Deliverable 2.9

D2.9: Accuracy and precision of earthquake forecasts using the new generation catalogues for open dissemination.

Deliverable information	
Work package	[WP2: Exploiting innovation, technology advances and opportunities of big data for earthquake loss reduction]
Lead	[INGV]
Authors	Simone Mancini (UniNa, BGS), Marcus Herrmann (UniNa) and Warner Marzocchi (UniNa) Margarita Segou (BGS) Max Werner (Uni-Bristol) Giuseppe Falcone, Lauro Chiaraluce and Maddalena Michele (INGV)
Other contributors	[Tom Parsons (USGS), Greg Beroza (Stanford University)]
Reviewers	[Ian Main]
Approval	[Management Board]
Status	[Final]
Dissemination level	[Public]
Delivery deadline	[28.02.2023]
Submission date	[28.02.2023]
Intranet path	[DELIVERABLES]

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 821115

Table of contents

Summary

This Deliverable benefits from the availability of a suite of high-resolution earthquakes catalogues generated for the 2016-17 Central Italy seismic sequence (Figure 1; Chiaraluce et al. 2022), to produce a set of retrospective earthquake forecasts, including physics-based models such as Coulomb Rate-and-State (CRS) friction and purely statistical ones such as the Epidemic-Type Aftershock Sequence (ETAS) model.

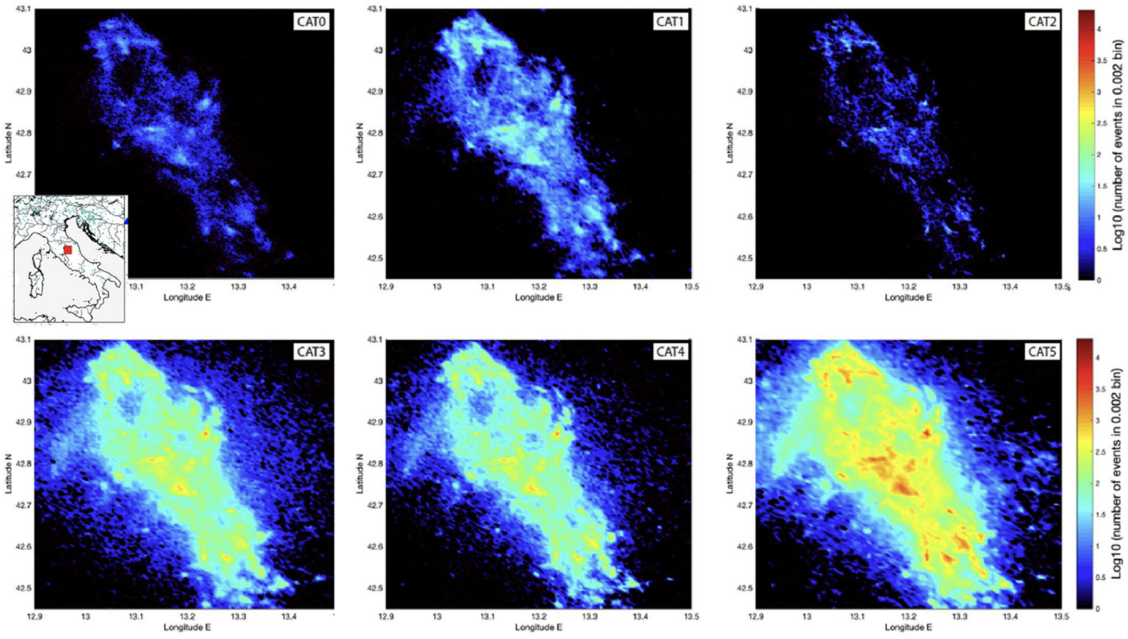

Figure 1. *Maps showing the event density of each catalogue describing the 2016-15 Central Italy seismic sequence, reported as Log10 of the number of events in 0.002 × 0.002 degrees (°) cells (modified from Chiaraluce et al., 2022).*

We provide an evaluation of the comparative performance of forecasts made using the same models but informed by different catalogues with increased resolution in space time and magnitude, including those that could be generated in real-time in a prospective scenario, and those generated by state-of-the-art machine learning techniques. The results allow us to isolate the most beneficial (or detrimental) features of these new catalogues (e.g., increased spatial clustering, event relocations, magnitude re-estimations) for the models' predictive skill, and to evaluate the relative performance of catalogues with gradually decreasing magnitude (and hence triggering) thresholds down to $M_{MIN} = 1$.

1. Introduction

Modern seismic catalogues based on advanced detection algorithms reveal the evolution of earthquake sequences in space and time at a dramatically higher resolution compared to those derived from standard processing workflows (e.g., routine detections, analyst-reviewed travel time measurements). Therefore, it does not come as a surprise that the scientific community devoted to the study of earthquake triggering mechanisms places high hopes on exploiting such datasets in real-

time conditions (Zhu et al., 2022) to ultimately produce improved probabilistic models (Beroza et al., 2021) for operational earthquake forecasting protocols (Jordan et al., 2011). However, how the additional information encoded in those catalogues should be integrated into state-of-the-art modelling strategies is still to be understood.

Over the last few years, prospective and pseudo-prospective experiments dedicated to the development and validation of short-term predictive models have led to a tangible step forward in recognizing which elements boost the performance of physics-based earthquake forecasts (e.g., Cattania et al., 2018; Mancini et al., 2019; 2020), including those based on continuum mechanics such as the Coulomb Rate-and-State (CRS) models. For example, Mancini et al. (2020) quantified how the incorporation of high-quality input data in CRS models makes them potentially as informative as standard statistical or empirical models like the Epidemic-Type Aftershock Sequence (ETAS; Ogata, 1988) model. They concluded that the most critical modelling elements to be considered are: (1) an optimized calibration of the model parameters, (2) the usage of finite-fault slip distributions, (3) the small-scale spatial variability of receiver faults with rupture kinematics informed from multiple data sources (e.g., past and unfolding focal mechanism solutions, smoothed regional stress inversions, mapped active faults, etc.), and (4) the effects of secondary triggering from smaller-magnitude events. The latter notion has also been supported in the case of purely statistical models (e.g., Werner et al., 2011) and in fully prospective CSEP tests (e.g., Bayona et al., 2022).

An impressive suite of enhanced earthquake catalogues has been recently developed and released for the 2016-2017 Amatrice-Visso-Norcia (hereinafter, AVN) seismic cascade in Italy within the jointly funded NERC-NSFGEO project "The Central Apennines Earthquake Cascade Under a New Microscope"; this makes it the most appropriate case study to test the efficacy of the new generation datasets for earthquake forecasting purposes.

We use the three high-resolution and deep-learning catalogues available for the AVN sequence, featuring a magnitude of completeness up to two units lower than the previous real-time catalogue, to produce a set of CRS and ETAS forecasts and to validate them retrospectively. Our aim is to: (*i*) compare their performance against that of the same models informed by real-time data only, (*ii*) evaluate which key features of these new catalogues (e.g., increased spatial clustering, event relocations, magnitude re-estimations) are the most beneficial or detrimental for the models' predictive skill, and (*iii*) assess whether the performance of the forecasts improve when the assumed minimum triggering magnitude gradually decreases until $M_{MIN} = 1$.

2. Employed datasets.

The forecasts presented in this report are informed by four earthquake catalogues for the first year of the AVN sequence, each of which is a result of a different workflow in terms of network geometries, detection, arrival time measurements, phase associations, locations, and magnitudes of increasing sophistication. The main features of the employed catalogues are reported in Table

1. For a more detailed description of the development process leading to each catalogue and a thorough comparative illustration of the datasets used in this study, see Chiaraluce et al. (2022).

