Question-Driven Ensembles of Flexible ETAS Models

Leila Mizrahi*[1], Shyam Nandan[1], William Savran[2], Stefan Wiemer[1], and Yehuda Ben-Zion[2]

Abstract

The development of new earthquake forecasting models is often motivated by one of the following complementary goals: to gain new insights into the governing physics and to produce improved forecasts quantified by objective metrics. Often, one comes at the cost of the other. Here, we propose a question-driven ensemble (QDE) modeling approach to address both goals. We first describe flexible epidemic-type aftershock sequence (ETAS) models in which we relax the assumptions of parametrically defined aftershock productivity and background earthquake rates during model calibration. Instead, both productivity and background rates are calibrated with data such that their variability is optimally represented by the model. Then we consider 64 QDE models in pseudoprospective forecasting experiments for southern California and Italy. QDE models are constructed by combining model parameters of different ingredient models, in which the rules for how to combine parameters are defined by questions about the future seismicity. The QDE models can be interpreted as models that address different questions with different ingredient models. We find that certain models best address the same issues in both regions, and that QDE models can substantially outperform the standard ETAS and all ingredient models. The best performing QDE model is obtained through the combination of models allowing flexible background seismicity and flexible aftershock productivity, respectively, in which the former parameterizes the spatial distribution of background earthquakes and the partitioning of seismicity into background events and aftershocks, and the latter is used to parameterize the spatiotemporal occurrence of aftershocks.

Cite this article as Mizrahi, L., S. Nandan, W. Savran, S. Wiemer, and Y. Ben-Zion (2023). Question-Driven Ensembles of Flexible ETAS Models, *Seismol. Res. Lett.* **94**, 829–843, doi: 10.1785/0220220230.

Introduction

Earthquake forecasting is one of the defining problems of seismology. To provide useful solutions, forecasting models use a wide range of approaches: Coulomb rate-and-state (CRS) models (Cocco *et al.*, 2010; Parsons *et al.*, 2012; Mancini *et al.*, 2019) calculate Coulomb stress changes and couple them with a lab-based constitutive friction law (Dieterich, 1994). On the other end of the spectrum are statistical models, with the epidemic-type aftershock sequence (ETAS) model being the best performing current statistical approach (Cattania *et al.*, 2018; Taroni *et al.*, 2018). First introduced by Ogata (1988), it models seismicity rate as the sum of background and aftershock events, where aftershocks are triggered according to regional empirical laws. In between the purely physics-based and purely statistics-based approaches are models such as the short-term earthquake probability (STEP) model (Gerstenberger *et al.*, 2005), the Inlabru model (Bayliss *et al.*, 2020), and hybrid Coulomb and statistical models (Steacy *et al.*, 2014). The STEP model combines clustering principles with fault information in a statistical model to produce time-dependent forecasts. The Inlabru model more generally allows the inclusion of diverse data sets as covariates to issue time-independent seismicity forecasts. A hybrid Coulomb/statistical model redistributes seismicity forecasted by STEP according to Coulomb stress changes.

While physics-based models aim to describe the processes and mechanisms underlying seismogenesis, statistical models are generally more empirical and data driven. Ultimately, "all models are wrong, but some are useful," to cite the famous statistician George Box (1979). Usefulness can be viewed from different perspectives. Different forecasting models can be useful for gaining new scientific insight, for producing the most accurate forecasts, or for producing forecasts that are most suited for operational earthquake forecasting (OEF), given the trade-off between accuracy and computational cost. Cattania *et al.*

1. Swiss Seismological Service, ETH Zurich, Switzerland, https://orcid.org/0000-0002-5262-3168 (LM); https://orcid.org/0000-0002-4954-5314 (SN); https://orcid.org/0000-0002-4919-3283 (SW); 2. University of Southern California, Los Angeles, California, U.S.A., https://orcid.org/0000-0001-5404-0778 (WS); https://orcid.org/0000-0002-9602-2014 (YB-Z)

*Corresponding author: leila.mizrahi@sed.ethz.ch

(2018) found in a pseudoprospective forecasting experiment for the 2010–2012 Canterbury, New Zealand, earthquake sequence that hybrid Coulomb/statistical models have a similar forecasting skill as CRS models, at a lower computational effort. Mancini *et al.* (2019, 2020) conducted pseudoprospective experiments for the 2016 central Italy and the 2019 Ridgecrest, California, sequences, comparing CRS models of different complexity with ETAS forecasts. In both studies, the forecasting skill of CRS models increases with their complexity, with the most complex CRS model performing similarly to ETAS. Hardebeck (2021) investigated possible reasons for the general underperformance of the physics-based models relative to statistical models and suggested that understanding and incorporating heterogeneities in background conditions into physical forecasting models may be key in improving their skill.

Having been tested thoroughly and systematically (Woessner *et al.*, 2011; Ogata *et al.*, 2013; Strader *et al.*, 2017; Taroni *et al.*, 2018; Nandan *et al.*, 2019b; Savran *et al.*, 2020), ETAS models meanwhile remain the state-of-the art of earthquake forecasting and are being used or considered for OEF at various locations (Marzocchi *et al.*, 2014; Rhoades *et al.*, 2016; Field *et al.*, 2017; Kamer *et al.*, 2021; Nandan, Kamer, *et al.*, 2021; van der Elst *et al.*, 2022). Besides using the most basic formulation of ETAS, modelers also commonly refine the model. For instance, Bach and Hainzl (2012) enhanced ETAS with fault information, ShakeMaps, ground-motion models, or Coulomb stress changes. Seif *et al.* (2017) assessed the biasing effects of data incompleteness and model assumptions on the estimated ETAS parameters. Several techniques have been proposed to address the effects of short-term aftershock incompleteness (Mizrahi *et al.*, 2021b; Grimm *et al.*, 2022; Hainzl, 2022) or the assumption of isotropic aftershock triggering (Grimm *et al.*, 2022; Page and van der Elst, 2022). Other studies focus on deriving spatial variations of ETAS parameters or background seismicity (Enescu *et al.*, 2009; Nandan *et al.*, 2017; Nandan, Ram, *et al.*, 2021), also relating parameter variations with physical quantities such as heat flow. Others have refined the standard ETAS model with a relationship between magnitudes of triggered and triggering earthquakes and a magnitude-dependent Omori kernel and found the resulting models to possess improved forecasting performance (Nandan *et al.*, 2019; Nandan, Kamer, *et al.*, 2021). A recent framework for modeling seismicity with an invariant Galton–Watson stochastic branching process provides a generalization of ETAS that is invariant with respect to various common deficiencies of earthquake catalogs (Kovchegov *et al.*, 2022). However, this framework has not yet been used for forecasting seismicity.

A related forecasting topic, which has recently received attention, is ensemble modeling (Rhoades and Gerstenberger, 2009; Marzocchi *et al.*, 2012; Taroni *et al.*, 2014; Bird *et al.*, 2015; Akinci *et al.*, 2018; Llenos and Michael, 2019; Bayona *et al.*, 2021). The idea, widely used for decades in the meteorological and climate forecasting community (Tracton and Kalnay, 1993; Leutbecher and Palmer, 2008; Eyring *et al.*, 2016), is to combine different models in an overarching ensemble model to obtain more robust forecasts. Commonly, an ensemble is a linear or multiplicative combination of ingredient models (e.g., Bird *et al.*, 2015), and the challenge is to optimize the weights given to each model. In a recent study, Bayona *et al.* (2021) found that the time-independent ensemble models WHEEL and GREAR1 (Bird *et al.*, 2015) outperform the ingredient models of which they consist. Akinci *et al.* (2018) found that their time-independent ensemble model outperforms its ingredients and performs similarly to the best-performing time-independent model tested in the 2009 CSEP experiment (Schorlemmer, Zechar, *et al.*, 2010; Zechar *et al.*, 2010) for Italy. In the context of time-dependent models, Taroni *et al.* (2014) and Gerstenberger *et al.* (2014) used ensemble approaches, and Llenos and Michael (2019) found that ensembles of ETAS models perform best for the 2015 San Ramon, California, Swarm. Shebalin *et al.* (2014) proposed an iterative method to combine forecasting models and found the resulting models to have advantageous properties compared to the ingredient models or traditional linear combinations thereof. The emerging consensus across the mentioned studies is that ensemble modeling is a promising path to use for earthquake forecasting; this is also demonstrated by the fact that they are currently implemented in Italy's OEF system (Marzocchi *et al.*, 2014). Yet, a breakthrough of ensemble models as established in the meteorological forecasting community is still pending.

