DEEP

Innovation for:
De-Risking Enhanced
Geothermal Energy Projects

Deliverable

Deliverable 3.1 - Extensive comparison of existing forecasting induced seismicity models with existing datasets

Report information	
Work package	WP3 Innovation in forecasting models and uncertainty quantification
Lead	ETH
Authors	Antonio Pio Rinaldi & Luigi Passarelli
Reviewers	Federica Lanza
Approval	Stefan Wiemer
Status	Draft
Dissemination level	Internal
Will the data supporting this document be made open access?	Yes
If No Open Access, provide reasons	-
Delivery deadline	31.08.2021
Submission date	31.08.2021
Intranet path	[DOCUMENTS/DELIVERABLES/Deliverable3.1.docx]

This project has been subsidized through the Cofund GEOTHERMICA, which is supported by the European Union's HORIZON 2020 programme for research, technological development and demonstration under grant agreement No 731117.

Table of Contents

1.	**Introduction**	**3**
2.	**Seismicity models**	**4**
2.1	**Empirical Model 1 with Maximum Likelihood Estimation (EM1_MLE)**	**4**
2.2	**Empirical model 1 within a Bayesian hierarchical framework (EM1_BH)**	**5**
2.3	**Analytical Hydro-Mechanical Model (HM0)**	**5**
3.	**Model comparison via information gain**	**6**
3.1	**Uncertainties estimation and model simulations**	**7**
3.2	**Probability gain as a model comparison tool**	**7**
4.	**Results**	**8**
4.1	**Bedretto Lab November 2020 injection experiment.**	**8**
4.2	**Basel 2006**	**11**
5.	**Conclusions and Outlook**	**12**
Reference List		***15***

Summary

In this deliverable we report on a standardized method to compare induced seismicity models. Here, we discuss how a comparison based on empirical distribution is more reliable than Log-Likelihood based on the Poissonian approximation as standardize approach in several earthquake forecasting exercises and practices (e.g., the Collaboratory for the Study of Earthquake Predictability – CSEP). Starting from the approach developed in the "Induced Seismicity Testbench" by Kiraly et al. (2016), we improve the model comparison by accounting for the full empirical distribution of forecasted rate. We derive the probability of reproducing the observed rates directly from the full empirical distribution. We introduce the concept of Probability Gain as a measure of the model performance with respect to a null model. The approach is compatible with the previous formulation, and comparison is still possible by the assumption of a Poissonian distribution when the empirical distribution is not available. Results shows that with this method, the uncertainties are better taken into account, in particular for the case of small sample datasets. We compare both models with and without empirical distribution, and demonstrate how models with apparent similar average prediction in reality perform quite differently from a statistical perspective.
This deliverable constitutes a first report on the current work, and in the future, it will be improved by accounting for more models and substantially more dataset. The approach will be then implemented in the Adaptive Traffic Light System (ATLS) for real-time evaluation of model performances.

1. Introduction

The assessment of performances of models in forecasting induced seismicity in real-time is essential for a correct implementation and weighting of model results in hazard calculation.In addition, developing model comparison tools can guide to the implementation of ensemble models that can also be constructed following a multi-component approach (*1*). Forecast tests can be run in retrospective, pseudo-prospective and also fully prospectively. The testbench in that sense is one element of the ATLS workflow, taking care of model performance evaluation and model weighting.

To compare the performance of different computational models, we need to adopt a standardize testbench approach. As a starting point, we build on the approaches developed as part of the Collaboratory for the study of earthquake predictability (CSEP , www.cseptesting.org) and on the work for the '*Induced Seismicity Testbench*' developed by Kiraly-Proag et al. (2016) (*2*). In their work (*2*), the authors compared both empirical/statistical and numerical models, and show how models may perform differently in the different stages of a stimulation procedure. Their approach is mainly based on standards as defined by the Collaboratory for the Study of Earthquake Predictability – CSEP. One drawback of this formulation is the assumption that the metrics for comparison is always a Poissonian distribution (*2*): the simulated seismicity rate is fed as mean for a Poisson's distribution, and the Poissonian Log-Likelihood is calculated accordingly. This method may introduce strong biases: first, the induced seismicity models often do not respect the Poissonian definition of having forecasting space-time bins being independent from each other; secondly it will favor models that are Poissonian by definition. The latter was demonstrated in the context of earthquake forecasting by Nandan et al. (*3*) who showed how by assuming for the full probability distribution constitutes a far better and fairer comparison metrics.

In this deliverable, we developed a more general and reliable approach to compare models. We modified the output of some of the models to be able to simulate not a single forecast, but a full empirical distribution, and hence better define with what probability a given model can reproduce the observations. The knowledge of such probability enables the calculation of the Probability Gain as a measure of the model performance with respect to a null model. Regarding models with a single forecast, the comparison is still possible by accounting for the Poissonian distribution and its Log-Likelihood.

Here we show the results of such approach for three different models: two variations of an empirical model (*4*, *5*) and a simplified hydromechanical model. The comparison is performed for two datasets. The first one was collected during a recent injection experiment at the Bedretto Underground Laboratory for Geonergy and Geosciences (*6*) and it serves as an example for a case of limited dataset. The second dataset is the injection experiment of Basel in 2006 (*7*), which serves as a classical dataset for model development and it has been extensively used in the past.

After introducing the three seismicity models, we present the approach for comparison via the Probability Gain and the results of the application to the two selected dataset. We then conclude with some outlook for future development.

2. Seismicity models

We have chosen three seismicity models that all share key characteristics useful for real-time data fitting, data assimilation and forecast of possible seismic hazard scenarios. The models are simple but robust enough to include the first order physical processes controlling induced seismicity during injection experiments. Two models (EM1_MLE and EM1_BH) are based on an empirical law that links the seismicity rate with the injected flow rate history and contains a set of parameters describing the geological and seismological characteristics of the stimulated rocks. Both models have been extensively tested on data of past injection experiments. Their statistical formulation allows to simulate forecasting scenarios and quantify probability associated to the expected seismicity rate changes as well as probability of occurrence of earthquakes in time bin horizon. The third model (HM0) belongs to a different class of simulator, linking the 1D analytical solution for fluid flow to a geomechanical-stochastic approach where the earthquake rate directly depends on the pressure profile simulated for the given injection operation. All three models can be efficiently executed in real-time applications and produce multiple forecasting scenarios, fundamental to advice operators during injection experiments. Computational efficiency is an important feature which guided us in selecting these rather simple models; nevertheless we are aware that these models can capture only the first order features of the complex processes behind induced seismicity. Moreover, we have implemented a simple and fast procedure to estimate uncertainties of the parameters of the two statistical models in order to introduce the expected aleatoric variability via Monte Carlo simulation when producing forecasting scenarios. The same approach was considered redundant for HM0 given its stochasticity nature, although limiting, for the time being, the output of such model to an average forecasted rate.

