Seismological Research Letters

The forecasting skill of physics-based seismicity models during the 2010-2012 Canterbury, New Zealand, earthquake sequence

--Manuscript Draft--

Manuscript Number:	SRL-D-18-00033R2
Full Title:	The forecasting skill of physics-based seismicity models during the 2010-2012 Canterbury, New Zealand, earthquake sequence
Article Type:	Focus Section - CSEP: New Results and Future Directions
Corresponding Author:	Camilla Cattania Stanford University Stanford, CA UNITED STATES
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Stanford University
Corresponding Author's Secondary Institution:	
First Author:	Camilla Cattania
First Author Secondary Information:	
Order of Authors:	Camilla Cattania
	Maximilian J Werner
	Warner Marzocchi
	Hainzl Sebastian
	David Rhoades
	Matthew Gerstenberger
	Maria Liukis
	William Savran
	Annemarie Christophersen
	Agnès Helmstetter
	Abigail Jiménez Lloret
	Sandy Steacy
	Thomas H Jordan
Order of Authors Secondary Information:	
Manuscript Region of Origin:	GERMANY
Suggested Reviewers:	Tom Parson, PhD Researcher, USGS tparsons@usgs.gov Expertise in earthquake triggering and forecasting.
	Bruce Shaw, PhD Professor, Columbia University shaw@ldeo.columbia.edu Expertise in earthquake triggering (incl. physics-based models).
	Ross Stein, PhD Research Geophysicist, USGS rstein@usgs.gov Expertise in physics-based models of earthquake triggering

	Shinji Toda, PhD Professor, Tohoku Univerisity toda@irides.tohoku.ac.jp Expertise in physics-based models of earthquake triggering
	Massimo Cocco, PhD Chief Scientist, Istituto Nazionale di Geofisica e Vulcanologia cocco@ingv.it Expertise in earthquake forecasting (including physics-based models similar to those presented here).
	Margarita Segou Expertise on aftershock forecasting; and she reviewed the initial submission.
Opposed Reviewers:	

The forecasting skill of physics-based seismicity models during the 2010-2012 Canterbury, New Zealand, earthquake sequence

Camilla Cattania*[1], Maximilian J. Werner[2], Warner Marzocchi[3], Sebastian Hainzl[1], David Rhoades[4], Matthew Gerstenberger[4], Maria Liukis†[5], William Savran[5], Annemarie Christophersen[4], Agnès Helmstetter[6], Abigail Jimenez[7], Sandy Steacy[8], and Thomas H. Jordan[5]

[1]GFZ German Research Centre for Geosciences, Potsdam, Germany

[2]School of Earth Sciences & Cabot Institute, University of Bristol, Bristol, UK

[3]Instituto Nazionale di Geofisica e Vulcanologia, Rome, Italy

[4]GNS Science, Lower Hutt, New Zealand

[5]Southern California Earthquake Center, University of Southern California, Los Angeles, USA

[6]University of Grenoble, ISTerre, CNRS, Grenoble, France

[7]Departamento de Computación e Inteligencia Artificial, Universidad de Granada, Spain

[8]University of Adelaide, Adelaide, Australia

*now at Department of Geophysics, Stanford University, Stanford, USA
†now at Jet Propulsion Laboratory, Pasadena, USA

Abstract

The static Coulomb stress hypothesis is a widely known physical mechanism for earthquake triggering, and thus a prime candidate for physics-based Operational Earthquake Forecasting (OEF). However, the forecast skill of Coulomb-based seismicity models remains controversial, especially in comparison to empirical statistical models. A previous evaluation by the Collaboratory for the Study of Earthquake Predictability (CSEP) concluded that a suite of Coulomb-based seismicity models were less informative than empirical models during the aftershock sequence of the 1992 $M_w 7.3$ Landers, California, earthquake. Recently, a new generation of Coulomb-based and Coulomb/statistical hybrid models were developed that account better for uncertainties and secondary stress sources. Here, we report on the performance of this new suite of models in comparison to empirical Epidemic Type Aftershock Sequences (ETAS) models during the 2010-2012 Canterbury, New Zealand, earthquake sequence. Comprising the 2010 $M7.1$ Darfield earthquake and three subsequent $M \geq 5.9$ shocks (including the February 2011 Christchurch earthquake), this sequence provides a wealth of data (394 $M \geq 3.95$ shocks). We assessed models over multiple forecast horizons (1-day, 1-month and 1-year, updated after $M \geq 5.9$ shocks). The results demonstrate substantial improvements in the Coulomb-based models. Purely physics-based models have a performance comparable to the ETAS model, and the two Coulomb/statistical hybrids perform better or as well as the corresponding statistical model. On the other hand, an ETAS model with anisotropic (fault-based) aftershock zones is just as informative. These results provide encouraging evidence for the predictive power of Coulomb-based models. To assist with model development, we identify discrepancies between forecasts and observations.

Introduction

Recent earthquakes in Italy, New Zealand, Japan and Nepal have demonstrated that forecasts of the space time evolution of seismic sequences provide information that can expand seismic risk reduction strategies beyond building codes, and enhance preparedness and resilience. This is the main goal of Operational Earthquake Forecasting (OEF) introduced by the International Commission on Earthquake Forecasting (ICEF; Jordan et al., 2011) appointed by the Italian government after the 2009 L'Aquila, Italy, earthquake.

For these applications, forecasts should be consistent with future seismicity and they should be the most skilful amongst alternatives (i.e. perform better than other forecasts, according to well defined quantitative measures such as the information gain). The evaluation of consistency and skill of forecast models is the main goal of the Collaboratory for the Studies of Earthquake Predictability (CSEP, Jordan, 2006; Zechar et al., 2010).s

To date, the first results of a prospective CSEP experiment (Nanjo et al., 2012), retrospective CSEP experiments (Woessner et al., 2011; Rhoades et al., 2015), and applications to ongoing earthquake sequences (Marzocchi et al., 2017; Kaiser et al., 2017; Christophersen et al., 2017) showed that statistical models of clustered seismicity like the epidemic-type aftershock sequence models (ETAS, Ogata 1998) and the short-term earthquake probability models (STEP; Gerstenberger et al., 2005) provide informative forecasts of future seismicity. In our view, these represent the first generation of earthquake forecasting models, and a benchmark for measuring any improvements in forecasting capability.

Ongoing model development aims to improve the skill of the forecasts (e.g. *Field et al.*, 2015; *Segou et al.*, 2013). One of the most promising approaches is based on Coulomb stress transfer, the most widely accepted mechanism for aftershock triggering (e.g. *Stein et al.*, 1992; *King et al.*, 1994; *Toda et al.*, 1998). The predictive power of this hypothesis, however, remains a subject of debate (*Hardebeck et al.*, 1998; *Marsan*, 2003). To date, most evaluations of the Coulomb hypothesis are retrospective, with stress changes often calculated at the locations of subsequent events without considering locations which experienced positive Coulomb stress changes without an increase in seismicity. There is a need to rigorously evaluate the Coulomb hypothesis (*Strader and Jackson*, 2014; *Toda and Enescu*, 2011). When coupled with Dieterich's rate-state friction formulation (or another framework for converting stress to seismicity), Coulomb-based models can generate probability forecasts, enabling evaluations of forecast reliability and skill against alternative models.

A previous CSEP evaluation of the predictive skills of forecast models during the 1992 $M_w7.3$ Landers earthquake sequence found that the Coulomb-based models performed worse than statistical models (*Woessner et al.*, 2011), even though they were comparable at short times after the mainshock. Subsequent studies have confirmed that physical models have a lower overall performance than statistical ones, but they can be comparable for at short times after the mainshock, and beyond the near-source region (*Segou et al.*, 2013). To increase our understanding of the physics of triggering, it is important to understand whether the poor performance is due to a failure of the Coulomb stress hypothesis - i.e., static stress changes are not an important mechanism for aftershock triggering - or whether the implementations of the hypothesis involved inappropriate model choices. For instance, large uncertainties exist in Coulomb stress calculations, due to errors in the slip models and receiver fault orientations (e.g. *Steacy et al.*, 2005; *Hainzl et al.*, 2010, 2009). Here, we test recently developed Coulomb models designed to address some of these issues (*Cattania et al.*, 2014).

We investigate the forecasting consistency and skill of this new generation of physics-based forecasting models, as well as new non-parametric models and hybrid Coulomb/statistical models, during the 2010-2012 Canterbury earthquake sequence. The September 3, 2010, $M7.1$ Darfield earthquake initiated a vigorous and damaging aftershock sequence, including the damaging $M6.2$ Christchurch earthquake in February 2011 (Fig. 1).

CSEP Experiment Design

Before submitting models, participants agreed on forecast formats, target data and performance measures. Three forecast horizons were considered (1-day, 1-month, 1-year); models update their forecasts at the end of each forecast horizon, and after each of the four $M \geq 5.9$ earthquakes of the sequence (Fig. 1). We test the effect of data quality with three data-availability scenarios. In the first scenario, most interesting scientifically, models were provided best-available data (a reviewed earthquake catalog, focal mechanisms, and published slip models) to generate forecasts. In the second scenario, the slip models were provided with a 10-day delay to mimic delays in finalising a slip model; no slip models were provided in the first 10 days. In the third scenario, only preliminary data were made available, namely preliminary slip models and catalogs, to mimic the real-time situation of operational earthquake forecasting. All scenarios were evaluated against the best available earthquake catalog data. For brevity, here we focus on the results from the two extreme setups (scenario 1 and 3), and present all results in the electronic supplement.

Forecasts were specified as numbers of earthquakes in space and magnitude bins (*Schorlemmer et al.*, 2007). The spatial region extends between 170.5° and 174.0° longitude, and $-44.5°$ and $-42.5°$ latitude, and a single layer extending to 40 km depth. Spatial cells are 0.05° by 0.05° wide. Magnitude bins are 0.1 units wide, starting from M 3.95; the last bin has no upper bound.

Data

The data sets associated with the three data-input scenarios (best-available, delayed best-available, and near real-time) are summarized in Table: 1 and shown in Fig. 2 and Fig.S1, S2 of the electronic supplement. The target data set comprises 394 $M \geq 3.95$ earthquakes between the September 3, 2010 (UTC), M7.1 Darfield earthquake and the end of the experiment at midnight on February 29, 2012 (the last date of reviewed data available at experiment conception). The catalog was later reviewed by GeoNet; magnitudes were initially given as local magnitudes, and later replaced by moment magnitudes when available. The input data sets for the scenario with best available data comprise (i) the reviewed GeoNet catalog, (ii) published slip models of the four large earthquakes (*Beavan et al.*, 2012), and (iii) a GeoNet focal mechanism catalog. The same data sets are provided in the second scenario, except that slip models are provided to models 10 days after each of the four largest quakes. The data sets of the near-real-time data scenario include (i) a very preliminary GeoNet catalog that was downloaded intermittently by one of us during the sequence (*Christophersen, private communication*) and (ii) preliminary slip models (*Holden et al.*, 2011, ; Beaven and Holden, private communication). The preliminary model and best slip model were computed from the same dataset of near-source strong motion data. The preliminary models are based on a single fault inversion, while the best models invert for kinematic parameters for three newly-defined fault planes. These models provide more details about the overall rupture process and better overall waveform fits (see Fig. S1 in the electronic supplement).

Evaluation Metrics

We evaluated the model forecasts with several CSEP methods (*Rhoades et al.*, 2011; *Schorlemmer et al.*, 2007; *Zechar et al.*, 2010; *Werner et al.*, 2011), which test for consistency of the observations with the probabilistic forecasts and compare the predictive skills of the models. We focus here on the comparison of the forecast skills and on a qualitative consistency check between the numbers of observed and forecast earthquakes. The electronic supplement contains remaining results.

We measure the skill of forecasts with the information gain per earthquake, which compares a model's predictive skill against a benchmark (*Rhoades et al.*, 2011). The benchmark is a time-independent and spatially-uniform Poisson (SUP) process with a Gutenberg-Richter magnitude distribution ($b = 1$). The SUP model is updated at each time step, so that the total forecast rate for the next time step matches the average rate over the past catalog in the test region. The information gain per earthquake calculates the average difference, per earthquake, of the log-likelihood scores of a model and the benchmark. We use 95% confidence bounds estimated by *Rhoades et al.* (2011) to assess statistical significance.

Models

Modelers submitted a total of sixteen models as software to the CSEP testing center, which generated and evaluated forecasts. Due to a bug in STEP-cff, the first 1-day forecast was produced offline. Here, we focus on the results of eight representative models (table 2), described next.

Non-parametric kernel smoothing models: K2, K3

Models K2 and K3 represent statistical end-members on the spectrum of competing models. None of the usual assumptions about earthquake clustering are explicitly included, such as the Omori law or the Utsu-Seki clustering law (large earthquakes generate exponentially more aftershocks). Instead, the models employ Gaussian kernels to estimate seismicity as a function of time, space and magnitude (*Helmstetter and Werner*, 2014). K2 does assume a Gutenberg-Richter magnitude distribution with $b = 1$, while K3 uses kernels to estimate the (space-time dependent) magnitude distribution. The widths of the kernels adapt to the activity level: sparse seismicity (in space and time) widens kernels; concentrated seismicity narrows kernels. The models thereby adjust to the current seismicity rate, which is extrapolated over the forecast horizon. These non-parametric kernel models offer maximum flexibility, at the cost of dispensing with commonly observed empirical laws. Further details can be found in the Supplementary Material.

ETAS implementations: ETAS, ETAS-fault, ETAS-cff

The empirical ETAS model and its hybrid model versions (ETAS-fault, ETAS-cff) are implemented in an identical framework for the setup and parameter estimation which is explained in detail in the Supplementary Material. Any difference in the performance is therefore directly related to the ignorance or use of additional source information and stress calculations. In particular, the only difference is the spatial triggering kernel which is in the case of the ETAS model one or a sum of isotropic power-law kernels centered at the location of the preceding events. In contrast, for events with available slip models, ETAS-fault uses an anistropic power-law kernel as a function of the nearest distance to the mainshock fault plane and ETAS-cff uses a probability distribution based on calculated Coulomb stress changes (*Bach and Hainzl*, 2012).

STEP and STEP-cff

The hybrid model proposed by *Steacy et al.* (2014) is based on the STEP (Short-Term Earthquake Probability) model, a purely statistical approach proposed by *Gerstenberger et al.* (2005). STEP is a weighted sum of models with increasing spatial complexity, including background seismicity and Omori decay. In addition, STEP-cff redistributes seismicity according to the sign of the Coulomb stress change: 93% and 7% of events in regions of positive and negative stress changes respectively. More details can be found in the electronic supplement.