For practical operational forecasting, especially in regions that are less studied due to a lack of data or resources, a balance must be achieved between model accuracy and simplicity. With this in mind, we relax some of the assumptions behind ETAS. We allow aftershock productivity and background seismicity to be described nonparametrically, providing event-specific productivity and background rates. This aims to better capture the real behavior of seismicity without making any choices on resolution, parametric form, and so on. Using pseudoprospective forecasting experiments in southern California and Italy, we evaluate whether these flexible ETAS (flETAS) models provide superior forecasts.

We also propose a novel approach for QDE modeling, fundamentally different from traditional ensemble modeling approaches. In the QDE approach, models are combined in the parameter space as opposed to the solution space. Several ETAS-like models are fit to the observed data, yielding an individual set of parameters for each model. A QDE model is then created by defining a new set of parameters based on a combination of the ingredient model parameters. The rules to combine parameters are defined by dividing the forecasting problem into several subproblems. Each subproblem addresses a question regarding the number of forecasted events or the spatiotemporal distribution of either background earthquakes or aftershocks. A QDE model can be viewed as a model that

addresses different questions with different ingredient models. This approach allows the combination of ETAS variants but can be extended to combining more general types of seismicity models.

By including such QDE models in the forecasting experiments, we assess their forecasting capability in comparison with their ingredient models, standard ETAS, and flETAS. At the same time, the QDE approach helps to understand which ingredient models are best suited to solve different forecasting subproblems, thus, making it useful from the perspective of gaining new scientific insight.

The remainder of this article is structured as follows. We describe flETAS models and the QDE approach in the next section flETAS models. The setup for the forecasting experiments, the data analyzed, and the metrics used to evaluate forecasting performance are described in the Forecasting experiments section. We present and discuss our results in the Results and discussion section and finally provide our Conclusions.

flETAS Models

The following subsections describe flETAS models and explain the QDE modeling. We begin by explaining the algorithm used to estimate the parameters of the ETAS model. Then, we describe how to relax some parametric assumptions of the ETAS model. Finally, we introduce a framework to create QDEs of flETAS models.

Expectation maximization algorithm

Consider an earthquake catalog:

$$C = \{e_i = (m_i,t_i,x_i,y_i), i \in \{1,\ldots,n\}\}, \tag{1}$$

consisting of events e_i of magnitudes m_i, which occur at times t_i and locations (x_i,y_i).

The ETAS model describes earthquake rate as

$$\lambda(t,x,y|\mathcal{H}_t) = \mu + \sum_{i:t_i<t} g(m_i,t-t_i,x-x_i,y-y_i). \tag{2}$$

That is, the sum of background rate μ and the rate of all aftershocks of previous events e_i. The aftershock triggering rate $g(m,\Delta t,\Delta x,\Delta y)$ describes the rate of aftershocks triggered by an event of magnitude m, at a time delay of Δt and a spatial distance $(\Delta x,\Delta y)$ from the triggering event. We use here the definition:

$$g(m,\Delta t,\Delta x,\Delta y) = \frac{k_0 \times e^{a(m-m_{\text{ref}})} \times e^{-\Delta t/\tau}}{\left((\Delta x^2+\Delta y^2)+d\times e^{\gamma(m-m_{\text{ref}})}\right)^{1+\rho} \times (\Delta t+c)^{1+\omega}}, \tag{3}$$

as in Nandan, Kamer, *et al.* (2021) and Mizrahi *et al.* (2021a). This formulation differs from other, more commonly used formulations of ETAS models in that it uses an Exponentially Tapered Omori Kernel (ETOK). In their article, Nandan, Kamer, *et al.* (2021) compare the ETAS model with ETOK to a more general version thereof, MDOK, which allows a magnitude dependency, finding that the more general version allows better forecasts. This indicates that including an exponential taper does lead to improved forecasts when compared to the commonly used Omori kernel. Besides allowing less heavy tails in the temporal distribution of aftershocks, this formulation of the Omori kernel makes it possible for the parameter ω to attain negative values, which is not possible in the traditional formulation. Also, our choice of this base model does not impact the main conclusions that can be drawn from comparing it to modified versions of itself.

To calibrate the ETAS model, the nine parameters to be optimized are the background rate μ and k_0, a, c, ω, τ, d, γ, ρ, which parameterize the aftershock triggering rate $g(m, t, x, y)$ given in equation (3). Implicitly, the model assumes that only earthquakes with magnitudes larger than or equal to m_{ref} can trigger aftershocks. Most applications of the method define m_{ref} as equal to the constant value of m_c.

We build on the expectation maximization (EM) algorithm to estimate the ETAS parameters (Veen and Schoenberg, 2008). In this algorithm, the expected number of background events $\hat{n}$ and the expected number of directly triggered aftershocks $\hat{l}_i$ of each event e_i are estimated in the expectation step (E step), along with the probabilities p_{ij} that event e_j was triggered by event e_i, and the probability p_j^{ind} that event e_j is independent. Following the E step, the nine parameters are optimized to maximize the complete data log likelihood in the maximization step (M step). E and M steps are repeated until convergence of the parameters. The usual formulation of the EM algorithm defines:

$$\hat{n} = \sum_j p_j^{\text{ind}}, \tag{4}$$

$$\hat{l}_i = \sum_j p_{ij}, \tag{5}$$

and st

$$p_{ij} = \frac{g_{ij}}{\mu + \sum_{k:t_k<t_j} g_{kj}}, \tag{6}$$

$$p_j^{\text{ind}} = \frac{\mu}{\mu + \sum_{k:t_k<t_j} g_{kj}}, \tag{7}$$

with $g_{kj} = g(m_k,t_j-t_k,x_j-x_k,y_j-y_k)$ being the aftershock triggering rate of e_k at location and time of event e_j. For a given target event e_j, equations (6) and (7) define p_{ij} to be

proportional to the aftershock occurrence rate g_{ij}, and p_j^{ind} to be proportional to the background rate μ. As an event must be either independent or triggered by a previous event, the normalization factor $\Lambda_j := \mu + \sum_{k:t_k<t_j} g_{kj}$ in the denominator of equations (6) and (7) stipulates that $p_j^{\text{ind}} + \sum_{k:t_k<t_j} p_{kj} = 1$.

Introducing flexibility

In the formulation of the ETAS model given in Equation (2), the rate of background earthquakes is described by the parameter μ, which does not vary with space nor time. During the maximization step of the EM algorithm, μ can be estimated independently from the other parameters as

$$\mu = \frac{\hat{n}}{A_R \times T}, \tag{8}$$

in which A_R and T denote the area of the study region and the length of the considered time window, respectively. In some approaches, the region of interest is divided into several subregions, which can have their own values for μ (Veen and Schoenberg, 2008). An iterative algorithm to estimate spatial variations of background rate based on maximum-likelihood estimation (Zhuang, 2012) uses a Gaussian kernel smoothing applied to the catalog event locations, weighted by their estimated independence probability, to obtain an estimate of $\mu(x,y)$. Here, we present a similar approach using EM, which has been shown to be more stable with respect to the initial conditions compared to maximum-likelihood approaches (Veen and Schoenberg, 2008). Our approach is similar yet not identical to the one described by Nandan, Ram, *et al.* (2021), which uses a regularized inverse power law for smoothing the locations. We define the background rate at a location (x, y) as

$$\mu(x,y) = \frac{1}{T} \times \sum_j p_j^{\text{ind}} \times k(\Delta x_j, \Delta y_j), \tag{9}$$

in which $k(\Delta x_j, \Delta y_j)$ is the Gaussian kernel with bandwidth σ applied to the distance $(\Delta x_j, \Delta y_j)$ of event e_j to the location (x,y),

$$k(\Delta x, \Delta y) = \frac{1}{2\pi\sigma^2} \times \exp\left(-\frac{1}{2} \times \frac{\Delta x^2 + \Delta y^2}{\sigma^2}\right). \tag{10}$$