Table 1. General features of the four catalogues used in this study

Name	Start	End	Total number of events	Type of analysis	Magnitude type	Mc
CAT0	24-08-2016	31-08-2017	73,009	Real-time	ML	2.3
CAT3	24-08-2016	31-08-2017	440,727	Offline	ML	0.4
CAT4	24-08-2016	31-08-2017	390,336	Offline	ML	0.4
CAT5	15-08-2016	15-08-2017	900,058	Offline	Mw	0.2

CAT0 is the real-time catalogue obtained by the Italian National Institute of Geophysics and Volcanology (INGV) monitoring system from data collected by the permanent Italian National Seismic Network (ISIDe Working Group, 2007). We choose not to consider the CAT1 and CAT2 catalogues as they are different realizations of the real-time catalogue, featuring the same CAT0 detections and magnitudes. The successively released catalogues all benefit from a much denser seismic network of 155 permanent and temporary stations deployed in the affected area following the M6.0 Amatrice earthquake (Moretti et al., 2016), and are here categorized as the 'enhanced' catalogues. In CAT3, (Spallarossa et al., 2020), detections are generated by an automated picker (Spallarossa et al., 2014) while the absolute hypocentres are obtained by a non-linear location algorithm (Lomax et al., 2000). CAT3 also features an automated re-evaluation of local magnitudes. Starting from the CAT3 locations, Waldhauser et al. (2021) applied a double-difference relocation algorithm with cross-correlation-based arrival time measurements to reduce the location error to only a few tens of meters, obtaining the high-resolution CAT4 catalogue. Finally, we use the deep-learning-derived CAT5 by Tan et al. (2021) which is the largest catalogue released so far for the sequence because of the efficiency with which this machine learning approach detects numerous small events. In CAT5, earthquakes are detected using the PhaseNet picker (Zhu & Beroza, 2019) based on a deep-neural network, and relative locations are obtained by means of the hypoDD double-difference method (Waldhauser & Ellsworth, 2000), but without the benefit of cross-correlation-based arrival time measurements of CAT4. Since both CAT4 and CAT5 feature high-precision relative relocations, we refer to them as the 'high-resolution' catalogues.

To calibrate the models parameters, we use the same data as Mancini et al. (2019) to ensure consistency. They fit the rate-and-state and ETAS parameters on the M3+ pre-sequence catalogue ('learning phase catalogue') of the Italian Seismological Instrumental and Parametric Database (1990-2016 and 2005-2016 time periods) for CRS and ETAS models, respectively. Likewise, our CRS models employ their set of finite-fault slip models (Chiaraluce et al., 2017). To define the receiver-fault matrix of the CRS models, we use their combination of kinematic parameters of large-scale fault structures of the Central Apennines as described by the Database of Individual Seismogenic Sources (DISS Working Group, 2018), and focal mechanisms for the CRS learning phase reported in the Italian centroid moment tensor catalogue.

3. Experimental setup

Our forecasts adhere to the following rules:

- A 6-month forecast horizon (24 August 2016 – 24 February 2017), except for the CRS model developed using CAT5 along with all ETAS models that are limited to the high-rate period of the first 3 months (24 August 2016 – 24 November 2016) for reasons of computational limitations.
- A 2D testing region of about ~150 x 150 km centred on the M6.0 Amatrice earthquake, subdivided in 0.02° (~2 km) wide square bins.
- A model update frequency of 24 hours or at the occurrence of a M5.4+ events.
- For the stress-based models, static stress changes are calculated between 0-12 km of depth in cubic bins.

We generate six new forecast versions based on the newly available catalogues: CRS-CAT3, ETAS-CAT3, CRS-CAT4, ETAS-CAT4, CRS-CAT5, and ETAS-CAT5. Supported by the improvements in the completeness of the enhanced earthquake catalogues, these models allow secondary triggering over a wide range of minimum triggering magnitudes than before, down to the completeness threshold, here $M_{MIN} = 1$ (Figure **2**b-d). These new models are benchmarked versus the best-performing forecasts by Mancini et al. (2019), that we rename here as CRS-CAT0 and ETAS-CAT0 since they were developed using the real-time catalogue as 'input seismicity'. The main features of the CRS models considered in this study are summarized in Table 2. The ETAS-CAT*x* models (with x=3,4,5) present the same ETAS-CAT0 parameterization (with parameters fixed for the whole forecast horizon) but reduce their minimum triggering magnitude until $M_{MIN} = 1$.

Table 2. Main attributes of the CRS models. M_{MIN} = minimum magnitude for stress sources; USD = uniform slip distribution on a synthetic fault with empirically-derived dimensions and random selection of nodal plane from its moment tensor solution; FFM = finite-fault rupture model; I = magnitude-dependent isotropic stress field; SUP = spatially uniform receiver planes; SVP = spatially variable planes derived from focal mechanisms included in the CRS learning phase catalogue and from the DISS database; μ'= coefficient of effective friction.
Note: the background rate (r_0) is estimated using the CRS 'learning phase' catalogue.

Model name	Input catalogue	Stress Calculations					Rate-and-State Parameters (Optimised on learning catalogue)		
		Secondary Triggering	M_{MIN}	Slip Distribution	μ'	Receiver faults	r_0	$A\sigma$ (MPa)	$\dot{\tau}$ (MPa/yr)
CRS-CAT0	CAT0	Yes	3.0	FFM (M ≥ 5.4) USD (M ≥ 4.0) I (M ≥ 3.0)	0.4	SVP	Spatially heterogeneous	0.015	0.00019
CRS-CAT3	CAT3	Yes	1.0	FFM (M ≥ 5.4) USD (M ≥ 4.0) I (M ≥ 1.0)	0.4	SVP	Spatially heterogeneous	0.015	0.00019
CRS-CAT4	CAT4	Yes	1.0	FFM (M ≥ 5.4) USD (M ≥ 4.0) I (M ≥ 1.0)	0.4	SVP	Spatially heterogeneous	0.015	0.00019
CRS-CAT5	CAT5	Yes	1.0	FFM (M ≥ 5.4) USD (M ≥ 4.0) I (M ≥ 1.0)	0.4	SVP	Spatially heterogeneous	0.015	0.00019

Each forecast is formally evaluated against the M3+ seismicity reported in each catalogue, here named as the 'target catalogue' or 'target seismicity' (Figure 2e-h). We use standard metrics, such as the likelihood-based S-test (Zechar et al., 2010) to assess the absolute spatial consistency betweeen the forecasts with the outcome as well as the information gain per earthquake (IG; Rhoades et al., 2011) for the relative model ranking as established by standard practice in CSEP (Collaboratory for the Evaluation of Earthquake Predictability).

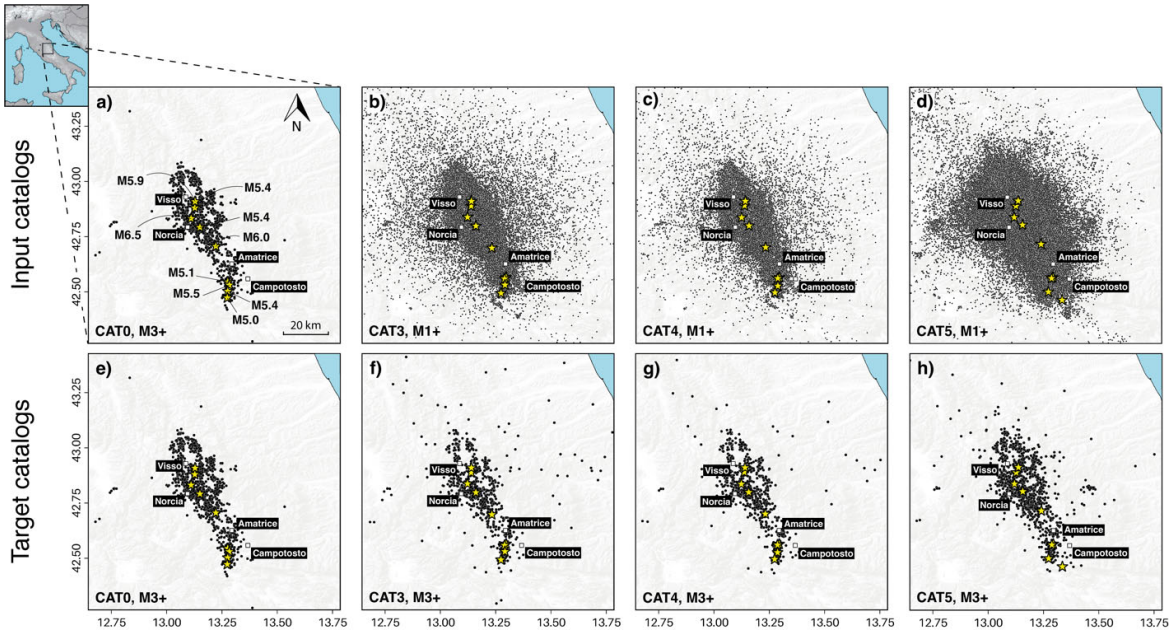

Figure 2. *Input and target earthquake catalogues used in this study. Input catalogues inform the development of the models, targets catalogues are used to assess the performance of the forecasts. Yellow stars indicate the location of M5+ earthquakes.*

4. Forecast results and validation

A preliminary visual inspection of Figures 3 and 4, which respectively show the expected rates of the CRS and ETAS forecasts, reveals a good agreement between the aftershock areas projected by the models and those outlined by the M3+ target seismicity in each of the four catalogues. Overall, the CRS and ETAS models developed with CAT0 and the counterparts employing the enhanced catalogues present only subtle differences, mostly located in the near-source area of the testing region. However, some of the off-fault seismicity patterns forecast from the enhanced catalogues are not well captured by the CRS/ETAS-CAT0 models, especially those in CAT5 (Figure 3b-d and Figure 4b-d). On the other hand, the CRS/ETAS-CAT0 models expect heightened seismicity rates to the north-east of the main fault system where no M3+ triggered events were reported in real time. This could be interpreted as a failure of that model, but the new CAT5 data include such events, which is consistent with that model (Figure 3d).