Coulomb-rate-state models

We focus here on two of the submitted Coulomb/rate-state (CRS) models, with the following features:

- CRS-oop: uses Coulomb stresses imparted by the mainshocks ($M \geq 5.95$) on planes optimally oriented with respect to the total Coulomb stress (optimally oriented planes, OOPs).

- CRS-unc: in addition to mainshocks, this model includes stress changes from smaller earthquakes. Instead of using OOPs, CRS-unc accounts for the variability of receiver fault orientations by resolving stress changes on a set of faults from the regional focal mechanisms catalog (electronic supplement and *Cattania et al.* (2014))

For both model versions, seismicity rates are calculated by considering the response of a population of faults with rate-and-state dependent friction (*Dieterich*, 1994), where parameter setting and estimations are done in an identical manner. Both models use an internal grid with a higher resolution than the output grid. CRS-oop is similar to earlier implementations (*Woessner et al.*, 2011), except for the use of an internally refined grid, while the additional features in CRS-unc are new.

Results

Temporal performance

Most models successfully forecast the total number of events and the main features of day-by-day evolution (Fig. 3a). All ETAS models, the STEP models and CRS-unc forecast the total event number to within (Poissonian) uncertainty. Models K2 and K3 underpredict by a factor of about two, while CRS-oop strongly overpredicts. K2 and K3 heavily underestimate the number of triggered earthquakes during the first day of each sequence, but they otherwise forecast the rates well. Since these models do not include mainshock magnitude, but estimate aftershock number from the observed seismicity, starting the forecast exactly at the time of each mainshock (before aftershocks have occurred) hinders their performance on the first day. This is a weakness in the present experimental design.

All models underestimate the number of events triggered by the three large quakes that followed Darfield. The STEP models and, more so, the CRS models, forecast a slower decay after the large shocks than is observed. We note that both CRS models tend to select high values of the aftershock duration t_a (close to 27 yrs, the upper end of the parameter search range): lower t_a values, which would give a faster decay, give a worse fit during the inversion period and are not selected. CRS-oop severely overestimates the number of shocks after the Darfield earthquake, and it does not predict any aftershocks of the last mainshock: this is because the model only considers stress sources from events above a user-defined minimum magnitude, which was set to a value 5.95 (greater than the last large shock's magnitude $M5.9$). The use of a predefined "mainshock" magnitude has been eliminated in later versions of the code (*Cattania and Khalid*, 2016).

The ETAS models match the temporal evolution most closely. They forecast identical numbers because they differ only in their spatial densities. Their success is a result of an Omori p-value $p > 1$ (one of the models' free parameters) and a rather high α value, which reduces the relative importance of secondary triggering by smaller events.

Spatio-temporal performance

The largest differences between model forecasts occur during the first day after each mainshock (e.g. Fig. 4). The expression of the Coulomb component appears remarkably different between CRS-unc and the hybrid models ETAS-cff and STEP-cff, illustrating the sensitivity of these forecasts to the specific implementation of the hypothesis. ETAS-cff and STEP-cff display the more commonly expected Coulomb lobes of a predominantly strike-slip earthquake, while CRS-unc displays much smoother lobes. For ETAS-cff, seismicity rates are linearly related to stress changes; STEP-cff considers only the sign of the stress change, and hence it presents sharp transitions along the nodes, not seen in ETAS-cff (Fig. 4); CRS models are strongly nonlinear in stress, due to the rate-state equations. Moreover, the different treatment of uncertainties (such as receiver faults and subgrid variability) introduces additional differences. The overall pattern for model CRS-oop (not shown) is similar to CRS-unc. The four Coulomb-based models forecast the first day of seismicity much more successfully along the Darfield rupture than the three statistical models; ETAS-fault model forecasts the seismicity about as well, and indeed better after the first few days (Fig. 3).

As already discussed, K2 and K3 do not use the mainshock magnitude to forecast aftershocks and therefore forecast very low seismicity on the first day (Fig. 4). On the second day, however, they forecast a spatial pattern similar to the ETAS model and consistent with observations. This highlights the ability of these models to adapt quickly once enough quakes have occurred (about 10 events).

ETAS-based models are the most successful at reproducing the spatial distribution of seismicity with distance from the fault integrated over the entire time period (Fig. 5). Models K2 and K3 underestimate seismicity, but they have an overall trend similar to the catalog, with most seismicity within cells centered at 0.5-10km from the mainshock fault. (Fig. 3). Both CRS-models underestimate seismicity rates within the first few kilometers from the fault, and overestimate rates beyond 5 km from the fault. ETAS-cff also tends to overestimate rates beyond 10 km from the mainshock faults, while ETAS and ETAS-fault predict a faster spatial decay. In contrast, the difference between STEP and STEP-cff is minimal.

Model ranking

The STEP models and the hybrid models ETAS-cff and ETAS-fault generated the most informative forecasts across all three (1-day, 1-month, 1-year) forecast horizons (see Fig. 6, best-available data scenario). CRS-unc and CRS-oop performed slightly less well, but they were quite close to the hybrid models, and better than the simple ETAS model over longer forecast horizons. Nonetheless, the Coulomb component as implemented in STEP-cff affected its performance very slightly (lowering the information gain), and the ETAS-cff model did not provide additional skill over the ETAS-fault model. CRS-oop consistently performed slightly worse than CRS-unc, to some extent because CRS-oop did not use the last large shock as a stress source (see Fig. 3b).

K2 and K3 presented the lowest information gains, because they performed poorly during the first time window after each mainshock (Fig. 4). Because the log-likelihood score is dominated by earthquake occurrences rather than empty bins, the slower-than-observed decay predicted by most models did not affect their ranking significantly.

Most models (except the STEP models, and the 1-month forecasts of K2 and K3) performed better when they were provided the best-available input data, due to either a more complete and accurate catalog (K2, K3 and ETAS) and also to better slip models (CRS and hybrid models). We found that even the CRS models were more sensitive to the quality of the catalog than to the slip models (fig. S10). Models ETAS-cff and ETAS-fault performed identically to the simple ETAS model in the near-real-time data scenario: in the absence of preliminary slip models (not provided until day 10), these models reverted to simple ETAS models, and the first 10 days heavily dominated the information gain. For a few models, the difference in information gain with best and preliminary data is smaller than 95% confidence intervals (Fig. 6); and even with preliminary data, all models do significantly better than the SUP model, as previously observed for Japanese sequences (*Omi et al.*, 2016).

We can gain some insight into the model performance from the spatial distribution of information gains (Fig. 7). Near the Darfield fault, ETAS-cff was the best performing model, followed by the CRS models. ETAS-cff also better forecasted the few aftershocks to the north-west of the Darfield earthquake, but it overpredicted in the remainder of this enhanced Coulomb lobe (region (1)). ETAS performed worse than its hybrid counterparts, except for the aforementioned lobe of ETAS-cff and a small region near the epicenter of the Darfield earthquake (since its isotropic kernel leads concentrates the forecast for the first day in this area; region (2) in Fig. 7). STEP and STEP-cff present small differences in information gains, indicating that the Coulomb mask has only a subtle effect; this occurs because most of the events forecasts by the STEP model already occur in regions where stress changes resolved on OOPs are positive, so that the redistribution of events does not change the rates significantly. We verified that few points of negative information gains for STEP-cff fall into cells where STEP-cff calculated negative stress changes (region (4) in Fig. 7), near a node of the stress field; in contrast, ETAS-cff does not present a stress shadow and performed better than ETAS in the same cells. This can be due to two reasons: ETAS-cff considers stresses at multiple depth layers, and resolves it on a set of receiver faults; and since STEP-cff only considers the sign of the stress change, it overestimates its effects near nodes of the stress field, where the absolute value is low.

While CRS-unc outperformed STEP along much of the Darfield fault, STEP better captured the Christchurch and Pegasus Bay sequences. As noted above, CRS-oop did not consider the Pegasus Bay $M5.9$ earthquake as a stress source, and it therefore predicted no aftershocks (Fig. 3a). Both CRS models did poorly at intermediate distances ($\gtrsim 10km$ from the mainshock faults), where higher seismicity rates were forecasted than observed. The good performance of the CRS models along the Darfield fault may seem surprising since the CRS models predicted lower near-fault rates than others (Fig. 5): this occured because the log-likelihood is space and time dependent, and CRS models predicted higher seismicity rates than others on the first day of the forecast, when about a third of the aftershocks took place.

Discussion and Conclusions

The ranking of the models indicates that including physical information, such as fault geometry or Coulomb stress changes, can lead to better overall model performance. This is particularly clear from the comparison ETAS to ETAS-cff and ETAS-fault, in agreement with a retrospective case study for California mainshocks (*Bach and Hainzl*, 2012).

Coulomb rate-state models, and in particular CRS-unc, have a performance comparable to the hybrid models, in stark contrast to a previous retrospective evaluation (*Woessner et al.*, 2011), and in agreement with a comparative study of seismicity in Northern California (*Segou et al.*, 2013). This result indicates that when resolving Coulomb stresses in more detail, by using an internally refined grid and including uncertainties and secondary stress sources, the overall performance of physics-based models greatly improves. On the other hand, their spatial and temporal fit indicate that some aspects of the triggering mechanism are not yet captured, as discussed below.

Most of the ETAS-based models (ETAS, ETAS-fault and STEP) prescribe a functional form for the spatial decay of seismicity from the mainshock sources (a power law), and they reproduce the observed decay reasonably well. The inclusion of Coulomb stress changes in ETAS-cff leads to overestimation of off-fault seismicity, analogous to the CRS models. STEP-cff, on the other hand, only considers the sign of the stress change and therefore preserves the power-law decay prescribed by STEP, so that the two models exhibit a similar decay (Fig. 5). Models K2 and K3, which estimate the spatial distribution directly from the catalog itself, also provide a good fit when accounting for the fact that rates are underestimated everywhere due to the lack of information on the first day.

A major simplification in the CRS models was to assume spatially uniform background rate. With this assumption, the model does not distinguish between areas with fault structures capable of hosting seismicity, and areas without pre-existing faults, leading to overestimation in the far-field. Moreover, in the rate-state formulation, weakly stressed regions contribute to seismicity later in the sequence (*Dieterich*, 1994; *Helmstetter and Shaw*, 2006), so that assuming a uniform background rate leads to a slower decay (*Cattania et al.*, 2015), consistent with Fig. 3. One of the models submitted for testing (electronic supplement) is a variation of CRS-unc, including heterogeneous background rate derived from smoothed seismicity (*Helmstetter et al.*, 2007). However, this model has a poorer performance than CRS-unc, because before the Darfield earthquake the seismic activity was dominated by the Alpine fault system, with relative little seismicity in the area of the Darfield-Canterbury sequence (see model K2 in Fig. 4, first day). Estimating the spatially variable background seismicity rate, especially when the mainshock hits relatively quiescent regions, remains one of the challenges of Coulomb rate-state models (e.g. *Bhloscaidh et al.*, 2014; *Cocco et al.*, 2010). Another challenging aspect in modeling Coulomb stress triggering is the heterogeneity in stress, especially in the near field. In addition to the variable orientation of receiver faults, a source of stress heterogeneity is the small scale variability of seismic slip, generating locally high stresses on the fault plane and seismicity within the rupture area, where the average stress is negative (*Helmstetter et al.*, 2007). We note that considering multiple fault orientations has a similar effect in terms of stress shadow reduction (*Cattania et al.*, 2014), leading to reasonable information gains even near the mainshock faults (Fig. 6); however, underestimation of near-field stresses may contribute to the overall undererstimation of seismicity in these regions (Fig. 5).

The better performance of CRS-unc over CRS-oop is consistent with previous studies (*Cattania et al.*, 2014, 2015), and was due to the inclusion of secondary triggering and uncertainties due to receiver fault orientation (for a comparison with models including only one of these aspects, see electronic supplement). The use of OOPs instead of a fixed receiver fault typically leads to a better performance in the near field and short time after the mainshock (e.g. *Hainzl et al.*, 2009; *Woessner et al.*, 2011; *Segou et al.*, 2013), since they can reproduce high rates near the mainshock. Here we find that model CRS-unc has a better performance than CRS-oop across all temporal and spatial scales (Fig. 3, 5). This result suggests that using known information on the local fault geometry (from focal planes, as done here; or from mapped faults, when available) may be the optimal forecasting strategy, as long as the variability of fault orientations is also modeled.

We note that CRS-oop has better scores, relative to statistical models, than a similar model tested in the Landers retrospective experiment (*Woessner et al.*, 2011). This is most likely due to the use of a refined grid for Coulomb stress calculations. Considering multiple depth layers, for example, accounts for the fact that stress changes (resolved on the mainshock fault plane) are negative within the rupture area and positive above and below it; therefore, calculating stresses at a single intermediate depth will likely result in underestimation of on-fault rates. Since we did not test a CRS model without grid refinement, we can not directly measure the improvement due to this aspect. An indirect test, however, comes from the study by *Steacy et al.* (2014), who compared STEP, STEP-cff and a classic CRS model (using OOPs, and no grid refinement) for the Canterbury sequence. While the different time window and target magnitude prevent us from comparing information gains exactly, we note that the overall performance of the CRS model was significantly lower than the STEP model, with a difference in information gains per event of about $2-3$. The relative performance between STEP and STEP-cff, on the other hand, is close to what we find here: a small difference in information gain per event (< 0.1), with the STEP model performing slightly better.

ETAS-cff shares certain aspects of model implementation with CRS-unc: vertical grid refinement and consideration of receiver fault variability (even though the set of receiver faults was different; see Electronic Supplement). The improvement of ETAS over ETAS-cff, in contrast with models STEP-cff and STEP, confirms that these aspects have a first-order effect on information gains.

The relatively good performance of CRS models is encouraging in terms of our physical understanding of earthquake triggering. Like earlier versions (*Woessner et al.*, 2011), these models are based on two widely accepted concepts: that aftershocks are mainly caused by static stress changes, and that time-dependence of their nucleation is controlled by rate-state friction (*Dieterich*, 1994). The drawback of physical models is that several of the quantities involved in Coulomb stress calculations and rate-state seismicity evolution are not known precisely. The improvement in performance compared to earlier studies suggests that the main issues with physics-based models was not in the fundamental process but rather in specific details of model implementation.

There are still multiple ways in how physical models can be refined: for example, we identified the spatial dependence of background seismicity as a particularly challenging aspect. Other improvements such as the inclusion of aseismic stresses or consideration of the spatial variability in receiver fault orientations can in the future be tested in the context of CSEP. Another important question to address is the practical use of these models in the context of operational earthquake forecasting: we note that currently, hybrid models with a similar performance require less computation time, making them more suitable for real-time applications.