The bandwidth σ determines the smoothness of the background event density. In principle, σ could be calibrated itself, but we choose to fix it to 5 km for simplicity. Our next modification to the standard ETAS model is to allow flexibility of the aftershock probability. The number of directly triggered aftershocks $\hat{l}_j$ is estimated during the expectation step of the EM algorithm as described in equation (5). We can, thus, replace the term $k_0 \times e^{a(m-m_{\text{ref}})}$ in equation (3) with κ_j, in which κ_j is stipulated to be proportional to $\hat{l}_j$. Instead of parameterizing aftershock productivity to be exponentially increasing with the magnitude of the triggering event, we allow each event to have its own productivity. This yields:

$$g_{j_{\theta,\kappa_j}}(m,\Delta t,\Delta x,\Delta y) = \frac{\kappa_j \times e^{\Delta t/\tau}}{\left((\Delta x^2 + \Delta y^2) + d \times e^{\gamma(m-m_{\text{ref}})}\right)^{1+\rho} \times (\Delta t + c)^{1+\omega}}, \tag{11}$$

for given parameters $\theta = (c,\omega,\tau,d,\gamma,\rho)$ and κ_j. The EM algorithm is adapted as follows:

1. Define initial estimates of κ_j as $\kappa_j = e^{a(m_j - m_{\text{ref}})}$ with a random guess for a;
2. define initial estimates of independence probability $p_j^{\text{ind}} \equiv 0.1$. The inversion result is not sensitive to this choice;
3. define random initial guesses for the parameters $\theta = (c,\omega,\tau,d,\gamma,\rho)$;
4. expectation step: calculate $\hat{n},\hat{l}_j,p_{ij},p_j^{\text{ind}}$ using the current estimates of κ_j,θ, and p_j^{ind}. p_{ij},p_j^{ind} are calculated using equations (6) and (7), but using the flexible definitions of g_{ij} and $\mu(x,y)$ of equations (9) and (11);
5. maximization step: optimize the parameters θ to minimize the complete data log likelihood (see Mizrahi *et al.*, 2021a for details), given the current estimates of $\hat{n},\hat{l}_j,p_{ij},p_j^{\text{ind}}$;
6. update κ_j^{new} to be $\kappa_j^{\text{old}} \times \frac{\hat{l}_j}{G_{j_{\theta,\kappa_j^{\text{old}}}}}$, in which $G_{j_{\theta,\kappa_j^{\text{old}}}}$ is the expected total number of aftershocks of e_j, given θ and κ_j^{old}. This ensures that $\hat{l}_j = G_{j_{\theta,\kappa_j^{\text{new}}}}$. We calculate $G_{j_{\theta,\kappa_j}}$ as

$$G_{j_{\theta,\kappa_j}} = \iint_R \int_0^{t_{\text{end}}-t_j} g_{j_{\theta,\kappa_j}}(m_j,t,x,y)\, dt\, dx\, dy, \tag{12}$$

   in which t_{end} is the end time of the considered time window, and we assume the spatial region R to extend infinitely in space, allowing a facilitated, asymptotically unbiased estimation of ETAS parameters (Schoenberg, 2013), and
7. repeat from step 4 until convergence of θ, that is, until $\sum_{a_i \in \theta} |a_i^{\text{new}} - a_i^{\text{old}}| < 10^{-3}$.

After the inversion, we calibrate an overall productivity law for the flETAS models with free productivity to avoid over fitting with event-wise productivity. From the individually estimated productivities κ_j of magnitude m_j events, we calibrate a law of the form:

$$\kappa(m) = k_0 \times e^{a(m-m_{\text{ref}})}, \tag{13}$$

by minimizing the sum of absolute residuals between the observed $\bar{\kappa}(m) = \frac{1}{n(m)} \sum_{i:m_i=m} \kappa_i$ and the theoretical $\kappa(m) = k_0 \times e^{a(m-m_{\text{ref}})}$, in which $n(m)$ is the number of events with magnitude m.

Then, productivity is treated the same way as in the case of standard ETAS. In this way, the variability of productivity is

only accounted for during the parameter inversion process and may lead to more accurate estimators of the productivity as well as the remaining ETAS parameters.

QDE modeling

We propose a novel approach for QDE modeling, in which a forecast is created by combining model parameters of different ingredient models. The rules for how parameters can be combined are defined by questions that divide the forecasting problem into several subproblems: *How many background events are expected? Where are they expected? When are they expected? How many aftershocks are expected? Where are they expected? When are they expected?*

By answering each of these questions with different ingredient models, we create a suite of ensembles. The remainder of this section establishes rules to combine parameters based on the questions.

Consider a collection of ETAS or flETAS ingredient models, $(M_i)_{i=0,\dots,n_M}$. As they are sufficiently defined through their parameters, we can write:

$$M_i = (\mu_i,\kappa_i,c_i,\omega_i,\tau_i,d_i,\gamma_i,\rho_i). \tag{14}$$

In case M_i is a flETAS model, $\mu_i = \mu_i(x,y)$ can vary with space. For simplicity, we denote with κ_i the function that assigns to each event its appropriate value to replace the term κ_j in equation (11). In our case, this means that we define $\kappa_i(m) = k_{0_i} \times e^{a_i(m-m_{\text{ref}})}$, in which k_{0_i} and a_i are either obtained during parameter inversion directly, or afterward in case M_i is a flETAS model with free productivity. We chose the notation of κ_i instead of (k_{0_i},a_i) to emphasize this possible distinction. We can then generally describe the aftershock triggering kernel g as

$$g_i(m,\Delta t,\Delta x,\Delta y) = \frac{\kappa_i \times e^{\Delta t/\tau_i}}{\left((\Delta x^2+\Delta y^2)+d_i \times e^{\gamma_i(m-m_{\text{ref}})}\right)^{1+\rho_i} \times (\Delta t + c_i)^{1+\omega_i}}. \tag{15}$$

Let us now revisit the earlier questions.

1. *How many background events are expected?*
   More precisely, what we want to ask here is how many background events do we expect in total in the region R and forecasting horizon $[T_0,T_1]$ we are issuing a forecast for. The answer to this question, given out of the perspective of model M_i, is

$$N_{B_i} = \iint_R \int_{T_0}^{T_1} \mu_i(x,y)\, dt\, dx\, dy. \tag{16}$$

2. *Where and when are they expected?*
   We address for now these two questions jointly. The spatiotemporal density of background events is given by

$$f_{B_i}(x,y,t) = \frac{\mu_i(x,y)}{\iint_R \int_{T_0}^{T_1} \mu_i(x,y)\, dt\, dx\, dy} = \frac{\mu_i(x,y)}{N_{B_i}}, \tag{17}$$

   which is effectively time independent due to our choice of a time independent $\mu(x,y)$.
3. *How many aftershocks are expected?*
   Again, what we want to ask here is how many aftershocks do we expect in total in the region R and forecasting horizon $[T_0,T_1]$ we are issuing a forecast for. For an individual event e_j, we expect it to have n_A aftershocks, in which

$$n_{A_i}(e_j) = \iint_R \int_{T_0}^{T_1} g_i(m_j,t-t_j,x-x_j,y-y_j)\, dt\, dx\, dy. \tag{18}$$

   The total number of aftershocks N_{A_i} is then given as the sum of aftershocks of all events

$$N_{A_i} = \sum_{j:t_j<T_1} n_{A_i}(e_j). \tag{19}$$

4. *Where and when are they expected?*
   We again answer these two questions jointly. If we define

$$G_i(x,y,t) := \sum_{j:t_j<T_1} g_i(m_j,t-t_j,x-x_j,y-y_j), \tag{20}$$

   as the total rate of aftershocks at time t and location (x,y), consisting of the sum of aftershock rates of all events that occurred prior to the end T_1 of the forecasting horizon, the spatiotemporal density of aftershocks is given by

$$f_{A_i}(x,y,t) = \frac{G_i(x,y,t)}{\iint_R \int_{T_0}^{T_1} G_i(x,y,t)\, dt\, dx\, dy} = \frac{G_i(x,y,t)}{N_{A_i}}. \tag{21}$$

We now construct a QDE model E^{klm} as follows. The number questions (1) and (3) are answered with model M_k, the background density question (2) is answered with model M_l, and the aftershock density question (4) is answered with model M_m. Questions (1) and (3) are addressed with the same model. This is a choice made to avoid unrealistic event numbers. If one model interprets the majority of events as background, and another model interprets the majority of events to be aftershocks, answering the two questions with two different models would lead to exceptionally high or low total event numbers, which is not intended by the two ingredient models.