2.1 Empirical Model 1 with Maximum Likelihood Estimation (EM1_MLE)

The basis for the EM1_MLE model has been originally proposed by Shapiro et al. 2007 (*8*). The model is constructed to simulate the observed piecewise evolution of the seismicity rate during the two phases of the injection strategy, namely during and after the termination of the injection operation. It has been observed that the seismicity rate closely follows the injected flow rate but it is also dependent on the characteristics of the seismogenic volume stimulated during the injection. Mignan et al. (*4*) improved then such a model to account for the post shut-in decay, and defined seismicity rate as follow:

$$\lambda(t, m \geq m_0, \theta) = \begin{cases} 10^{a_{fb} - b m_0} \dot{V}(t) & t \leq t_s \\ 10^{a_{fb} - b m_0} \dot{V}(t) exp\left(-\frac{t - t_s}{\tau}\right) & t > t_s \end{cases} \quad (1),$$

where $\dot{V}(t)$ is the injection flow rate as a function of time t measured in $m^3/[t]$ and $\theta = [a_{fb}, b, \tau]$ are the model parameters describing the characteristics of the seismogenic volume stimulated, m_0 is the magnitude of completeness and t_s is the shut-in time, a_{fb} is the activation feedback, b is the earthquake size ratio, and τ is the mean relaxation time of the medium after the injection ends.

The model in Eq. 1 assumes that the seismicity rate λ is modulated by the injected flow rate during the operational phase and bears the underlying assumption that $\dot{V}(t)$ scales linearly with the overpressure induced in the rock volume. After the shut-in phase ($t > t_s$), the exponential decay of the rate is a first order approximation of a fluid diffusion process. Furthermore, Eq. 1 - despite the intrinsic simplicity compared with more complex fluid modeling - can be calculated in real-time and under data assimilation schemes, and the exponential term can capture the response of the fluid-stimulated seismogenic volume after shut-in operation. Finally, Mignan et al., (*4*) successfully fitted Eq.1 on different database of induced seismicity related to: six enhanced geothermal systems (EGS), an initial stage of a long-term brine sequestration and one fracking at an oil field. For details on the model performance and statistical tests of

goodness-of-fit, please refer to the original paper (*4*). Following Broccardo et al. (*5*), the model in Eq.1 is fitted to the data through a Maximum Likelihood Estimation method.

2.2 Empirical model 1 within a Bayesian hierarchical framework (EM1_BH)

The EM1_BH model consists in the implementation of Eq. 1 in a fully Bayesian hierarchical model with the assumption that the earthquake occurrence and hence the seismicity rate λ can be well approximated by a Non-Homogeneous Poisson Process (NHPP) (*5*). Mathematical details of the Bayesian hierarchical model are in Broccardo et al. (*5*), that we do not report to avoid redundancy. We limit to only discuss the main characteristics and flexibility that the Bayesian formulation introduces with respect to the EM1_MLE model.

The fit of Eq. 1 with a statistical frequentist approach, as for EM1_MLE, returns a punctual estimation of the model parameters which leads to a "deterministic" value of the seismicity rate λ. The implication is that the uncertainties in the parameter estimation is by definition only aleatoric. However, the uncertainty in the occurrence of induced earthquakes during injection operation can be divided in two contributions: one coming from the intrinsic aleatoric nature of the earthquake nucleation process, and the other arising from our limited knowledge of the physics controlling earthquake genesis in rocks stimulated by fluid-driven overpressure. The latter contribution is referred as epistemic uncertainty and can be reduced as our understanding of the physical processes leading to induced seismicity are better understood. In addition, information and data coming from past injection experiments can be used to further reduce or constrain the epistemic part of the total uncertainties.

This is the rationale behind the development of the Bayesian model of Broccardo et al. (*5*), where the authors modelled Eq.1 through a NHPP as a likelihood of the model. In addition, they added a layer called prior model, where they modelled the parameters $\theta = [a_{fb}, b, \tau]$ of Eq.1 as random variables using probability distribution functions (pdfs). The support of the pdfs of the parameters θ in the prior model is constrained by tapping from information of past injection experiments and thus reducing the epistemic uncertainties on the parameters. The prior model is combined with the model likelihood through the Bayes theorem, and the inference produces the posterior distribution of the parameter θ and given the injected flow rate the seismicity rate λ can be inferred. The added value of the Bayesian approach is that the parameters estimation is not a single value but a full posterior distribution from which it is straightforward to derive descriptive statistics, i.e., mean, median, mode and their uncertainty as for example percentile range. The uncertainties of the posterior distributions can be easily propagated in the estimation of seismicity rate from Eq.1, as we will discuss later. This model has been successfully tested on the induced seismicity dataset of Basel 2006 injection experiments as a probabilistic forecasting tool proving also to be computationally efficient for real-time application.

2.3 Analytical Hydro-Mechanical Model (HM0)

We implemented a hybrid model by loosely coupling a 1D analytical solution for fluid flow with an analytical geomechanical-stochastic seismicity simulator. The latter was developed in the framework of the COSEISMIQ project. The 1D fluid flow simulator accounts for the solution of the linear pore-pressure diffusion inside the cylinder when hydraulic properties are constant. An analytical solution exists for this problem and at each time moment *t* the overpressure at distance *r* is the superposition of the kernel for every rate change in the injection rate Δq_j that happened at $t_j<t$ and thus overpressure equals the summation of these kernels:

$$\Delta P(t,r) = -\frac{1}{4\pi T}\left(\sum_{j=0}^{N_j} \Delta \mathrm{q}_j E_i\left(-\frac{r^2}{4D(t-t_j)}\right)\right) \quad (2),$$

where $D = \kappa/s\eta$ is diffusivity, $T = \kappa h/\eta$ is the transmissibility, $E_i(\cdot)$ is the exponential integral function where $E_i(-x) \approx \ln x$ for any large positive *x*. *k*, *h*, *s*, η are the permeability, the width of the cylinder, the effective compressibility and the viscosity, respectively.
Practically the solution is the superposition of many simpler solutions. The computational cost scales with ($N_j \cdot N_r \cdot N_t$), where N_r and N_t are the size of the radii and the time moments at which the solution needs to be found.