The incorporation of secondary triggering from M1+ events in CRS forecasts (Figure 3e-j) can now locally explain the occurrence of isolated aftershocks within the stress shadows cast by the mainshocks (e.g., CRS-CAT4, Figure 3h,i). However, some triggered earthquakes reported in the enhanced catalogues continue to occur in regions of expected seismicity suppression. In CRS-CAT4 we observe some differences with respect to CRS-CAT3 that are likely an effect of its relocation process, namely increased expected rates at (1) the edges of main aftershock region, and (2) the near-epicentral area of the Visso and Norcia events. By contrast, the 3-month CRS-CAT5 does not present striking differences with respect to CRS-CAT4 by visual inspection. Similarly, there are only minor large-scale visual differences between the preliminary ETAS (Figure 4a-d) and the updated ETAS models implementing M1+ parent events (Figure 4e-j).

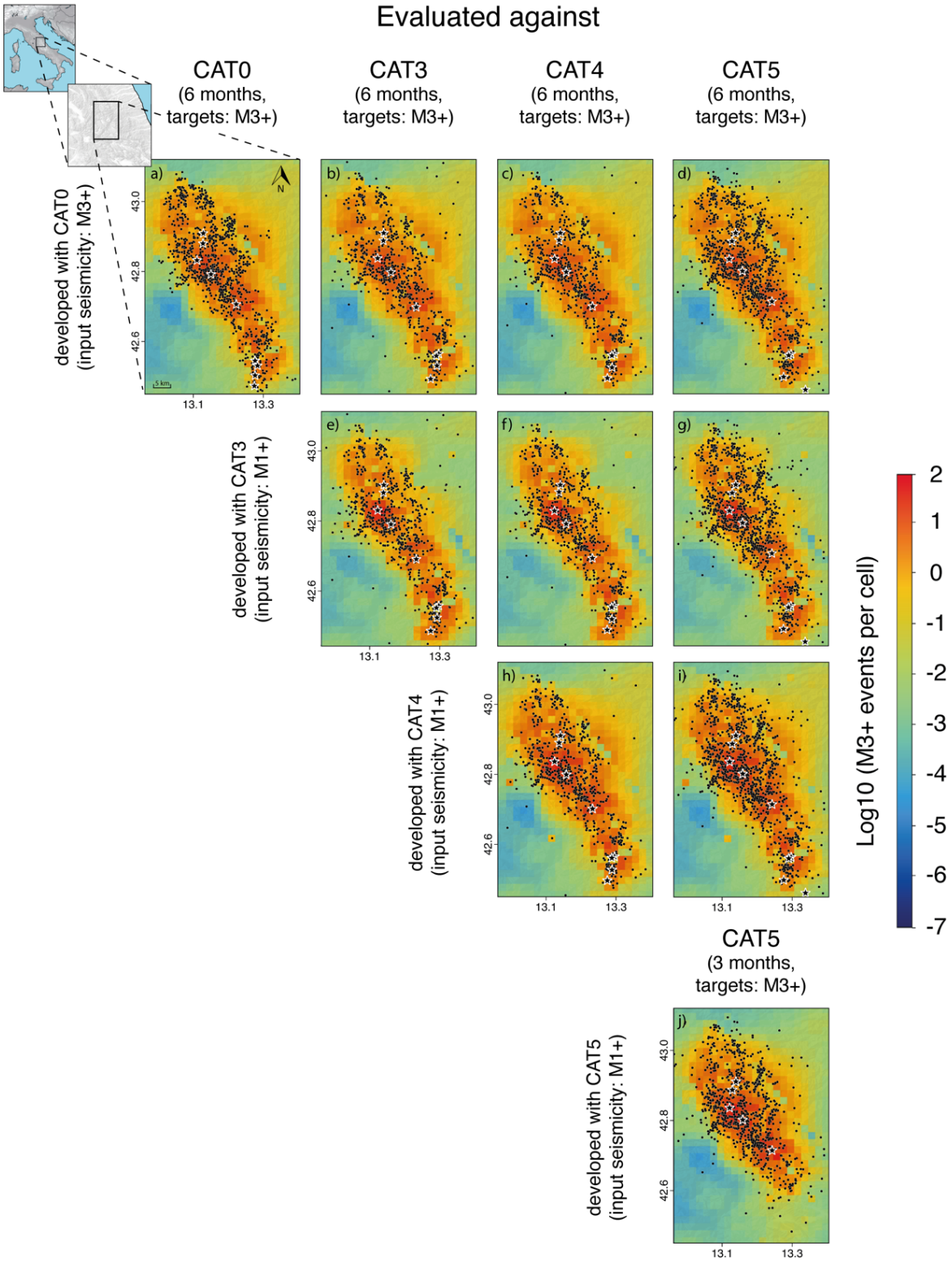

Figure 3. *Maps of expected seismicity rate for the Coulomb Rate-State (CRS) models developed with and evaluated against the four catalogues. CAT0, CAT3 and CAT4 models cover a 6-month forecast period, while CRS-CAT5 has a 3-month horizon. Each rate map is overlain with the corresponding target seismicity for the periods of interest: black stars for the M5+ earthquakes and black dots for the 3≤M<5 events.*