In the earlier notation, which identifies a model with its parameters, this would give us:

$$E^{klm} = \left(\mu_l \times \frac{N_{B_k}}{N_{B_l}},\kappa_m \times \frac{N_{A_k}}{N_{A_m}},c_m,\omega_m,\tau_m,d_m,\gamma_m,\rho_m\right). \tag{22}$$

Forecasting Experiments

To test whether flETAS models and QDE models, which consist of ETAS and flETAS models, provide better forecasts, we conduct pseudoprospective forecasting experiments for southern California and Italy.

Competing models

In these experiments, we consider the following four competing ingredient models.

- M_0: standard ETAS.
- M_1: flETAS with free productivity and standard background.
- M_2: flETAS with standard productivity and free background.
- M_3: flETAS with free productivity and free background.

Out of these, $4^3 = 64$ QDE models can be constructed.

M_2 is conceptually close to the models described by Zhuang (2012) and Nandan, Ram, *et al.* (2021).

Evaluation metric

We use interevent time horizons: whenever an event occurs, a forecast is issued, which is valid until the occurrence of the next event. A pseudoprospective model evaluation then aims to capture how well a forecast issued using data until event e_{j-1} can describe the occurrence of the next event e_j.

An ETAS forecast always consists of the forecasted background seismicity rate plus the forecasted aftershock seismicity rate. With this flexible definition of forecasting horizon, our ETAS forecast can be calculated and evaluated analytically.

Consider $\lambda_i(t,x,y|\mathcal{H}_{t_{j-1}})$, the event rate under model M_i as of time t_{j-1} of the $(j-1)$th earthquake. This formulation of λ_i is valid for times $t \in (t_{j-1},t_j]$ between the occurrence of event e_{j-1} and event e_j, and hence, this is the forecasting horizon we consider.

For the traditional experiment settings for which one is interested in the seismicity forecast of the next days, months, or years, such an analytical description of the forecasted seismicity is not possible. As soon as an event occurs during the forecasting period, its aftershocks are not part of the background seismicity, nor of the aftershock seismicity that was calculated at the start of the forecasting period. For this reason, ETAS forecasts for fixed forecasting horizons are usually produced through the simulation of a large number of possible continuations of the catalog.

In our case of flexible forecasting horizons, the log likelihood of observing e_j under model M_i is analytically defined (see Ogata *et al.*, 2013; Daley and Vere-Jones, 2003) as

$$\ln \mathcal{L}_i(e_j) = \ln \lambda_i(t_j,x_j,t_j|\mathcal{H}_{t_{j-1}}) - \iint_R \int_{t_{j-1}}^{t_j} \lambda_i(t_j,x_j,t_j|\mathcal{H}_{t_{j-1}})\, dt\, dx\, dy. \quad (23)$$

We then define the information gain $IG_j^{i_1,i_2}$ of model i_1 over model i_2 during the jth forecasting period $(t_{j-1},t_j]$ as

$$IG_j^{i_1,i_2} = \ln \frac{\mathcal{L}_{i_1}(e_j)}{\mathcal{L}_{i_2}(e_j)} = \ln \mathcal{L}_{i_1}(e_j) - \ln \mathcal{L}_{i_2}(e_j). \quad (24)$$

The information gain per event (IGPE) over forecasting periods $j_1,\ldots,j_K$ is defined as

$$\frac{1}{K} \sum_{k=1,\ldots,K} IG_{j_k}^{i_1,i_2}, \quad (25)$$

the average of IGs over those testing periods.

Compared to evaluation techniques based on the simulation of large numbers of possible catalog continuations such as in Nandan *et al.* (2019a) and Mizrahi *et al.* (2021a), which are encouraged by CSEP (see Savran *et al.*, 2022), this approach allows us to compare models much faster, accelerating the development and testing process. To apply these models operationally, in which forecasts are required for a fixed time horizon, simulations would still be required. This evaluation approach allows us to save time when developing and selecting the model to be used operationally and is especially useful for evaluating a large suite of QDE models.

Data

For southern California, we consider the Advanced National Seismic System (ANSS) comprehensive earthquake catalog (ComCat), in the polygon given by the vertices in Table A1. We consider earthquakes of magnitude $M \geq 2.0$ from 1 January 2010 until 1 January 2022. The first two years serve as an auxiliary period in the ETAS and flETAS parameter inversion, and thus, the start of the primary catalog is 1 January 2012. This means that the events between January 2010 and January 2012 can act as triggering events during the inversion but not as triggered events. Using the method described by Mizrahi *et al.* (2021b), we find that the overall catalog is complete at this threshold, although there are likely periods during which the catalog is incomplete due to short-term aftershock incompleteness (STAI). Although Mizrahi *et al.* (2021a) have proposed a method to account for STAI in the ETAS model, we do not address this issue here.

For Italy, we consider the Italian Seismological Instrumental and Parametric Data-Base catalog (ISIDe, Group, 2007), in the area defined for the first CSEP experiment (Schorlemmer, Christophersen, *et al.*, 2010, vertices given in Table A2). We consider earthquakes of magnitude $M \geq 2.5$ from 16 April 2005 until 1 July 2021. This is the time horizon available to modelers in the upcoming prospective CSEP forecasting experiment in Italy, and the estimated magnitude of completeness provided in the experiment description. The start of the primary catalog is 1 January 2010.

Experiment setting

For southern California, we consider 5 yr of testing, with the start of the first forecasting period at the occurrence of event