We implemented a fit function that finds the best fitting pair of T, D for the N_j injection steps Δq_j found in the training dataset. For fitting, we minimize the Least Square with the differential evolution, which is an evolutionary algorithm for minimizing arbitrary non-linear cost functions.

The fluid flow simulator is then coupled to a simplified, analytical approach considering two-dimensional Mohr-Coulomb circles for potential hypocenters in space, with two principal stresses $\sigma_3 < \sigma_1$ following normal distributions. Then, the analytical Cumulative Density Function (CDF) and Probability Density Function (PDF) of a hypocenter being reactivated at a certain pressure P_f is found for the hydrostatic conditions of each radius. These analytical solutions are truncated to the lowest positive value P_f for which reactivation of fractures is expected and a numerical integration with depth may be needed. Besides the principal stresses, the solutions are conditioned on the orientation of the fracture and to the frictional properties of the fracture (i.e. friction and cohesion).

The failure over-pressure P_f according to the Mohr-Coulomb criterion equals: $P_f = c/\mu + \sigma_n - P_p + \tau/\mu$, where P_p is the formation's pressure, C and μ are the cohesion and the friction coefficient of the fracture, and σ_n and τ are the normal and the shear stress, respectively. Important angle property in 2D Mohr Coulomb Circles is the angle θ of the fracture with the two principal stresses S_1 and S_3, with $S_1 > S_3$. The approach then assumes normal distributions for the principal stresses such that $\sigma_1 \sim \aleph(\sigma_{\sigma 1}, \sigma_{\sigma 1}^2)$ and $\sigma_3 \sim \aleph(\sigma_{\sigma 3}, \sigma_{\sigma 3}^2)$, where $\sigma_{\sigma i}, \sigma_{\sigma i}^2$ are the mean and standard deviation for each stress component. It is easy to show that P_f would also follow a Normal distribution with mean value and standard deviation as:

$$P_f = \left(\frac{C}{\mu} - P_p\right) + \sigma_1\left(\frac{1}{2}\left(1 + \underbrace{\left(\cos 2\theta - \frac{\sin 2\theta}{\mu}\right)}_{\Theta}\right)\right) + \sigma_3\left(\frac{1}{2}\left(1 - \underbrace{\left(\cos 2\theta - \frac{\sin 2\theta}{\mu}\right)}_{\Theta}\right)\right) \quad (3),$$

$$\sigma_{P_f}^2 = \sigma_{\sigma_1}^2 \frac{(1+\Theta)^2}{4} + \sigma_{\sigma_3}^2 \frac{(1-\Theta)^2}{4}$$

The model assumes a truncation at P_0, the minimum possible failure pressure to be expected. This value should theoretically be at least equal to the hydrostatic pressure.

The seismicity model is calibrated by matching, for each spatial bin, the cumulated evolution of seismicity with a Least Square minimization method via differential evolution algorithm. Calibrated parameters are the standard deviation of the maximum principal stress ($\sigma_{\sigma_1}^2$), the minimum failure pressure (P_0), and the density of the normal distributions.

3. Model comparison via information gain

The ultimate goal during injection experiments is to produce reliable forecast of seismicity in order to advice operators and decision makers about plausible hazard and risk scenarios related to the induced seismicity. One of the key ingredients in hazard assessment is the choice of a reliable seismicity model that is able to not only fit/reproduce the seismicity, but provide meaningful forecast for the evolution of the seismicity. The seismicity model should provide a quantification of uncertainties related to the parameter's estimation that in turn can be propagated through the model in the forecast phase. The final goal is to produce multiple scenarios that can capture the intrinsic variability of the induced seismicity and being meaningful to advice injection operators about actions to undertake during injection experiments. Therefore, the forecast performance of the seismicity models needs to be evaluated before their use in real-time application throughout an objective procedure that can enable to quantify the forecast ability of each model against the others.

Here we present a strategy we have developed for model comparison. The algorithm is based on the information gain theory and is designed under a general framework which allows to test forecast performance of a model against any other. We apply this model selection tool to the three seismicity models discussed in Section 2 using data from Bedretto Undeground Lab 2020 and Basel 2006 injection experiment. We want to remark that the strategy is general and any seismicity model can be tested under the proposed strategy. We will make available the code to DEEP consortium partners to further test their seismicity models.

3.1 Uncertainties estimation and model simulations

The uncertainties on the parameters of both statistical models presented in Section 2 is described in the following and it is different for the two models as a result of the different statistical framework they are built on.

For model EM1_MLE (subsection 2.1), the Maximum Likelihood Estimation is used to infer the best fit parameters. In order to calculate the uncertainties on the model parameters $\theta = [a_{fb}, b, \tau]$ we apply a non-parametric test based on the Likelihood Ratio (LR) concept. The LR test allows to calculate the confidence interval (CI) of each parameter in the model, in other words we find all values of the parameter θ (unidimensional for this example) within a given interval of the maximum value $l(\theta_{max}, data)$ of the log-likelihood function. Under the normal assumption of the MLE, if θ_0 is the true value of the parameter θ then the log-likelihood ratio statistics $LRS = 2Log(L(\theta_{max}) / L(\theta_0))$ is approximately distributed like a χ_1^2 distribution with one degree-of-freedom. We can therefore construct the LR test under the null hypothesis that $H_0: \theta = \theta_0$ and we reject the null hypothesis at the α-level if the LRS exceeds the $100(1-\alpha)th$ percentile of the χ_1^2 distribution, i.e., for $\alpha = 0.05$ we reject H_0 if LRS >3.84. The test allows to construct a CI for the model parameters in Eq.1 via the well know procedure called likelihood profiling. Specifically, we retain all θ_k for which $LRS = 2Log(L(\theta_{max}) / L(\theta_k)) < 3.84$ holds and obtain the 95% CI of each parameter.

For the Bayesian model EM1_BH the estimation of uncertainties associated to each parameter is straightforward since the output of the model is already in the form of probability density function. We therefore simply calculated the 95% CI, i.e. all simulated values of $\theta = [a_{fb}, b, \tau]$ falling in the 2.5 and 97.5 percentiles from the posterior distribution to be consistent with the LR test approach described above for EM1_MLE. We use the mode evaluated from the posterior distributions as best-fit parameters θ_{max}.

