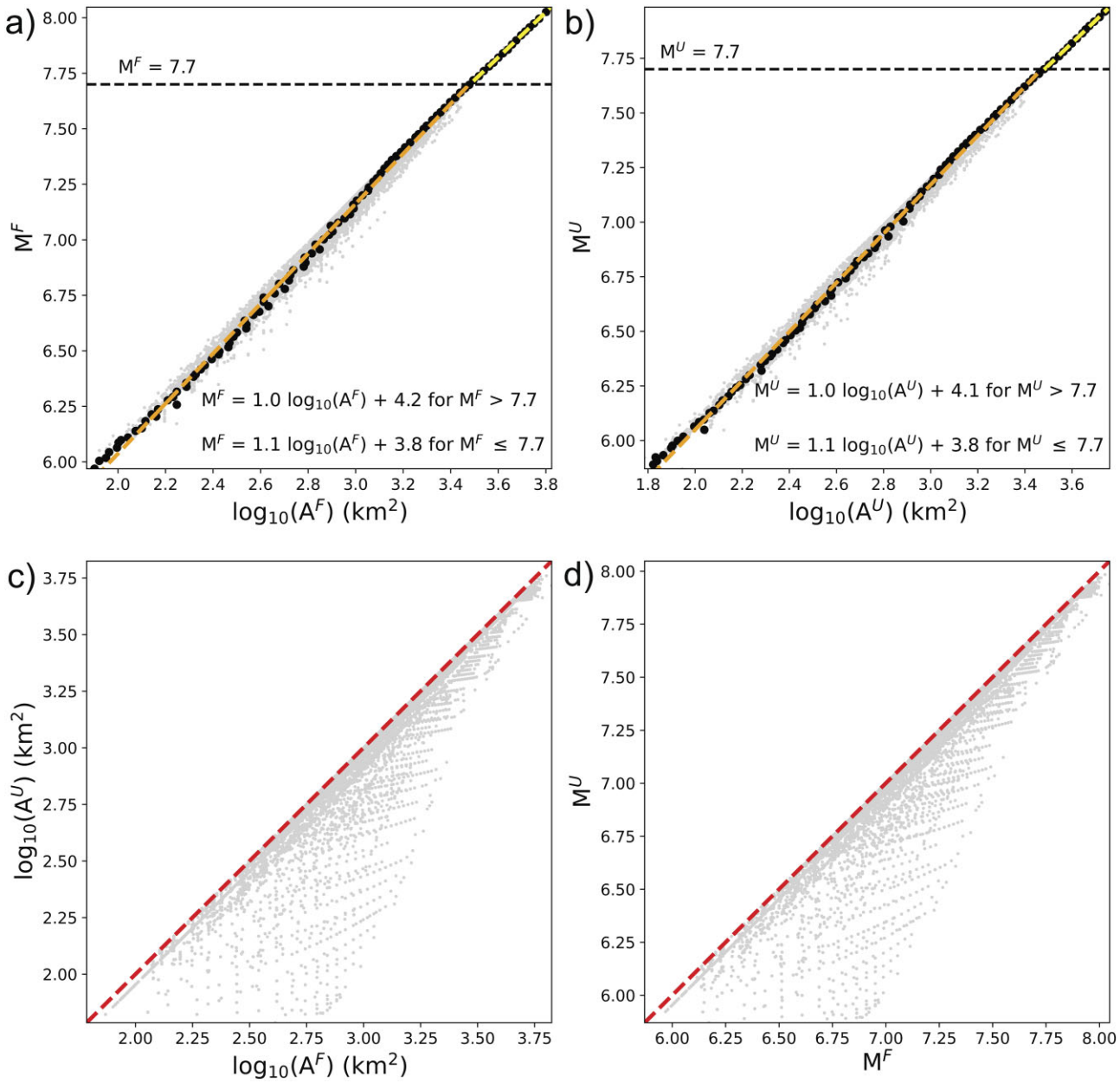

Figure 3. Scatter plots of the two magnitude–area distributions for UCERF3, (a) $\boldsymbol{M_i^F}$ *versus* $\boldsymbol{A_i^F}$ and (b) $\boldsymbol{M_i^U}$ *versus* $\boldsymbol{A_i^U}$. Both can be approximated by a bilinear scaling relation of the form $\boldsymbol{M} \sim \boldsymbol{C_n^U} \log \boldsymbol{A}$, where the slope transitions from $\boldsymbol{C_1^U} = 1.1$ for $\boldsymbol{M} \leq 7.7$ to $\boldsymbol{C_2^U} = 1.0$ for $\boldsymbol{M} > 7.7$. These lines fit the binned medians (dots centred in bins of 0.02 magnitude units). Lower panels show $\boldsymbol{A_i^U}$ *versus* $\boldsymbol{A_i^F}$ (c) and $\boldsymbol{M_i^U}$ *versus* $\boldsymbol{M_i^F}$ (d), where dashed lines represent equality.

4.3 Association algorithm

For a fixed value of η, we associate each RSQSim realization R_j with a unique UCERF3 rupture $U_{I(j)}$ by minimizing the differences between the subsection sets. The result is an index function $i = I(j)$ that maps the j^{th} RSQSim rupture to the i^{th} UCERF3 rupture. We construct this mapping by measuring the set differences by two non-negative integers: the number of subsections in R_j that are not in U_i, which we call the 'R-excess' and denote by r_{ij}, and the number of subsections in U_i that are not in R_j, which we call the 'U-excess' and denote by u_{ij}. Using set-theoretic notation, in which the differencing of one set from another is expressed by a backslash, the difference sets and excesses can be written,

$$R_j \backslash U_i := \{s \,:\, s\varepsilon R_j \wedge s \notin U_i\}\,,\; r_{ij} := \left|R_j \backslash U_i\right|\,, \tag{6a}$$

$$U_i \backslash R_j := \{s \,:\, s\varepsilon U_i \wedge s \notin R_j\}\,,\; u_{ij} := \left|U_i \backslash R_j\right|\,. \tag{6b}$$

(i) The U-excess measures the area of U_i not covered by R_j. An example is a long, single-fault RSQSim rupture in which the rupture area of one and only one subsection near the middle of the fault falls below the overlap threshold, $a_{sj}^{\mathrm{R}} < \eta a_s^{\mathrm{U}}$. No UCERF3 rupture contains such a gap, owing to the contiguity constraint, but there are UCERF3 ruptures that can completely span the

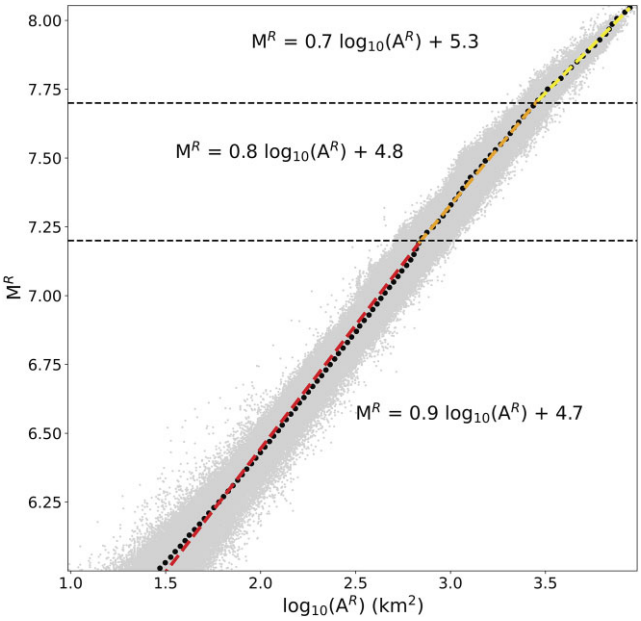

Figure 4. Magnitude-area scatter plot of the 547 563 RSQSim ruptures in the Shaw18 catalogue. The binned medians (dots centred in bins of 0.02 magnitude units) can be approximated by a trilinear scaling relation, $\boldsymbol{M}^R \sim \boldsymbol{C}_{\boldsymbol{n}}^R \log \boldsymbol{A}^R$. The slope decreases from $\boldsymbol{C}_1^R = 0.9$ at lower magnitudes to $\boldsymbol{C}_3^R = 0.7$ at higher magnitudes, as expected from stress-drop models (Shaw 2013). The magnitude range $7.2 \leq \boldsymbol{M}^R < 7.7$ is fit by an intermediate value $\boldsymbol{C}_2^R = 0.8$.

RSQSim rupture. Hence, the most similar UCERF3 rupture overlaps the entire RSQSim rupture but has one extra subsection: $r_{ij} = 0$ and $u_{ij} = 1$.

(ii) The R-excess measures the area of R_j not covered by U_i. An example is a two-fault rupture R_j that jumps across an 8-km fault gap and releases approximately equal moment on both sides of the gap. Because UCERF3 ruptures are not allowed to jump a gap greater than 5 km, the most similar UCERF3 ruptures can only cover about half of the RSQSim rupture ($r_{ij} \approx \frac{1}{2}|R_j|$), although they may be completely covered by the RSQSim rupture ($u_{ij} = 0$).

As these examples suggest, the U-excess is limited to small integer values by the general continuity of typical RSQSim ruptures, whereas the R-excesses can reach much larger rupture fractions, particularly for the RSQSim ruptures that fail the UCERF3 plausibility filter. Fig. 5 confirms this asymmetry. For $\eta = 0.2$ and $M_i^{\mathrm{F}} \geq 6.7$, we find $u_{ij} > 2$ in only 1.8 per cent of the associations, whereas $r_{ij} > 2$ in 7.5 per cent, and $r_{ij} > 10$ in 0.5 per cent.

We construct the association index function $I(j)$ using a three-step optimization algorithm:

1. For each R_j, we minimize the total excess $t_{ij} = r_{ij} + u_{ij}$.

(i) When the minimum is zero, the subsection sets are identical, $U_{I(j)} = R_j$. For $\eta = 0.2$, about 59 per cent of the associations involve identical ruptures.

(ii) For some ruptures, the index $I(j)$ is unique, but the subsection sets are not identical ($t_{I(j)} \geq 1$). About 21 per cent of the associations for $\eta = 0.2$ have non-zero excesses and are unique.

2. For others, the total-excess minimum is achieved by more than one UCERF3 rupture (28 per cent for $\eta = 0.2$), in which case we select those that minimize r_{ij} as well as t_{ij}. In 16 per cent of the cases, the R-excess minimum uniquely determines $I(j)$.

3. For the remaining 3.6 per cent, the multiple UCERF3 ruptures have the same values of r_{ij} and u_{ij}. In these cases, we choose $U_{I(j)}$ to maximize the rupture rate λ_i; that is, we associate the RSQSim realization with the most probable UCERF3 rupture in the qualifying set.

This sequence of optimizations ensures that the association index function $I(j)$ is complete; that is, every of the RSQSim realizations is associated with a unique UCERF3 rupture.

Reducing η allows more connectivity within an RSQSim rupture, which decreases the number of gaps and thus the U-excess, whereas increasing η reduces the small-area outliers, making the RSQSim ruptures more compact and decreasing the R-excess. This trade-off is not large, however, and we opt for a small overlap threshold, $\eta = 0.2$—the same value used by Shaw *et al.* (2018)—in order to better preserve the RSQSim magnitudes.

Shaw *et al.* (2018) limited their analysis to RSQSim realizations that had exact UCERF3 matches; that is, those satisfying criterion (1a). Requiring the total excess to be zero excludes 41 per cent of the realizations. In this study, we focus on large-magnitude ($M^{\mathrm{F}} \geq 6.7$) ruptures and allow ruptures with U-excesses as large as two and R-excesses as large as ten. These criteria include 97 per cent of all ruptures and 97.8 per cent of the large-magnitude ruptures (Fig. 5).

4.4 Area-magnitude relationships

Fig. 6(a) is a scatter plot of the RSQSim area A_j^{R} versus the unreduced UCERF3 area $A_{I(j)}^{\mathrm{F}}$ for ruptures associated using an overlap threshold of $\eta = 0.2$. Fig. 6(b) is the corresponding plot of M_j^{R} versus $M_{I(j)}^{\mathrm{F}}$. These comparisons ignore the UCERF3 fault-area reductions due to aseismic slip, which RSQSim does not attempt to model. The associations excluded from our analysis owing to large excesses ($u_{I(j)} > 2$, $r_{I(j)} > 10$) are highlighted in red. They show positive biases in area and magnitude that largely account for the upper tail of the joint magnitude distribution.