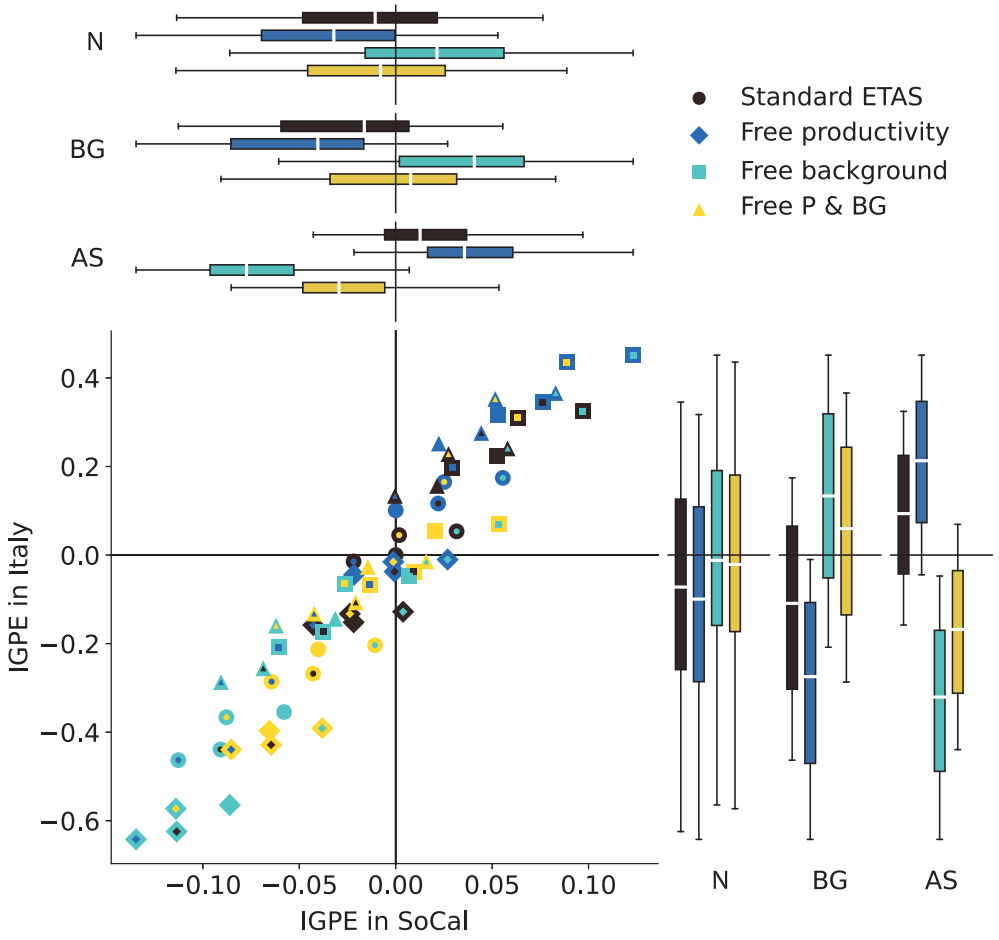

Figure 1. Scatter plot of information gain per event (IGPE) over standard epidemic-type aftershock sequence (ETAS) of the 64 question-driven ensemble (QDE) models in Italy and southern California. The symbol shape, fill color, and edge color describe the composition of the QDE. The shape, fill color, and edge color represent the ingredient model used to answer the background density (BG), number (N), and aftershock density (AS) questions, respectively. Box plots on top (for southern California) and to the right (for Italy) of the scatter plot: for N, BG, and AS questions, the four boxes represent the IGPE of four groups of QDE models. Each group contains the 16 QDE models that use a specific ingredient model (indicated by box color) to answer the indicated question.

e_0, the first event on or after 1 January 2017. In Italy, we consider 3 yr of testing, starting at the occurrence of the first event on or after 1 July 2018. The idea of the pseudoprospective experiments is to only use data that would have been available at the time the forecast is issued to calibrate the models. One could, thus, recalibrate the model at the start of each forecasting period, whenever one more event becomes part of the catalog. To limit the number of computationally expensive parameter inversions for these experiments, we re-estimate the model parameters every 7 days in southern California, and every day in Italy, and use the latest available set of parameters at the start time of each forecasting interval. This does not mean that events between the calibration time and forecasting start are ignored. Their aftershocks are still considered in the calculated aftershock rate. We chose a shorter parameter updating interval for Italy to mimic the conditions of the CSEP experiment, and a longer one for southern California to limit computational cost.

We then calculate $IG_j^{i_1,i_2}$ for all j, and for all pairs of models M_{i_1}, M_{i_2}. If the IGPE over all forecasting periods of one model to another is positive, we consider the model to produce superior forecasts.

As one could argue that generating a large number of models and then selecting the best performing ones somewhat invalidates the pseudoprospective nature of our experiments, we consider the following additional model. At the start of the jth forecasting period, the total information gain of all QDE models during the last n forecasting periods, that is, periods $j-(n+1)$ to $j-1$, is compared. The model with the highest IG is selected to produce the forecast for the jth forecasting period. We call this model QDE-S_n.

This type of model, if capable of producing a powerful forecast, would be well suited to be used in an OEF context.

Results and Discussion

The parameters that were obtained using the flETAS inversion algorithm are described in the Inverted parameters section. Here, we present the results of the Forecasting Experiments.

Experiment results

Figure 1 compares the IGPE over the standard ETAS null model ($M_0 = E^{000}$) of all 64 QDE models in Italy and southern California. The IGPE varies between −0.64 and 0.45 in Italy, and between −0.13 and 0.12 in southern California. The best and worst performing QDE models are E^{221} and E^{112}, respectively, for both regions. The best performing model E^{221} uses the free background model M_2 to answer the number and background density questions, and the free productivity model M_1 to answer the aftershock density question. Vice versa, the worst performing model E^{112} uses M_1 to answer the number

and background density questions, and model M_2 to answer the aftershock density question. Generally, the models that perform well or poorly in Italy are also performing similarly in southern California.

The symbol shape, fill color, and edge color in the scatter plot of Figure 1 represent the ingredient model used to answer the background density (BG), number (N), and aftershock density (AS) questions, respectively. Models that perform well tend to answer the BG question with the free background ingredient model and the AS question with the free productivity model. Conversely, models that address the BG question with the free productivity model, and those that address the AS question with the free background model, tend to perform poorly.

This is highlighted in the box plots of Figure 1. There, for each question, the distribution of IGPE of the 64 QDE models is given per possible answer. Although for the number questions, no clear trend can be inferred, it is evident that the free background model serves well at answering the BG question and the free productivity model serves well at answering the AS question. These trends are qualitatively very similar in southern California and Italy.

These results emphasize the added value generated by the flETAS approach, although most flETAS models individually do not outperform standard ETAS. Apparently, a model that gives full flexibility to the background rate during parameter inversion is more informative than others when addressing the background density question. And a model that is flexible at identifying aftershocks is more informative than others when answering the aftershock density question. These observations are made for both considered regions.

While conceptually it makes sense that a model that can more flexibly capture one particular aspect of seismicity is particularly successful at answering questions about this very aspect of seismicity, this is simultaneously a somewhat counterintuitive result. If flETAS with free background is more successful than other models at identifying background events, one would expect it, due to the self-consistent nature of parameter inversion, to also be more successful at identifying aftershocks, and thus at describing their occurrence times and locations.

A possible interpretation of the observation that E^{221}, E^{220}, and even E^{223} can so clearly outperform E^{222}, is the following. Compared to the null model M_0, model $M_2 = E^{222}$ allows the background seismicity to be free and, therefore, interprets a higher fraction of events in the training catalog to be background earthquakes, which manifests in a much higher background rate. M_2 can, thus, explain the spatial distribution of background events well, as well as the partitioning of seismicity into background events and aftershocks. Possibly, M_2 overestimates the background portion of the training catalog due to "too much freedom." The level of overestimation may be small enough so that M_2 still captures the fraction and locations of background earthquakes better than the other ingredient models do. Overestimation of the background seismicity comes with underestimation of the fraction of aftershocks in the training catalog. Although this underestimation may have a minor biasing effect on the number of background earthquakes and aftershocks, the spatiotemporal distribution of aftershocks can be affected in a more harmful way. Aftershocks that occur in the tails of the spatial or temporal distributions have higher chances to be falsely identified as background events compared to aftershocks that are close to their parent event. This leads to a distorted characterization of the aftershock triggering behavior of model M_2, which can be fixed using the triggering parameters from models M_0 or M_1, as indicated by the good performance of models E^{221} and E^{220}.

Another noteworthy observation is that model M_3, which in principle has all the flexibility necessary to encompass the parameterization of model E^{221}, is clearly outperformed by E^{221}. We interpret this to be a consequence of the fact that the information that is optimized during model calibration and the information used for forecasting are not the same. This does not indicate a flaw in the method presented, but rather illustrates a complexity of the forecasting problem to which the QDE approach offers an apparently useful solution.

Figure 2a shows the cumulative information gain (CIG) over the standard ETAS model over time of the three flETAS ingredient models and the three best performing QDE models. The CIG of model i_1 over model i_2 at time t is given as the sum of IGs of all forecasting periods ending prior to time t:

$$\sum_{j:t_j<t} IG_j^{i_1,i_2}. \tag{26}$$

In southern California, the flETAS ingredient models have a negative information gain following the Ridgecrest events in July 2019, meaning that during this time, the standard ETAS model (M_0) is better performing. The free background model M_2 outperforms M_0 immediately after the onset of the sequence and suffers from information loss later during the sequence. The other two ingredient models do not exhibit the initial information gain. Among the flETAS models, only M_2 can compensate for the information loss during the course of the 5 yr of testing and ends up with a positive overall information gain.

Among the QDE models presented, models E^{221} and E^{220} show an initial information gain after the onset of the Ridgecrest sequence, followed by a period of information loss. In contrast to the ingredient models, the information loss during the sequence is smaller than the gain at the beginning of the sequence, such that these models show positive information gain during the Ridgecrest sequence. The three QDE models in Figure 2a also show a rapidly accumulating information gain throughout the testing period, arriving at an overall IGPE of 0.12, 0.10, and 0.09.