Once we have calculated the best-fit parameters and their 95% CI for both models, we produce synthetic catalogs as a forecast of the models in the validation phase by randomly drawing triplets of parameters $\theta = [a_{fb}, b, \tau]$ from independent Gaussian distributions. The Gaussian distributions for each parameter are set with mean equal to the best-fit value of parameters and standard deviation equal to 95% CI divided by four. We run the models to produce a single forecast in the validation phase using the flow rate $\dot{V}(t)$ as the only input parameter. The flow rate is taken deterministic as it is planned before the start of the injection experiment.

For the HM0 model instead, estimating the uncertainty for the model parameters is not straightforward, given the more complex nature of this hydromechanical model. Therefore, we assume for now this model to be fully deterministic. Similarly to the statistical models, we fit HM0 in the training phase with flow rate/pressure and seismicity data and we produce a single forecast seismicity catalog in the validation phase using the hydraulic data as input. Finally, we are currently working on developing an efficient algorithm to simulate stochastically the HM0 input parameters and assess the variability of the output parameters in order to estimate model parameter uncertainties similarly with the methodology used for the two statistical models (see Outlook section).

3.2 Probability gain as a model comparison tool

In the previous section we have introduced the procedure to estimate uncertainties of the models and to simulate synthetic catalogs for the forecast in the validation phase given the parameter uncertainties. Now we will introduce and discuss our new model comparison approach.

We use information gain theory and based on the seminal work proposed by Kagan and Knopoff (*9*), we follow the strategy proposed by Passarelli et al. (*10*) to compare frequentists and Bayesian models in real-time prospective. Kagan and Knopoff propose to use the probability gain (information gain in their terminology) as a measure of the predictive performance of a seismicity model against a Poisson model for earthquake occurrence. They calculate the probability gain as the difference between the log-likelihoods of any model and the Poisson one (*9*) scaled by a logarithmic conversion constant. The rationale behind the choice of testing again Poissonian earthquake occurrence is that this model is stationary and can describe the mainshock rate for a given large seismogenic zone when aftershocks are removed. However, induced seismicity is dominated by transient processes and involves fluid-rock interaction for the earthquake nucleation that make time and space evolution of seismicity far to be Poissonian. Induced seismicity falls in

the realm of so-called fluid-induced tectonic earthquake swarms which depart from mainshock-aftershocks sequences and do not have yet a general governing law to describe them (*11*). We therefore cannot use Poisson model as a benchmark for the model comparison test. We instead develop the model comparison test for two generic models, however in some situations one particular model can be set as benchmark for the comparison test (see our strategy in Section 4).

Let's suppose we have model A and B to test against each other and assess the forecast performance. The classical approach would be to calculate the probability gain (PG) or the predictive performance of the model as (*9*):

$$PG = l_A(\hat{\theta}|H) - l_B(\hat{\theta}|H)/ln2 \quad (2),$$

where $l_i(\hat{\theta}|H)$ is the log-likelihood of model i given the best estimate of parameters $\hat{\theta}$ in the training phase conditioned to the data, and H are the seismicity data in the catalog. PG is positive when model A performs better than model B, and for negative PG B performs better than A, while for $PG = 0$ the two models equally perform. The absolute value of PG gives an estimation on the magnitude of the gain.

For our purposes, we do not have a classical likelihood function for the Bayesian model, instead we have a posterior distribution of the parameters from which we can simulate synthetic catalogs and calculate seismicity rates from Eq. 1. In addition, we are interested in evaluate the PG for the forecasts on the validation phase, i.e. "future observation" $l_A(\hat{\theta}|H_{future})$. Passarelli et al. (*10*) modified the strategy to calculate PG fulfilling the above conditions, directly calculating the probability of a given model to reproduce the H_{future} observation.

We follow the latter approach and calculate the terms in Eq. 2 as $l_i(\hat{\theta}|H_{future}) = ln(\Pr(model_i \mid \hat{\theta}, H_{future}))$ where we take the natural logarithm of the probability for $model_i$ to reproduce the observed H_{future} that in our case is the observed seismicity rate $r_{obs} = N_{obs}/\Delta t$. N_{obs} is the number of earthquakes registered in the time window Δt, which can be the total time of the validation phase or a finer time bin discretization. In the following we name the probability $l_i(\hat{\theta}|H_{future}) = ln(\Pr(model_i \mid \hat{\theta}, H_{future}))$ as "*log-likelihood*" for simplicity, although strictly speaking this is not a log-likelihood in the statistical frequentist term and Eq. 2 is not anymore a log-likelihood ratio. We calculate $\Pr(model_i \mid \hat{\theta}, \Delta r_{obs})$ from the empirical distribution of the simulated rate r_{sim}^j (where j is the number of synthetics catalogs from EM1_MLE and EM1_BH see Subsection 3.1) and evaluate the probability of the small interval $\Delta r_{obs} = [r_{obs} - r_-, r_{obs} + r_+]$, where is $r_\pm = (N_{obs} \pm 1)/\Delta t$. For HM0 model, where a single seismicity catalog is used, we calculated instead the probability from a Poissonian Log-Likelihood as in previous approaches (*2*).

We finally calculated the PG_k in k time binned intervals on the validation set of data and call it "*punctual PG"* and the *"total PG"* as $PG_t = \sum_k PG_k$. The former will give an estimation of the forecasting performance in the single time bins, whereas the latter gives information on the overall forecasting horizon of the model.

4. Results

4.1 Bedretto Lab November 2020 injection experiment.

We used the hydraulics and seismicity data recorded in the Bedretto Lab during the November 2020 injection experiment. The time of occurrence, magnitudes and 3D distribution of earthquakes is shown in Fig. 1 together with the injected flow rate and pressure. The data from the experiment spans nearly 35 hours and seismicity is recorded for the first 27 hours from the start of the injection operation. A total of 115 earthquakes were recorded with magnitudes ranging from -3.2 to -1.8. We split the dataset at time 15 hours after the start of the injection in a training phase which contains 80 earthquakes with the rest belonging to the validation phase (Fig. 1). The seismicity rate follows reasonably well the injection flow rate corroborating the use of the models presented in Section 2.