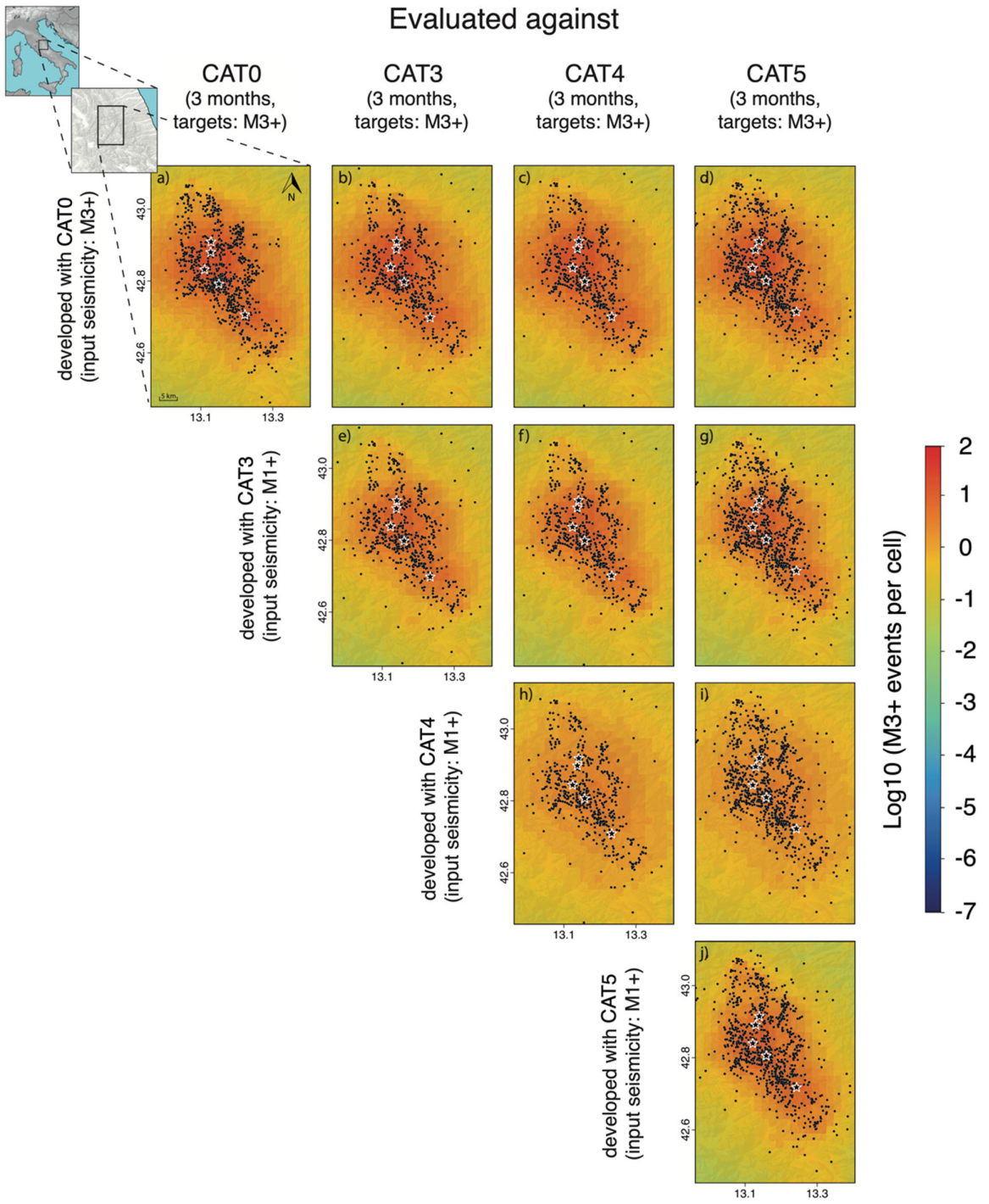

Figure 4. *Maps of expected seismicity rate for the Epidemic-Type Aftershock Sequence (ETAS) models developed with and evaluated against the four catalogue generations for a 3-month period. Each rate map is overlaid with the corresponding target seismicity for the periods of interest: black stars for the M5+ earthquakes and black dots for the 3≤M<5 events.*

Figures 5 and 6 provide maps of the S-test's log-likelihood (LL_S) scores for the testing region, aggregated over multiple daily forecast windows. Overall, we observe that for all input-target catalogue combinations, the ETAS joint log-likelihood (jLL_S) scores are higher than the CRS counterparts (see the values reported at the bottom-right corner of each panel), indicating a better spatial consistency. We find that most CRS models developed with enhanced catalogues (*i.e.*, incorporating M1+ stress sources) are not able to overcome the low LL_S values of CRS-CAT0 in the critical high-clustering region around Mt. Bove (Figure 4a), which is the northern termination of the Mt. Vettore fault system activated by the M6.5 Norcia mainshock. Furthermore, we find that

all CRS models evaluated using the enhanced catalogues (Figure 5b-j) suffer from the presence of the now revealed, sparse off-fault target seismicity which was instead undetected in CAT0. Models validated against CAT5 observations present the lowest jLL_S scores, presumably because the likelihood values are locally altered by tightly clustered target seismicity in off-fault regions (Figure 5j). On the other hand, the highest jLL_S scores are obtained when the CRS-CAT0 model is evaluated vs. CAT3 and CAT4 catalogues (Figure 5b,c), suggesting that the near real-time forecast solely incorporating M3+ stress sources had already a satisfactory spatial performance. This latter observation is also confirmed in the case of the ETAS models, where the ranking based on the LL_S values reveals that the near real-time ETAS-CAT0 evaluated against the CAT3 and CAT4 catalogues is among the best-performing models (Figure 6).

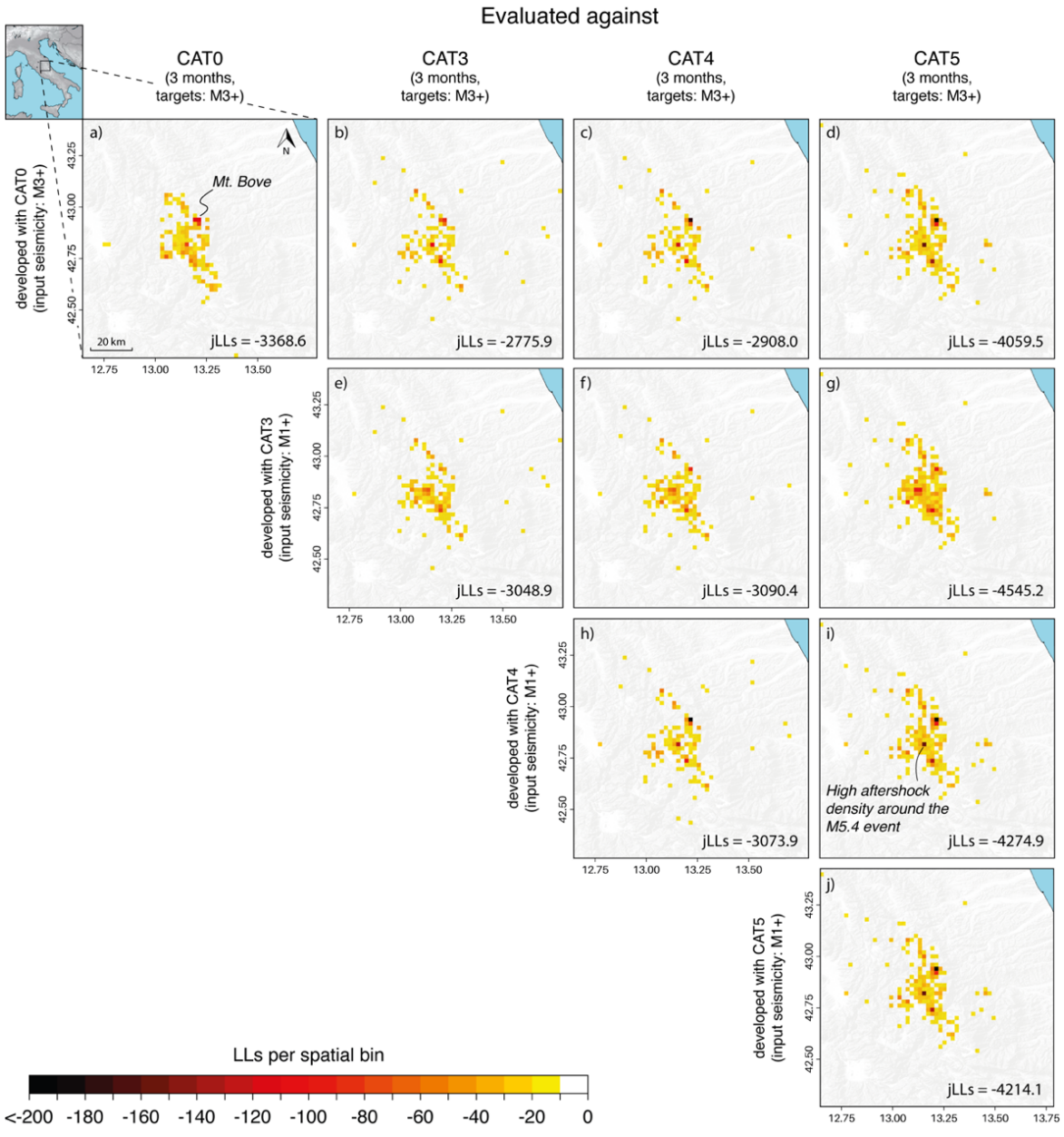

Figure 5. *S-test's log-likelihood (LL_S) maps for the CRS models developed with and evaluated against the four generations of catalogue. In each cell, LL_S values are aggregated over single daily forecast windows for a total period of 3 months. For each model, we report its joint log-likelihood value when it is validated vs. catalogues that are either equal or more evolved than the one used for its development.*