The median of A_j^{R} is offset from the median of $A_{I(j)}^{\mathrm{F}}$ by the ratio of 18 to 13 km, the value expected from the difference in rupture depths (Fig. 6a). The median of M_j^{R} is offset from the median of $M_{I(j)}^{\mathrm{F}}$ on a locus approximated by combining the rupture-depth difference with the multilinear magnitude–area relations in Figs 3(a) and 4.

5 ONTOLOGICAL TESTING

In the following sections on ontological testing (Section 5) and Bayesian updating (Section 6), we restrict our analysis to 'large' UCERF3 ruptures, defined to be those with unreduced magnitudes $M^{\mathrm{F}} \geq 6.7$. This target set has a size $N_{\mathrm{U}} = 245\,042$, comprising 96.9 per cent of all UCERF3 ruptures. Applying the association algorithm to this set maps the RSQSim realizations ($N_{\mathrm{R}} = 147\,065$) onto a UCERF3 subset of size $N_{\mathrm{I}} = 10\,696$. The index function $I(j)$ uniquely associates the j^{th} RSQSim rupture with the I^{th} UCERF3 rupture. At fixed I, summing $I(j)$ over all j yields the number of realizations associated with a particular UCERF3 rupture—the 'hit count' n_i for $T = 1$ My. By these definitions, $N_{\mathrm{R}} = \sum_{i=1}^{N_{\mathrm{U}}} n_i$.

5.1 Ontological null hypotheses

We consider two distinct expressions of the ontological null hypothesis. The first deals with model consistency on a rupture-by-rupture basis:

$\mathbb{O}_\lambda$: *RSQSim counts for individual rupture events* $\{n_i\}$ *are statistically consistent with rates predicted by UCERF3 model* $\{\lambda_i\}$*, given*

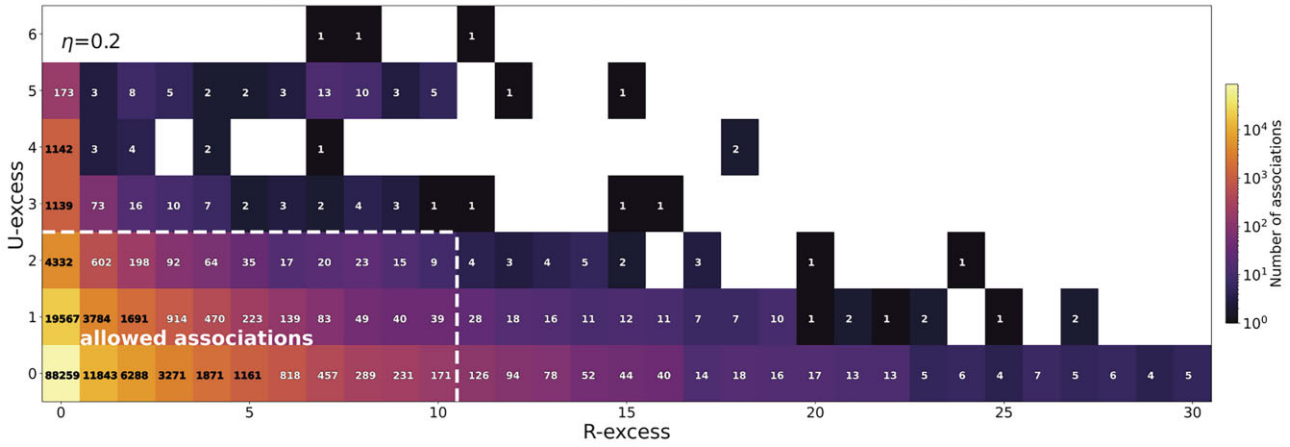

Figure 5. Rupture association counts for R-excess (horizontal axis) and U-excess (vertical axis) for $\eta = 0.2$ and $M_i^F \geq 6.7$. Associations with $u_{I(j)} \leq 2$ and $r_{I(j)} \leq 10$ are allowed (dashed white region). These criteria disallow 3400 of the 150 406 associations (2.3 per cent).

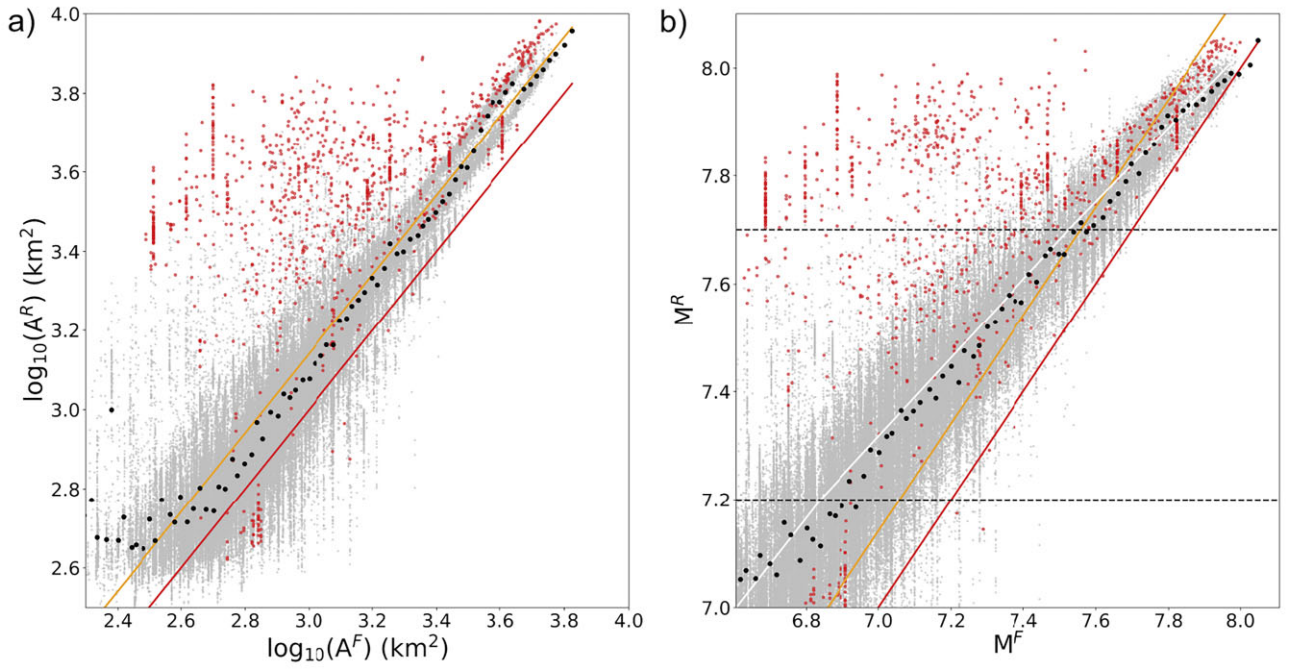

Figure 6. (a) Scatter plot of the true RSQSim area A_j^R versus unreduced UCERF3 area A_j^F for $M^F \geq 6.7$, $M^R \geq 7.0$ using an overlap threshold of $\eta = 0.2$. Lower line is $A_j^R = A_j^F$; upper line is $A_j^R = \frac{18}{13} A_j^F$, which accounts for the mean difference in rupture depths (18 km versus 13 km). (b) Scatter plot of the RSQSim magnitude M_j^R versus unreduced UCERF3 magnitude $M_j^F \geq 6.7$. Lower line is $M_j^R = M_j^F$; middle line is $M_j^R = M_j^F + \log \frac{18}{13}$. Upper line is the locus obtained by combining the bilinear magnitude-area relations in Fig. 3(a), the trilinear magnitude–area relations in Fig. 4, and the linear area–area relation in panel (a). In both panels, the dots for each rupture association are represented according to their R-excesses: grey for $r_{I(j)} = 0 - 10$, and red for $r_{I(j)} > 10$. A total of 150 406 rupture associations are displayed on each of the plots.

the sampling uncertainty of the Shaw18 catalogue and the epistemic uncertainty of UCERF3.

This hypothesis applies to the entire collection of large UCERF3 ruptures. We consider individual rupture rates to be features of the model that can be independently tested under a rupture-specific null hypothesis $\mathbb{O}_{\lambda_i}$. Our objective is not to accept or reject the collective null hypothesis $\mathbb{O}_\lambda = \{\mathbb{O}_{\lambda_i}\}$ per se, because we already know that models as complex as UCERF3 and RSQSim are very unlikely to agree on all rupture rates, even accounting the sampling and epistemic uncertainties. Instead, we seek to assess which specific features of the model are inconsistent with the data, that is, which subsets of $\mathbb{O}_\lambda$ can be rejected owing to the statistical discrepancies between UCERF3 and RSQSim. Our assessments borrow from the techniques of multiple-hypothesis testing developed for genomewide studies in biology, where thousands of features in genomewide data sets collected by DNA microarray technology are tested against a collective null hypothesis (Storey & Tibshirani 2003).

Constructing the tests and assessing the results require a number of statistical assumptions. The most salient is the *experimental concept* that underlies the statistical test of the ontological null hypothesis (Marzocchi & Jordan 2014). The experimental concept identifies one or more sets of observations that are judged to be exchangeable,

that is, have joint probability distributions that are invariant to any reordering of the data set. Exchangeability is the key to predictability, because it posits that the statistics of future observations will be the same as those of past observations. In this case, there are N_{U} exchangeability judgments, one for each target rupture. The data for testing $\mathbb{O}_{\lambda_i}$ are the number of i^{th}-rupture occurrences per unit time, and the experimental concept asserts that (1) this number is the output of a time-independent (stationary) Poisson process with a fixed but unknown rate $\hat{\lambda}_i$ and (2) the occurrence times of any two ruptures are independent (e.g. uncorrelated)—precisely the assumptions underlying the time-independent UCERF3 model (Field *et al.* 2014). In other words, we assume that the true rate is the limiting value of an arbitrarily long RSQSim catalogue, $\hat{\lambda}_i = \lim_{T\to\infty} n_i/T$.

Our goal is to identify as many significant discrepancies as possible, while incurring a low 'false discovery rate' (FDR), defined to be the proportion of false positives among all positives. We employ the q-values introduced by Storey (2003) for this purpose. The ability to distinguish true discrepancies in rupture rates is compromised by the sample sizes available from the Shaw18 catalogue. Only a small fraction (4.4 per cent) of the UCERF3 ruptures have one or more RSQSim associations. There are no associations ($n_i = 0$) for the vast majority of ruptures, and, in most of these cases, the predicted rates are so low that no associations are to be expected (Section 2). At such low rates, the significance tests perform poorly because the measures of statistical inconsistency (p-values) have essentially no power to discriminate against $\mathbb{O}_{\lambda_i}$. In the language of Mayo (2018), the tests of low-rate ruptures are not *severe* because they cannot ferret out discrepancies, were they present.

The sampling problems are ameliorated in the second test, which examines the model consistency on a subsection-by-subsection basis; that is, the tested features are the rates at which fault subsections participate in large-magnitude ruptures. Subsection participation rates are more useful and robust measures of the hazard than individual rupture rates, especially when communicating hazard levels to end-users (see e.g. the UCERF3 fact sheet by Field *et al.* 2015a). By definition, the UCERF3 participation rate of subsection s in ruptures of $M^{\mathrm{F}} \geq 6.7$ is the sum over the rupture rates (Field *et al.* 2014):

$$\mu_s = \sum_{i=1}^{N_{\mathrm{U}}} 1_{U_i}(s)\,\lambda_i,\ \mathrm{s} = 1, \cdots, N_s. \tag{7}$$

In this expression, $1_{U_i}(s)$ is the indicator function, which is unity if $s \varepsilon U_i$ and zero otherwise, and the sum is taken over all UCERF3 $M^{\mathrm{F}} \geq 6.7$ ruptures. Hence, μ_s is the rate at which subsection s participates in all large-magnitude UCERF3 ruptures, not just those associated with the RSQSim realizations. The experts' ensemble for each subsection can be written $\mathcal{E}(\mu_s) = \{\mu_s^{(l)}, w_l\}$, where $\mu_s^{(l)}$ is compiled from $\{\lambda_i^{(l)}\}$ according to eq. (7) and $\{w_l\}$ is the common set of weights used in all of the experts' ensembles. The RSQSim participation count associated with each subsection is,

$$m_s = \sum_{j=1}^{N_{\mathrm{R}}} 1_{U_{I(j)}}(s),\ \mathrm{s} = 1, \cdots, N_s. \tag{8}$$

The second null hypothesis is cast in terms of these participation measures:

$\mathbb{O}_{\mu}$:*The RSQSim participation counts for individual fault subsections* $\{m_s\}$ *are statistically consistent with the UCERF3 participation rates* $\{\mu_s\}$*, given the sampling uncertainty of the Shaw18 catalogue and the epistemic uncertainty of UCERF3.*

This version of the ontological null hypothesis is again collective, comprising many subhypotheses, in this case one for each of the 2606 subsections: $\mathbb{O}_{\mu} = \{\mathbb{O}_{\mu_s}\}$. The experimental concept for the test states that the number of large ruptures in which subsection s participates is governed by a time-independent Poisson process with a fixed but unknown rate $\hat{\mu}_s$, and the occurrence times of subsection ruptures are independent. The latter assumption is clearly untrue (see below), but we ignore such correlations in our statistical testing.