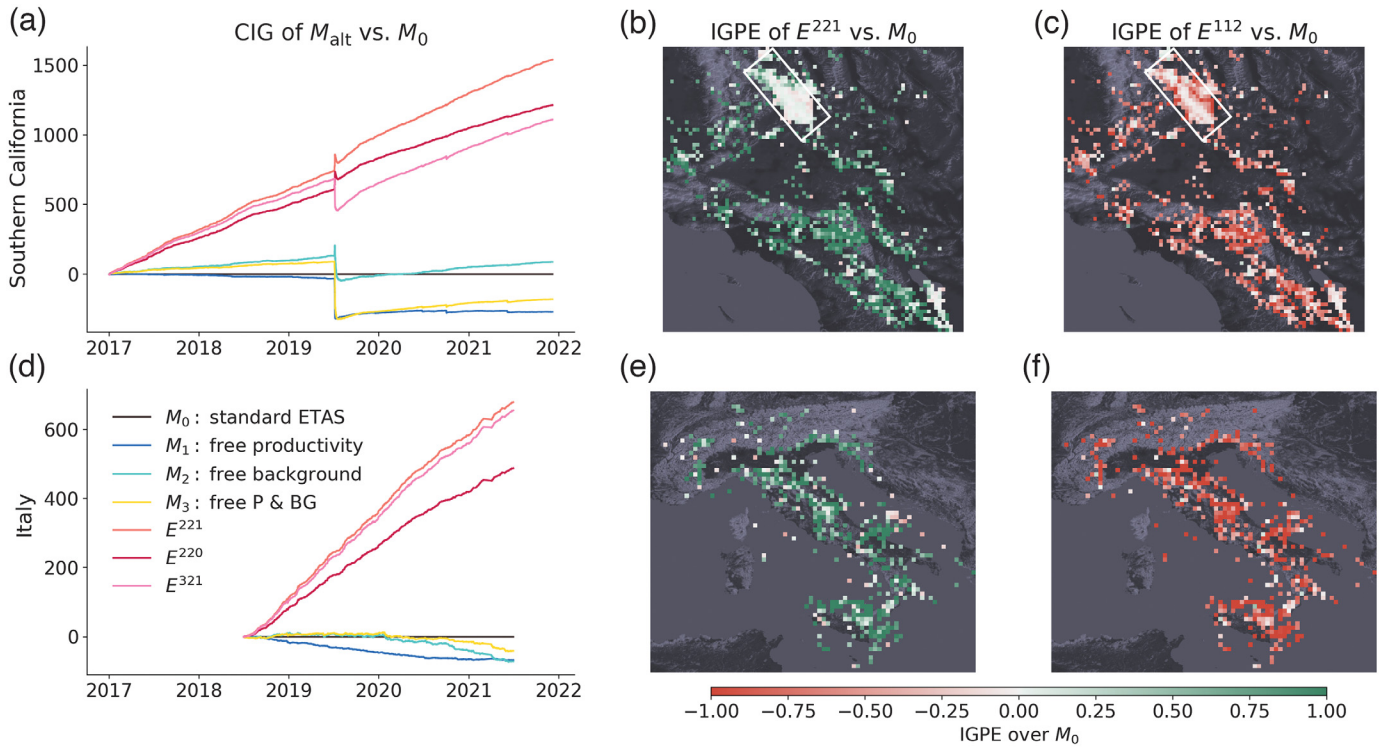

Figure 2. (a,b,c) Results for southern California and (d,e,f) results for Italy. Panels (a) and (d) Cumulative information gain (CIG) over time of the ingredient models and the three QDE models best performing in southern California, compared to the standard ETAS model indicated by the black horizontal line. Panels (b, c) and (e,f): Information gain per earthquake (IGPE) per spatial grid cell of the best performing QDE model (E^{221}, panels b and e) and the worst performing QDE model (E^{112}, panels c and f), compared to standard ETAS ($M_0 = E^{000}$). Grid cell resolution is 0.05° × 0.05° in southern California, and 0.2° × 0.2° in Italy, chosen for best visibility. The white rectangle in panels (b, c) highlights the region of the Ridgecrest sequence in 2019.

From Figure 2b, it is clear that the IGPE is relatively close to zero in the Ridgecrest area, and the positive IG during the sequence must come from a few specific locations. In the rest of southern California, higher IGPE values are achieved, with a median grid-cell-wise IGPE of 0.66 for model E^{221} shown in Figure 2b. Conversely, the median grid-cell-wise IGPE for the worst performing model E^{112} shown in Figure 2c is −0.54. Generally, it performs poorly where E^{221} performs well.

In Italy, all flETAS models have negative total information gain over M_0. Nevertheless, two of the top three QDE models that perform best in southern California are also among the top three in Italy, with overall IGPE values of 0.45 and 0.44 for E^{221} and E^{321}. The second best model of southern California, E^{220}, ranks sixth in Italy with an IGPE of 0.32. Similar to what can be observed in southern California, the regions in Italy in which the best performing model E^{221} performs well coincide with the areas in which model E^{112} shown in Figure 2f performs poorly. The median grid-cell-wise IGPE of the two models are 0.76 and −0.82, respectively. Although these grid-cell-wise IGPE values cannot directly be compared between Italy and southern California due to the different size of the grid cells, the results suggest a qualitatively more similar model performance between the two regions than what is shown by the overall IGPE shown in Figure 1. The lower IGPE in southern California is likely caused by a relatively small IG during the Ridgecrest sequence when a large fraction of events occurred.

Pseudoprospective model selection

Figure 3 illustrates the composition and performance of QDE-S$_n$ models. The number n of past forecasting periods considered when selecting the forecasting model for the next period is in $\{1 = 2^0, 2, 4, 8, 16, 32, 64, 128, 256, 512, 1024 = 2^{10}\}$ for southern California, and $n \in \{1 = 2^0, \ldots, 512 = 2^9\}$ for Italy. We do not consider n = 1024 for Italy, as this would reduce the number of testing periods in which QDE-S$_n$ is defined by more than half compared to the QDE models. The top, middle, and bottom parts of Figure 3a,b show the ingredient model used by QDE-S$_n$ to answer the N, BG, and AS questions over time. Within each part, n increases from top to bottom. As expected, the composition of QDE-S$_n$ is more stable as n increases and is almost always defined via E^{221} for large n, in both regions.

In southern California, a change in composition can be observed after the onset of the Ridgecrest sequence in July 2019. Specifically, the number questions are best answered by standard ETAS, free productivity flETAS, and free productivity and background flETAS, in this order, before moving back to answering with free background flETAS. The