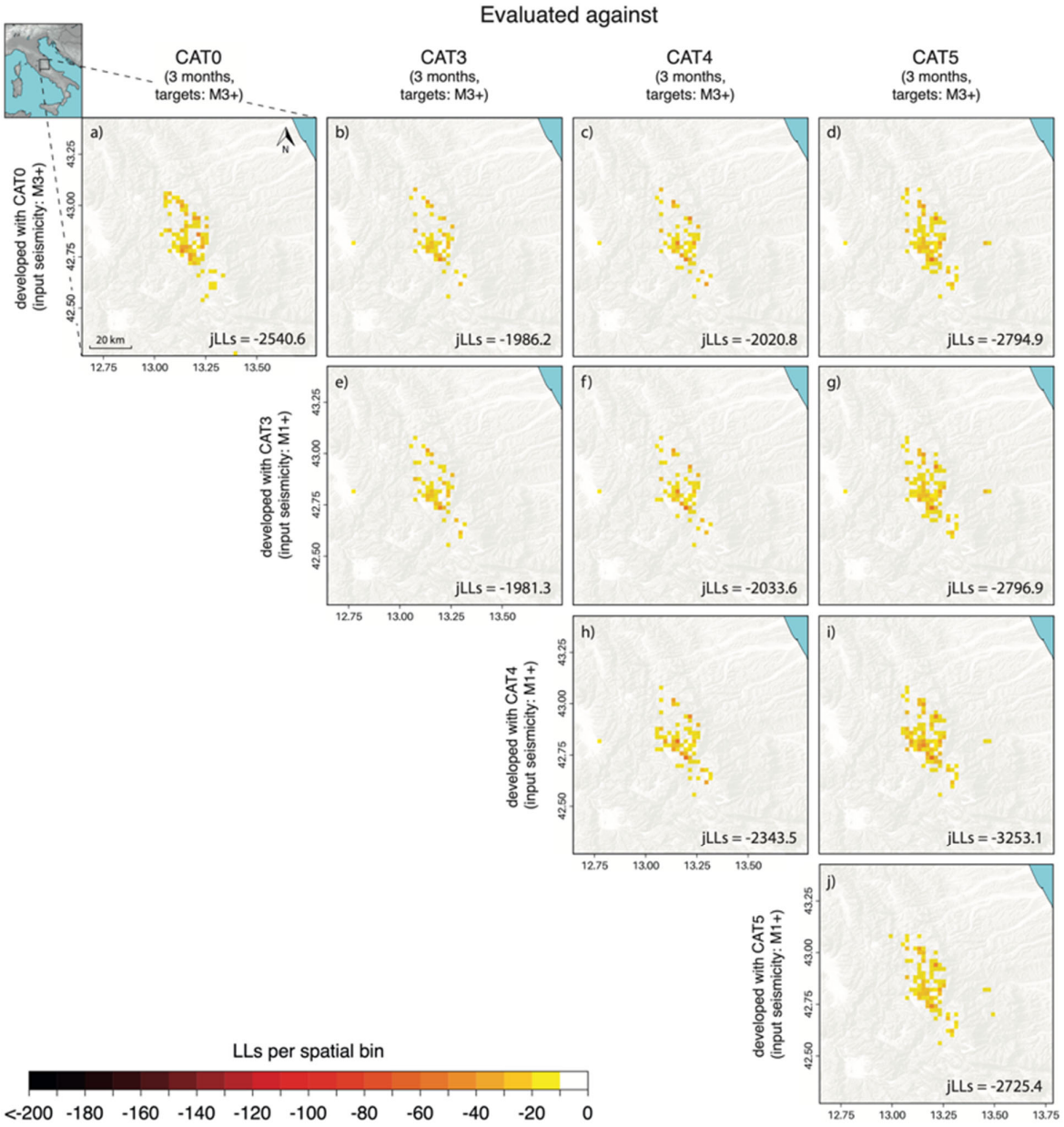

Figure 6. *S-test's log-likelihood maps for the ETAS models developed with and evaluated against the four catalogue generations for a 3-month period. For each model, we report its joint log-likelihood value when it is validated vs. catalogues that are either equal or more evolved than the one used for its development.*

To assess the overall skill of the forecasts generated by the new models (*i.e.*, considering model performance in both space and time domains) compared to the near real-time forecasts, we calculate their information gain per earthquake (IG) on the respective "CAT0" realization. To standardize this test and to facilitate the interpretation of the results, we select a common 3-month testing phase for all models and employ a consistent testing catalogue for each model-benchmark couple (*i.e.*, the likelihoods of model CRS/ETAS-CAT_i, with i = 0,3,4,5, and of the benchmark CRS/ETAS-CAT0 are both calculated against CAT_j, with j = 4,5). Furthermore, to quantify the effect of incorporating gradually more complete input catalogues in the forecasting protocols, for each model we illustrate how the IG scores vary when we consider different M_{MIN} thresholds (Figure 7).

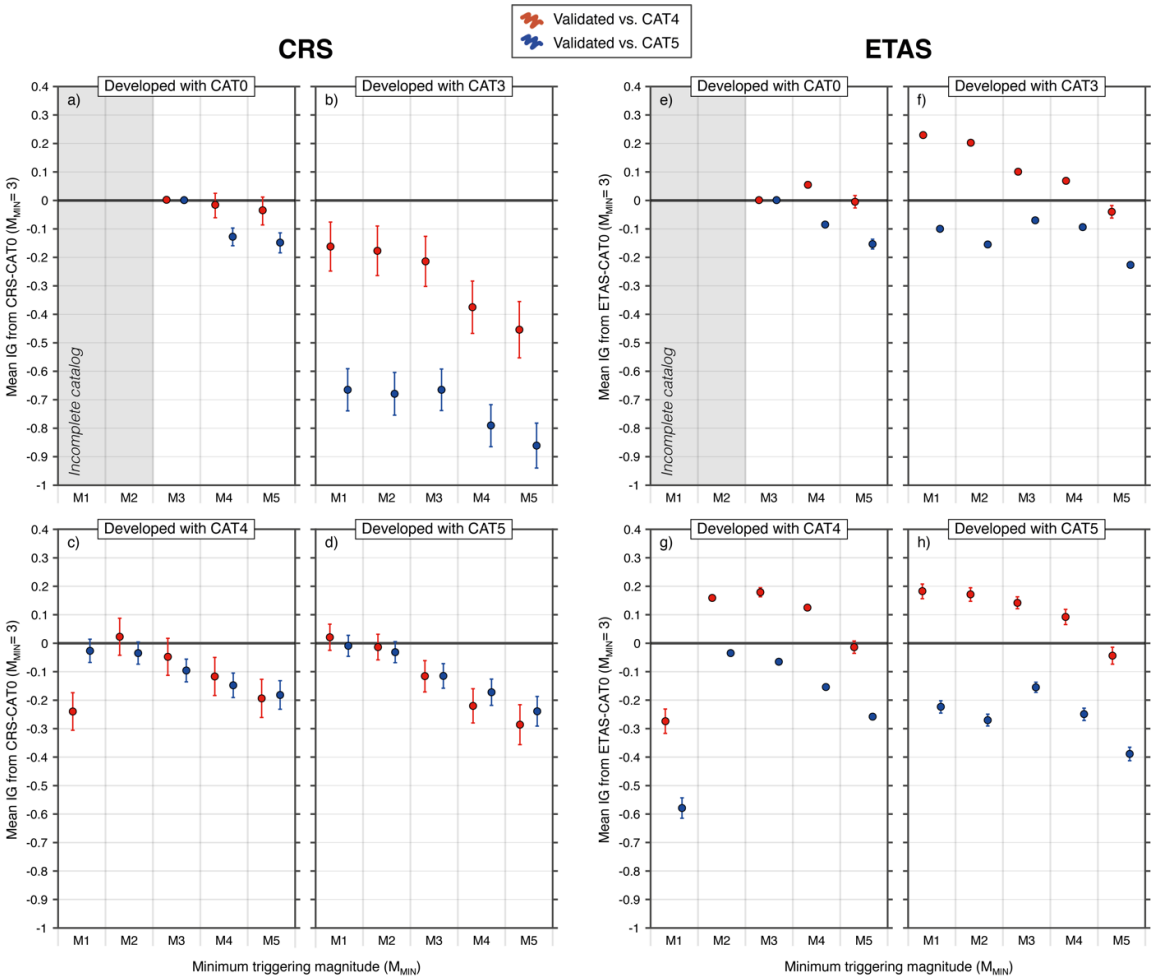

Figure 7. *(a-d) Average daily information gain (IG) per earthquake from CRS-CAT0 of the whole set of CRS models for a cumulative 3-month forecast horizon. Each of the CRS models developed with enhanced catalogues is presented in five versions implementing a different minimum triggering magnitude (M_{MIN}) from M5 to M1. The M_{MIN} values for CRS-CAT0 range from M5 to M3 due to the more limited completeness of the real-time catalogue (grey shaded area). A model is deemed more informative than the reference if its mean IG is positive and if its error bars do not cross the IG=0 line. Red and blue symbols indicate models validated vs. CAT4 and vs. CAT5, respectively. (e-h) Same as the left panels but for the set of ETAS realizations.*

For M_{MIN} = 3 (that is, the same value featured by the CRS-CAT0 benchmark), no new CRS forecast outperforms the model developed with the monitoring room catalogue (Figure 7a-d), with CRS-CAT4 being the only model comparable in performance to CRS-CAT0 when CAT4 is set as target seismicity (Figure 7c). In any case, the information losses are no greater than 0.9 IG units; by comparison, Mancini et al. (2019) found that the CRS-CAT0 model that we use here as benchmark reached information gains up to 8 units over simplistic CRS forecast models using real-time data. When we extend our analysis to a wider spectrum of minimum triggering magnitudes (M_{MIN} = 1-5), we still find that no model is genuinely more informative than CRS-CAT0. However, within any single model IG values gradually rise as forecasts incorporate secondary stress triggering from progressively smaller events, both when CRS models are validated vs. CAT4 (red symbols in Figure 7) and vs. CAT5 (blue symbols). Nevertheless, for $M_{MIN} \leq 2$ the IG tends to plateau (Figure 7a,b,d) and in some cases drops at M_{MIN} = 1 (CRS-CAT4, Figure 7c). Therefore, the question arises of whether this outcome is because M<2 earthquakes do not contribute towards the local M3+ aftershock triggering and patterns, or instead reflects the limits of an insufficient model spatial resolution to describe stress changes at a sub-kilometric level.