The UCERF3 participation rates μ_s and RSQSim participation counts m_s are larger than individual rupture rates λ_i and counts n_i, respectively, so that the tests of $\mathbb{O}_{\mu_s}$ are more severe than those of $\mathbb{O}_{\lambda_i}$. Only 5.3 per cent (139) of the subsections have zero counts, compared to 95.6 per cent of the ruptures. The mean count for the sampled UCERF3 ruptures, is $\bar{n} = N_{\mathrm{R}}/N_{\mathrm{I}} = 14$, whereas the mean count for the sampled subsections is $\bar{m} = N_{\mathrm{R}}/(N_{\mathrm{S}} - 139) = 60$. In both cases, the count distribution is strongly skewed to higher values ($\max n_i = 1864$; $\max m_s = 8552$).

5.2 Extended experts' distribution

UCERF3 is a complete probabilistic forecasting model that characterizes the aleatory variability of the forecast by a set of mean rupture rates $\{\bar{\lambda}_i\}$ and represents the epistemic uncertainty of the forecast by the experts' ensembles, $\mathcal{E}(\lambda_i)$. From this model, we can compute the mean subsection participation rates $\{\bar{\mu}_s\}$ and the corresponding experts' ensembles $\mathcal{E}(\mu_s)$. In general, the rupture and subsection participation rates are co-dependent random variables that have a complex covariance structure imposed by the inversion algorithm and the logic-tree structure (Field *et al.* 2014; Baker *et al.* 2021). The rates of two ruptures are likely to be positively correlated, for example, if they share a large fraction of fault subsections, whereas negative correlations are imposed on parallel ruptures by the balancing of seismic moment release across the California plate-boundary fault system. Similarly, the participation rates of two nearby subsections on the same fault will be positively correlated owing to the ruptures they share.

In this exploratory study, we will ignore these co-dependencies in the ontological testing and cast the experts' ensembles as the marginal distributions $\mathcal{E}(\lambda_i)$ and $\mathcal{E}(\mu_s)$. Assuming the rates are independently distributed is convenient both theoretically and computationally, because it allows us to test for RSQSim/UCERF3 consistency one rupture at a time using analytical models. Accounting for model correlations is a difficult problem that we have not yet addressed.

$\mathcal{E}(\lambda_i)$ is represented by the 720 weighted rate values $\lambda_i^{(l)}$ given by the UCERF3 logic tree (eq. 1). In PSHA, $\{\lambda_i^{(l)}\}$ is often considered to be a mutually exclusive and collectively exhaustive (MECE) partition of the sample space (Bommer & Scherbaum 2008), and the weightings $\{w_l\}$ are interpreted to be a probability mass function (PMF) on this partition. However, the random variable λ_i has a continuous sample space, so the probability that it will equal any value in the finite set $\{\lambda_i^{(l)}\}$ is zero. This discretized description of the epistemic uncertainty has led some practitioners to interpret w_l as the probability that the single-branch rate $\lambda_i^{(l)}$ is the 'one that should be used' (Scherbaum & Kuhen 2011) or 'the best of those available' (Musson 2012).

Marzocchi & Jordan's (2014, 2018) probabilistic framework, adopted here, removes the unrealistic MECE requirement by interpreting $\mathcal{E}$ as a weighted ensemble of samples from an *extended experts' distribution* (EED), which we take to be continuous on

$[0, \infty)$. In other words, we consider the experts' ensemble to be a weighted set of likelihoods rather than a set of probabilities, and we use these likelihoods to infer the parameters of the continuous EED.

An appropriate two-parameter form of the EED for both types of rate variables is the gamma distribution, $\lambda_i \sim \text{Gamma}(\alpha_i, \beta_i)$, $\mu_s \sim \text{Gamma}(\alpha_s, \beta_s)$. The probability density function (PDF) is

$$p_{\text{gamma}}(\lambda \mid \alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}, \ 0 < \lambda \langle \infty, \ \alpha, \beta \rangle 0. \tag{9}$$

The gamma distribution can be reparameterized in terms of its mean $\bar{\lambda} = \alpha/\beta$ and variance $\sigma^2 = \alpha/\beta^2$ or, equivalently, its mean and coefficient of variation,

$$c := \frac{\sigma}{\bar{\lambda}} = \frac{1}{\sqrt{\alpha}}. \tag{10}$$

We obtain the parameters $\bar{\lambda}_i$ and c_i by maximizing the log-likelihood summed over the weighted logic-tree values for the i^{th} rupture. The maximum-likelihood estimate (MLE) of $\bar{\lambda}_i$ equals the mean rupture rate of the discrete experts' ensemble (eq. 2) and thus preserves the mean hazard, which is the hazard measure most used in PSHA practice (Baker *et al.* 2021). The MLEs of the participation rate parameters, μ_s and c_s, are computed in the same way.

Examples of the EED fitting procedure are illustrated in Fig. 7 for representative sets of UCERF3 ruptures (panel a) and subsections (panel b). In each example, the blue and grey histograms are the experts' ensembles, and the red curve is the PDF obtained by fitting a gamma distribution to this set of samples. For comparison, we also fit the experts' ensemble by a log-normal distribution (dashed black curve). The log-normal fit is marginally better than the gamma fit—for the rupture rates, the likelihood ratios range from 0.96 to 1.10 with an average about 1.03—but the cumulative distributions for both models show significant under-dispersion relative to the experts' ensembles, primarily owing to outliers in latter.

Employing gamma or log-normal EEDs thus tends to reduce the range of the epistemic uncertainties represented by the UCERF3 experts' ensembles. Whether this distributional narrowing is good or bad depends on whether the outliers of the experts' ensemble truly indicate the uncertainty range or are statistically spurious, for example, due to implausible combinations of UCERF3 alternative models.

The use of gamma distributions in this exploratory study is practically motivated by two theoretical advantages:

(i) To conduct the ontological tests, we compute an EED-weighted mixture of Poisson distributions. If the EED is represented by a gamma distribution, the mixture has a simple analytic form—a negative binomial distribution.
(ii) To assimilate the RSQSim data, we recalibrate the UCERF3 prior forecast with the RSQSim data using standard Bayesian procedures. Because the gamma distribution is conjugate to the Poisson distribution, the posterior forecast is also gamma distributed, so that the Bayesian updating can also be done analytically.

The log-normal model may have advantages because it can represent the correlations among rates in terms of multivariate Gaussian distributions, but its use in testing and updating requires heavy numerical calculations that are not attempted here.

5.3 Test distribution

To test the ontological null hypotheses $\mathbb{O}_\lambda$ and $\mathbb{O}_\mu$ against observed data—or, in our case, the set of model-generated counts $\{n_i\}$ and $\{m_s\}$—we construct test (null) distributions that account for the epistemic uncertainty in the ERF model as well as the finiteness of the ERS sample. The test distributions for $\mathbb{O}_\lambda$ and $\mathbb{O}_\mu$ have the same form.

According to a time-independent ERF, the probability that k realizations of the rupture with rate λ will occur during a catalogue interval T is given by the Poisson distribution, $k \sim \text{Poisson}(\lambda)$, which has the PMF,

$$p_{\text{poisson}}(k \mid \lambda) = \frac{(\lambda T)^k e^{-\lambda T}}{k!}, \ k = 0, 1, \cdots, \ 0 < \lambda < \infty. \tag{11}$$

This sampling variability is conditional on knowing the rate λ. We can account for the epistemic uncertainty in λ by averaging the Poisson sampling uncertainty over the EED (Marzocchi & Jordan 2014). If we choose the EED to be a gamma distribution, the requisite Poisson mixture is a negative binomial (Polya) distribution, $k \sim \text{NB}(\alpha, \beta)$, which has the analytic PMF,

$$\begin{aligned} p_{\text{test}}(k \mid \alpha, \beta) &= \int_0^\infty p_{\text{poisson}}(k \mid \lambda) \ p_{\text{gamma}}(\lambda \mid \alpha, \beta) \ d\lambda \\ &= \frac{\Gamma(k+\alpha)}{k!\Gamma(\alpha)} \left(\frac{1}{\beta+1}\right)^k \left(\frac{\beta}{\beta+1}\right)^{\alpha}. \end{aligned} \tag{12}$$

We have simplified this expression by setting the unit of time to be 1 My, so that the Shaw18 catalogue interval is $T = 1$. The Negative-Binomial (NB) distribution has the same mean as the gamma distribution, $\alpha/\beta = \bar{\lambda}$, and its variance is related to the mean by $\bar{\lambda}(1 + 1/\beta)$. The over-dispersion relative to Poisson reflects the variance increase due to the epistemic uncertainty in the rate. For our purposes, it will be useful to reparametrize the NB distribution in terms of its mean $\bar{\lambda} = \alpha/\beta$ and coefficient of variation $c = \alpha^{-1/2}$:

$$p_{\text{test}}(k \mid \bar{\lambda}, c) = \frac{\Gamma\left(k + c^{-2}\right)}{k!\Gamma\left(c^{-2}\right)} \left(\frac{c^2\bar{\lambda}}{c^2\bar{\lambda} + 1}\right)^k \left(\frac{1}{c^2\bar{\lambda} + 1}\right)^{c_i^{-2}}. \tag{13}$$

We could further refine our statistical analysis by integrating over the estimation uncertainties in $\bar{\lambda}$ and c; for example, by treating them as nuisance parameters in a Bayesian hierarchical model. However, in our exploratory analysis, we ignore this second-order uncertainty and adopt eq. (13) as the test distribution. Dropping the dependence on the parameters, the cumulative distribution function (CDF) can be written,

$$P_{\text{test}}(k) = \sum_{k'=0}^{k} p_{\text{test}}(k'). \tag{14}$$

The test distribution PMFs for the nine UCERF3 ruptures in Fig. 7 are plotted in Fig. 8.

5.4 *P*-values

For each rupture or fault subsection, we can test the null hypothesis against one of three alternative hypotheses about how the data might differ from the model (Table 1). For each test, we can measure the consistency of the RSQSim datum, n or m, with the UCERF3 EED by a p-value, which is defined to be the probability of outcomes at least as extreme as the observation assuming the null hypothesis is true.