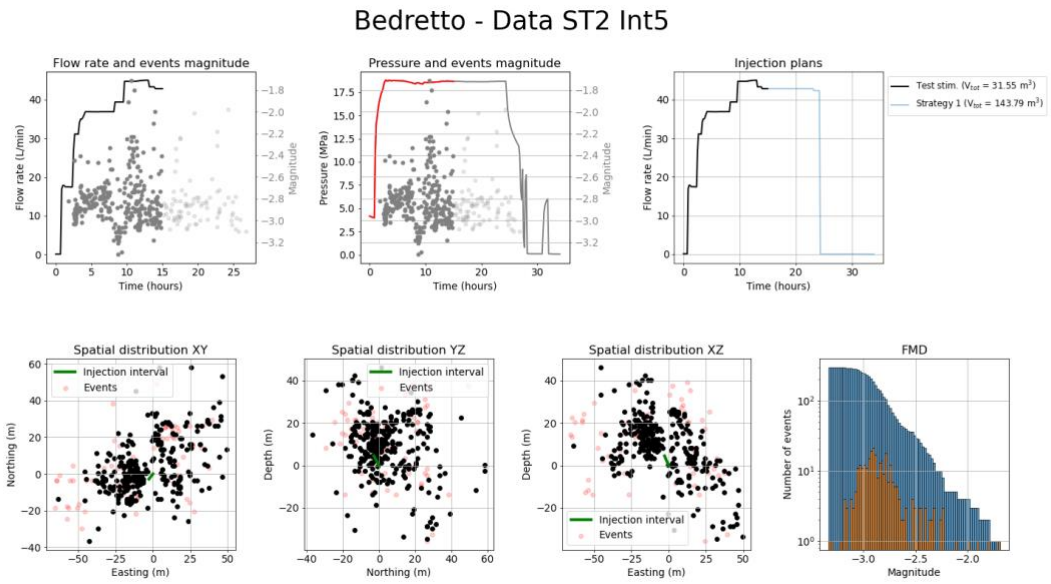

Figure 1: Data collected during the November 2020 injection experiment at the Bedretto Underground Lab. From top right to bottom left: Panel 1) Plot of seismicity and flow rate vs time. The solid line is the cumulative flow rate, circles are earthquakes. The dark color for line and circle indicates the training phase, light colors are validation phase. Panel 2) Injection plan. Panel 3) Pressure evolution and seismicity. Panel 4-6) Spatial distribution of seismicity, black circle represents earthquakes during the training phase, red during the validation phase. Cartesian reference system is with respect the top edge of the injection interval in the well. Panel 5) Frequency magnitude distribution of all earthquakes, orange absolute and blue cumulative frequencies.

As discussed in the previous sections, we fit the model and parameter uncertainties in the training phase for both EM1 and HM0 models (Fig. 2). After time 15 hours, we use the model in a predictive way to forecast the validation phase by only using hydraulic data as input (Fig. 2). For EM1 models we produced 1000 synthetic seismicity catalogs by tapping the parameters from their uncertainty distributions (see subsection 3.1) while for HM0 we simulated a single realization. The results are reported in Fig. 2, where we present for EM1 models the median as best estimate of the number of events together with the variability of the simulated catalogs and parameters (shaded areas in Fig. 2 left and middle panel), while for HM0 we also show the single simulated catalog along with the input pressure profile used (Fig. 2 right panel). The median forecast catalogs of EM1 models overestimate the observed seismicity rate (thin black line in Fig. 2, left panel). However, the larger variability of forecast catalogs of EM1_BH can well reproduce the observed earthquake rate (see blue shaded area in Fig. 2 left panel). This large variability in the forecast arises from the larger uncertainties associated with this EM1_BH (see Fig. 2 middle panel) compared with EM1_MLE. This implies that even if EM1_BH has larger uncertainties associated to the model parameter and apparently less precise, this model is more accurate in the simulation of forecast catalogs and better describe the temporal evolution of the observed seismicity of the validation phase. HM0 instead slightly underestimates the observed seismicity in the validation phase suggesting that Monte Carlo simulation of this model can be a promising path to explore. We however, postpone any discussion of the results of HM0 to after the implementation a stochastic simulation of this model within their uncertainties (see Outlook Section).

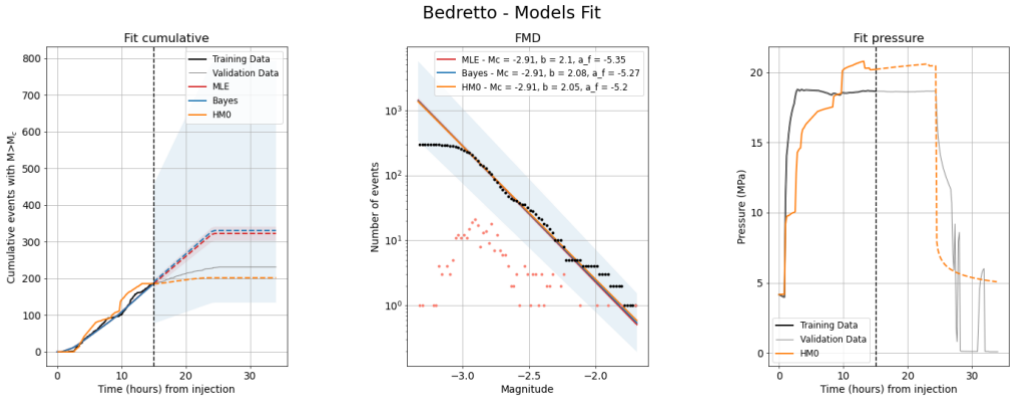

Figure 2: From left to right: panel 1) Cumulative number of events for observed data and simulated by the three seismicity models used in main text. Vertical dashed line indicate separation in training and validation data. The light blue and red shaded areas represent the 25th and 75th percentiles plotted around the median value of the N synthetics catalogs simulated from EM1_MLE and EM1_BH., respectively.. Panel 2) Frequency magnitude distribution and fit, shaded area (light blue and red) indicates the uncertainties in the parameters of Eq.1. Panel 3) Pressure profile (black/gray) and fit (orange) for model HMO.

We have calculated the information gain as punctual and total PG using EM1_MLE as a reference model in time bins of 1800 sec. We present the results in Fig. 3 as plots of the terms in Eq. 2, namely the cumulative evolution of the log-likelihood values (left panel) for the three models and the PG curves form EM1_B H and HMO against EM1_MLE (right panel) in the form of cumulative plots. As expected from the forecast results presented in Fig. 2, the best forecast performance is of EM1_BH which outperforms the other two models. The second model is HM0 for which unfortunately we do have only a single forecast catalog to test and not many scenarios.