Acknowledgements

This research was carried out in the framework of REAKT Project (Strategies and tools for Real-Time EArthquake RisK ReducTion) funded by the European Community via the Seventh Framework Program for Research (FP7), contract no.282862.Forecasts were generated and evaluated by the SCEC CSEP testing center. This research was supported by the Southern California Earthquake Center (Contribution No. 8045). SCEC is funded by NSF Cooperative Agreement EAR-1033462 & USGS Cooperative Agreement G12AC20038. We thank Margarita Segou and an anonymous reviewer for careful and constructive reviews.

Data and Resources

The GeoNet catalog and the focal mechanism catalog can be accessed at: https://www.geonet.org.nz/

References

Bach, C., and S. Hainzl (2012), Improving empirical aftershock modeling based on additional source information, *J. Geophys Res.*, *117*(4), 1–12, doi:10.1029/2011JB008901.

Beavan, J., M. Motagh, E. J. Fielding, N. Donnelly, and D. Collett (2012), Fault slip models of the 2010-2011 Canterbury, New Zealand, earthquakes from geodetic data and observations of postseismic ground deformation, *New Zealand J. of Geology and Geophysics*, *55*(3), 37–41, doi:10.1080/00288306.2012.697472.

Bhloscaidh, M. N., J. McCloskey, and C. J. Bean (2014), Response of the San Jacinto Fault Zone to static stress changes from the 1992 Landers earthquake, *J. Geophys Res.*, *119*, 8914–8935, doi:10.1002/2014JB011164.

Cattania, C., and F. Khalid (2016), A parallel code to calculate rate-state seismicity evolution induced by time dependent, heterogeneous Coulomb stress changes, *Computers and Geosciences*, *94*, 48–55, doi:10.1016/j.cageo.2016.06.007.

Cattania, C., S. Hainzl, L. Wang, F. Roth, and B. Enescu (2014), Propagation of Coulomb stress uncertainties in physics-based aftershock models, *J. Geophys Res.*, *119*(10), 7846–7864, doi:10.1002/2014JB011183.

Cattania, C., S. Hainzl, L. Wang, B. Enescu, and F. Roth (2015), Aftershock triggering by postseismic stresses: A study based on Coulomb rate-and-state models, *J. Geophys. Res*, *120*(4), 2388–2407, doi:10.1002/2014JB011500.

Cocco, M., S. Hainzl, F. Catalli, B. Enescu, A. M. Lombardi, and J. Woessner (2010), Sensitivity study of forecasted aftershock seismicity based on Coulomb stress calculation and rate- and state-dependent frictional response, *J. Geophys Res.*, *115*(B05307), doi:10.1029/2009JB006838.

Dieterich, J. H. (1994), A constitutive law for rate of earthquake production and its application to earthquake clustering, *Journal of Geophysical Research*, *99*(B2), 2601–2618.

Field, E. H., G. P. Biasi, P. Bird, T. E. Dawson, K. R. Felzer, D. D. Jackson, K. M. Johnson, T. H. Jordan, C. Madden, A. J. Michael, K. R. Milner, M. T. Page, T. Parsons, P. M. Powers, B. E. Shaw, W. R. Thatcher, R. J. Weldon, and Y. Zeng (2015), Long-term time-dependent probabilities for the third uniform California earthquake rupture forecast (UCERF3), *Bulletin of the Seismological Society of America*, *105*(2), 511–543, doi:10.1785/0120140093.

Gerstenberger, M. C., S. Wiemer, L. M. Jones, and P. a. Reasenberg (2005), Real-time forecasts of tomorrow's earthquakes in California., *Nature*, *435*(7040), 328–31, doi:10.1038/nature03622.

Hainzl, S., B. Enescu, M. Cocco, J. Woessner, F. Catalli, R. Wang, and F. Roth (2009), Aftershock modeling based on uncertain stress calculations, *J. Geophys Res.*, *114*(B5), 1–12, doi:10.1029/2008JB006011.

Hainzl, S., G. Zöller, and R. Wang (2010), Impact of the receiver fault distribution on aftershock activity, *J. Geophys Res.*, *115*(B5), 1–12, doi:10.1029/2008JB006224.

Hardebeck, J. L., J. J. Nazareth, and E. Hauksson (1998), The static stress change triggering model: Constraints from two southern California aftershock sequences, *J. Geophys Res.*, *103*(B10), 24,427–24,437.

Helmstetter, A., and B. E. Shaw (2006), Relation between stress heterogeneity and aftershock rate in the rate-and-state model, *J. Geophys Res.*, *111*(B7), 1–12, doi:10.1029/2005JB004077.

Helmstetter, A., and M. J. Werner (2014), Adaptive Smoothing of Seismicity in Time, Space, and Magnitude for Time-Dependent Earthquake Forecasts for California, *Bull. Seism. Soc. Am.*, *104*(2), 809–822, doi:10.1785/0120130105.

Helmstetter, A., Y. Y. Kagan, and D. D. Jackson (2007), High-resolution Time-independent Grid-based Forecast for M >= 5 Earthquakes in California, *Seism. Res. Lett.*, *78*(1).

Holden, C., J. Beavan, B. Fry, M. Reyners, J. Ristau, R. V. Dissen, P. Villamor, and M. Quigley (2011), Preliminary source model of the M w7.1 Darfield earthquake from geological, geodetic and seismic data, *Proceedings of the Ninth Pacific Conference on Earthquake Engineering Building an Earthquake-Resilient Society*, *164*, 1–7.

King, G. C. P., R. S. Stein, and J. Lin (1994), Static stress changes and the triggering of earthquakes, *Bull. Seismol. Soc. Am.. Soc. Am.*, *84*(3), 935–953.

Marsan, D. (2003), Triggering of seismicity at short timescales following Californian earthquakes, *J. Geophys Res.*, *108*, 1–14, doi:10.1029/2002JB001946.

Ogata, Y. (1988), Statistical Models for Earthquake Occurrences and Residual Analysis for Point Processes, *J. Am. Stat. Ass.*, *83*, 9–27.

Omi, T., Y. Ogata, K. Shiomi, B. Enescu, K. Sawazaki, and K. Aihara (2016), Automatic aftershock forecasting: A test using real-time seismicity data in Japan, *Bulletin of the Seismological Society of America*, *106*(6), 2450–2458, doi:10.1785/0120160100.

Rhoades, D., D. Schorlemmer, M. Gerstenberger, A. Christophersen, J. Zechar, and M. Imoto (2011), Efficient testing of earthquake forecasting models, *Acta Geophysica*, *59*(4), doi:10.2478/s11600-011-0013-5.

Schorlemmer, D., M. C. Gerstenberger, S. Wiemer, D. D. Jackson, and D. A. Rhoades (2007), Earthquake Likelihood Model Testing, *Seismological Research Letters*, *78*(1).

Segou, M., T. Parsons, and W. Ellsworth (2013), Comparative evaluation of physics-based and statistical forecasts in Northern California, *J. Geophys Res.*, *118*(12), 6219–6240, doi:10.1002/2013JB010313.

Steacy, S., S. Nalbant, J. McCloskey, O. Scotti, and D. Baumont (2005), Onto what planes should Coulomb stress perturbations be resolved?, *J. Geophys Res.*, *110*, 1–14, doi:10.1029/2004JB003356.

Steacy, S., M. Gerstenberger, C. Williams, D. Rhoades, and A. Christophersen (2014), A new hybrid Coulomb/statistical model for forecasting aftershock rates, *Geophys. J. Int.*, *196*, 918–923, doi:10.1093/gji/ggt404.

Stein, R. S., G. C. P. King, J. Lin, and N. Palm (1992), Change in failure stress on the southern San Andreas fault system caused by the 1992 Magnitude = 7 . 4 Landers earthquake, *Change*, *4*, 2–9.

Strader, A., and D. Jackson (2014), Near-prospective test of Coulomb stress triggering, *J. Geophys Res.*, *119*, 3064–3075, doi:10.1002/2013JB010780.

Toda, S., and B. Enescu (2011), Rate/state Coulomb stress transfer model for the CSEP Japan seismicity forecast, *Earth, Planets and Space*, *63*(3), 171–185, doi:10.5047/eps.2011.01.004.

Toda, S., R. S. Stein, P. A. Reasenberg, J. H. Dieterich, and Y. Akio (1998), Stress transferred by the 1995 Mw=6.9 Kobe, Japan, shock: Effect on aftershocks and future earthquake probabilities, *Journal Geophys. Res.*, *103*, 24,543–24,656.

Werner, M. J., A. Helmstetter, D. D. Jackson, and Y. Y. Kagan (2011), High-resolution long-term and short-term earthquake forecasts for California, *Bulletin of the Seismological Society of America*, *101*(4), 1630–1648, doi:10.1785/0120090340.

Woessner, J., S. Hainzl, W. Marzocchi, M. J. Werner, a. M. Lombardi, F. Catalli, B. Enescu, M. Cocco, M. C. Gerstenberger, and S. Wiemer (2011), A retrospective comparative forecast test on the 1992 Landers sequence, *J. Geophys Res.*, *116*(B5), 1–22, doi:10.1029/2010JB007846.

Zechar, J. D., M. C. Gerstenberger, and D. a. Rhoades (2010), Likelihood-Based Tests for Evaluating Space-Rate-Magnitude Earthquake Forecasts, *Bulletin of the Seismological Society of America*, *100*(3), 1184–1195, doi:10.1785/0120090192.

Table 1: Overview of data sets.

	data type	use	features
Geonet-Cat	catalog	target data/model input (mode 1/2)	$M \geq 3.95$
Preliminary-Cat	catalog	model input (mode 3)	Captured by NZ CSEP Testing Center
Best-Slip	slip models	model input (mode 1/2)	Beaven et al. 2012
Preliminary-Slip	slip models	model input (mode 3)	Holden et al. 2011 & personal comm.
Focal Mechanisms	focal mechanisms	model input (mode 1-3)	GeoNet

Table 2: Overview of models participating in the test with reference whether or not they use the Gutenberg-Richter distribution (GR), the Omori-Utsu decay function (OU), an exponential productivity function ($N(M)$), fault information (fault), Coulomb Failure Stress (CFS), rate- and state-dependent frictional response (RS), or focal mechanisms (FM).

		empirical relations			fault information & physics				
index	model name	GR	OU	$N(M)$	Fault	CFS	RS	FM	reference
0	SUP	x							Rhoades et al, 2018, this issue
1	K2	x							*Helmstetter and Werner* (2014)
2	K3								*Helmstetter and Werner* (2014)
3	ETAS	x	x	x					*Ogata* (1988)
4	ETAS-fault	x	x	x	x				*Bach and Hainzl* (2012)
5	ETAS-cff	x	x	x		x			*Bach and Hainzl* (2012)
6	STEP	x	x	x	x				*Gerstenberger et al.* (2005)
7	STEP-cff	x	x	x	x	x			*Steacy et al.* (2014)
8	CRS-oop	x				x	x		*Cattania et al.* (2014)
9	CRS-unc	x				x	x	x	*Cattania et al.* (2014)

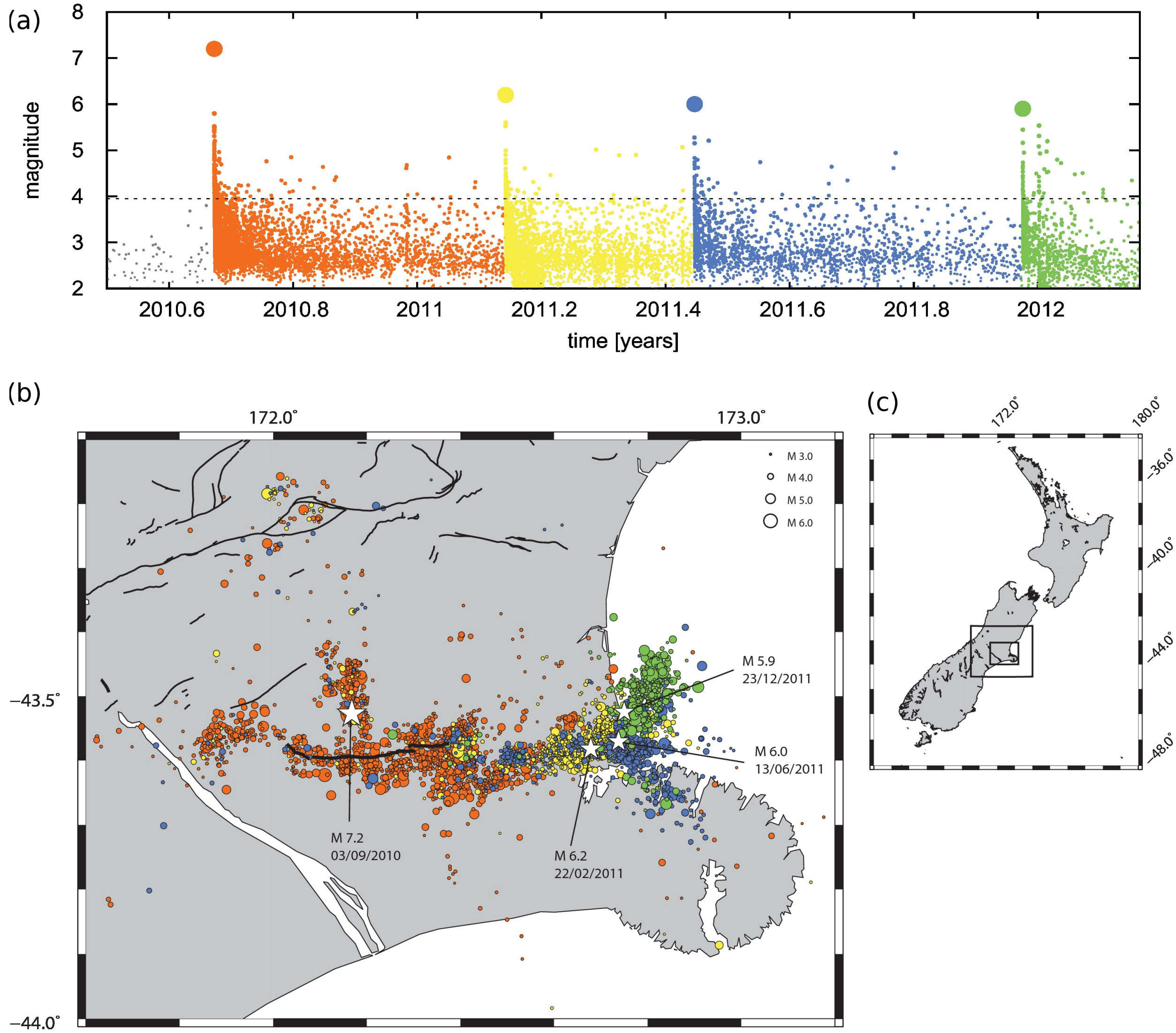


Figure 1: (a) Magnitude vs. time from the reviewed GeoNet catalog. The different colors indicate the sequences of events with $M \geq 5.9$. (b) Map view of the seismic sequence, colorcoded in agreement with the top panel. Fault lines are from the New Zealand Active Fault Database, and the thicker line is the Greendale fault. (c) map of new Zealand, with the boxes marking the forecast area (larger box) and the map on the left.