Expressions for the three types of p-value are listed in Table 1. If the test distribution were continuous, then $p^{(\text{left})}$ would equal the null CDF (eq. 14) evaluated at $k = n$, and $p^{(\text{right})}$ would equal the corresponding survival distribution function (SDF), SDF = 1–CDF. Because the test distribution is discrete, however, this specification overweights the left tail. Consider, for example, a rupture whose model rate is so low that no counts are expected ($p_{\text{test}}(0) \approx 1$) and

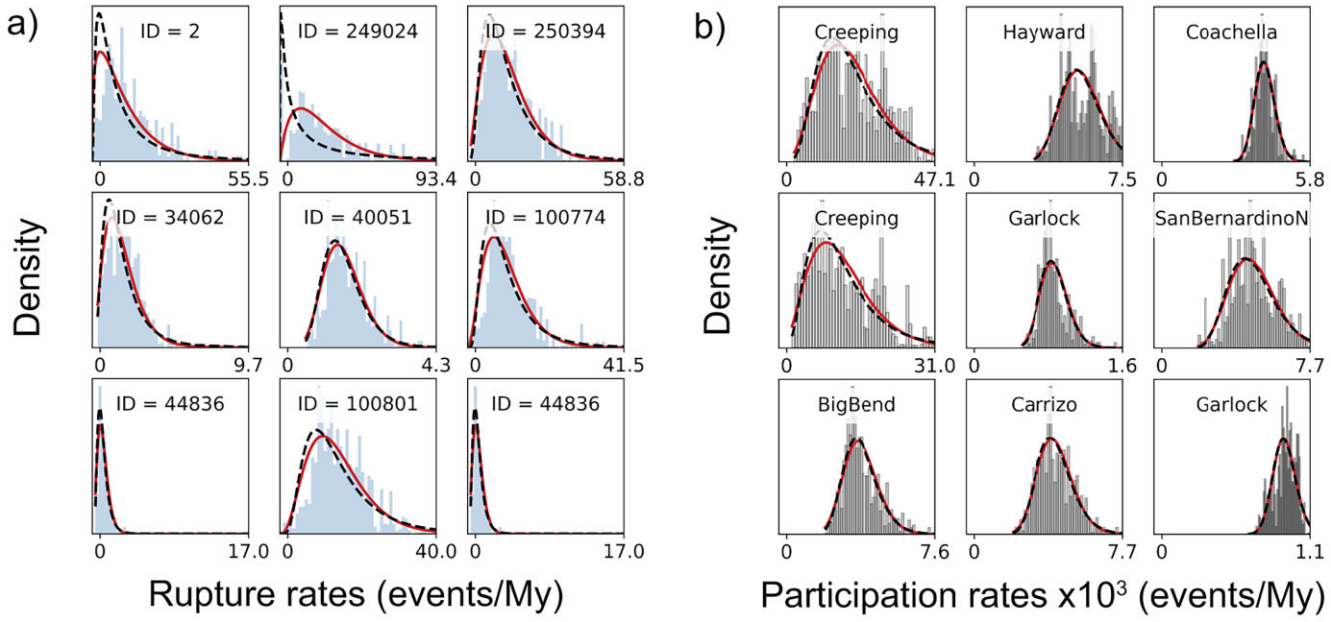

Figure 7. Representative examples of continuous EEDs obtained by fitting the discrete experts' ensemble from the UCERF3 logic tree. Panel (a) shows nine examples of individual ruptures, identified by their UCERF3 rupture numbers (histograms). Panel (b) shows nine examples of individual subsections, identified by parent fault sections (histograms). Solid lines are gamma PDFs; dashed lines are log-normal PDFs. The maximum-likelihood fitting procedures assumed the experts' ensembles consist of statistically independent samples.

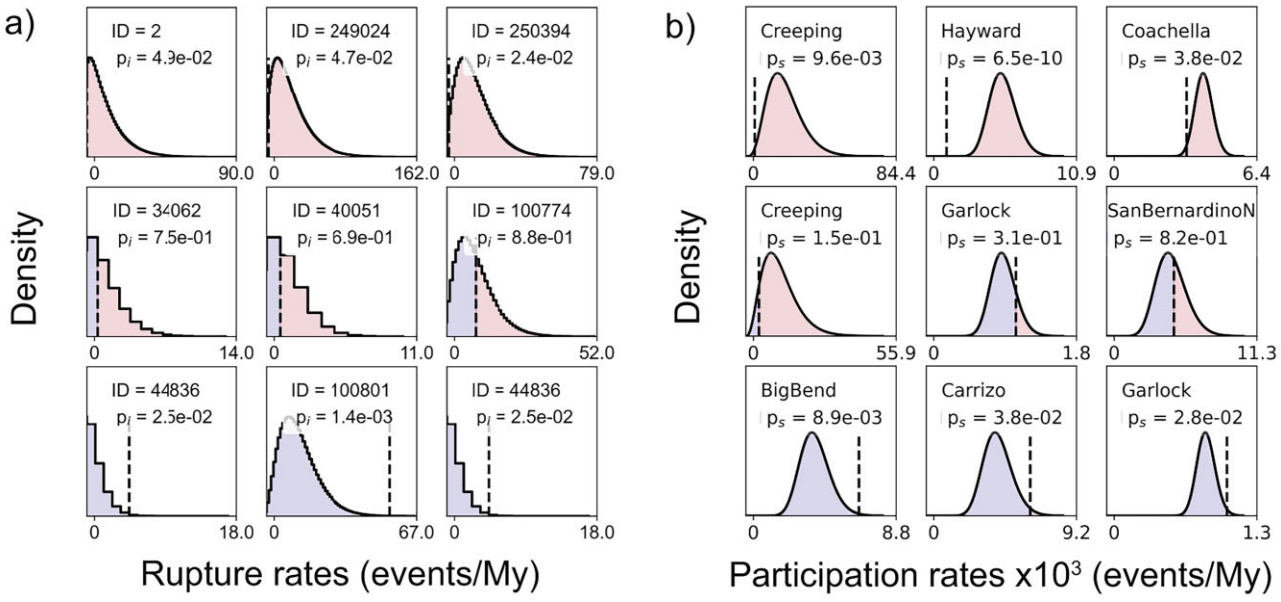

Figure 8. Negative binomial test distributions (black) for the nine examples shown in Fig. 7. The vertical, dashed black lines show $\boldsymbol{n_i}$ or $\boldsymbol{m_s}$, the number of RSQSim associations for the ruptures of panel (a) or the subsections of panel (b). For example, the UCERF3 rupture number 100 774 was associated with 10 RSQSim ruptures, which yields a p-value of 0.878.

Table 1. p-value definitions for testing alternate hypotheses.

Alternate hypothesis	Test type	p-value
RSQSim rate < UCERF3 rate	left-sided	$p^{(\text{left})} = P_{\text{test}}(n) - \frac{1}{2}p_{\text{test}}(n)$
RSQSim rate > UCERF3 rate	right-sided	$p^{(\text{right})} = 1 - P_{\text{test}}(n) + \frac{1}{2}p_{\text{test}}(n)$
RSQSim rate ≠ UCERF3 rate	two-sided	$p^{(\text{either})} = 2\min[p^{(\text{left})},\ p^{(\text{right})}]$

no counts are observed ($n = 0$). The CDF evaluated at n is close to unity and the corresponding SDF is close to zero, whereas the lack of information about the relative rates requires the left-side and right-side p-values to be approximately equal. This type of discretization bias can be reduced by making a 'continuity correction' that subtracts half the value of the probability $p_{\text{test}}(n)$ from the CDF and adds half to the SDF (Table 1) (Berry & Armitage 1995).

We are interested in assessing the possibility that some RSQSim rates might be significantly lower than the UCERF3 rates while others might be higher. The appropriate test is therefore two-sided, allowing for either possibility, and the p-value used in the test is $p^{(\text{either})}$ of Table 1, hereafter abbreviated p (subscripted with i or s as appropriate). When we compute p, we keep track of which is

smaller, $p^{(\mathrm{left})}$ or $p^{(\mathrm{right})}$, allowing us to classify the testing failures as left-sided (RSQSim rate < UCERF3 rate; denoted $R < U$) or right-sided (RSQSim rate > UCERF3 rate; denoted $R > U$), respectively.

5.5 Significance testing

We employ the standardized protocol for significance testing: an individual null hypothesis represented by $P_{\mathrm{test}}(k)$ flunks the test if $p \le \alpha$, where p is the observed p-value and α is a prescribed significance level. The choice of α sets the fraction of failures (positives) expected when the null hypothesis is true—the Type I (false positive) error rate. Examples of the tests are presented in Fig. 8, and the results are summarized for $\alpha = 0.05$ & 0.01 in Tables 2 and 3.

5.5.1 Rupture rates

To evaluate the testing results across the entire rupture set, we examine the distribution of p-values (Fig. 9). If the test distribution were continuous (as it becomes in the large-rate limit), and if $\mathbb{O}_\lambda$ were true, then the random variables $\{p_i\}$ would be uniformly distributed on (0,1). A nearly uniform p-value would thus imply that the UCERF3 forecast is well calibrated against the RSQSim data (Gneiting & Katzfuss 2014).

The distribution of p-values for the entire set of UCERF3 ruptures is far from uniform, however (Fig. 9a). The peak at small p_i can be ascribed to the discrepant ruptures—those less than α are the ones we formally reject—but the swooping increase in rupture counts for $p_i > 0.2$ is entirely due to the discreteness of the counting statistics. If $n_i = 0$, the smallest p-value attained by the test is the zero-count probability,

$$p_{\mathrm{test}}(0) = \left(\frac{1}{c_i^2 \bar{\lambda}_i T + 1}\right)^{c_i^{-2}}. \tag{15}$$

At $\bar{\lambda}_i T = 1$, the EED coefficients of variation cover the range $0.3 \le c_i \le 3$, implying $0.4 \le p_{\mathrm{test}}(0) \le 0.8$. Therefore, ruptures with low rates and zero counts will never fail the test (e.g. they can never be smaller than 0.05), which biases p_i to high values. This bias can be seen in Fig. 9(a), where the p-value density for the entire UCERF3 rupture set rises by almost three orders of magnitude from $p_i = 0.2$ to 1. All of the rise is due to the increase in zero-count p-values (black bars).

To reduce this bias, we excluded from the testing any UCERF3 ruptures with mean rates less than 10^{-6} yr^{-1} ($\bar{\lambda}_i T < 1$); that is, we only tested ruptures expected to have one or more realizations in a million-year catalogue. Imposing this rate cut-off eliminates 95 per cent of the ruptures, leaving a reduced set of size $N'_{\mathrm{U}} = 11\,473$ (Table 2). Among these, 4060 are associated with one or more in the set of RSQSim realizations, which is reduced in size to $N'_{\mathrm{R}} = 97,634$. This cut-off removed 97 per cent of the zero-count samples, including nearly all with p-values above 0.8, eliminating the peak at $p_i \approx 1$ (Fig. 9b). However, a bulge associated with the zero-count samples still dominates for $p_i < 0.8$.

Applying the test to the reduced set of UCERF3 ruptures with $\bar{\lambda}_i T \ge 1$, we found that 688 (6.0 per cent) fail the test at the 5 per cent significance level and 386 (3.4 per cent) at the 1 per cent level. A sizeable fraction of these failures can be attributed to false positives, indicating that there is little evidence against $\mathbb{O}_\lambda$ from the Shaw18 catalogue. The weakness of the test is further elucidated by analyses of the *a priori* power in Appendix A and the *a posteriori* false discovery rate (FDR) in Appendix B.

5.5.2 Participation rates

To test $\mathbb{O}_\mu$, we computed $\mathcal{E}(\mu_s)$ from the experts' ensembles of the large-magnitude ruptures ($M^{\mathrm{F}} \ge 6.7$) from eq. (7), and we counted the RSQSim hits m_s according to eq. (8). From m_s and the test distribution, we calculated the continuity-corrected p-value, $p_s = p_s^{(\mathrm{either})}$, for the two-sided test of Table 1.

Tests of the subsection participation rates show much more statistical power in distinguishing rate variations than tests of individual ruptures (Appendix A), and the UCERF3/RSQSim discrepancies uncovered by the tests are much more evident. The p-value histogram for the 2606 subsections is relatively uniform for $p_s > 0.6$, but the counts increase as p_s become smaller, spiking up in the lowest one percentile (Fig. 10). There are 643 failures for $p_s \le 0.05$ and 406 for $p_s \le 0.01$ (Table 3), which is much larger than the expected number of Type I errors (130 and 26, respectively).

Another error statistic to be considered is the FDR, which can be measured by the q-value of Storey & Tibshirani (2003). We define and compute the q-values of the subsection participation tests in Appendix B. The q-value at $\alpha = 0.05$ is 0.114, so that our estimate of the expected number of false discoveries is only 11.4 per cent of 643, or about 73, only about half of the number of false positives. At $\alpha = 0.01$, the q-value decreases to 0.036, yielding an estimated FDR of 15 out of 406. Hence, the FDR is well controlled, and we can be confident that most of the failures are correct rejections; that is, the same subsections would likely fail if tested against an independent million-year catalogue generated by the same RSQSim model.

At both significance levels, about 70 per cent of the failures are left-sided, indicating that the RSQSim empirical rates are substantially lower than the UCERF3 model rates ('$R < U$ failures'). This can be seen in Fig. 11(a), which displays the $\alpha = 0.05$ test results for the subsections of the UCERF3 fault model; the $R < U$ failures (in blue) outnumber the $R > U$ failures (in red). North of the San Francisco Bay Area, all but a few of the failures are $R < U$, including those along the Hayward–Rodgers Creek–Maacama and Calaveras–Green Valley–Bartlett Springs fault systems. The subsections south of the Bay Area show a more equal mixture of the two types. $R > U$ failures are more common in central California west of the San Andreas fault (SAF), where they concentrate on a system of faulting that includes the northern San Gregorio, Zayante–Vergeles, Reliz, Rinconada, South Cuyama and East Huasna faults. $R < U$ failures predominate among the thrust faults of the Western Transverse Ranges (e.g. Ventura, Pitas Point, Red Mountain faults) and along the North Frontal Thrust of the Eastern Transverse Ranges. Failures east of the SAF are largely $R < U$, although the western Garlock fault, showing $R > U$ failures, is a notable exception.