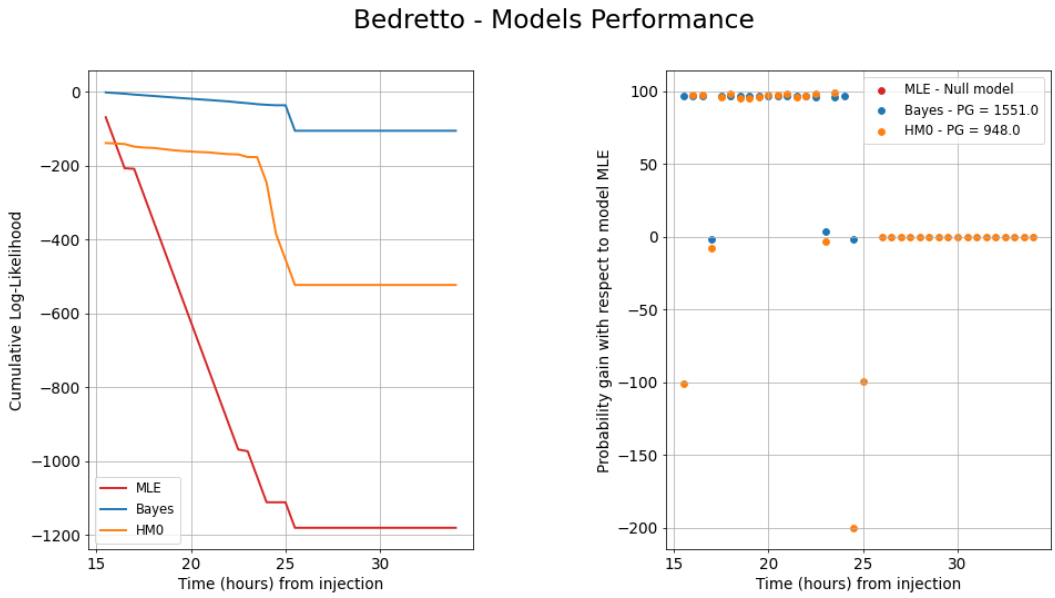

Figure 3: From left to right: Panel 1) Cumulative value of the log-likleihood calculated from eq. 2 (see subsection 3.2 for details). Panel 2) Cumulative plot of the punctual probability gain (PG); the total PG for each model is indicated in the legend. The PG is calculated against the EM1_MLE model and indicates that EM1_BH and HMO perform almost always better in forecasting seismicity than EM1_MLE.

In summary, for the dataset from the Bedretto Underground Lab, where the number of induced earthquakes is relatively small, the best model is the one of larger parameter uncertainties, i.e. EM1_BH. The EM1_MLE model, where the model parameters in Eq.1 are well constrained by the data, has an intrinsic smaller variability in the forecast and it is not able to reproduce the observation in the validation phase. We want to point out that both EM1_MLE and the median model EM1_BH (blue and red dashed lines in Fig. 2 left panel) are very similar in the forecast performance. This arises from

the use of the small amount of data in the likelihood function of the two models. The difference with the Bayesian approach is that can incorporate additional information in the prior distribution model that combined with the likelihood model (data model) produce a better estimation of the epistemic uncertainties when a small sample dataset is used to fit the model. This is a well-known advantage of using Bayesian hierarchical modeling in cases where only the data per se are not enough to capture the variability of the process. In essence, the larger uncertainties in the model parameters are not always indicative of a poor fit, rather they indicate that data alone cannot suffice to fully capture the epistemic uncertainties in the studied process.

4.2 Basel 2006

We used the hydraulics and seismicity data recorded during the Basel 2006 injection experiment. The time of occurrence, magnitudes and distribution of earthquakes are shown in Fig. 4 together with the injected flow rate and pressure. The full dataset spans 12 days and a total of about 2000 earthquakes were recorded with magnitudes ranging from 0 to 3. As before, we split the dataset at time 100 hours after the start of the injection with training phase containing about 600 earthquakes and the rest belong to the validation phase (Fig. 2). The seismicity rate follows reasonably well the injection flow rate corroborating the use of the models presented in Section 2.

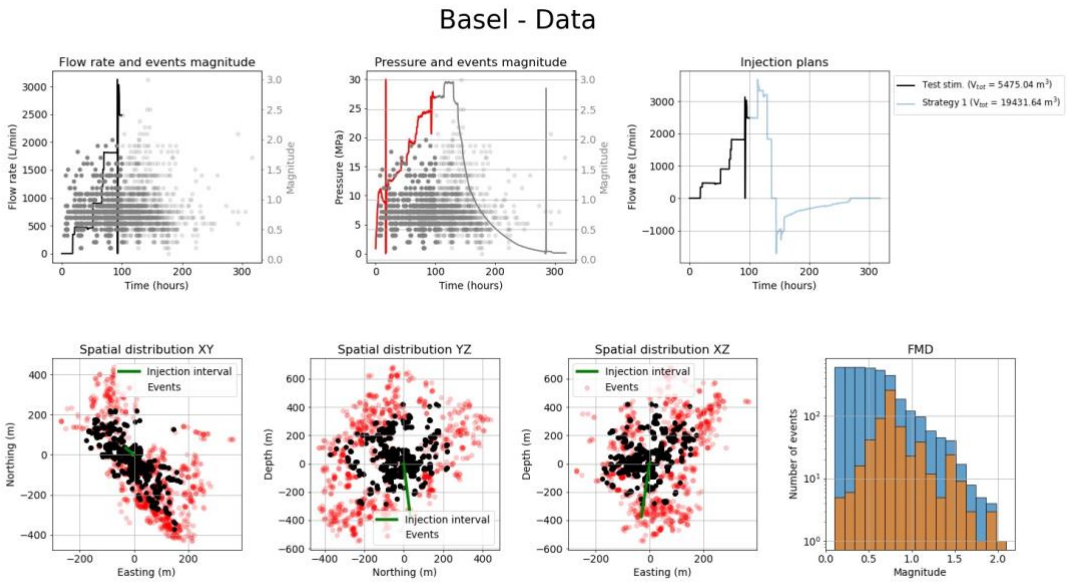

Figure 4: Data collected during the stimulation of the Basel EGS in 2006. From top right to bottom left: Panel 1) Plot of seismicity and flow rate vs time. The solid line is the cumulative flow rate, circles are earthquakes. The dark color for line and circle indicates the training phase, light colors are validation phase. Panel 2) Injection plan. Panel 3) Pressure evolution and seismicity. Panel 4) Spatial distribution of seismicity, black circle represents earthquakes during the training phase, red during the validation phase. Panel 5) Frequency magnitude distribution of all earthquakes, orange absolute and blue cumulative frequencies.