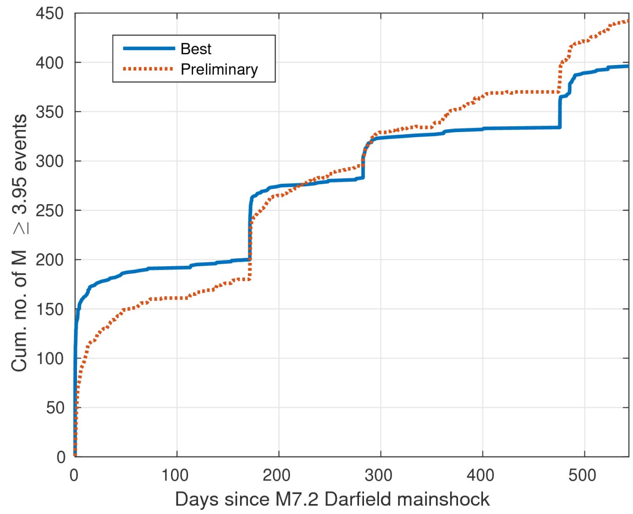


Figure 2: Cumulative number of $M \geq 3.95$ events in the reviewed reviewed catalog and real time data. For the real time data, we report the total number of events in the catalog used on each day: as the catalog is revised, the number of events may vary because the catalog becomes more complete, or because magnitudes are revised. Magnitude were initially reported as M_L, and later replaced by M_w; since M_L were systematically overestimated until the end of 2011, the number of events in the real time catalog can exceed the reviewed catalog.

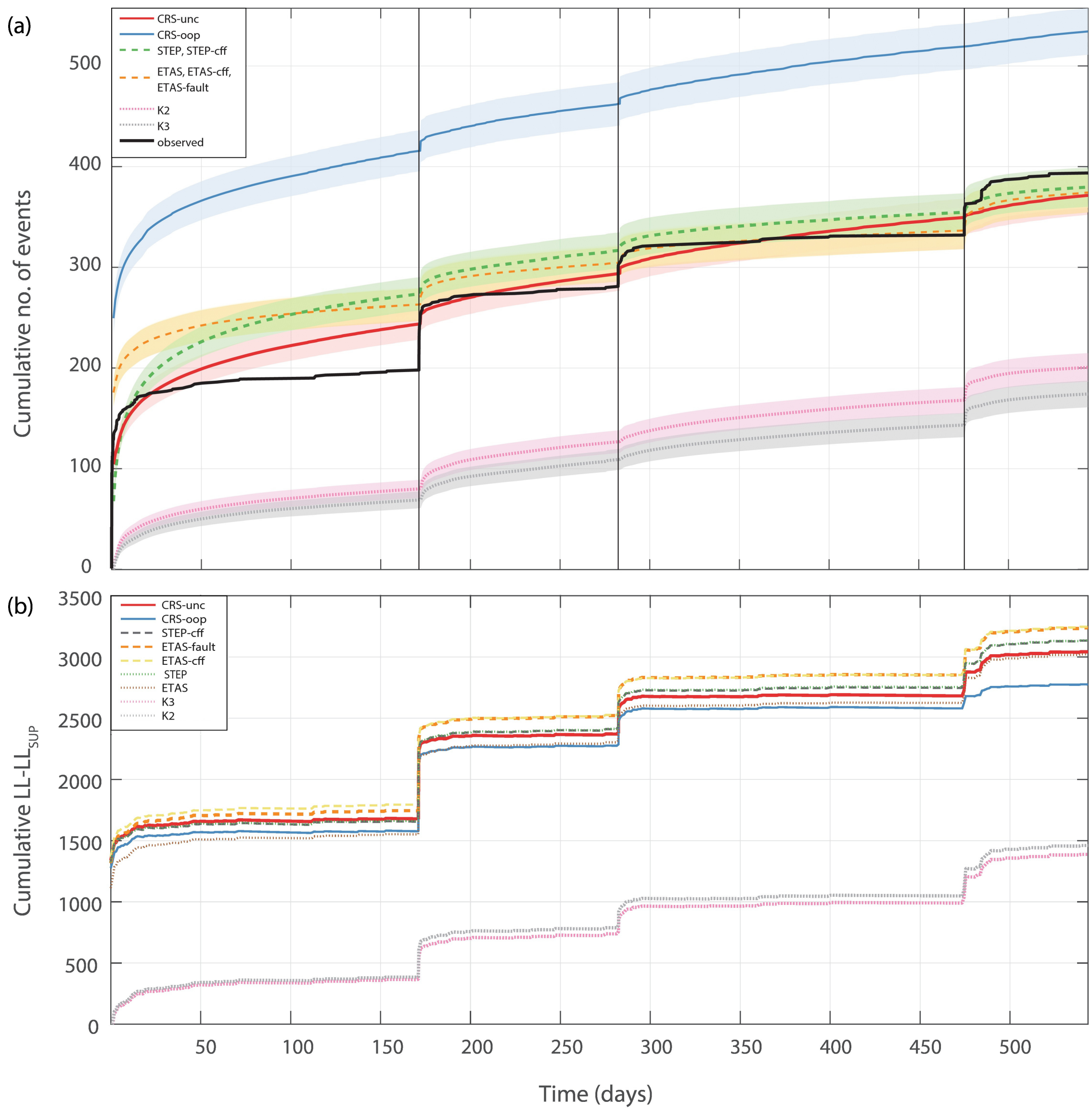


Figure 3: (a) Forecasted and observed temporal evolution. Shaded areas indicate Poissonian errors; vertical lines are events with $M \geq 5.9$. (b) Cumulative difference in log-likelihood with respect to the SUP model, obtained from the sum of the log-likelihood calculated for (space, time, magnitude) bins.

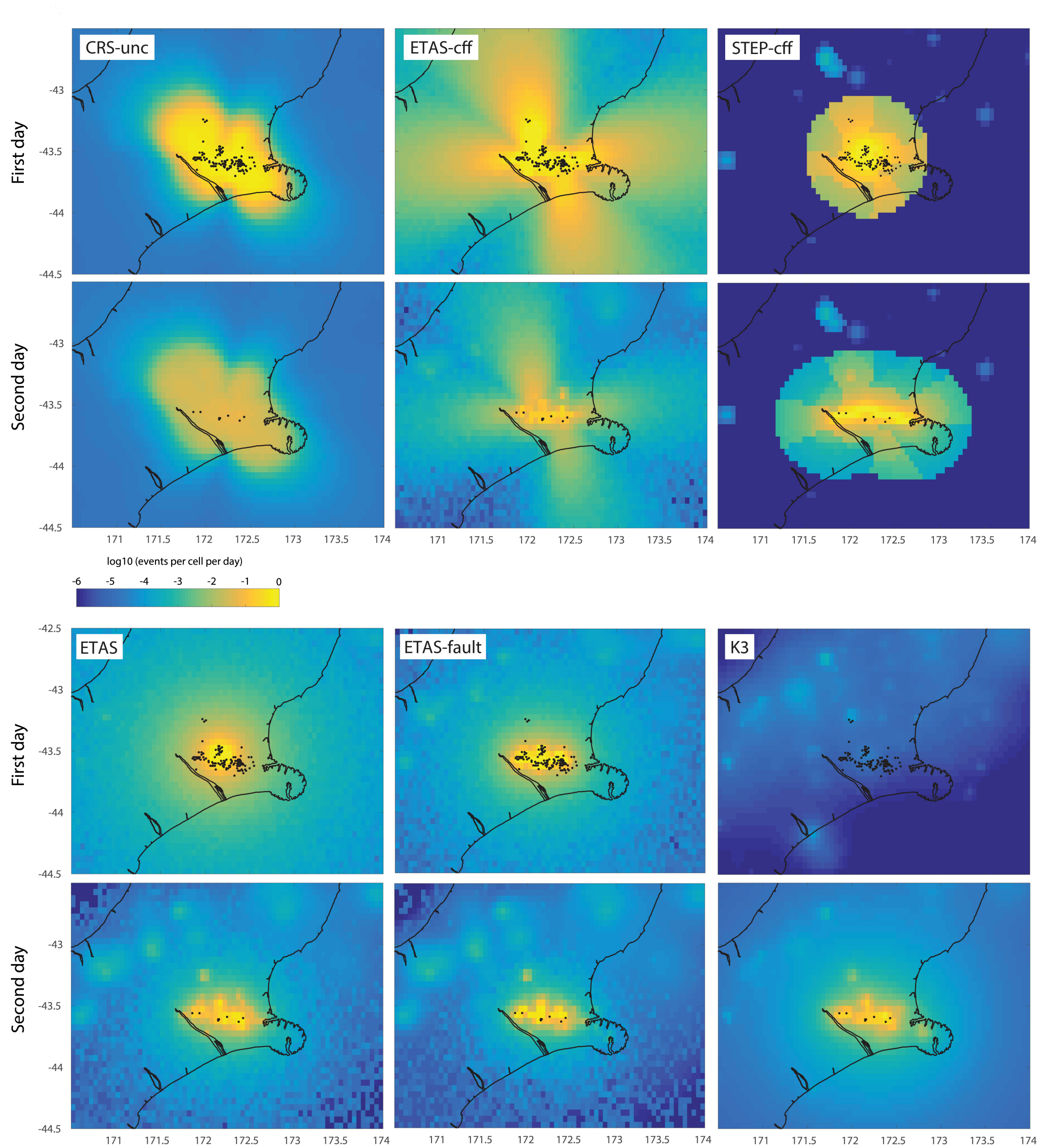


Figure 4: Examples of 1-day forecasts for selected models. The top row is the forecast starting at the time of the Darfield mainshock, and the second line forecasts are the second day: the difference highlights how each model incorporates information from the early part of the sequence. The dots are the observed events in each time period.

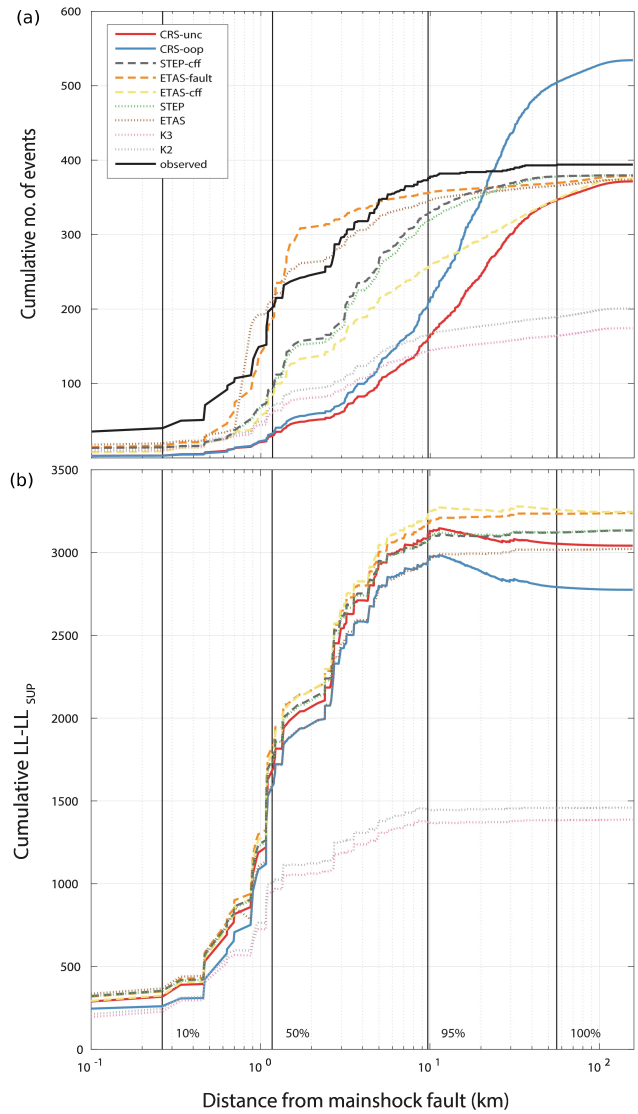


Figure 5: (a) Cumulative number of earthquakes as a function of distance from the nearest mainshocks fault trace, based on the location of the cell center. For consistency, the catalog was also binned into the forecast cells, so that distances does not reflect exact earthquake locations but rather the cell to which they are assigned. (b) cumulative difference in log-likelihood from the SUP forecast, obtained from the sum of the log-likelihood calculated for (space, time, magnitude) bins. The vertical lines indicate the percentage of earthquakes within cells at a given distance from the faults.