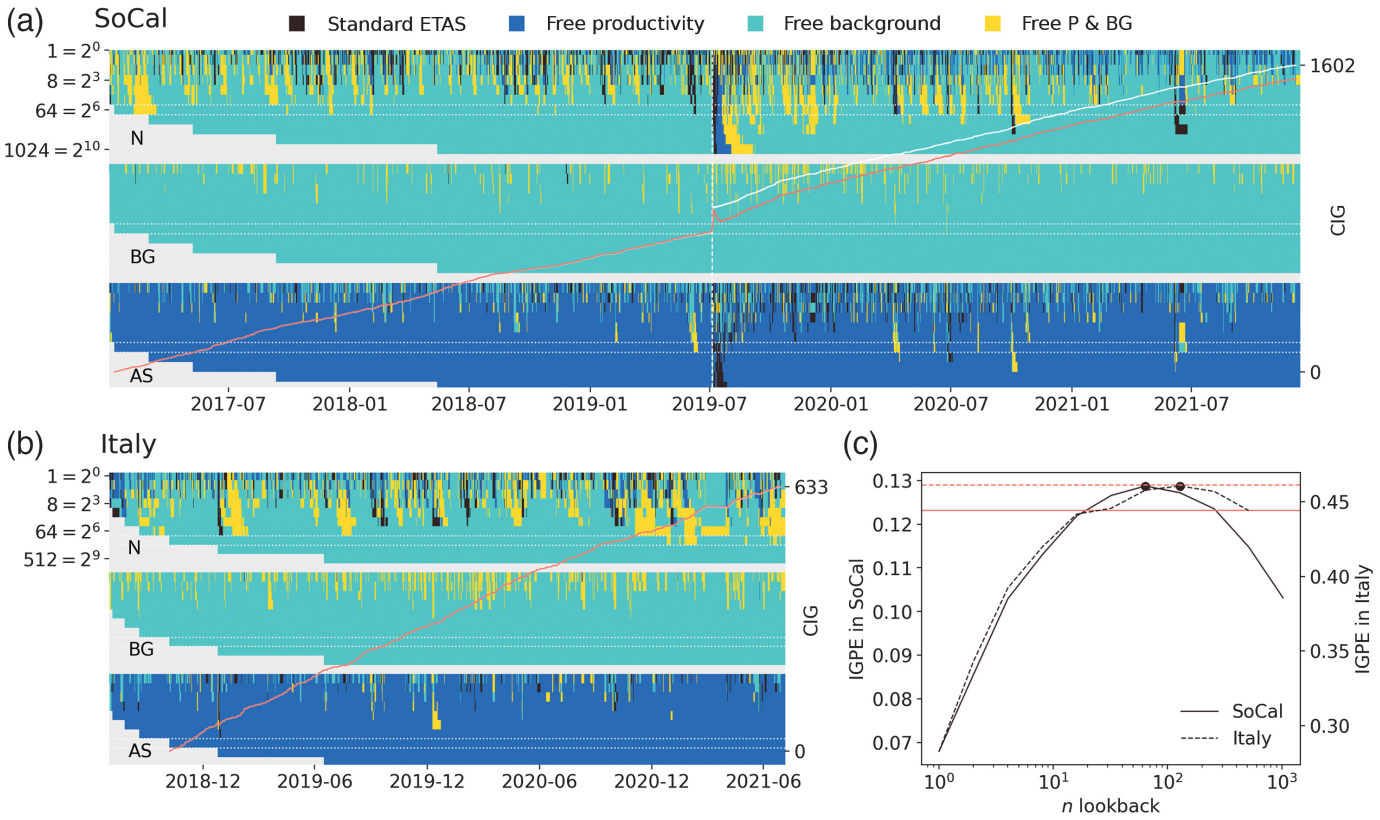

Figure 3. Composition and performance of QDE-S_n models. (a,b) For southern California and Italy: composition of QDE-S_n, in which n takes values of powers of 2. Top, middle, and bottom part represent the ingredient model used to answer the number (N), background density (BG), and aftershock density (AS) questions. Within each part, n increases from top to bottom. The dotted white lines highlight the best performing QDE-S_n. The solid white and orange line show the cumulative information gain (CIG) of the best QDE-S_n and best QDE (E^{221}), respectively, for the period in which both are defined. The white line is barely visible for Italy because it coincides with the orange line. The vertical dashed line indicates the occurrence time of the **M** 6.4 Ridgecrest event on 4 July 2019. (c) IGPE of different QDE-S_n (black lines), for different values of n. The horizontal orange lines indicate IGPE of E^{221} for the period in which the best QDE-S_n is defined. The solid lines represent southern California; dashed lines represent Italy.

aftershock question is intermittently best answered by standard ETAS during the sequence. It is interesting to note here that the performance of E^{221} and QDE-S_{64} are almost identical throughout the 5 yr of testing, with the difference that QDE-S_{64} does not show the information loss after the initial information gain after the onset of the sequence. This results in an overall IGPE of 0.13 and 0.12 for QDE-S_{64} and E^{221}, during the period in which both are defined, as is shown in Figure 3c. Thus, the QDE-S_n model, which was originally designed to avoid a biased selection of the winning model after knowing the experiment outcome, is capable of outperforming the winning QDE model for good choices of n, and clearly outperforms all ingredient flETAS models for any tested choice of n.

In Italy, the best performing QDE-S_n model is QDE-S_{128}. It is almost always using E^{221} to issue a forecast for the next period and, thus, unsurprisingly achieves the same IGPE. As in southern California, all tested choices of n yield a model that clearly outperforms all ingredient flETAS models. The simplest QDE-S_n model, QDE-S_1, which always selects the best QDE model of the previous forecasting period to issue the next forecast, already achieves a very high IGPE of 0.28.

Conclusions

We describe an adapted ETAS EM algorithm that allows a nonparametric inversion of aftershock productivity and/or background rate. Further, we introduce a novel approach of QDE modeling, which combines ingredient models by using them to answer different forecasting subproblems. In pseudoprospective forecasting experiments for southern California and Italy, we compare the forecasting skill of three flETAS models and a total of 60 nontrivial QDEs of flETAS and ETAS models, to that of the standard ETAS null model.

We find that the best models tend to use flETAS with free background to model the number of events and locations of background earthquakes and flETAS with free productivity to model the times and locations of aftershocks. The best model is the same in both regions and achieves an IGPE over standard ETAS of 0.12 in southern California and 0.45 in Italy.

To address the possible concern of a biased selection of the winning model after knowing the experiment outcome, we also test the forecasting skill of a model that pseudoprospectively selects the currently best performing QDE model to issue the forecast for the next testing period. Depending on the criteria to identify the best QDE model, we find that the forecasting skill can be greater than that of the overall best QDE model. This approach thus provides a promising candidate for an operational earthquake forecast.

During the 2019 Ridgecrest sequence in southern California, different ingredient models are best suited to model the number of events during different stages of the sequence. The idea of operationally selecting different QDE models (i.e., selecting different ETAS model parameters) based on their recent performance is in this case related to the idea of Page *et al.* (2016). They considered sequence-specific parameters to be sampled from an underlying distribution and described a Bayesian approach to update this distribution as aftershock data become available.

Our results can also be viewed as a first step toward developing a potentially fruitful branch of earthquake forecasting research. Several key questions remain open and are to be addressed in future studies: *Why do QDE models outperform ingredient models that were inverted in a self-consistent way? What drives the success of different QDE models during different phases of the Ridgecrest sequence? How does QDE performance increase when further ingredient models are considered? And what does all of this teach us about the dynamics of seismicity?*

Data and Resources

The Advanced National Seismic System (ANSS) Comprehensive Earthquake Catalog (ComCat) provided by the U.S. Geological Survey (USGS) was searched using https://earthquake.usgs.gov/data/comcat/ (last accessed January 2022). The Italian Seismological Instrumental and Parametric Data-Base (ISIDe) was used as provided by the organizers of the upcoming CSEP experiment in Italy and can be accessed via http://terremoti.ingv.it/en/search (last accessed March 2022).

Declaration of Competing Interests

The authors acknowledge that there are no conflicts of interest recorded.

Acknowledgments

This study has been funded by the Eidgenössische Technische Hochschule (ETH) research grant for Project Number 2018-FE-213, "Enabling dynamic earthquake risk assessment (DynaRisk)," the European Union's Horizon 2020 research and innovation program under Grant Agreement Number 821115, real-time earthquake risk reduction for a resilient Europe (RISE), the National Science Foundation (Grant Number EAR-2122168), and the Southern California Earthquake Center (based on NSF Cooperative Agreement Number EAR-1600087 and USGS Cooperative Agreement Number G17AC00047). The article benefited from constructive comments by the Associate Editor and two anonymous referees.

References

Akinci, A., M. P. Moschetti, and M. Taroni (2018). Ensemble smoothed seismicity models for the new Italian probabilistic seismic hazard map, *Seismol. Res. Lett.* **89,** no. 4, 1277–1287.

Bach, C., and S. Hainzl (2012). Improving empirical aftershock modeling based on additional source information, *J. Geophys. Res.* **117,** no. B4, doi: 10.1029/2011JB008901.

Bayliss, K., M. Naylor, J. Illian, and I. G. Main (2020). Data-driven optimization of seismicity models using diverse data sets: Generation, evaluation, and ranking using Inlabru, *J. Geophys. Res.* **125,** no. 11, e2020JB020226, doi: 10.1029/2020JB020226.

Bayona, J., W. Savran, A. Strader, S. Hainzl, F. Cotton, and D. Schorlemmer (2021). Two global ensemble seismicity models obtained from the combination of interseismic strain measurements and earthquake-catalogue information, *Geophys. J. Int.* **224,** no. 3, 1945–1955.

Bird, P., D. D. Jackson, Y. Y. Kagan, C. Kreemer, and R. Stein (2015). Gear1: A global earthquake activity rate model constructed from geodetic strain rates and smoothed seismicity, *Bull. Seismol. Soc. Am.* **105,** no. 5, 2538–2554.