We use the same fitting strategy as the Bedretto Underground Lab dataset: we train the models in the first 100 hours of data and forecast 1000 synthetic catalogs for the validation phase (Fig. 5), while we calibrate the HM0 with pressure and simulate a single forecast catalog (Fig. 5). The results reported in Fig. 5 indicate a good match between the forecast median EM1_BH and EM1_MLE while HM0 systematically overestimate the observed seismicity rate. The uncertainties in the model parameter for EM1 models are smaller compared with the previous case study, and this is because of the larger dataset of Basel 2006 (Fig. 5 middle panel). As a consequence, the variability of the simulated forecast catalogs is smaller in Basel 2006 than in Bedretto (Fig. 5 left panel). In this case both EM1 models can produce reasonably well the data in forecast for the validation phase.

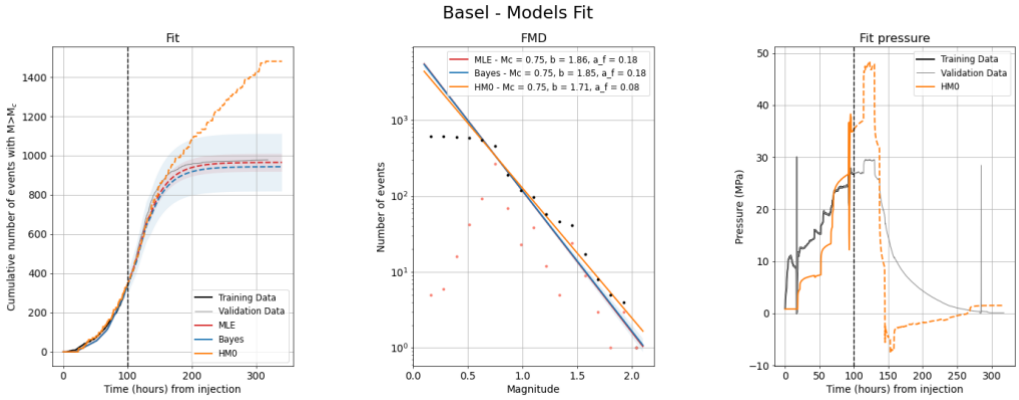

Figure 5: From left to right: panel 1) Cumulative number of events for observed data and simulated by the three seismicity models used in main text. Vertical dashed line indicate separation in training and validation data. The light blue and red shaded areas represent the 25th and 75th percentiles plotted around the median value of the N synthetics catalogs simulated from EM1_MLE and EM1_BH. Panel 2) Frequency magnitude distribution and fit, shaded area (light blue and red) indicates the uncertainties in the parameters of Eq.1. Panel 3) Pressure profile (black/gray) and fit (orange).

We have calculated the information gain as punctual and total PG using EM1_MLE as a reference model in time bins of 3600 sec. We present the results in Fig. 6 as plots of the term in Eq. 2, namely the log-likelihood values (left panel) for the three models and the PG curves form EM1_B H and HMO against EM1_MLE (right panel) in the form of cumulative plots. For the model comparison performance, EM1_BH performs best compared to EM1_MLE although it is the model with largest uncertainties associated to the model parameters.

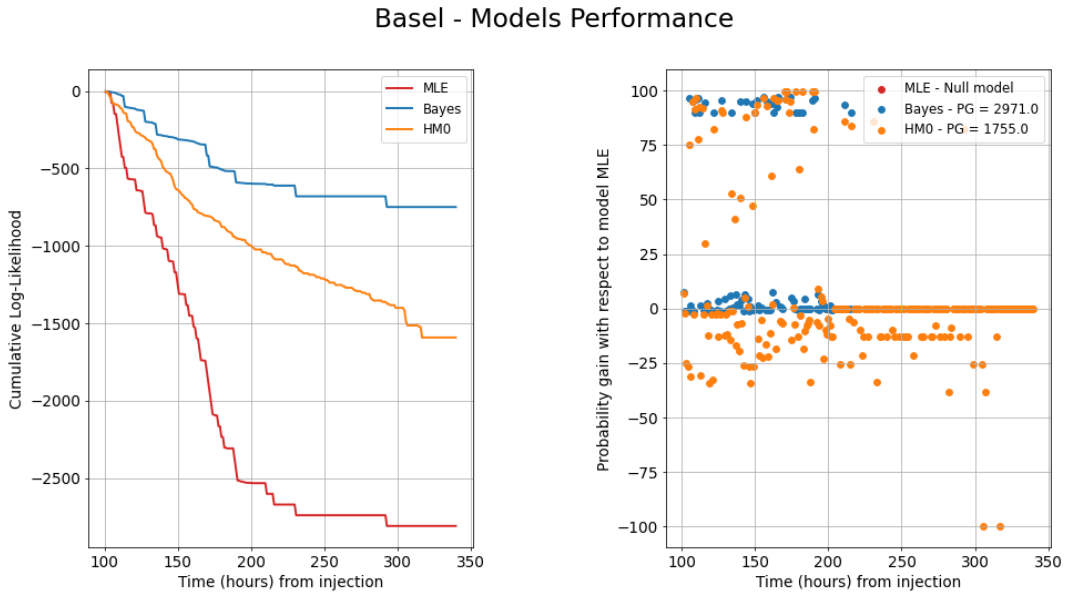

Figure 6. From left to right: Panel 1) Cumulative value of the log-likleihood calculated from eq. 2 (see subsection 3.2 for details). Panel 2) Cumulative plot of the punctual probability gain (PG), the total PG for each model is indicated in the legend. The PG is calculated against the EM1_MLE model and indicates that EM1_BH and HMO perform almost always better in forecasting seismicity than EM1_MLE.

5. Conclusions and Outlook

In the current deliverable we have successfully implemented a probabilistic model comparison tool. While the CSEP community is highly active in developing standards for comparing seismicity forecasting model, they rely on a general Poissonian assumption that may not be reliable for induced seismicity models. Results show that with the currently

developed method, based on empirical distribution, the uncertainties are better taken into account, in particular for the case of little amount of data collected. We compare both models with and without empirical distribution, and demonstrate how models with apparent similar average prediction in reality perform quite differently from a statistical perspective.