To isolate the distribution of failures on subsections with high rupture rates (a proxy for seismic hazard), we re-indexed the subsections in descending order of their UCERF3 rupture rates, $\mu_{[1]} \ge \mu_{[2]} \ge \cdots \ge \mu_{[2606]}$, and we defined a fractional cumulate $\Sigma_{[s]}$ by summing the rates up to $[s]$ and normalizing by the total sum:

$$\Sigma_{[s]} = \sum_{1}^{[s]} \mu_{[t]} \Big/ \sum_{1}^{2606} \mu_{[t]}, \tag{16}$$

Fig. 12 compares this cumulative rate with the cumulative fractions of $R < U$ and $R > U$ failures as a function of the reordered index $[s]$. For $[s] < 2000$, the number of test failures scales in rough proportion to the number of subsections; that is, the relative failure rates (the chord slopes of the red and blue lines in Fig. 12) do

Table 2. Count statistics for $\mathbb{O}_\lambda$ significance testing.

				$p_i \leq 0.05$		$p_i \leq 0.01$	
Rate minimum	UCERF3 ruptures	RSQSim associations	Zero-count ruptures	$R < U$ failures	$R > U$ failures	$R < U$ failures	$R > U$ failures
10^{-5} My^{-1}	245 042	147 065	234 346	4291	219	2890	92
1 My^{-1}	11 473	97 634	7413	469	219	294	92

Table 3. Count statistics for $\mathbb{O}_\mu$ significance testing.

				$p_s \leq 0.05$		$p_s \leq 0.01$	
Rate minimum	UCERF3 subsections	RSQSim associations	Zero-count subsections	Left-side failures	Right-side failures	Left-side failures	Right-side failures
2.2 My^{-1}	2606	147 065	139	468	175	286	120

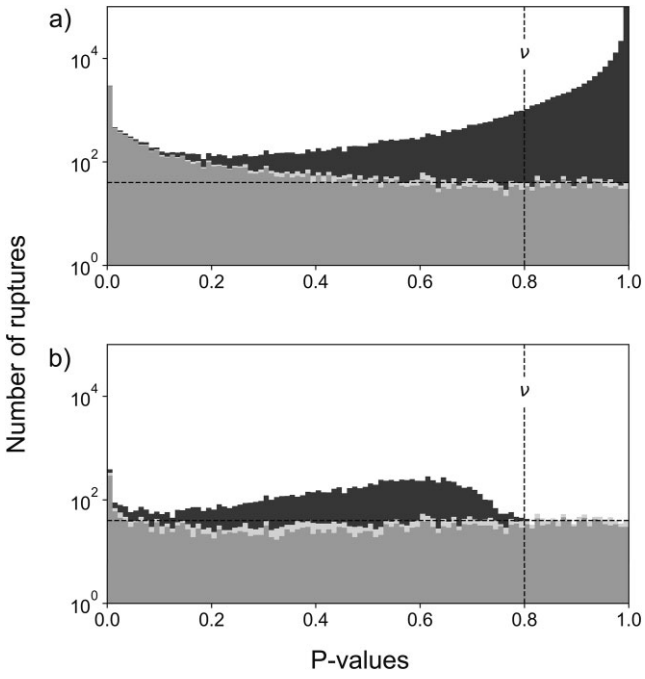

Figure 9. Histograms of p-values for rupture rates. (a) All UCERF3 ruptures. (b) UCERF3 ruptures with mean recurrence intervals less than one million years ($\bar{\lambda}_i < 10^{-6}$ yr^{-1}). Black bars contain ruptures with zero RSQSim counts ($n_i = 0$). Light grey bars contain the $R < U$ (left-side) p-values with non-zero counts ($n_i > 0$). Dark grey bars contain the $R > U$ (right-side) p-values with non-zero counts. Vertical dashed line marks $\nu = 0.8$, which is the p-value threshold we used to determine $\tilde{N}_0(\nu)$ in calculating q-values.

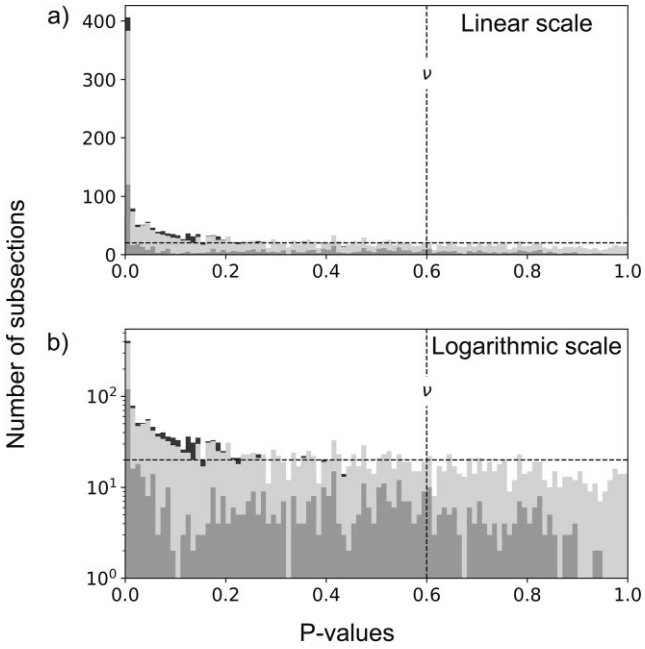

Figure 10. Histograms of p-values for subsection participation rates: (a) linear scale and (b) logarithmic scale. Participation rates are for UCERF3 ruptures with $M_i^{\mathrm{F}} \geq 6.7$. Light grey bars contain the $R < U$ (left-side) p-values with non-zero counts ($n_i > 0$). Dark grey bars contain the $R > U$ (right-side) p-values with non-zero counts. Vertical dashed line marks $\nu = 0.6$, which is the p-value threshold we used to determine $\tilde{N}_0(\nu)$ in calculating q-values.

not vary much as $\Sigma_{[s]}$ increases. The failure fraction of high-rate subsections is about the same rate as low-rate subsections.

The cumulative rate distribution $\Sigma_{[s]}$ rises rapidly from zero in a concave curve that flattens as $[s]$ approaches 2606 (Fig. 12). The 223 highest-rate subsections account for half the cumulative rate ($\Sigma_{[223]} = 50$ per cent), and only 457 are needed to reach 70 per cent. This latter group of fault subsections defines a high-rate 'San Andreas spine' that comprises the SAF proper plus the Hayward–Rodgers Creek–Maacama and Calaveras–Green Valley–Bartlett Springs fault systems in northern California and the San Jacinto system, Brawley seismic zone, and Imperial fault in southern California. The SAF spine is annotated with test failures in Fig. 13. The failures, 156 in total, are concentrated on the two fault systems east of the northern SAF, on the Creeping Section, and on the southernmost SAF (Fig. 13); 147 are $R < U$ failures and 11 are $R > U$ failures, similar to the bias seen for the model as a whole (Fig. 11).

All failures on the Creeping Section are $R < U$. In this case, and in many others, the discrepancies can be attributed to the differences in the slip rates assumed to drive the models, as illustrated by the rate comparisons for the SAF in Fig. 14. The RSQSim achieved rates are, on average, 46 per cent lower than the achieved UCERF3 rates. As noted by Field *et al.* (2014), the UCERF3 inversions give models with achieved rates twice as large as the target slip rates, whereas the achieved slip rates for RSQSim are more in line with the slip rate reduction imposed by the creep constraints.

Other examples where $R < U$ failures can be attributed to lower RSQSim slip rates include the Brawley seismic zone, Southern San Bernardino segment of the SAF, the offshore Carlsbad fault, the Compton thrust beneath Los Angeles, the Green Valley and Bartlett Springs faults in northern California, and the Red Mountain,

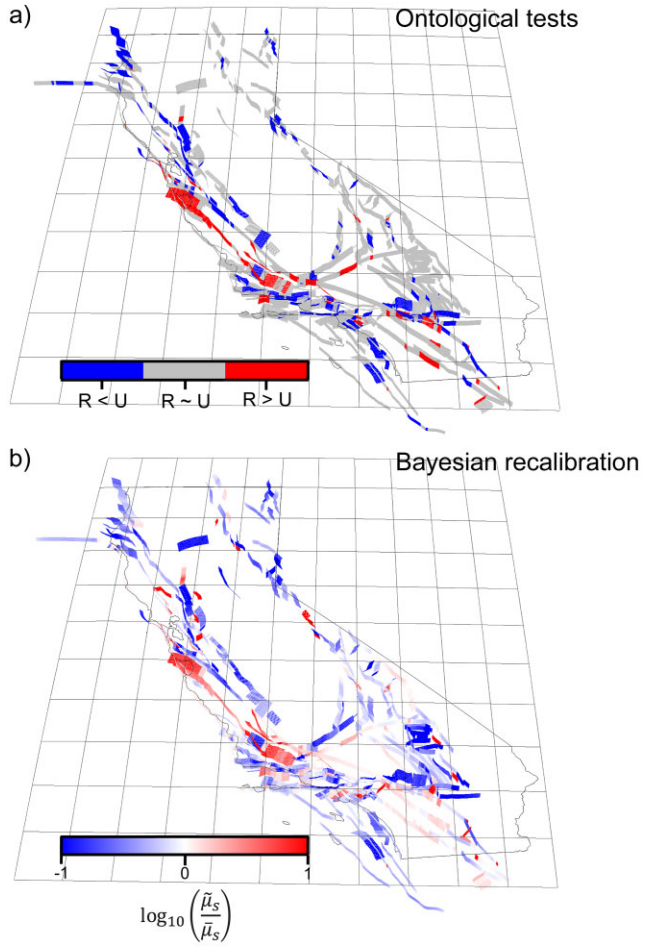

Figure 11. (a) Map of the UCERF3 fault system (FM3.1) showing subsections that fail the ontological test at $\boldsymbol{\alpha} = 0.05$. Blue panels mark $R < U$ (left-side) failures; red panels mark $R > U$ (right-side) failures. (b) Map showing changes in subsection participation rates derived by Bayesian recalibration. Deep blue panels are subsections where the recalibrated rate is a factor of 10 or more smaller than the UCERF3 rate; bright red panels are where the recalibrated rate is a factor of 10 or more larger than the UCERF3 rate.

Ventura-Pitas Point, and Malibu faults of the western Transverse Ranges.

Examples where $R > U$ failures can be attributed to higher RSQSim slip rates include Pilarcitos and Silver Creek faults in the eastern the Bay Area, the San Gregorio, Zayante-Vergeles, Reliz, Rinconada, South Cuyama, and East Huasna system in central California, and the Northern Elsinore, Southern Elysian Park, San Gabriel and Northern San Jacinto faults in southern California.

There are also discrepancies that are not explained solely by slip-rate differences. The Hayward and Calaveras faults show $R < U$ failures, even though the model slip rates are similar. Large ruptures simulated by RSQSim tend to be more frequent on the westwards side of the Bay Area than predicted by UCERF3. A second interesting discrepancy involves the Big Bend and Mojave sections of the SAF near its intersection with the Garlock fault, where again the slip rates are comparable, but the RSQSim participation rates are more frequent. The UCERF3 rupture plausibility filter disallowed ruptures propagating through fault junctures with an azimuth change of more than 60°, but an exception was made at the SAF-Garlock juncture, where RSQSim simulations showed a preference

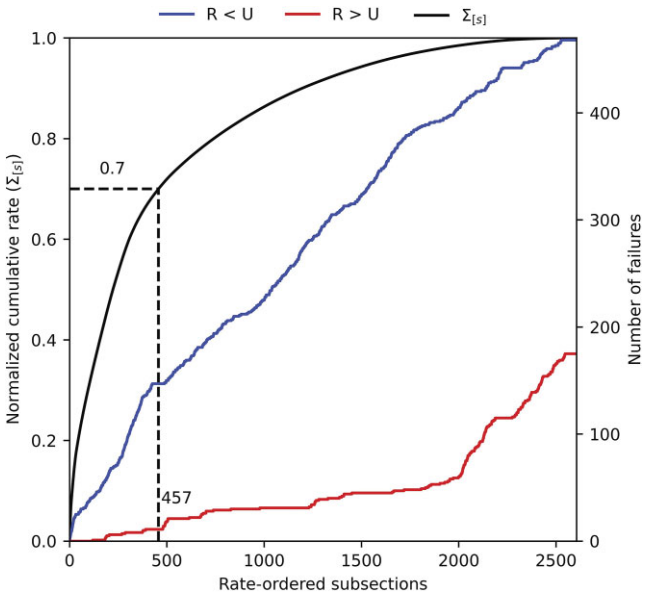

Figure 12. Red and blue lines show the cumulative number of rate-ordered subsections with $R > U$ and $R < U$ failures, respectively (right-side axis), plotted against rate-ordered subsection number. Black line is the normalized cumulative rate $\Sigma_{[s]}$ (left-side axis). The SAF spine is defined to be the 457 subsections that account for 70 per cent of total cumulative participation rate (dashed lines).

of Garlock co-rupturing with the North Mojave section (132° azimuth change) rather than with the Big Bend section (40° azimuth change) (Field *et al.* 2014). Despite this modification, RSQSim generates ruptures through this juncture at higher rates than the UCERF3 estimate.