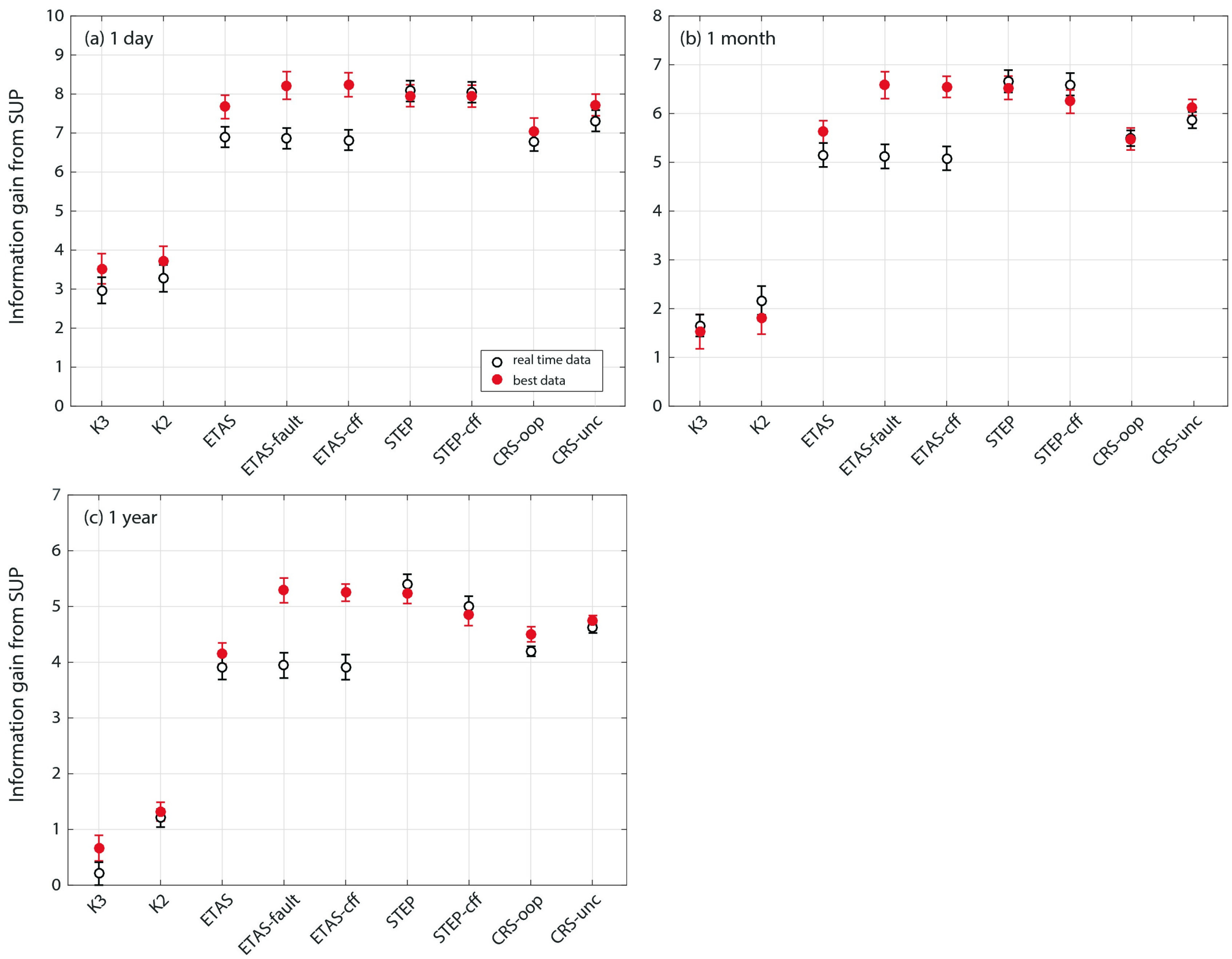


Figure 6: Information gain per earthquake relative to the SUP model, for real time and best available data. Each panel shows a forecast horizon. Error bars represent 95% confidence levels from a paired student-T test (*Rhoades et al.*, 2011).

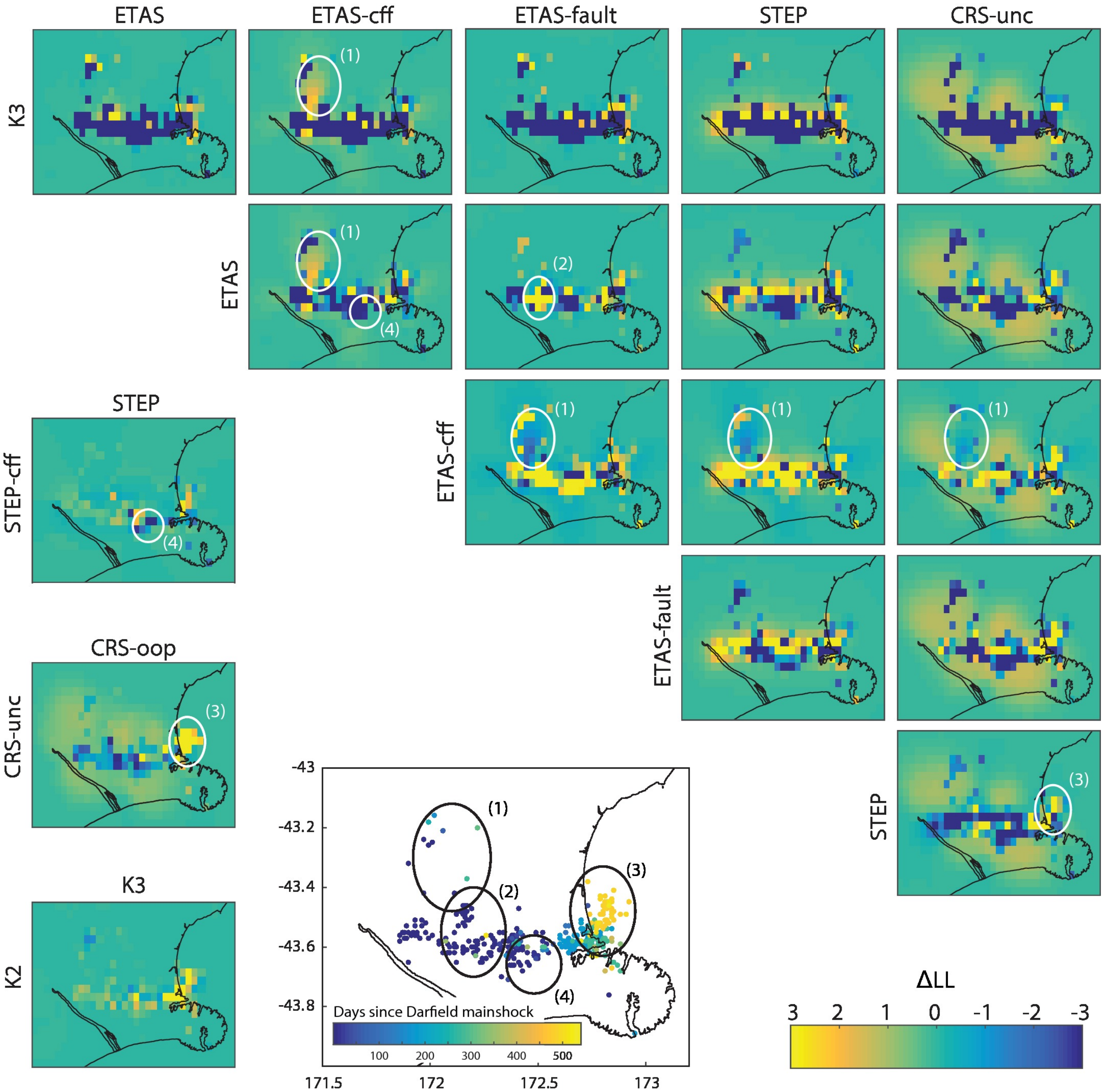


Figure 7: Map of log-likelihood differences between pairs of models, for a subset of the model domain near the mainshock faults. The color indicates $\Sigma_n LL_{n,i} - \Sigma_n LL_{n,j}$, where i and j are the row and column index, and the summation is performed over all time steps and magnitude bins. Note that the values are capped at ± 3 for clarity. Positive values along a row indicate good model performance, and along a column they indicate poor performance. The ellipses mark the following features, discussed in the main text: (1) a Coulomb stress lobe; (2) the area near the $M7.2$ Darfield epicenter; (3) the aftershocks of December 23^{th} $M5.9$ Pegasus Bay earthquake; (4) few cells near a node of the Coulomb stress field, where STEP-cff predicts a stress shadow.

Electronic Supplement to

The forecasting skill of physics-based seismicity models during the 2010-2012 Canterbury, New Zealand, earthquake sequence

by Camilla Cattania, Maximilian J. Werner, Warner Marzocchi, Sebastian Hainzl, David Rhoades, Matthew Gerstenberger, Maria Liukis, William Savran, Annemarie Christophersen, Agnès Helmstetter, Abigail Jimenez, Sandy Steacy and Thomas H. Jordan

Part 1 – Model Description

This section includes a description of how the following models are implemented: K2/K3 (section 1), ETAS models (Section 2), the STEP models (Section 3) and the CRS models (Section 4). Table 1,2 contains model parameters for the K2/K3 and ETAS models, while table 3 lists the CRS models submitted and their relative performance.

Part 2 – Comparison of real time and best available data

In this section we present the preliminary and finalized slip models (Fig.S1) and a comparison between the preliminary and best available catalog (Fig.S2)

Part 2 – Results of CSEP tests for all models and testing modes

In this section we report all the test results: consistency tests (L-test, M-test, N-test and S-test), as well as the T-test and W-test for 15 of the models submitted and 3 testing modes ("real time" data, with a preliminary catalog and slip models provided with a 10 days delay; best data, with a reviewed catalog and slip models provided without delay; and an intermediate mode with the reviewed catalog, and slip models provided with a 10 days delay). Figure S3 summarizes the L-test, M-test, S-test. Figure S4 demonstrates why certain models do better than others in the consistency tests, despite a lower performance (as measured by the T-test). Figures S5-10 shows the results of the N-test on each day. Figure S11 shows the W-test, and Figure S12 shows the T-test. Due to a bug in model STEP-cff, the tests could not be performed in time and we are not including this model.

Figures

Figure S1. Preliminary (left) and final (right) slip models for all mainshocks. Note the different colorscale for each mainshock.

Figure S2. Comparison of real time and final catalog. Since the real time data was continuously updated, we present two representative days: the 10th day after the Darfield earthquake, and the last day. Left: catalogs provided on day 10. Note that the real time catalog is more complete, and the median difference in horizontal locations is comparable to the cell dimension (about 5km). Right: comparison of the catalogs on the last day of the experiment. The median distance between events in the catalogs is lower, since some events have been already relocated. The larger number of events in the real time catalog is due to the systematic underestimation of magnitudes, which can be seen in the lower plot.

Figure S3. Results of the consistency tests implemented in CSEP (Schorlemmer et al, 2007). Each test compares the observed likelihood score with the distribution of likelihood for that model, obtained from a set of 100 Poissonian simulations drawn from the forecast itself. The L-test compares log-likelihoods in space-time-magnitude bins; the S-test evaluates the spatial distribution of the forecasts; the M-test evaluates the magnitude distribution. Each row indicates a testing mode; the second and third row correspond to the "real time" and "best" data presented in the main text. A cross indicates that a model fails the test (at 95% confidence) on that day, and the number on the right is the total number of failures. Failure of the M-test for ETAS, RETAS and CRS models was due to an overestimation of the b-value on certain days, when using the best catalog. We note that most models fail the L-test and S-test more frequently when better data was provided (cf. 2nd and 3rd rows), even though better data leads to higher likelihood scores (Fig. 6 in the main text). This counter-intuitive result is due to the distribution of likelihood scores from the synthetic catalog (see Fig. S4).

Figure S4. Example of observed and simulated likelihood scores for model CRS-unc, for two testing modes: preliminary data (left) and best available data (right), for one day starting with the M 6.2 event on February 2nd (day 173 in Fig.S3). Top: entire forecast, on a log scale; middle: forecast for the area in the box, on a linear scale, indicating that the model produces a better forecast when the best data is provided (black dots are observed events). Bottom: distribution of spatial log-likelihood scores. The forecast on the left, produces lower simulated log-likelihood scores, so that the observed log-likelihood (red line) is within the 95% confidence level and the model passes the L-test on that day. The forecast on the right, which as expected has a higher log-likelihood (black line) also produces higher simulated log-likelihoods, and therefore fails the S-test.

Figure S5. N-test results for the ETAS, RETAS and K2,3 models, using the best catalog and slip models provided with a 10 days delay. Grey bars indicated the forecasted daily number of events. Green dots and red crosses are the observed number of events, with the color indicating that the model passes or fails the N-test at a 95% confidence level. Red circles at the top indicate days in which a model fails the N-test, and the number of events is off scale. The text reports the observed and forecasted number of events. On the top-right in each panel we indicate the total number of failures.

Figure S6. Same as Fig. S5, for the CRS, STEP and SUP models.

Figure S7. Same as Fig. S5, but with “real time” data (preliminary catalog; slip model provided with 10 days delay).

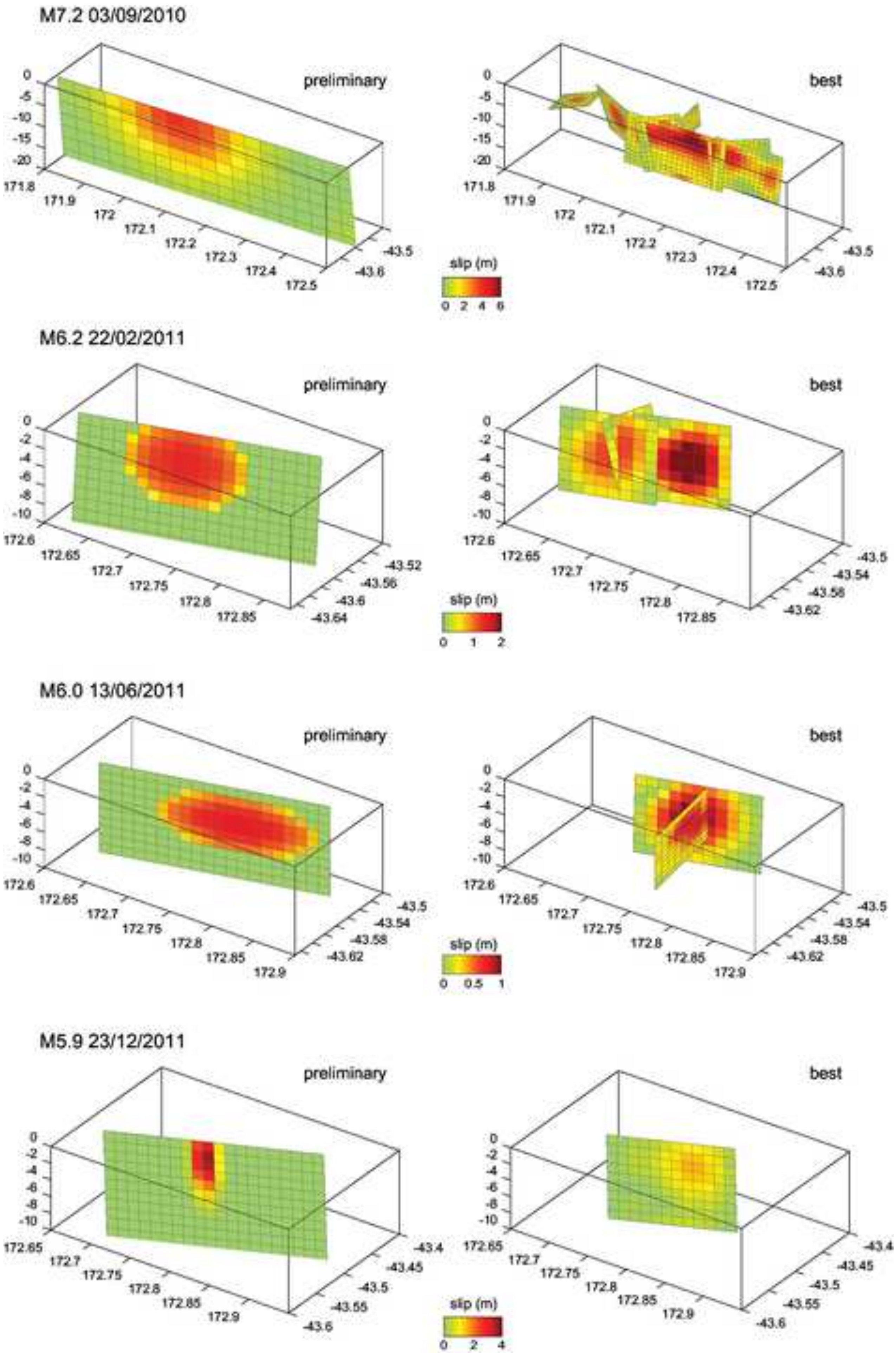
Figure S8. Same as Fig. S7, for the CRS, STEP and SUP models.