The model comparison tool allows a quick and immediate methodology to assess the forecast performances of two models in reproducing the data in a forecast perspective. At the same time, the approach is extremely flexible, and it can be implemented for any type of seismicity model, and it is comparable with the standard developed by the CSEP community.

In the future, we plan to further improve the current approach by introducing an additional test in order to assess the quality of fit of a single model against the data in the validation phase (e.g. the N test as developed by the CSEP community to assess the over- and under-fitting of a model to the data). Future work will also see the testing of the approach on fully pseudo-perspective test. In particular, we plan to introduce a data assimilation scheme, where each model will be updated using the "future" data of each time bin in the validation phase following a pseudo real-time strategy.

Ideally, all future models within DEEP will feature the simulation of N synthetic catalog, and for this reason we also plan to implement a stochastic simulation strategy for HM0 in order to obtain uncertainties in the parameters.

Crucial will be the definition on how to perform the comparison for real-time application. In the current deliverable, we compared the model forecast on known data for each given sequence (the validation or forecast dataset). In real-time such a dataset is not yet available at the time of running the models, hence there are two possible solution that will be investigated. For the first approach, the training dataset will be split to allow some data for validation, although this would mean that not all available information will be used for calibrating the models. This approach is certainly very similar to what it was shown in this deliverable, and relies on the fact that more time bins are simulated (to model both the data in the validation dataset and to run the actual forecast on unknow period). A second approach would rely on past simulations: the data collected at time T could be used to compare performances of models that run at time T-1. In this way for each time T, the entirely available dataset could be used to train the models, although the weighing is actually done for past data, and it may not be accurate for the next forecast.

While performing nicely, the proposed tool only compares forecasts in terms of seismicity, and does not account how well a given model can reproduce pressure (if simulated). Basically, a model with a completely off pressure solution, but with good matching for the seismicity evolution, will be considered "better" than others, while the eye of an expert modeler would reject such unphysical model. The current approach could be extended to work also for the pressure forecasting, with similar assumptions on the empirical distribution of the forecasted solution, but the comparison will never be fair between models of different classes (e.g. empirical and hydromechanical models). However, it should be noted that the model testbench is not meant to select one model, rather to assign weights to their forecasts to be used in hazard predictions. In this term, a fairer comparison would be to assign equal weights to the various model class, and perform comparison between models of the same class.

The standardized approach will be available in the future as open-source toolbox (Task 5.3), and it constitutes a solid foundation for future model development. New or improved models can be plugged into the testbench and compared against a range of past sequences, so that modelers have rapid feedback on their models performance, speeding up model optimisation tremendously. It also establishes a benchmark metrics to assess how 'good' a model is, and how it compares to others. This will provide a rational baseline also for operators and decision makers to establish how much trust they can have in certain models and their forecast, and a baseline to develop good practise guidelines (Task 5.1). Finally, the testbench is a metric that through time will establish how the community has progressed in their skill to forecast induced seismicity in the context of deep geothermal energy exploitation. The standardize testbench approach will be fed with models from DEEP (Tasks 3.2 and 3.4) but will be opened for other modellers, representing a highly valuable resource for the FORGE community.

Liability Claim

Reference List

1. E. Király-Proag, V. Gischig, J. D. Zechar, S. Wiemer, Multicomponent ensemble models to forecast induced seismicity. *Geophys. J. Int.* **212** (2018), doi:10.1093/gji/ggx393.

2. E. Király-Proag, J. D. Zechar, V. Gischig, S. Wiemer, D. Karvounis, J. Doetsch, Validating induced seismicity forecast models—Induced Seismicity Test Bench. *J. Geophys. Res. Solid Earth*. **121** (2016), doi:10.1002/2016JB013236.

3. S. Nandan, G. Ouillon, D. Sornette, S. Wiemer, Forecasting the full distribution of earthquake numbers is fair, robust, and better. *Seismol. Res. Lett.* **90** (2019), doi:10.1785/0220180374.

4. A. Mignan, M. Broccardo, S. Wiemer, D. Giardini, Induced seismicity closed-form traffic light system for actuarial decision-making during deep fluid injections. *Sci. Rep.* **7** (2017), doi:10.1038/s41598-017-13585-9.

5. M. Broccardo, A. Mignan, S. Wiemer, B. Stojadinovic, D. Giardini, Hierarchical Bayesian Modeling of Fluid-Induced Seismicity. *Geophys. Res. Lett.* **44** (2017), doi:10.1002/2017GL075251.

6. M. Hertrich, B. Brixel, K. Broeker, T. Driesner, N. Gholizadeh, D. Giardini, D. Jordan, H. Krietsch, S. Loew, X. Ma, H. Maurer, M. Nejati, K. Plenkers, M. Rast, M. Saar, A. Shakas, R. van Limborgh, L. Villiger, Q. C. Wenning, F. Ciardo, P. Kaestli, A. Obermann, A. P. Rinaldi, S. Wiemer, A. Zappone, F. Bethmann, R. Castilla, F. Christe, B. Dyer, D. Karvounis, P. Meier, F. Serbeto, F. Amann, V. Gischig, B. Valley, Characterization, Hydraulic Stimulation, and Fluid Circulation Experiments in the Bedretto Underground Laboratory for Geosciences and Geoenergies (2021).

7. M. O. Häring, U. Schanz, F. Ladner, B. C. Dyer, Characterisation of the Basel 1 enhanced geothermal system. *Geothermics*. **37** (2008), doi:10.1016/j.geothermics.2008.06.002.

8. S. A. Shapiro, C. Dinske, J. Kummerow, Probability of a given-magnitude earthquake induced by a fluid injection. *Geophys. Res. Lett.* **34** (2007), doi:10.1029/2007GL031615.

9. Y. Y. Kagan, L. Knopoff, Statistical short-term earthquake prediction. *Science (80-.).* **236** (1987), doi:10.1126/science.236.4808.1563.

10. L. Passarelli, L. Sandri, A. Bonazzi, W. Marzocchi, Bayesian Hierarchical Time Predictable Model for eruption occurrence: an application to {Kilauea Volcano}. *Geophys J Int*. **181**, 1525–1538 (2010).

11. L. Passarelli, E. Rivalta, S. Jónsson, M. Hensch, S. Metzger, S. S. Jakobsdóttir, F. Maccaferri, F. Corbi, T. Dahm, Scaling and spatial complementarity of tectonic

earthquake swarms. *Earth Planet. Sci. Lett.* **482**, 62–70 (2018).