6 BAYESIAN RECALIBRATION

The significance testing of Section 5 relies on frequentist error statistics, but a Bayesian methodology is more suitable for assimilating rupture-count data into complete probabilistic forecasting models, which almost always involve subjective probabilities. Combining frequentist methods of model testing with Bayesian methods of model development—a practice Gelman & Hennig (2017) have dubbed 'falsificationist Bayesianism'—is a key feature of the uncertainty assessment framework proposed by Marzocchi & Jordan (2014, 2017). As we show here, the results of the two statistical approaches are consistent and complimentary.

Because the tests of individual rupture rates show little power and high FDR, we focus on the Bayesian recalibration of the subsection participation probabilities, for which Bayes's rule can be written,

$$p\left(\mu_s|m_s\right) = \frac{p\left(m_s|\mu_s\right) p\left(\mu_s\right)}{p\left(m_s\right)} \tag{16}$$

The prior distribution is $p(\mu_s)$, the PDF of the subsection EED, which we take to be the same gamma distribution used in the significance testing. The likelihood function $p(m_s|\mu_s)$ is the Poisson counting distribution. Because the Poisson distribution is conjugate to the gamma distribution (George *et al.* 1993; Gelman *et al.* 2013), the integrations can be done analytically, and the solution is also gamma distributed,

$$p\left(\mu_s|m_s\right) = \frac{\left(1+\beta_s\right)^{\alpha_s+m_s}}{\Gamma\left(\alpha_s+m_s\right)} \mu_s^{\alpha_s+m_s-1} e^{-(1+\beta_s)\mu_s}. \tag{17}$$

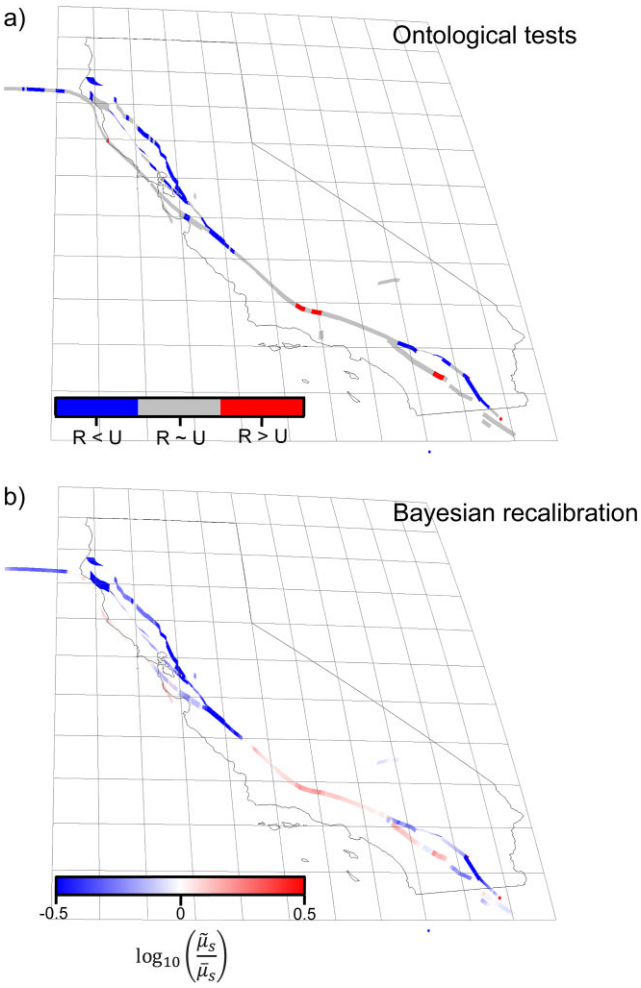

Figure 13. (a) Map of the SAF spine showing subsections that fail the ontological test at $\boldsymbol{\alpha} = 0.05$. The spine is defined as the 457 highest-rate subsections that account for 70 per cent of the cumulative rate (Fig. 12). Blue panels mark the 147 $R < U$ (left-side) failures; red panels mark the 11 $R > U$ (right-side) failures. (b) Map of the spine showing changes in subsection participation rates derived by Bayesian recalibration. Deep blue panels are subsections where the recalibrated rate is a factor of $\log 0.5 \approx 3.2$ or more smaller than the UCERF3 rate; bright red panels are where the recalibrated rate is a factor of 3.2 or more larger than the UCERF3 rate (a different colour scale than in Fig. 11). Viewing angles of the maps are different from those of Fig. 11.

The Bayesian recalibration can thus be written,

$$\mu_s|m_s \sim \text{Gamma}\,(\mu_s; \alpha_s + m_s, \beta_s + 1)\,. \tag{18}$$

The expected value of the recalibrated rate is

$$\tilde{\mu}_s := \text{E}\,[\mu_s|m_s] = \frac{m_s + \alpha_s}{1 + \beta_s} = \left(\frac{1}{1+\beta_s}\right) m_s + \left(\frac{\beta_s}{1+\beta_s}\right) \bar{\mu}_s. \tag{19}$$

where $\bar{\mu}_s$ is the prior mean and the catalogue duration is $T = 1$. For an arbitrary catalogue duration, $\tilde{\mu}_s = (m_s + \alpha_s)/(T + \beta_s)$. The parameter β_s, which has the dimension of time, governs the relative weighting of the counts and prior mean and can be interpreted as the catalogue pseudo-count. If $\beta_s = T$, the information from the RSQSim count data equals the prior information from the UCERF3 forecast. In the limits of large and small β_s, the posterior mean goes to $\bar{\mu}_s$ and m_s/T, respectively.

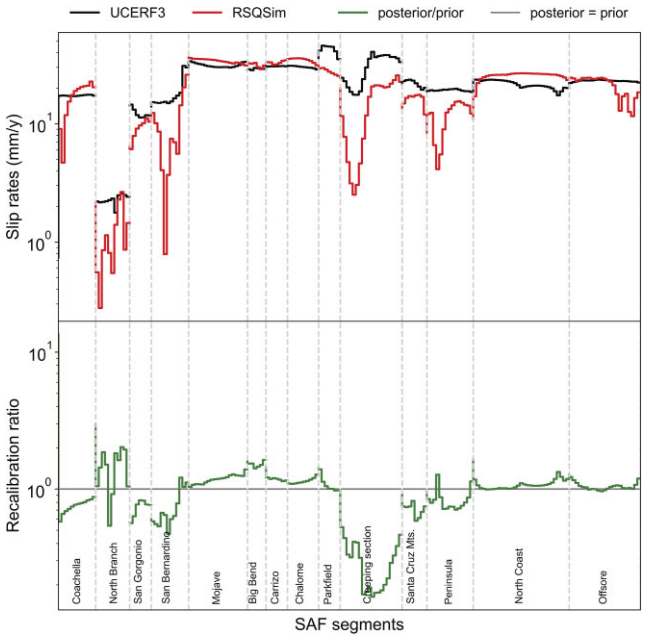

Figure 14. Comparison of slip-rate differences with the Bayesian recalibration results for the SAF. The SAF sections are arrayed from south to north. The achieved UCERF3 slip rates are shown in black and the achieved RSQSim slip rates are shown in red. The ratio of posterior to prior participation rates is shown as the green histogram; grey horizontal line is unity. Lower posterior rates are correlated with RSQSim rates lower than the UCERF3 rates.

Fig. 15 shows the Bayesian recalibration of the examples in Figs 7 and 8. For high-rate subsections that fail the $\alpha = 0.05$ test, the mean rates are often shifted outside the two-sigma zone, indicating that the posterior mean is controlled by the high-count likelihood, not the prior mean (i.e. $\beta \ll T$). High counts also yield sharp posterior distributions, as measured by the coefficient of variation $\tilde{c}_s = (\alpha_s + m_s)^{-1/2}$. Because the rates of individual ruptures are typically low, these coefficients decrease only modestly in the individual-rupture examples of Fig. 15(a). On the other hand, the posterior distributions for subsections with very high counts, such as the two on the Creeping Section shown in Fig. 15(b), are strongly peaked near m_s/T. This sharpness is a consequence of the TI experimental concept, which judges interevent times to be exchangeable and thus pegs the true rate as the limiting value, $\hat{\mu}_s = \lim_{T\to\infty} m_s/T$.

The map view of all subsections in Fig. 11(b) shows more blue shading (lower posterior rates) than red (higher posterior rates). The predominance of blue agrees with the $R < U$ bias seen in testing results of Fig. 11(a), revealing a left-side bias in the rates of subsections that passed the test. That is, within the grey-shaded subsections of Fig. 11(a) (those that passed the test), more fault area turns blue in Fig. 11(b) than turns red.

Along the SAF spine, the R/U discrepancies are smaller in amplitude, which allows us to reduce the logarithmic scale range in Fig. 13(b) by a factor of two. The $R < U$ bias is strong; the fraction of subsections with rate decreases is 70 per cent (318/457), compared to 30 per cent (139/457) (Fig. 13b). As noted in Section 5, many of the SAF failures can be explained by the slip-rate differences. Those that cannot reflect UCERF3 inversion constraints, such as paleoseismic data and regional MFDs and/or RSQSim constraints imposed by the dynamics of multicycle faulting.

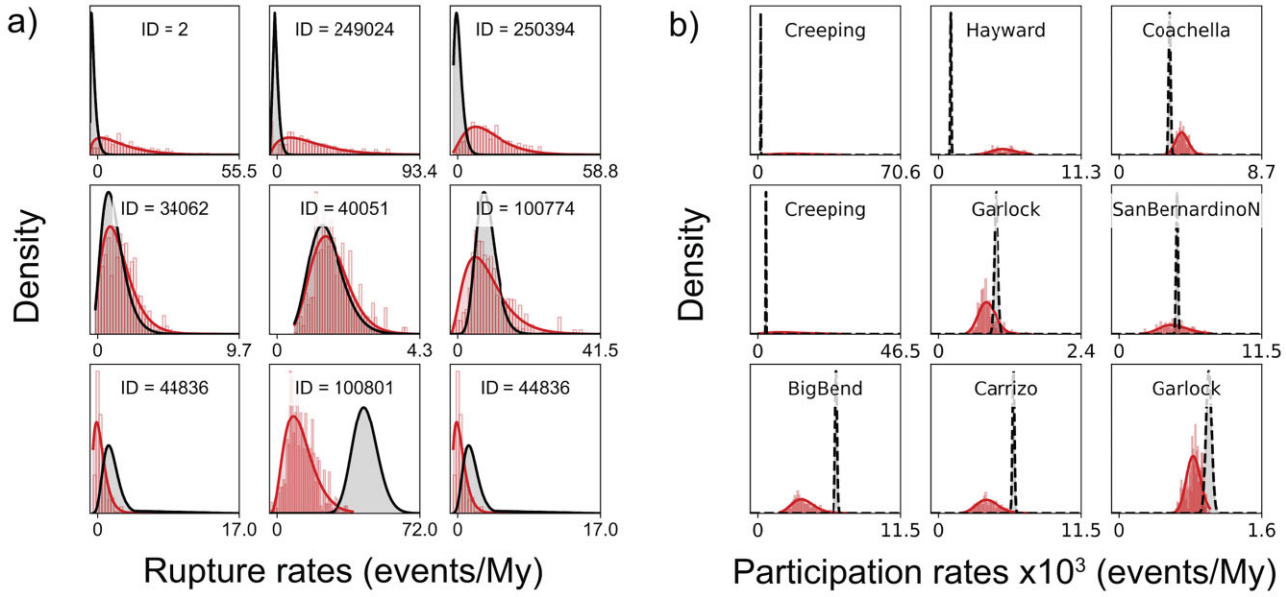

Figure 15. Bayesian recalibration of rupture rates (panel a) and subsection participation rates (panel b) for the examples of Figs 7 & 8. Red PDFs are the prior distributions, and black PDFs are the posterior distributions. Rates are in number of ruptures per year.