We can use Figure 7 also as a diagnostic tool to appreciate how the likelihood-based model ranking

is sensitive to the selection of the target seismicity, as alternate catalogues can describe the same sequence differently; while CRS models developed with CAT4/5 appear less sensitive to the choice of the target catalogue, CRS-CAT3 exhibits marked differences and performs sensibly worse when validated against CAT5 ($\Delta IG \approx 0.5$).

The IG trends of the ETAS forecasts (Figure 7e-h) mirror those of the physical models, confirming the benefit from secondary triggering processes for statistical models as well. Similarly to the CRS counterparts, the ETAS information gain values have the tendency to level out at $M_{MIN} \leq 2$ (Figure 7f,h), and in the case of ETAS-CAT4 they even fall at $M_{MIN} = 1$ (Figure 7g). Interestingly, ETAS realizations with $M_{MIN} < 5$ developed with the enhanced catalogues outperform ETAS-CAT0 when they are evaluated against CAT4, but they all rank worse when CAT5 is set as testing catalogue.

a. Sensitivity tests

To explore the reasons behind our findings, we perform some targeted sensitivity tests on three potentially critical elements of catalogue development that could influence the performance of forecast models: the magnitude estimation, the event locations, and the spatial discretization.

According to both CRS and ETAS formulations, the number of directly triggered earthquakes and the extension of the area over which they decay depend on the parent event's magnitude. Figure 8 shows the cumulative magnitude difference per spatial bin between the matching parent events of CAT4 (re-estimated M_L) and CAT0 (preliminary M_L) against the resulting cellwise log-likelihood differences when models developed using the two catalogue generations are evaluated against CAT4. We find that the magnitudes of parent events reported in CAT4 are typically lower than those of the real-time catalogue, and that this feature generally produces a poorer model performance (*i.e.*, a lower log-likelihood). The ETAS model (top row) presents a rough visual agreement between information loss and negative magnitude differences for ~60% of cells in the testing region. Conversely, such a behaviour is less evident in the CRS model (bottom row), where several cells present an almost constant performance especially until before the occurrence of the Norcia mainshock.

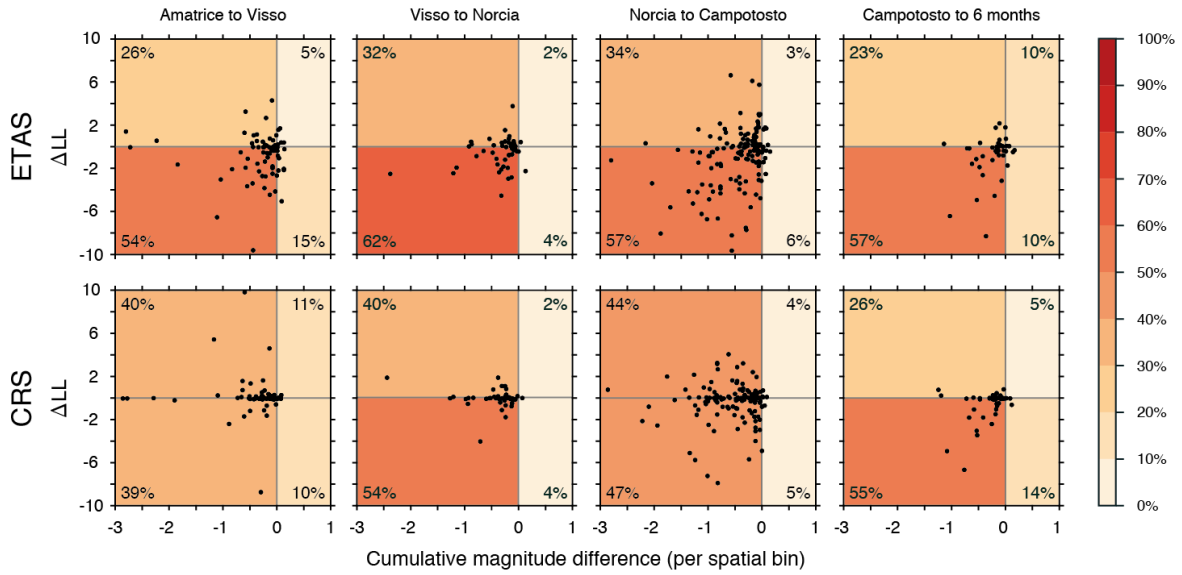

Figure 8. *Cell-wise differences in cumulative magnitudes between common source (parent) events of CAT0 and CAT4 vs. resulting log-likelihood differences in each spatial bin. Log-likelihood values are obtained by using CAT4 target seismicity for both models. Colours follow the relative percentage ranges in the four quadrants.*

One of the effects of earthquake relocation is to generate catalogues that aggregate highly clustered seismicity in a smaller number of spatial bins, generally leaving an increasing number of cells empty. The resulting modifications on the spatial distribution of target seismicity perturbs the likelihood-based scoring of the models. However, by visual inspection we see that the overall spatial differences arising from the relocation procedure in the clustering characteristics of CAT3 and CAT4 target seismicity are likely minimal at a regional scale (Figure 2f,g). Still, relocated events can locally redistribute the earthquake rates expected by a model. In our case, since CAT3 is simply the relocated version of CAT4, we quantify such an effect by calculating the information gain per earthquake of CRS-CAT4 from CRS-CAT3 when both models are evaluated against the relocated CAT4 catalogue (Figure 9). Not surprisingly, the influence of relocation fluctuates between periods of high and low seismicity rate, with a weak information loss during the high-rate period after the Amatrice earthquakes and a small gain at the lower rate period after the Campotosto events.

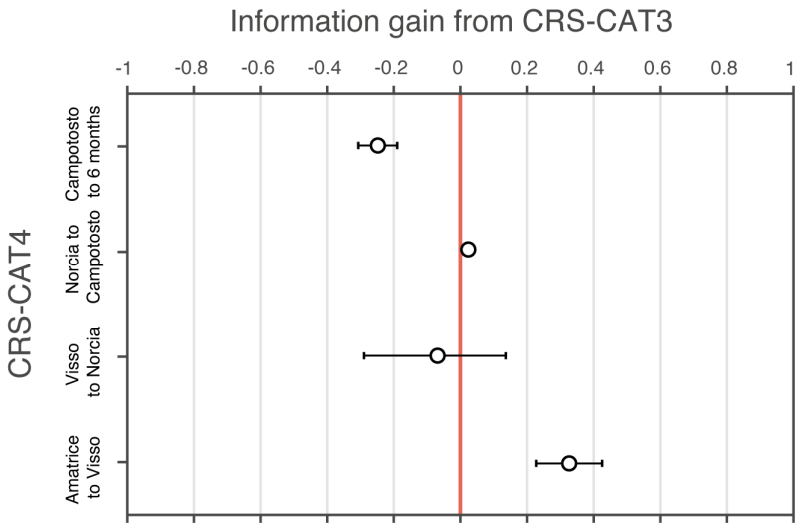

Figure 9. *Average daily information gain per earthquake (IG) of CRS-CAT4 (developed with a relocated catalogue) from CRS-CAT3 (developed with a non-relocated catalogue). We plot the IG values for each period between the first four mainshocks and from the Campotosto events until 6-months from the start of the sequence. CRS-CAT4 is more informative than the reference at 95% confidence interval if IG values are positive and the error bars do not cross the red no-gain line.*

Therefore, it could be argued that the effect of more precise hypocentral locations (*i.e.*, with average relative horizontal location error < 0.1 km in CAT4, about one order of magnitude smaller than in CAT3) is either negligible or not resolvable using our 2-km spatial resolution.