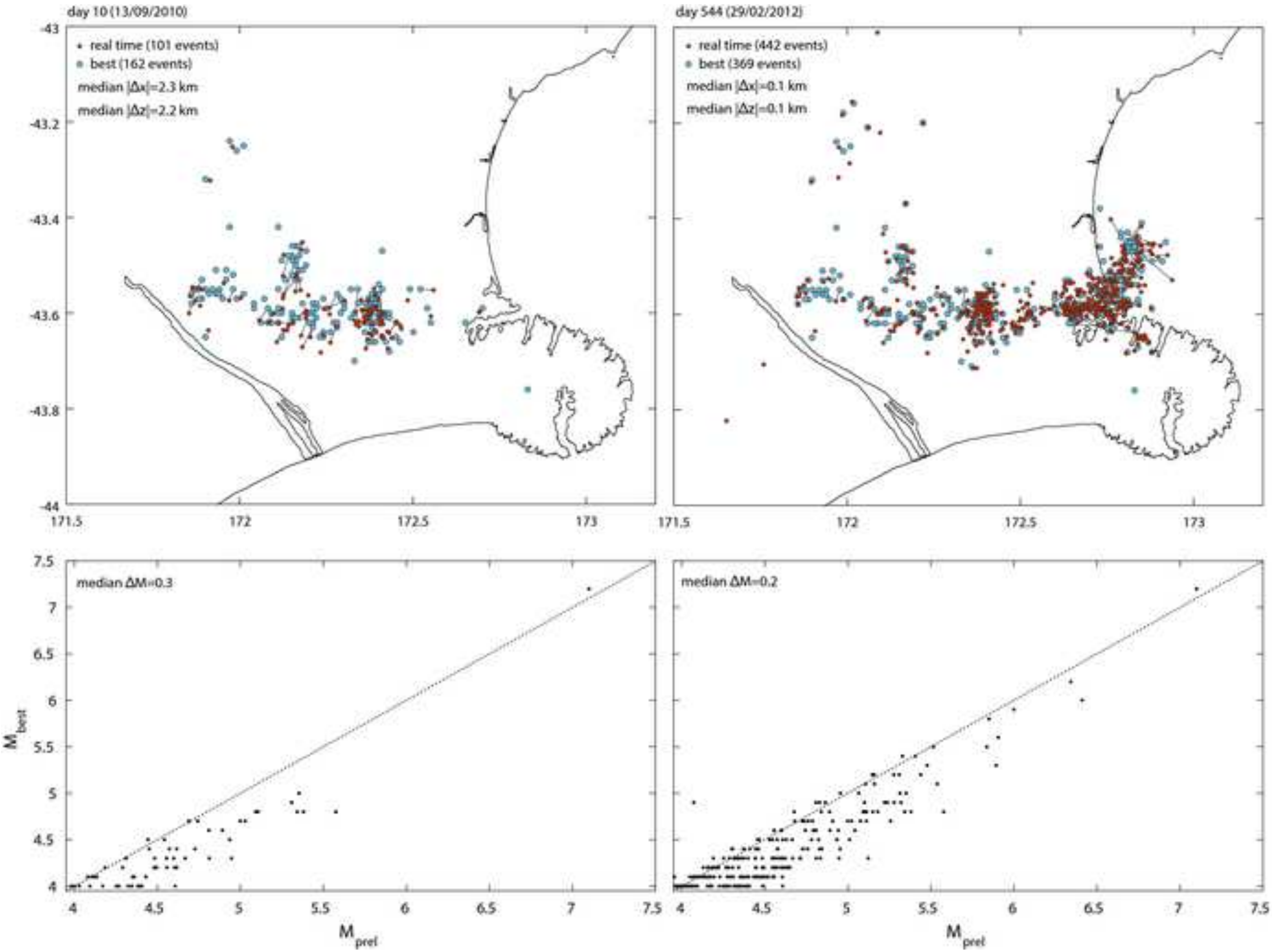
Figure S9. Same as Fig. S5, but with best available data (best catalog and slip model provided without delay).

Figure S10. Same as Fig. S9, for the CRS, STEP and SUP models.

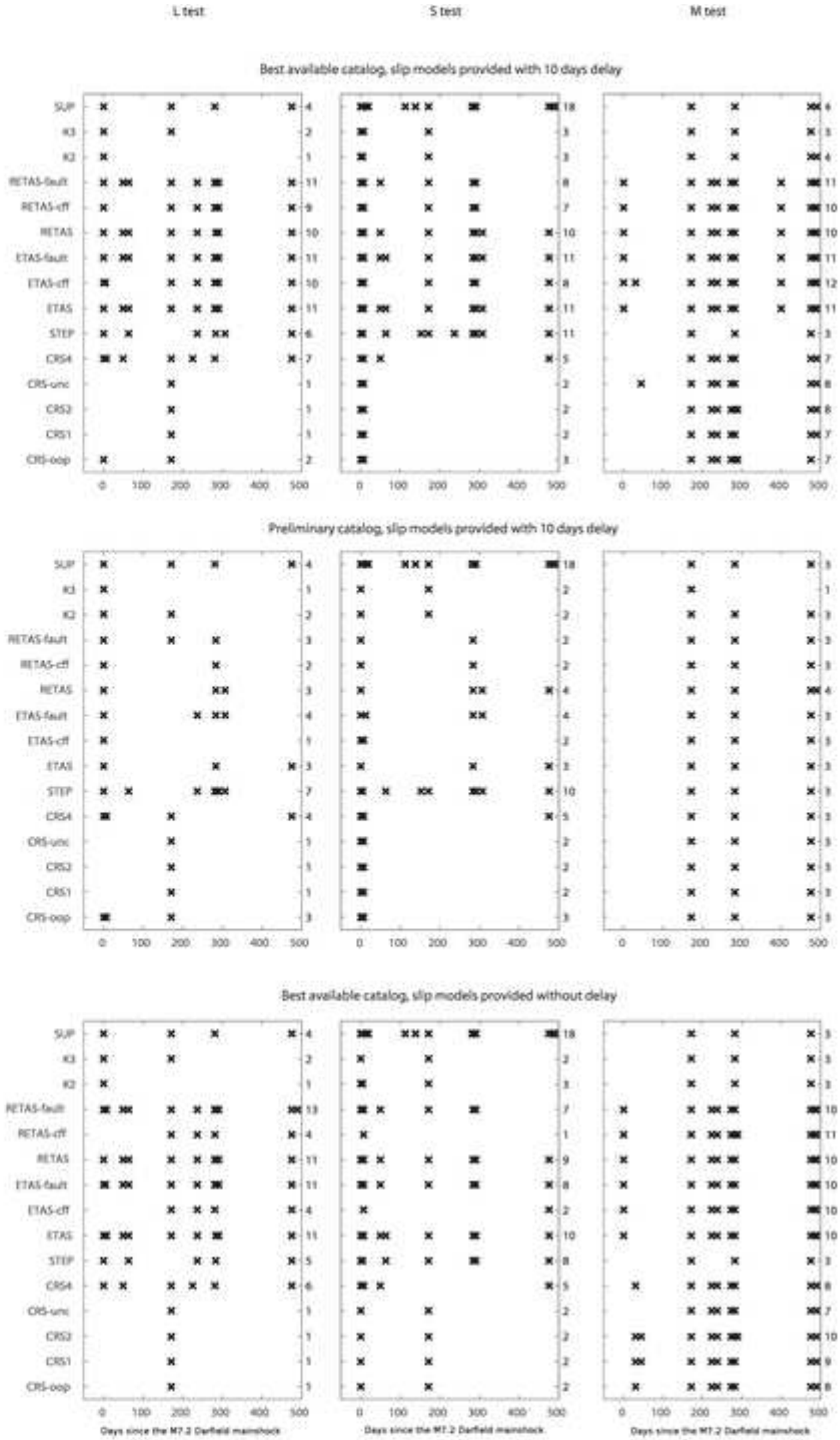
Figure S11. Results of the W-test (Rhoades et al, 2011), which measures whether the information gain between two models is significant at the 95% confidence level. The models on the left of the dotted lines are those discussed in the main text.

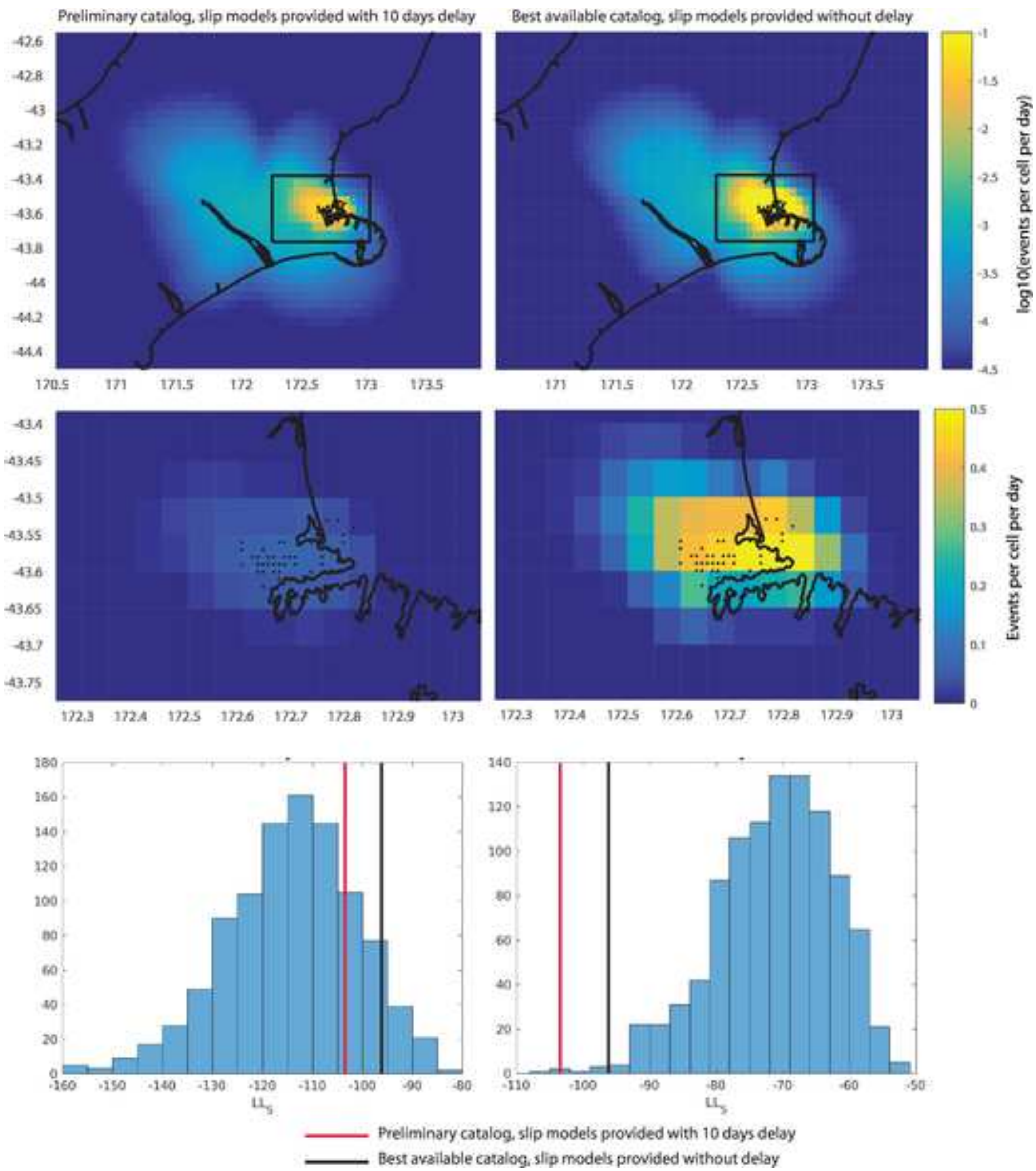
Figure S12. T-test results of all models against the SUP model, as in Fig. 6 (main text), for 15 of the models submitted and all 3 model classes.