7 DISCUSSION AND CONCLUSIONS

We have employed a frequentist methodology to test the UCERF model against RSQSim data under the well-defined experimental concept of TI forecasting, and we have used a Bayesian methodology to recalibrate the UCERF3 EED with RSQSim data under those same assumptions. The results of testing and recalibration, displayed in Figs 11 and 13, are mutually consistent and, taken together, quantify how well RSQSim rupture rates agree with, and differ from, the UCERF3 forecast rates. We have found that some of the discrepancies can be attributed to the differences in slip rates that drive the models, whereas others are governed by the RSQSim fault dynamics absent from UCERF3.

To illustrate how this quantification pertains to PSHA, we consider the hazard maps in fig. 2 of Shaw *et al.* (2018), which give the peak ground acceleration (PGA at 2 per cent in 50 yr) forecast by three California source models—UCERF2, UCERF3 and RSQSim. The latter is based on rupture counts from the same Shaw18 catalogue used in our study. As the authors pointed out, there are larger variations in the UCERF2/UCERF3 PGA ratio than in the RSQSim/UCERF3 PGA ratio. Here, we are primarily interested in explaining the latter, which is plotted in Shaw et al.'s fig 2(e). The same type of blue–white–red logarithmic scale allows an easy comparison with the Bayesian recalibration ratio of Fig. 11b. Accounting for the spatial smoothing inherent to the hazard maps, we see that the PGA and participation rate discrepancies are well correlated:

(i) $R < U$ (blue) discrepancies in both PGA and participation rates predominate north of San Francisco, in the California Continental Borderland, and in the southeastern part of the state.
(ii) $R > U$ (red) discrepancies in both maps are concentrated in central California along a locus of faulting west of the SAF that includes Reliz, Rinconada, South Cuyama, Zayante-Vergeles, La Panza and East Huasna faults, and also on subsections near the SAF-Garlock junction.
(iii) Along the San Andreas spine, the most significant $R < U$ discrepancies occur on the Creeping Section, the east-side faults of the Bay Area, and the San Bernardino section. The most significant $R > U$ discrepancies are on the Big Bend and North Mojave faults, the northwestern San Jacinto fault zone, and the west-side faults of the Bay Area.

The statistical apparatus applied in this paper allows us to identify the ERS–ERF discrepancies at the fault subsection level that map into significant differences in hazard, thus helping to document in a fault-system context which physical mechanisms of multicycle faulting govern the hazard.

Our study of the ERS-to-ERF assimilation problem has been motivated by the prospects for improving forecasting skill by incorporating more information about fault-system dynamics into probabilistic ERFs. Attempts to mine information from ERS–ERF comparisons have just begun (Console *et al.* 2018; Shaw *et al.* 2018). Simulators based on better physics than RSQSim are under rapid development (Tullis *et al.* 2012; Upholff *et al.* 2022; Zielke & Mai 2023), and the promise of more realistic synthetic catalogues has been a strong incentive for our development of ERS–ERF comparison methods. In particular, the statistical techniques described here furnish a framework for addressing the much more difficult problem of assimilating ERS data into time-dependent ERF models, such as the long-term model UCERF3-TD (Field *et al.* 2015b) and the short-term model UCERF3-ETAS (Field *et al.* 2017). The importance of multicycle simulations in the future development of time-dependent ERFs has been recently emphasized in the context of the NSHM by Field *et al.* (2023) and Jordan *et al.* (2023).

We highlight several technical achievements of this study. We have paid a lot of attention to the association problem (Section 4), because it is the bridge that connects the deterministic simulator to the probabilistic forecast. This problem is encountered in the real world when trying to score an observed rupture as an ERF hit or miss. Generalizing the association algorithms developed here may contribute to that capability.

We have investigated why RSQSim magnitudes are systematically higher than the magnitudes of the associated UCERF3

ruptures, and we show that these differences can be explained by two factors: the greater depths of RSQSim ruptures, and the differences in the magnitude-area scaling relations (Fig. 6).

We have attempted to be rigorous in applying the statistical techniques for ontological testing outlined by Marzocchi & Jordan (2014, 2018). One particular focus has been on the assessment of test severity by the *a priori* calculation of statistical power (Appendix A) and the *a posteriori* estimation of false discovery rates (Appendix B). We find that rupture-by-rupture testing is not informative, because it provides little evidence either for or against the collective null hypothesis $\mathbb{O}_\lambda$; the tests lack severity (Mayo 2018). This was known from the start, of course, owing to the structure of the UCERF3 rupture set, which includes many very large ruptures with recurrence intervals much longer than the million-year Shaw18 catalogue. We have shown how the severity of the test can be improved by restricting the collective to only those ruptures with high rupture rates or by generating longer catalogues.

Because the subsection participation rates are much larger than the individual rupture rates, the testing of $\mathbb{O}_\mu$ is more severe, and the discrepancies identified by the testing are more significant and physically interesting. Statistically significant discrepancies between RSQSim and UCERF3 are much more common at the subsection level than expected if $\mathbb{O}_\mu$ were true. We find 643 failures (25 per cent) for $\alpha = 0.05$ and 406 failures (15 per cent) for $\alpha = 0.01$. In comparison, the expected FDRs, based on the q-value calculations of Appendix B, are only 73/643 and 15/406, respectively. Hence, we can be confident that most of the failures are correct rejections; that is, those same subsections would likely fail if tested against an independent million-year catalogue generated by the same RSQSim model.

On the other hand, the power calculations of Appendix A show that the ability of the subsection tests to identify discrepancies is limited. According to Table A2, only 38 per cent of the tests have even moderate power ($\pi_s \geq 0.5$) in discriminating a factor-of-two upwards bias in participation rates ($b = 2$), and even fewer (26 per cent) achieve moderate power for a factor-of-two downwards bias ($b = 0.5$). The Type II errors in detecting factor-of-two discrepancies are expected to be common.

We conclude by pointing out that more stringent tests of model compatibility are possible under the same TI experimental concept used here. We can, for example, test individual samples from the EED against the participation counts for all subsections, thus using the logic-tree correlations to score the overall fit of the ERF to the ERS data. Such model-wide statistical tests, which we are in the process of implementing, should be much more powerful in discriminating ERF/ERS inconsistencies than the single-subsection tests described here.

ACKNOWLEDGMENTS

The authors thank Kevin Milner for his guidance in accessing and using the UCERF3 and RSQSim data sets and for his comments on the results of this study. We also thank the reviewers Rodolfo Console and Glenn P. Biasi for their constructive comments. The SCEC contribution number for this paper is 14124.

DATA AVAILABILITY

The UCERF3-TI model can be downloaded at http://opensha.usc.edu/ftp/kmilner/ucerf3/fss_csvs/2013_05_10-ucerf3p3-production-10runs_COMPOUND_SOL_FM3_1_MEAN_BRANCH_AVG_SOL.csv. RSQSim catalogue can be accessed at http://opensha.usc.edu/ftp/kmilner/simulators/catalogs/rundir2585_1myrs/.

REFERENCES

Baker, J., Bradley, B. & Stafford, P., 2021. *Seismic Hazard and Risk Analysis,* 1st edn., Cambridge University Press.

Berry, G. & Armitage, P., 1995. Mid-P confidence intervals: a brief review, *The Statistician,* **44,** 417–423.

Bommer, J.J. & Scherbaum, F., 2008. The use and misuse of logic trees in probabilistic seismic hazard analysis, *Earthq. Spectra,* **24,** 997–1009.

Chia-Cheng, H. & Hung-Yu, W., 2024. Using RSQSim to determine seismic sequence in Eastern Taiwan fault system, *Seismol. Res. Lett.,* **96,** 207–218.

Console, R., Chiappini, M., Minelli, L., Speranza, F., Carluccio, R. & Greco, M. 2018. Seismic hazard in southern Calabria (Italy) based on the analysis of a synthetic earthquake catalog, *Acta Geophys.,* **66,** 931–943.

Dieterich, J.H. & Richards-Dinger, K.B., 2010. Earthquake recurrence in simulated fault systems, *Pure appl. Geophys.,* **167,** 1087–1104.

Field, E.H., & 2014 Working Group on California Earthquake Probabilities, 2015a. *UCERF3: A New Earthquake Forecast for California's Complex Fault System,* U.S. Geological Survey Fact Sheet 2015-3009, 6.

Field, E.H., 2019. How physics-based earthquake simulators might help improve earthquake forecasts, *Seismol. Res. Lett.,* **90,** 467–472.

Field, E.H. *et al.* 2023. The USGS 2023 conterminous US time-independent earthquake rupture forecast, *Bull. seism. Soc. Am.,* **114**(1), 523–571.

Field, E.H. *et al.* 2014. Uniform California earthquake rupture forecast, version 3 (UCERF3)–the time-independent model, *Bull. seism. Soc. Am.,* **104,** 1122–1180.

Field, E.H. *et al.* 2015b. Long-term time-dependent probabilities for the third uniform California earthquake rupture forecast (UCERF3), *Bull. seism. Soc. Am.,* **105,** 511–543.

Field, E.H. *et al.* 2017. A spatiotemporal clustering model for the third uniform California earthquake rupture forecast (UCERF3-ETAS): toward an operational earthquake forecast, *Bull. seism. Soc. Am.,* **107,** 1049–1081.

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A. & Rubin, D.B., 2013. *Bayesian Data Analysis,* 3rd edn, Chapman and Hall/CRC.

Gelman, A. & Hennig, C., 2017. Beyond subjective and objective in statistics, *J. R. Statist. Soc. Ser. A: Stat. Soc.,* **180,** 967–1033.

George, E.I., Makov, U.E. & Smith, A.F.M., 1993. Conjugate likelihood distributions, *Scand. J. Stat.,* **20,** 147–156.

Gerstenberger, M.E. *et al.* 2022. New Zealand National Seismic Hazard Model 2022 revision : model, hazard and process overview, *GNS Sci.,* 106, doi:10.21420/TB83-7X19.

Gneiting, T. & Katzfuss, M., 2014. Probabilistic Forecasting, *Annu. Rev. Stat. Appl.,* **1,** 125–151.

Jordan, T.H. *et al.* 2023. Panel review of the USGS 2023 conterminous US time-independent earthquake rupture forecast, *Bull. seism. Soc. Am.,* **114**(1), 572–607.

Marzocchi, W. & Jordan, T.H., 2014. Testing for ontological errors in probabilistic forecasting models of natural systems, *Proc. Natl. Acad. Sci. USA.,* **111,** 11 973–11 978.

Marzocchi, W. & Jordan, T.H., 2017. A unified probabilistic framework for seismic hazard analysis, *Bull. seism. Soc. Am.,* **107,** 2738–2744.

Marzocchi, W. & Jordan, T.H., 2018. Experimental concepts for testing probabilistic earthquake forecasting and seismic hazard models, *Geophys. J. Int.,* **215,** 780–798.

Mayo, D.G., 2018. *Statistical Inference as Severe Testing: How to Get beyond the Statistics Wars,* Cambridge University Press.

Milner, K.R., Page, M.T., Field, E.H., Parsons, T., Biasi, G.P. & Shaw, B.E. 2013. Appendix T—Defining the inversion rupture set using plausibility filters, *US Geol. Surv. Open-File Rept. 2013-1165*.

Milner, K.R., Shaw, B.E. & Field, E.H., 2022. Enumerating plausible multifault ruptures in complex fault systems with physical constraints, *Bull. seism. Soc. Am.,* **112,** 1806–1824.

Milner, K.R. *et al.* 2021. Toward physics-based nonergodic PSHA: a prototype fully deterministic seismic hazard model for Southern California, *Bull. seism. Soc. Am.,* **111,** 898–915.

Musson, R.M.W., 2012. PSHA validated by quasi observational means, *Seismol. Res. Lett.,* **83,** 130–134.

Petersen, M.D. *et al.* 2015. The 2014 United States national seismic hazard model, *Earthq. Spectra,* **31,** S1–S30.

Petersen, M.D. *et al.* 2020. The 2018 update of the US National seismic Hazard Model: overview of model and implications, *Earthq. Spectra,* **36,** 5–41

Richards-Dinger, K. & Dieterich, J.H., 2012. RSQSim Earthquake Simulator, *Seismol. Res. Lett.,* **83,** 983–990.

Scherbaum, F. & Kuehn, N.M., 2011. Logic tree branch weights and probabilities: summing up to one is not enough, *Earthq. Spectra,* **27,** 1237–1251.