The way the testing region is spatially discretized might indeed hamper our ability to assess the actual local performance of models, particularly when forecasts consider stress perturbations from very small magnitude events whose fault lengths are smaller than the grid spacing. In Figure 10 we illustrate the cumulative LL_S trends of CRS-CAT4 and ETAS-CAT4 to isolate their absolute spatial performance. We produce three versions of each model using a 5-km, a 2-km, and a 500-m spatial binning for the first month of the sequence. Since more granular model resolutions imply lower probability of occurrence in any one cell, it is not surprising that absolute joint log-likelihood values drop with finer model discretizations. Instead, here we focus on the relative ΔLL_S between CRS-CAT4 and ETAS-CAT4 within each binning category. Interestingly, the 5-km binning models present the largest likelihood discrepancy between the two forecast classes (ΔLL_S = 150). When the spatial binning is reduced to 2 km, the difference in likelihood between the two models drops to half of the previous value and almost disappears at 500-m discretization. These results suggest that CRS forecasts are more affected by their spatial resolution than the simulation-based ETAS models, and that their performance improves when the stress field is resolved at a smaller scale allowing a better description of the small fault segments (< 1 km) contributing to the local evolution of the physical system.

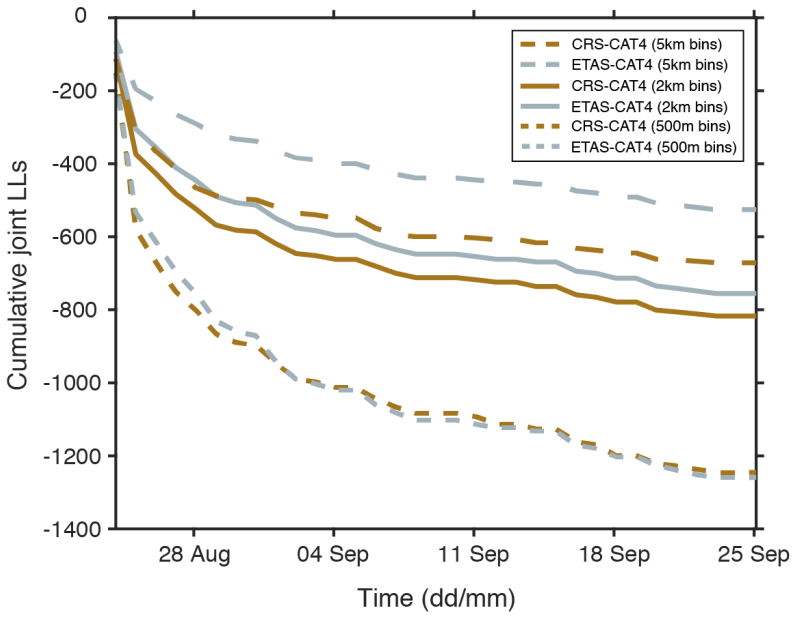

Figure 10. *Cumulative spatial joint log-likelihood (jLL_S) during the first month of the sequence for the CRS-CAT4 and ETAS-CAT4 models. We plot the spatial performance of the forecasts for three different spatial discretizations: 500m, 2km, and 5km. The jLL_S trends are obtained by summing the S-test log-likelihoods of each bin and 1-day time step.*

5. Discussion and conclusions

This study set out to explore the potential for improving standard ETAS and best-practice CRS forecasts for the M3+ earthquakes of the 2016-2017 Central Italy seismic cascade by incorporating additional information provided by enhanced, high-resolution and deep-learning earthquake catalogues released few years after the sequence. Overall, our results suggest that the new suite of models does not present a clear boost in predictive skills compared to the near real-time forecasts on more conventionally acquired catalogues, though it does explain some apparent shortcomings of the near-real time forecasts.

The near-fault aftershock patterns are largely dominated by the triggering effects from large to moderate events in all catalogues, and the incorporation of new small-magnitude triggering events ($1 \leq M < 3$) made available by enhanced catalogues exerts only a minor influence on the CRS and ETAS expected rates. On the other hand, accounting for the localized triggering effects of those events (at least until $M_{MIN} \approx 2$) improves the overall forecast performance, especially in off-fault regions. This should encourage catalogue developers to routinely produce more complete earthquake catalogues as seismic sequences unfold to allow testing of future generations of forecasts in operational applications. However, by further decreasing the minimum triggering magnitude to $M_{MIN} = 1$ the magnitude/location uncertainties of those events become comparable to their radius of influence; this likely negates their ability of improving the forecasts, occasionally leading to information losses.

Such an outcome raises a profound question for future model developments: is there a magnitude threshold below which physical fault-to-fault interactions become negligible, or are our current modelling strategies approaching an inherent limit of their skill at forecasting earthquakes for operational purposes?

The sensitivity tests that we performed do not let us rule out either hypothesis, because:

1. Commonly adopted forecast spatial discretizations are inadequate to resolve localized triggering patterns revealed by high-resolution catalogues.

   Indeed, we argue our standard binning (2 km) is the most limiting factor for properly resolving the triggering contributions of M<2 earthquakes at a local (*i.e.*, less than kilometric) scale. This is supported by the fact that the ETAS spatial consistency is superior to CRS at 5-km binning, but with a 500m discretization the two models are equally informative. This is a promising result as it shows the potential for improving physics-based model performance by resolving stress changes at a sub-cluster resolution of few hundred meters. A way forward will be to incorporate enhanced fault characterizations to capture the small-scale variability of the receiver-fault matrix. Moreover, future experiments on the adoption of enhanced catalogues for earthquake forecasting should consider testing 3D spatial models, potentially featuring locally variable or adaptive spatial discretizations. In this regard, current algorithms should become more computationally efficient, but the emerging cloud-based capabilities for catalogue development (e.g., QuakeFlow; Zhu et al., 2022) pave the way for real-time applications.

2. Seismic catalogues resulting from different workflows present remarkable differences even at moderate magnitudes (*i.e.*, M3+) that might only be reflected in models by *ad hoc* parameter calibrations.

   Our results show that forecasts developed with enhanced catalogues suffer from magnitude estimation resolution. The effect of magnitude inconsistencies on ETAS models' performance is not surprising as the magnitude of a parent event is directly related to the spatiotemporal distribution of triggered events. On the other hand, in physical models this translation is mediated through a series of operators (the slip distribution, the elastic dislocations, stress attenuation) that control the magnitude and spatial extent of the stress changes. We therefore stress the potential severe implications of magnitude inconsistencies in enhanced seismic catalogues (Herrmann & Marzocchi, 2020) on the performance of earthquake forecast with magnitude-dependent productivity. Although we find event relocations have a weak impact on information gains, we note that location uncertainties in relocated catalogues may perturb likelihood values, presumably at a cell-wise level. This experiment does not provide sound evidence for a systematic influence of input seismicity relocations on models' predictive skills, but the IG discrepancies between the CRS-CAT3 and CRS-CAT4 sets of models is surprising and begs the question on how stability of catalogues could be quantified during (or shortly after) their development. The fact that model ranking could be significantly influenced by the presence/absence of very few events in a small number of isolated cells (e.g., the cluster of earthquakes newly detected by CAT5 at the eastern off-fault region) underlines the necessity for more objectively defined testing regions in earthquake forecasting experiments.

3. The current likelihood-based validation metrics are extremely susceptible to the choice of input and target seismicity and to the extent and resolution of the grid used to evaluate models. Regarding the latter point, it should be also considered that catalogues of tightly clustered seismicity clearly illustrate the existence of strong small-scale space-time dependencies among earthquakes that are not accounted for in standard forecast evaluation protocols that assume independent Poisson distributions in each space-time-magnitude bin. The rapid development of deep-learning and other advanced techniques providing high-resolution catalogues, refocuses scientists on the development of sequence-specific earthquake forecasts evolving within shorter time frames (*i.e.*, daily, or even hourly) over spatial extents of few tens of kilometres when seismicity is understandably non-Poissonian in nature. Therefore, validation strategies will need to adapt to the experimental set-ups made possible by modern enhanced datasets. For the above reasons, a rigorous quantification of the added value of enhanced catalogues for short-term earthquake forecasts is challenging at present.