M7.2 03/09/2010
preliminary
best
slip (m)
0 2 4 6
M6.2 22/02/2011
preliminary
best
slip (m)
0 1 2
M6.0 13/06/2011
preliminary
best
slip (m)
0 0.5 1
M5.9 23/12/2011
preliminary
best
slip (m)
0 2 4


day 10 (13/09/2010)
real time (101 events)
best (162 events)
median |Δx|=2.3 km
median |Δz|=2.2 km
day 544 (29/02/2012)
real time (442 events)
best (369 events)
median |Δx|=0.1 km
median |Δz|=0.1 km
median ΔM=0.3
median ΔM=0.2
M best
M prel

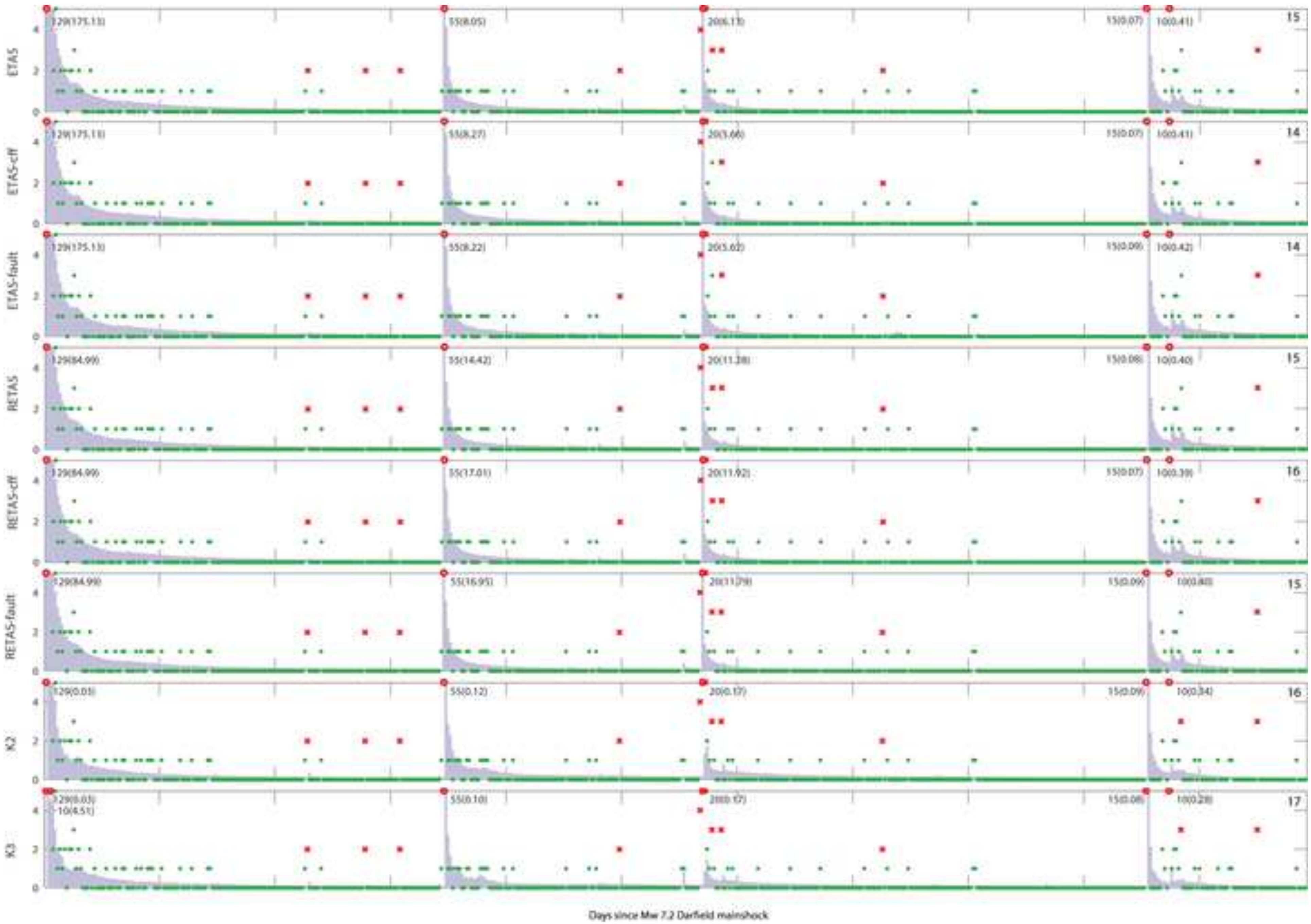
Click here to download Supplemental Material (Main Page, Tables, and Figures) figS3.jpg



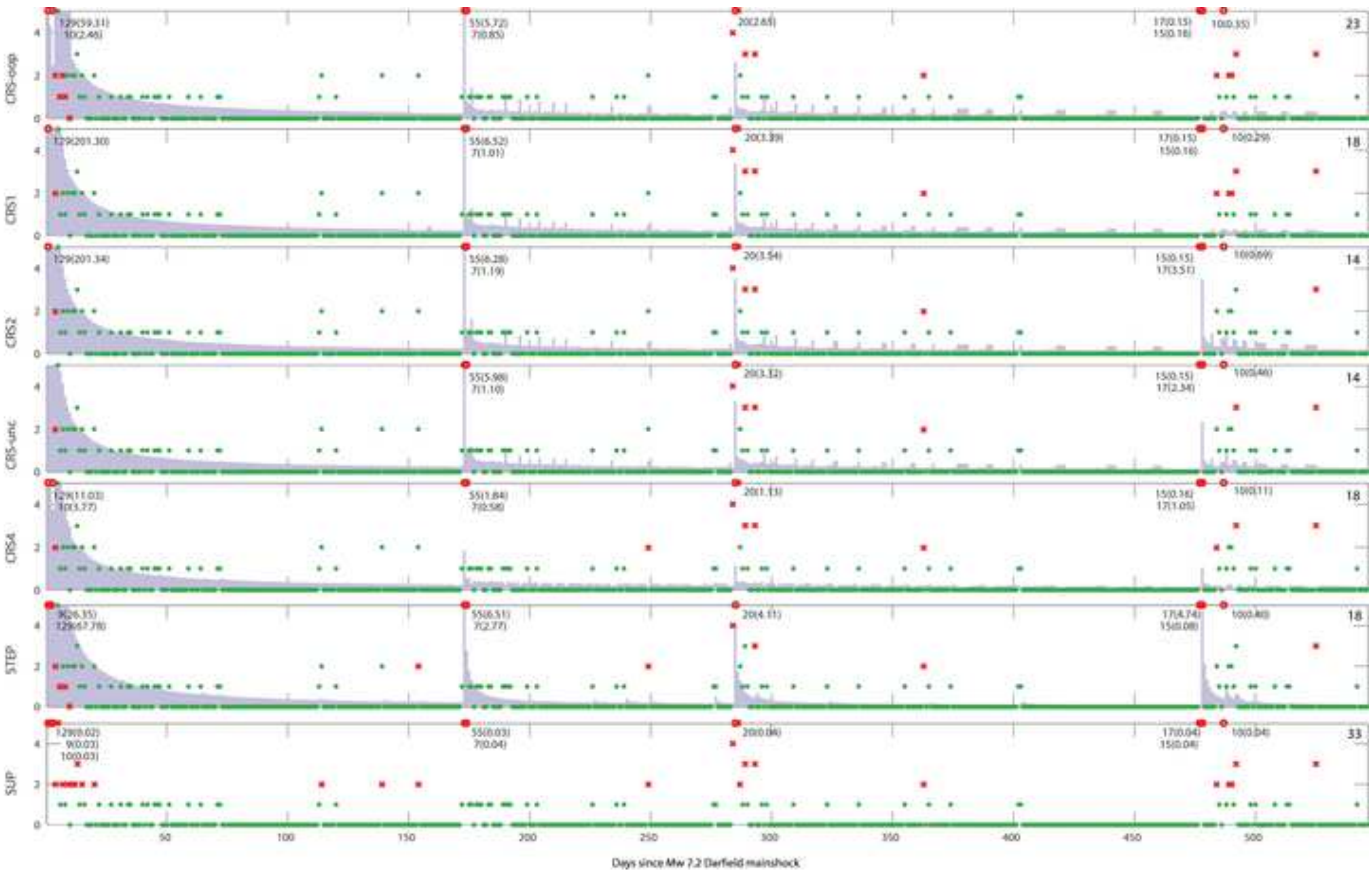


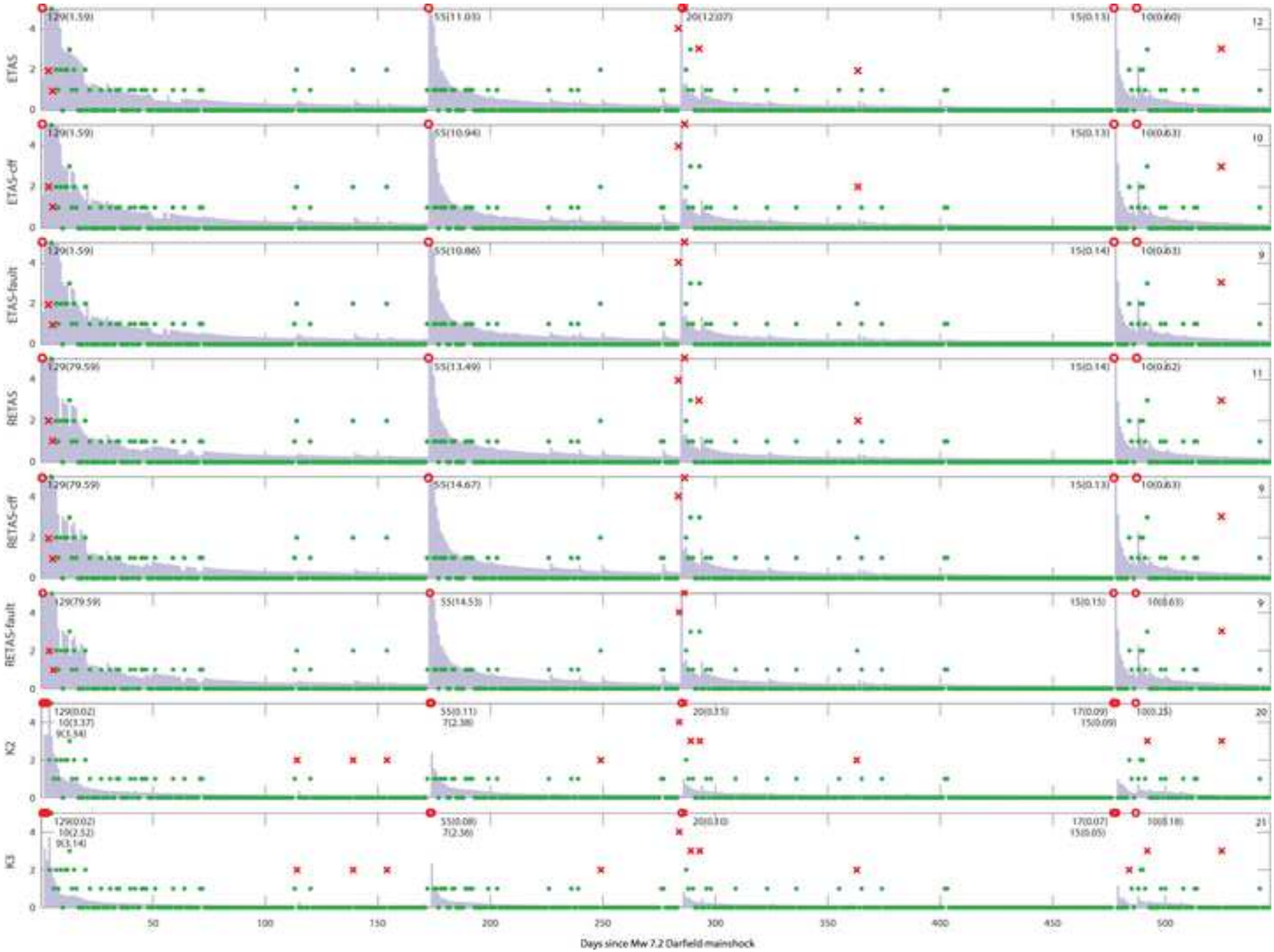
Preliminary catalog, slip models provided with 10 days delay
Best available catalog, slip models provided without delay
log10(events per cell per day)
Events per cell per day
LL_S
Preliminary catalog, slip models provided with 10 days delay
Best available catalog, slip models provided without delay

Click here to download Supplemental Material (Main Page, Tables, and Figures) figS5.jpg