Shaw, B.E., 2013. Earthquake surface slip-length data is fit by constant stress drop and is useful for seismic hazard analysis, *Bull. seism. Soc. Am.,* **103,** 876–893.

Shaw, B.E., 2019. Beyond backslip: improvement of earthquake simulators from new hybrid loading conditions, *Bull. seism. Soc. Am.,* **109,** 2159–2167.

Shaw, B.E., 2023. Magnitude and slip scaling relations for fault-based seismic hazard, *Bull. seism. Soc. Am.,* **113,** 924–947.

Shaw, B.E., Fry, B., Nicol, A., Howell, A. & Gerstenberger, M., 2022. An earthquake simulator for New Zealand, *Bull. seism. Soc. Am.,* **112,** 763–778.

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,** 1–9.

Storey, J.D., 2003. The positive false discovery rate: a Bayesian interpretation and the q-value, *Ann. Stat.,* **31,** 2013–2035.

Storey, J.D. & Tibshirani, R., 2003. Statistical significance for genomewide studies, *Proc. Natl. Acad. Sci. USA.,* **100,** 9440–9445.

Tullis, T.E. *et al.* 2012. A comparison among observations and earthquake simulator results for the allcal2 California fault model, *Seismol. Res. Lett.,* **83,** 994–1006.

Uphoff, C., May, D.A. & Gabriel, A.-A., 2022. A discontinuous Galerkin method for sequences of earthquakes and aseismic slip on multiple faults using unstructured curvilinear grids, *Geophys. J. Int.,* **233,** 586–626.

Zielke, O. & Mai, P.M., 2023. MCQsim: a multicycle earthquake simulator, *Bull. seism. Soc. Am.,* **113,** 889–908.

APPENDIX A: POWER OF THE TESTS

The power of a binary (accept/reject) hypothesis test, denoted $\pi = 1 - \beta$, is the probability that the test correctly rejects the null hypothesis when a specified alternate hypothesis is true; that is, β is the probability of making a Type II error by incorrectly failing to reject the null hypothesis. In cases where $\alpha = 0.05$, power is conventionally judged to be high if $\pi \geq 0.8$, which sets the Type II error rate at four times the Type I error rate. Assessing the power of a test requires the specification of alternative hypotheses that probe key aspects of the forecast. We illustrate the concept by considering alternate hypotheses regarding the rate bias,

$$\mathbb{H}_\lambda(b) : \textit{The UCERF3 rupture rates are uniformly biased relative to the RSQSim ratesby a factor } b > 0. \quad \text{(A1a)}$$

$$\mathbb{H}_\mu(b) : \textit{UCERF3 subsection participation rates are uniformly biased relative to RSQSim ratesby a factor } b > 0. \quad \text{(A1b)}$$

Under $\mathbb{H}_\lambda(b)$, the bias-corrected experts' ensemble becomes $\mathcal{E}_b = \{b\ \lambda_i^{(l)}, w_l\}$, which simply shifts the corresponding expected value of the EED to $b\ \bar{\lambda}_i$ without changing the coefficient of variation c_i. Similarly, under $\mathbb{H}_\mu(b)$, the mean participation rate is shifted to $b\ \bar{\mu}_s$.

The power calculated for order-of-magnitude bias factors is low for $\mathbb{H}_\lambda$ (Table A1). For $b = 10$, only 3.0 per cent of the ruptures have $\pi_i \geq 0.8$ and only 18 per cent have $\pi_i \geq 0.5$. For $b = 0.1$, the fractions drop to a meager 1.4 and 0.6 per cent, respectively. In the latter case, which measures power on the left-tail of the null distribution, only a small fraction of the ruptures (3.4 per cent) can be assessed because the minimum p-value exceeds $\alpha = 0.05$ for most zero-count ruptures, in which case the power at that significance level cannot be defined.

The power statistics for $\mathbb{H}_\mu$ are much better (Table A2). For $b = 10$ or 0.1, the power is high (≥ 0.8) for 77 and 64 per cent of the subsections, respectively. The higher power for participation rates is expected owing to the lower coefficients of variation and higher rates compared to the rupture-by-rupture tests. If we lower the bias to $b = 2$ or 0.5, the fractions with power greater than 80 per cent drop to 16 and 12 per cent, respectively.

Table A1. Count statistics for power computation of $\mathbb{H}_\lambda(b)$.

Number of ruptures		$\pi_i \geq 0.2$		$\pi_i \geq 0.5$		$\pi_i \geq 0.8$	
$b = 0.1$	$b = 10$	$b = 0.1$	$b = 10$	$b = 0.1$	$b = 10$	$b = 0.1$	$b = 10$
387	11 473	336	11 158	158	2047	64	349

Table A2. Count statistics for power computation of $\mathbb{H}_\mu(b)$.

Number of subsections		$\pi_i \geq 0.2$		$\pi_i \geq 0.5$		$\pi_i \geq 0.8$	
$b = 0.1$	$b = 10$	$b = 0.1$	$b = 10$	$b = 0.1$	$b = 10$	$b = 0.1$	$b = 10$
2 606	2606	2131	2524	1914	2246	1676	1999
$b = 0.5$	$b = 2$	$b = 0.5$	$b = 2$	$b = 0.5$	$b = 2$	$b = 0.5$	$b = 2$
2 606	2606	1303	1943	680	986	324	427

APPENDIX B: Q-VALUES

The significance tests and *a priori* power analysis show that a 1-My catalogue is much too short to meaningfully test UCERF3 on a rupture-by-rupture basis. This conclusion is reinforced by an *a posteriori* analysis of the FDR that adjusts the *p*-value thresholds using what are called '*q*-values' to control for FDR (Storey & Tibshirani 2003). To our knowledge, *q*-values have not been previously applied in earthquake forecast testing, so we provide a brief introduction to the Storey procedure in the context of our ontological testing.

We define $S(t)$ to be the total number of tests that are positive at the significance level t and $V(t)$ to be the number of these positives that are false, that is, resulting from ruptures or subsections for which the ontological null hypothesis is true. The FDR is the expected proportion of the false positives,

$$\mathrm{FDR}(t) = \mathrm{E}[V(t)/S(t)] \approx \mathrm{E}[V(t)]/\mathrm{E}[S(t)]. \tag{B1}$$

The approximation is good when the number of tests is large (Storey & Tibshirani 2003). $S(t)$ is estimated by re-indexing the *p*-values in ascending order, $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(N'_U)}$, and counting the number of values less than or equal to t,

$$\tilde{S}(t) = \max\left[k \ : p_{(k)} \leq t\right] \tag{B2}$$

$V(t)$ is estimated by assuming that the *p*-values for which the null hypothesis is true are uniformly distributed on (0,1) and their number totals $N_0 < N'_U$, so that $V(t) = t\ N_0$. We estimate N_0 by measuring the level of the *p*-value distribution above a threshold $\alpha < \nu < 1$, where it is presumed to be dominated by ruptures or subsections for which the null hypothesis is true:

$$\tilde{N}_0(\nu) = \frac{N'_U - \tilde{S}(\nu)}{1-\nu}. \tag{B3}$$

The estimate of the false discovery rate is thus $\widetilde{\mathrm{FDR}}(t, \nu) = t\ \tilde{N}_0(\nu)/\tilde{S}(t)$. The relationships of the count variables to the histograms of true and false null hypotheses are illustrated for a hypothetical distribution in Fig. B1.

The minimum value of $\widetilde{\mathrm{FDR}}$ above $p_{(k)}$ defines the *q*-value for the k^{th} rupture (Storey & Tibshirani 2003):

$$q_{(k)}(\nu) = \min_{t \geq p_{(k)}} \widetilde{\mathrm{FDR}}(t, \nu) \tag{B4}$$

The minimization ensures that $q_{(k)}$ increases monotonically with $p_{(k)}$. We can therefore modify the significance test to control the FDR rather than the Type I error rate by requiring the estimated FDR to be less than some chosen significance level, say 5 per cent. Among the set of positives $\{k \ : q_{(k)} \leq 0.05\}$, only 5 per cent or fewer are expected to be false. The FDR is different from the Type I error rate, which measures false positives in proportion to the entire test set, rather than the smaller subset of test failures.

The *p*-value histogram of Fig. 9(b) is distorted by the black bulge caused by zero-count discretization bias. However, eliminating ruptures with recurrence intervals greater than 1 My exposed a flat distribution above $\nu = 0.8$. We used this threshold to determine $\tilde{N}_0(\nu)$ and calculate $q_{(k)}$ versus $p_{(k)}$. For the rupture rates, we found that $q_{(k)} \leq 0.05$ corresponds to $p_{(k)} \leq 0.0082$. Hence, controlling the FDR at 5 per cent yields a much lower threshold for rejection than controlling the Type I error at 5 per cent. At the 5 per cent FDR level, only 779 ruptures out of a total of 11 473 (6.8 per cent) are formally rejected.

The testing of $\mathbb{O}_\mu$ yields more sizeable discrepancies between the UCERF3 subsection participation rates and the RSQSim participation counts. Because the counts are high enough, the *p*-value histogram (Fig. 10) shows little discretization bias and better resembles the idealized histogram in Fig. B1. The participation rates give a relatively a flat distribution for $\nu \geq 0.6$, which is the value we use to compute $\tilde{N}_0(\nu)$. We find $q_{(k)} \leq 0.05$ corresponds to $p_{(k)} \leq 0.015$; in other words, the FDR is controlled at 5 per cent by setting the critical value for the significance test at $\alpha = 0.015$. There are 448 subsections (18 per cent of the total) with *q*-values less than 5 per cent, which compares with 661 (25 per cent) with *p*-values less than 5 per cent. Within the former set, the false positives are expected to number only 22.

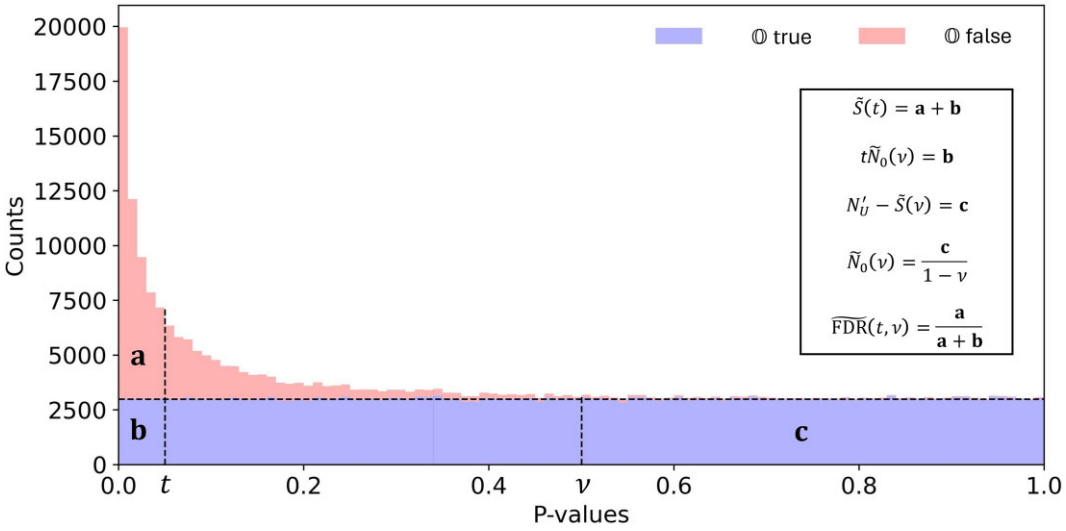

Figure B1. A hypothetical, idealized distribution of *p*-values for a very large ensemble of subhypotheses, showing rupture sets for which $\mathbb{O}$ is true (histogram section below horizontal dashed line) and false (histogram section above horizontal dashed line). The *p*-values from true hypotheses are uniformly distributed, whereas those from the false hypotheses are peaked at low values and tail off at high values. Inset box relates notation used in the text with counts a, b and c in the areas bounded by the dashed lines. $\tilde{\boldsymbol{S}}(\boldsymbol{t})$, the total number of positives with *p*-values less than the significance level $\boldsymbol{t}$, is the sum of counts a and b, and the estimated FDR is the ratio of count a to this sum. The threshold parameter $\boldsymbol{\nu}$ defines the count c, which is rescaled by $(1-\boldsymbol{\nu})$ to obtain $\tilde{\boldsymbol{N}}_0(\boldsymbol{\nu})$, the estimate of the number of true positives. By this construction, $\boldsymbol{t}\ \tilde{\boldsymbol{N}}_0(\boldsymbol{\nu})$ is the count b.