Despite its exploratory nature, this study offers valuable insights into the issues that modellers are likely to face soon. Notwithstanding the fact that current enhanced seismic catalogues provide

an unprecedented quality description of earthquake occurrence, each of them illustrates a different version of it and we cannot know a priori which catalogue, if any, more closely represents the ground truth. These catalogues are products of different choices in their serial components of detection, event association, and seismic parameters estimation that make it difficult to quantify the contribution of each choice towards improving seismicity forecasts. Finally, we believe that deep-learning-based forecasts shared with the community will promote detailed investigations in the wake of those presented in this study and motivate further research on probing on the actual power of enhanced catalogues for earthquake predictability.

Acknowledgements

This RISE research also received support from the project "The Central Apennines Earthquake Cascade Under a New Microscope" (NE/R0000794/1), funded by the UK National Environment Research Council (NERC) and the United States National Science Foundation, Directorate for Geosciences (NSFGEO). The authors would like to thank the participants in both projects for the fruitful discussion and the constructive comments that improved the quality of the manuscript.

Publication outputs:

Chiaraluce, L., Michele, M., Waldhauser, F., Tan, Y. J., Herrmann, M., Spallarossa, D. et al. (2022). A comprehensive suite of earthquake catalogues for the 2016-2017 Central Italy seismic sequence, Scientific Data. https://doi.org/10.1038/s41597-022-01827-z.

Mancini, S., M. Segou, M. J. Werner, T. Parsons, G. Beroza, and L. Chiaraluce (2022). On the use of high-resolution and deep-learning seismic catalogs for short-term earthquake forecasts: potential benefits and current limitations, *J. Geophys. Res. Solid Earth*, 127, e2022JB025202. https://doi.org/10.1029/2022JB025202.

References

Bayona, J. A., Savran, W. H., Rhoades, D. A., & Werner, M. J. (2022). Prospective evaluation of multiplicative hybrid earthquake forecasting models in California. *Geophys. J. Int.*, 229, 1736-1753. https://doi.org/10.1093/gji/ggac018.

Beroza, G. C., Segou, M., & Mousavi, S. M. (2021). Machine learning and earthquake forecasting–next steps. *Nat. Comm.*, 12. https://doi.org/10.1038/s41467-021-24952-6.

Cattania, C., Werner, M. J., Marzocchi, W., Hainzl, S., Rhoades, D., Gerstenberger, M., et al. (2018). The Forecasting Skill of Physics-Based Seismicity Models during the 2010–2012 Canterbury, New Zealand, Earthquake Sequence. *Seismol. Res. Lett.*, 89(4), 1238–1250. https://doi.org/10.1785/0220180033.

Chiaraluce, L., Di Stefano, R., Tinti, E., Scognamiglio, L., Michele, M., Casarotti, E., et al. (2017). The 2016 Central Italy seismic sequence: A first look at the mainshocks, aftershocks, and source models. *Seismol. Res. Lett.*, 88(3), 757–771. https://doi.org/10.1785/0220160221

Chiaraluce, L., Michele, M., Waldhauser, F., Tan, Y. J., Herrmann, M., Spallarossa, D. et al. (2022). A comprehensive suite of earthquake catalogues for the 2016-2017 Central Italy seismic sequence, *Scientific Data*. https://doi.org/10.1038/s41597-022-01827-z.

DISS Working Group (2018). Database of Individual Seismogenic Sources (DISS), Version 3.2.1: A compilation of potential sources for earthquakes larger than M5.5 in Italy and surrounding areas. *Istituto Nazionale di Geofisica e Vulcanologia (INGV).* https://doi.org/10.13127/diss3.3.0.

Herrmann, M., & Marzocchi, W. (2020). Inconsistencies and Lurking Pitfalls in the Magnitude–Frequency Distribution of High-Resolution Earthquake Catalogs. *Seismol. Res. Lett.*, 92 (2A): 909–922. https://doi.org/10.1785/0220200337.

ISIDe Working Group (2007). Italian Seismological Instrumental and Parametric Database (ISIDe). Istituto Nazionale di Geofisica e Vulcanologia (INGV). https://doi.org/10.13127/ISIDE.

Jordan, T. H., Chen, Y., & Main, I. (2011). Operational earthquake forecasting: State of knowledge and guidelines for utilization. *Ann. Geophys.* 54, no. 4. https://doi.org/10.4401/ag-5350.

Lomax, A., J. Virieux, P. Volant, & C. Berge-Thierry (2000). Probabilistic earthquake location in 3D and layered models: introduction of a Metropolis–Gibbs method and comparison with linear locations. In: *Advances in seismic event location*, ed. C. H. Thurber and N. Rabinowitz, 101–134. Dordrecht and Boston: Kluwer Academic Publishers.

Mancini, S., M. Segou, M. J. Werner, and C. Cattania (2019). Improving physics-based aftershock forecasts during the 2016-2017 Central Italy earthquake cascade, *J. Geophys. Res. Solid Earth* 124. https://doi.org/10.1029/2019JB017874.

Mancini, S., M. J. Werner, M. Segou, and T. Parsons (2020). The predictive skills of elastic Coulomb rate-and-state aftershock forecasts during the 2019 Ridgecrest, California, earthquake sequence, *Bull. Seismol. Soc. Am.* https://doi.org/10.1785/0120200028.

Moretti, M., Pondrelli, S., Margheriti, L. & SISMIKO Working Group (2016). Emergency network deployment and data sharing for the 2016 central Italy seismic sequence. *Ann. Geophys.*, 59 (5); https://doi.org/10.4401/ag-7212.

Ogata, Y. (1988). Statistical models for earthquake occurrences and residual analysis for point processes, *J. Am. Stat. Assoc.*, 83(401), 9–27.

Rhoades, D. A., Schorlemmer, D., Gerstenberger, M. C., Christophersen, A., Zechar, J. D., & Imoto, M. (2011). Efficient testing of earthquake forecasting models, *Acta Geophysica*, 59(4), 728–747. https://doi.org/10.2478/s11600-011-0013-5.

Spallarossa, D., G. Ferretti, D. Scafidi, C. Turino, & M. Pasta (2014). Performance of the RSNI-Picker. *Seismol. Res. Lett.* 85, 1243–1254.

Spallarossa, D., Cattaneo, M., Scafidi, D., Michele, M., Chiaraluce, L., Segou, M., Main, I. G. (2020). An automatically generated high-resolution earthquake catalogue for the 2016–2017 Central Italy seismic sequence, including P and S phase arrival times. *Geophys. J. Int.*, 225 (1), 555-571. https://doi.org/10.1093/gji/ggaa604

Tan, Y. J., Waldhauser, F., Ellsworth, W. L., Zhang, M., Zhu, W., Michele, M., Chiaraluce, L., Beroza, G. C., and Segou, M. (2021). Machine-Learning-Based High-Resolution Earthquake

Catalog Reveals How Complex Fault Structures Were Activated during the 2016–2017 Central Italy Sequence, *The Seismic Record* 1(1), 11–19.

Waldhauser, F., & W. L. Ellsworth (2000). A double-difference earthquake location algorithm: Method and application to the northern Hayward Fault, California, *Bull. Seismol. Soc. Am.* 90, 1353–1368.

Waldhauser, F., Michele, M., Chiaraluce, L., Di Stefano, R., & Schaff, D. P. (2021). Fault planes, fault zone structure and detachment fragmentation resolved with high-precision aftershock locations of the 2016-2017 central Italy sequence. *Geophys. Res. Lett.*, 48, e2021GL092918. https://doi.org/10.1029/2021GL092918

Werner, M. J., Helmstetter, A., Jackson, D. D., & Kagan, Y. Y. (2011), High-resolution long-term and short-term earthquake forecasts for California. *Bull. Seismol. Soc. Am.*, 101(4), 1630–1648. https://doi.org/doi:10.1785/0120090340

Zechar, J. D., Gerstenberger, M. C., & Rhoades, D. A. (2010). Likelihood based tests for evaluating space-rate-magnitude earthquake forecasts. *Bull. Seismol. Soc. Am.*, 100, 1184–1195. https://doi.org/10.1785/0120090192.

Zhu, W., & G. C. Beroza (2019). PhaseNet: A deep neural-network based seismic arrival-time picking method. *Geophys. J. Int.* 216, 261-273.

Zhu, W., Hou, A. B., Yang, R., Datta, A., Mousavi, S. M., Ellsworth, W. L., & Beroza, G. C. (2022). QuakeFlow: a scalable machine-learning-based earthquake monitoring workflow with cloud computing. *Geophys. J. Int.* 232(1), 684-693. https://doi.org/10.1093/gji/ggac355.