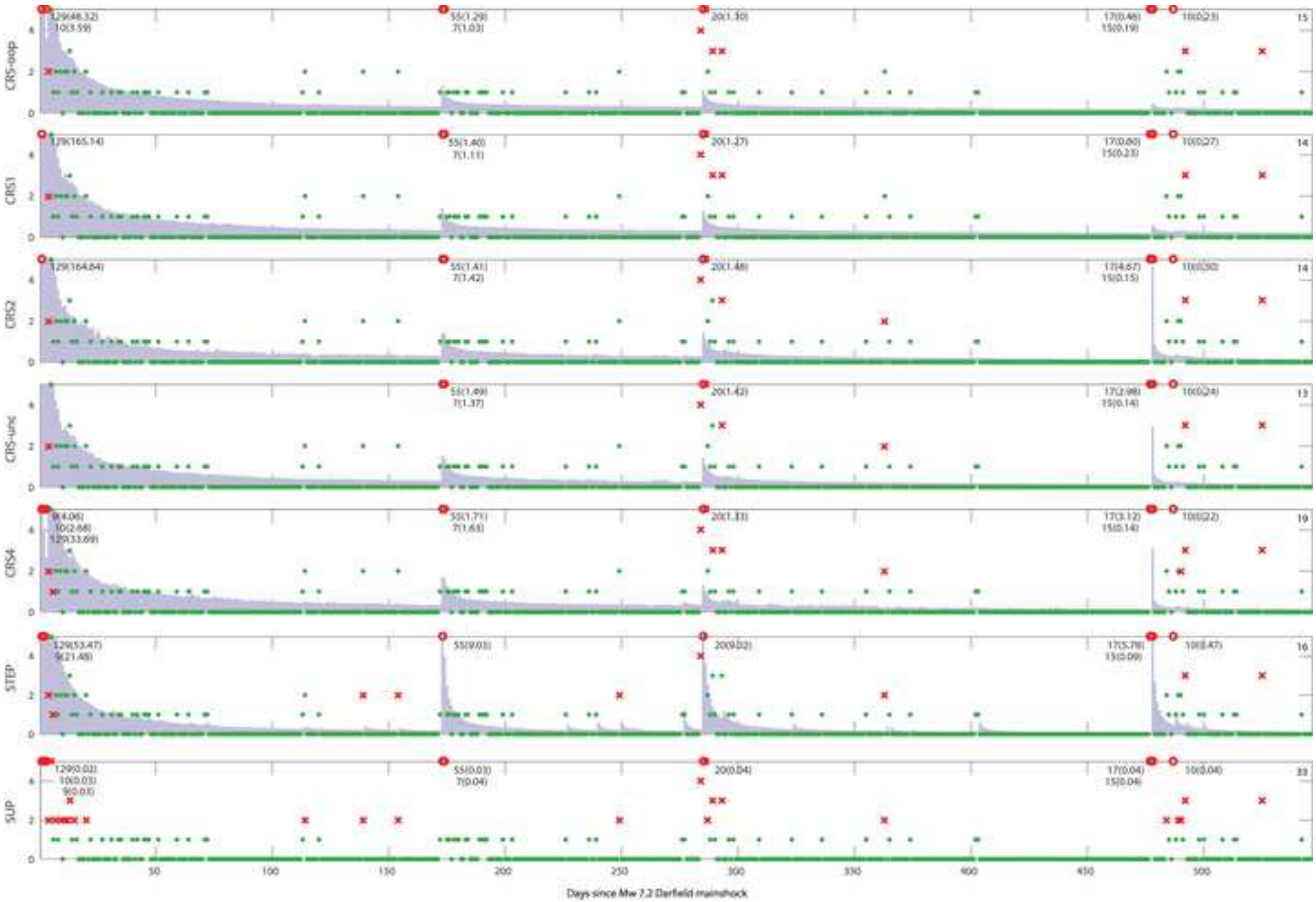


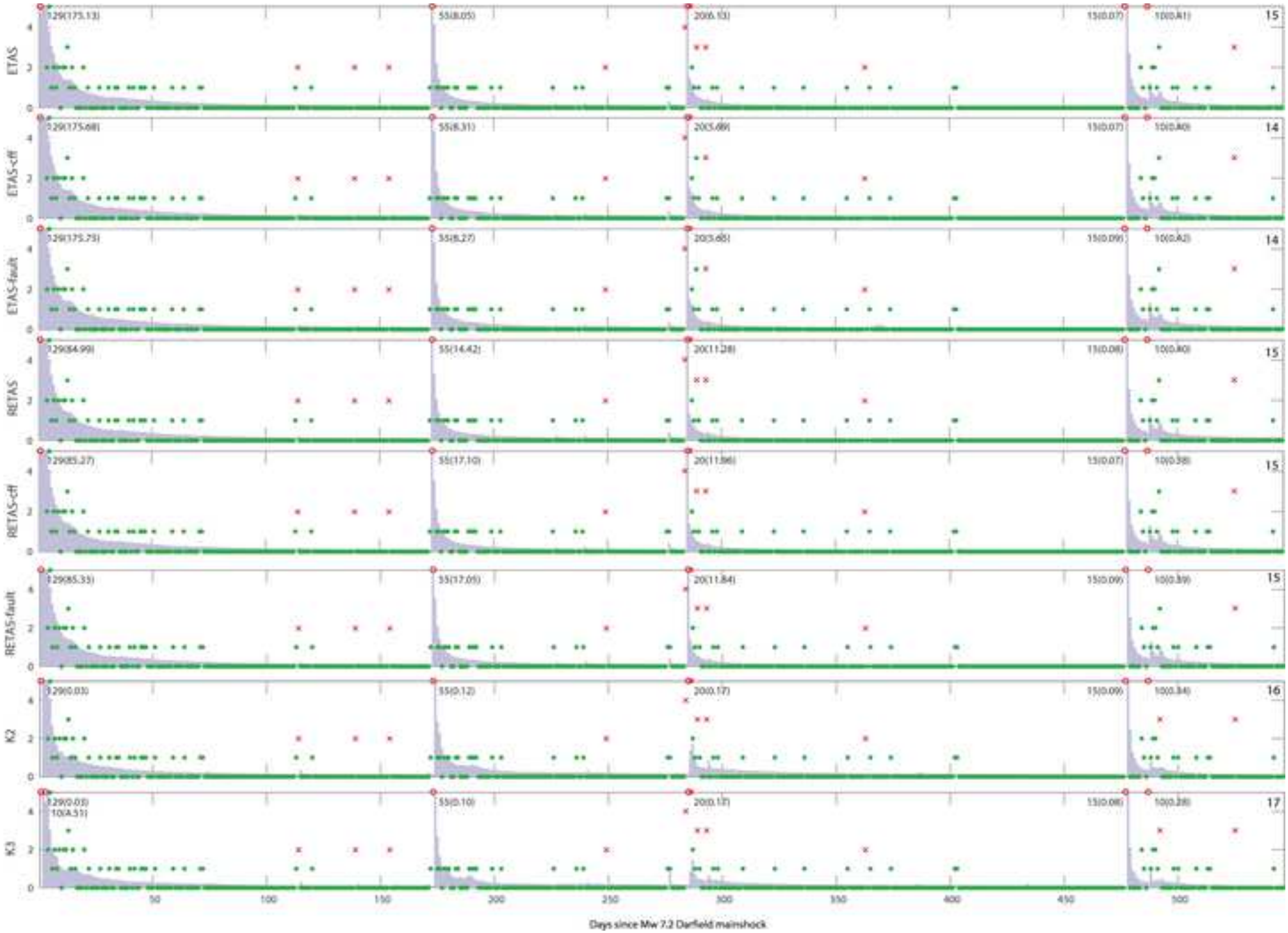
Click here to download Supplemental Material (Main Page, Tables, and Figures) figS6.jpg

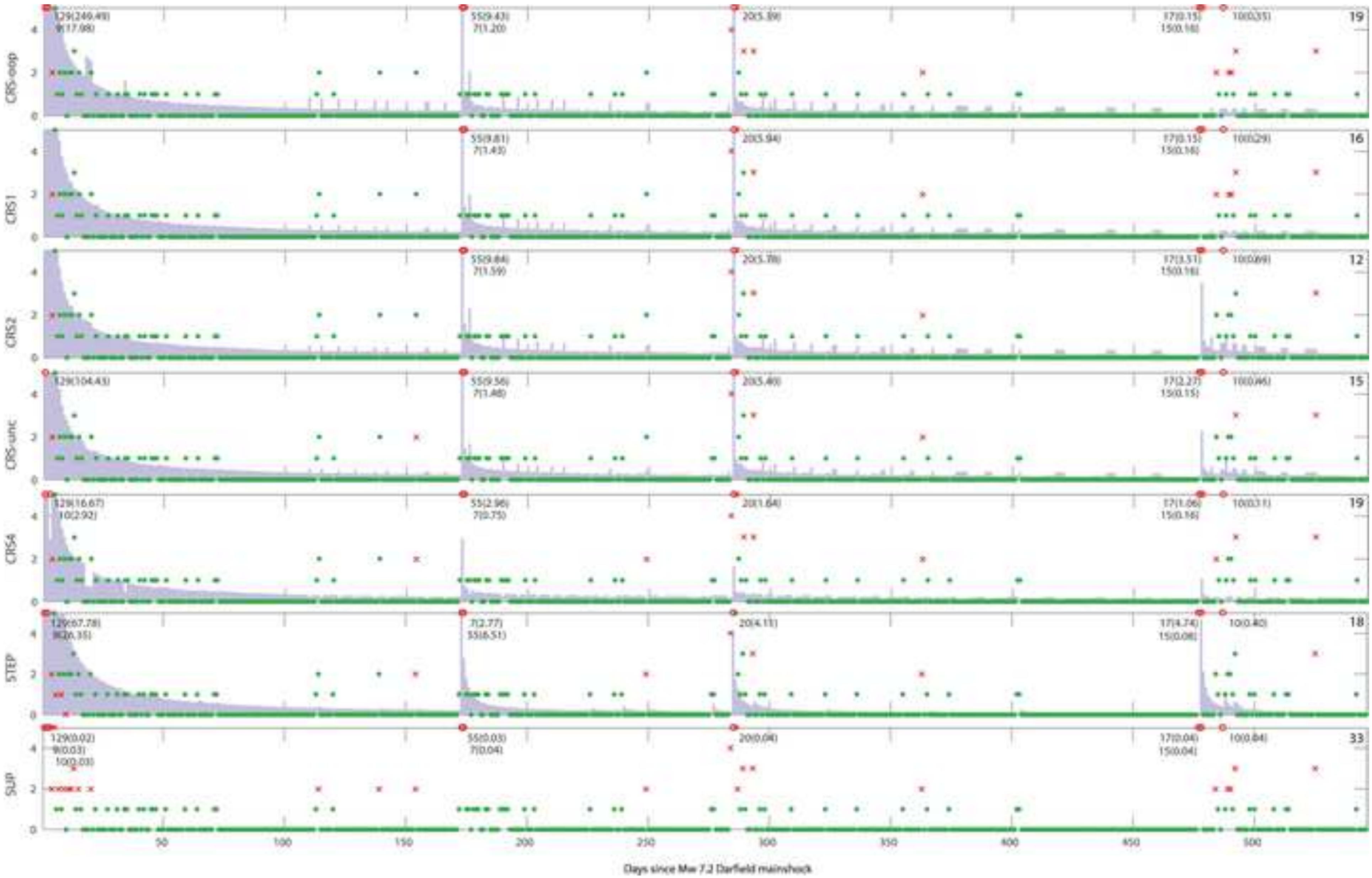



ETAS
ETAS-cff
ETAS-fault
RETAS
RETAS-cff
RETAS-fault
K2
K3
50
100
150
200
250
300
350
400
450
500
Days since Mw 7.2 Darfield mainshock

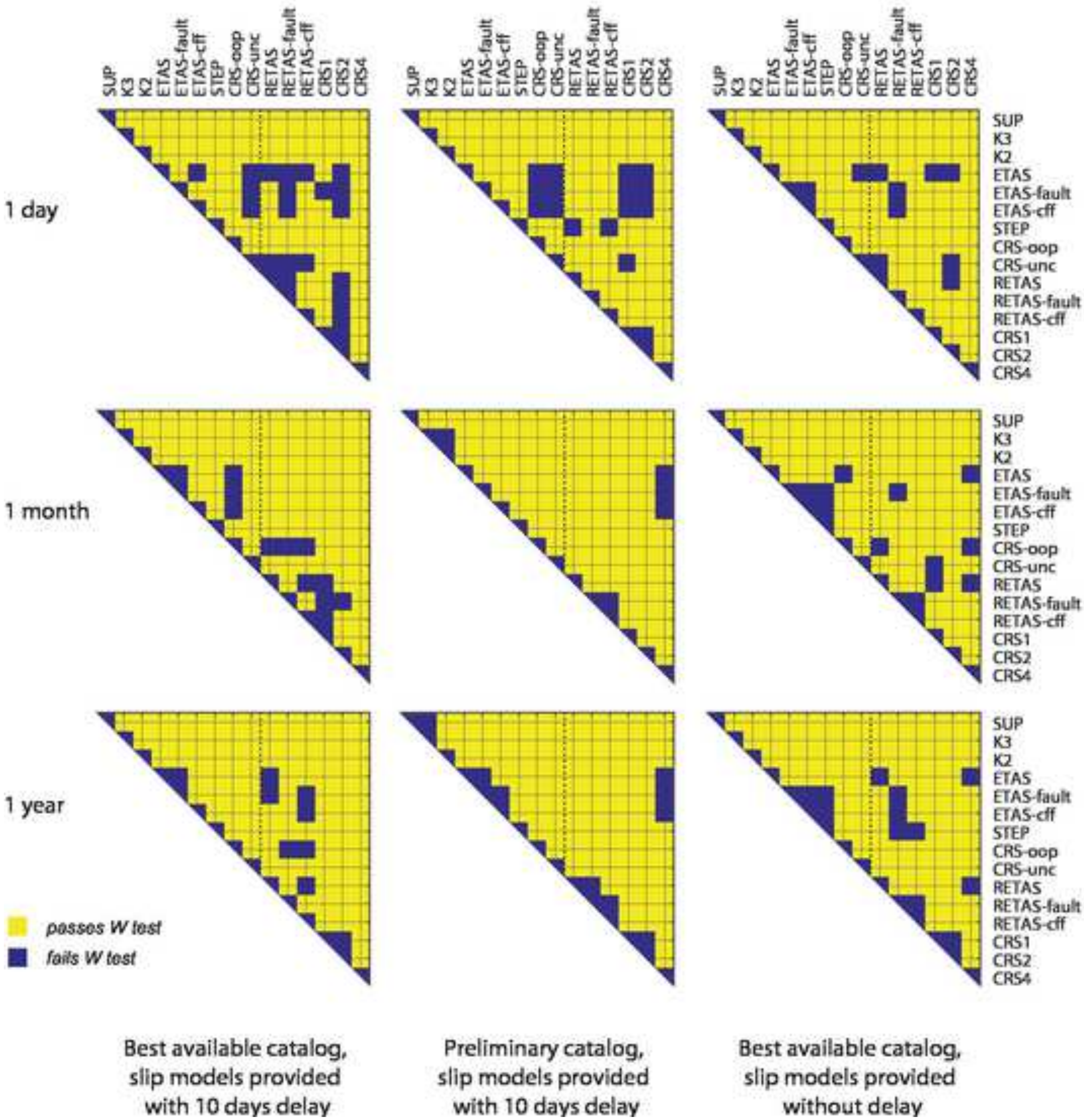
Click here to download Supplemental Material (Main Page, Tables, and Figures) figS8.jpg



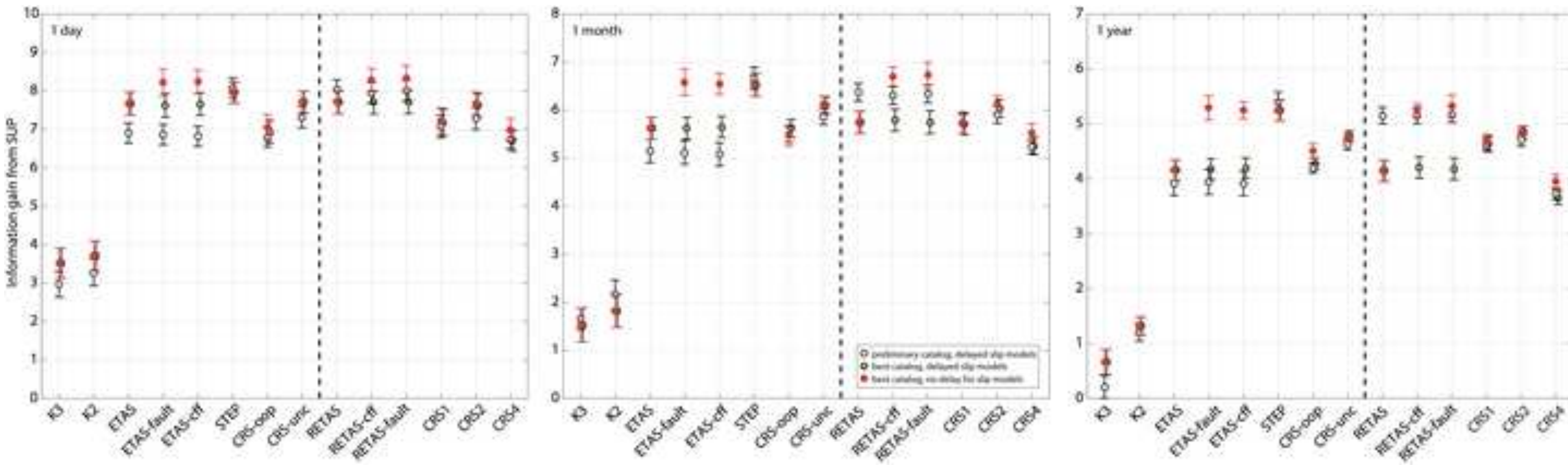

ETAS
ETAS-cff
ETAS-fault
RETAS
RETAS-cff
RETAS-fault
K2
K3
Days since Mw 7.2 Darfield mainshock



CRS-oop
CRS1
CRS2
CRS-unc
CRS4
STEP
SLIP
Days since Mw 7.2 Darfield mainshock



SUP
K3
K2
ETAS
ETAS-fault
ETAS-cff
STEP
CRS-oop
CRS-unc
RETAS
RETAS-fault
RETAS-cff
CRS1
CRS2
CRS4
1 day
1 month
1 year
passes W test
fails W test
Best available catalog, slip models provided with 10 days delay
Preliminary catalog, slip models provided with 10 days delay
Best available catalog, slip models provided without delay


1 day
1 month
1 year
Information gain from SUP
K3
K2
ETAS
ETAS-fault
ETAS-cff
STEP
CRS-oop
CRS-unc
RETAS
RETAS-cff
RETAS-fault
CRS1
CRS2
CRS4