Box, G. E. (1979). Robustness in the strategy of scientific model building, in, Elsevier, 201–236, doi: 10.1016/B978-0-12-438150-6.50018-2.

Cattania, C., M. J. Werner, W. Marzocchi, S. Hainzl, D. Rhoades, M. Gerstenberger, M. Liukis, W. Savran, A. Christophersen, A. Helmstetter, *et al.* (2018). The forecasting skill of physics-based seismicity models during the 2010–2012 Canterbury, New Zealand, earthquake sequence, *Seismol. Res. Lett.* **89,** no. 4, 1238–1250.

Cocco, M., S. Hainzl, F. Catalli, B. Enescu, A. Lombardi, and J. Woessner (2010). Sensitivity study of forecasted aftershock seismicity based on coulomb stress calculation and rate-and state-dependent frictional response, *J. Geophys. Res.* **115,** no. B5, doi: 10.1029/2009JB006838.

Daley, D. J., and D. Vere-Jones (2003). *An Introduction to the Theory of Point Processes: Volume I: Elementary Theory and Methods*, Springer, Heidelberg, Germany.

Dieterich, J. (1994). A constitutive law for rate of earthquake production and its application to earthquake clustering, *J. Geophys. Res.* **99,** no. B2, 2601–2618.

Enescu, B., S. Hainzl, and Y. Ben-Zion (2009). Correlations of seismicity patterns in southern California with surface heat flow data, *Bull. Seismol. Soc. Am.* **99,** no. 6, 3114–3123.

Eyring, V., S. Bony, G. A. Meehl, C. A. Senior, B. Stevens, R. J. Stouffer, and K. E. Taylor (2016). Overview of the coupled model intercomparison project phase 6 (cmip6) experimental design and organization, *Geosci. Model Dev.* **9,** no. 5, 1937–1958.

Field, E. H., T. H. Jordan, M. T. Page, K. R. Milner, B. E. Shaw, T. E. Dawson, G. P. Biasi, T. Parsons, J. L. Hardebeck, A. J. Michael, *et al.* (2017). A synoptic view of the third uniform California earthquake rupture forecast (UCERF3), *Seismol. Res. Lett.* **88,** no. 5, 1259–1267.

Gerstenberger, M., G. McVerry, D. Rhoades, and M. Stirling (2014). Seismic hazard modeling for the recovery of Christchurch, *Earthq. Spectra* **30,** no. 1, 17–29.

Gerstenberger, M. C., S. Wiemer, L. M. Jones, and P. A. Reasenberg (2005). Real-time forecasts of tomorrow's earthquakes in California, *Nature* **435,** no. 7040, 328–331.

Grimm, C., S. Hainzl, M. Käser, and H. Küchenhoff (2022). Solving three major biases of the etas model to improve forecasts of the 2019 Ridgecrest sequence, *Stochastic Environ. Res. Risk Assess.***36,** 2133–2152.

Group, I. W. (2007). Italian seismological instrumental and parametric database (ISIDe), available at https://www.earth-prints.org/handle/2122/5063 (last accessed December 2022).

Hainzl, S. (2022). ETAS-approach accounting for short-term incompleteness of earthquake catalogs, *Bull. Seismol. Soc. Am.* **112,** no. 1, 494–507.

Hardebeck, J. L. (2021). Spatial clustering of aftershocks impacts the performance of physics-based earthquake forecasting models, *J. Geophys. Res.* **126,** no. 2, e2020JB020824, doi: 10.1029/2020JB020824.

Kamer, Y., S. Nandan, G. Ouillon, S. Hiemer, and D. Sornette (2021). Democratizing earthquake predictability research: Introducing the Richterx platform, *Eur. Phys. J. Spec. Top.* **230,** no. 1, 451–471.

Kovchegov, Y., I. Zaliapin, and Y. Ben-Zion (2022). Invariant Galton-Watson branching process for earthquake occurrence, *Geophys. J. Int.* **231,** 567–583, doi: 10.1093/gji/ggac204.

Leutbecher, M., and T. N. Palmer (2008). Ensemble forecasting, *J. Comput. Phys.* **227,** no. 7, 3515–3539.

Llenos, A. L., and A. J. Michael (2019). Ensembles of etas models provide optimal operational earthquake forecasting during swarms: Insights from the 2015 san Ramon, California swarm ensembles of etas models provide optimal operational earthquake forecasting during swarms, *Bull. Seismol. Soc. Am.* **109,** no. 6, 2145–2158.

Mancini, S., M. Segou, M. Werner, and C. Cattania (2019). Improving physics-based aftershock forecasts during the 2016–2017 central Italy earthquake cascade, *J. Geophys. Res.* **124,** no. 8, 8626–8643.

Mancini, S., M. Segou, M. J. Werner, 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.* **110,** no. 4, 1736–1751.

Marzocchi, W., A. M. Lombardi, and E. Casarotti (2014). The establishment of an operational earthquake forecasting system in Italy, *Seismol. Res. Lett.* **85,** no. 5, 961–969.

Marzocchi, W., J. D. Zechar, and T. H. Jordan (2012). Bayesian forecast evaluation and ensemble earthquake forecasting, *Bull. Seismol. Soc. Am.* **102,** no. 6, 2574–2584.

Mizrahi, L., S. Nandan, and S. Wiemer (2021a). Embracing data incompleteness for better earthquake forecasting, *J. Geophys. Res.* **126,** no. 12, e2021JB022379, doi: 10.1029/2021JB022379.

Mizrahi, L., S. Nandan, and S. Wiemer (2021b). The effect of declustering on the size distribution of mainshocks, *Seismol. Res. Lett.* doi: 10.1785/0220200231.

Nandan, S., Y. Kamer, G. Ouillon, S. Hiemer, and D. Sornette (2021). Global models for short-term earthquake forecasting and predictive skill assessment, *Eur. Phys. J. Spec. Top.* **230,** no. 1, 425–449.

Nandan, S., G. Ouillon, and D. Sornette (2019). Magnitude of earthquakes controls the size distribution of their triggered events, *J. Geophys. Res.* **124,** no. 3, 2762–2780.

Nandan, S., G. Ouillon, D. Sornette, and S. Wiemer (2019a). Forecasting the full distribution of earthquake numbers is fair, robust, and better, *Seismol. Res. Lett.* **90,** no. 4, 1650–1659.

Nandan, S., G. Ouillon, D. Sornette, and S. Wiemer (2019b). Forecasting the rates of future aftershocks of all generations is essential to develop better earthquake forecast models, *J. Geophys. Res.* **124,** no. 8, 8404–8425.

Nandan, S., G. Ouillon, S. Wiemer, and D. Sornette (2017). Objective estimation of spatially variable parameters of epidemic type aftershock sequence model: Application to California, *J. Geophys. Res.* **122,** no. 7, 5118–5143.

Nandan, S., S. K. Ram, G. Ouillon, and D. Sornette (2021). Is seismicity operating at a critical point? *Phys. Rev. Lett.* **126,** no. 12, 128,501.

Ogata, Y. (1988). Statistical models for earthquake occurrences and residual analysis for point processes, *J. Am. Stat. Assoc.* **83,** no. 401, 9–27.

Ogata, Y., K. Katsura, G. Falcone, K. Nanjo, and J. Zhuang (2013). Comprehensive and topical evaluations of earthquake forecasts in terms of number, time, space, and magnitude, *Bull. Seismol. Soc. Am.* **103,** no. 3, 1692–1708, doi: 10.1785/0120120063.

Page, M. T., and N. J. van der Elst (2022). Aftershocks preferentially occur in previously active areas, *Seismol. Rec.* **2,** no. 2, 100–106.

Page, M. T., N. van Der Elst, J. Hardebeck, K. Felzer, and A. J. Michael (2016). Three ingredients for improved global aftershock forecasts: Tectonic region, time-dependent catalog incompleteness, and intersequence variability, *Bull. Seismol. Soc. Am.* **106,** no. 5, 2290–2301.

Parsons, T., Y. Ogata, J. Zhuang, and E. L. Geist (2012). Evaluation of static stress change forecasting with prospective and blind tests, *Geophys. J. Int.* **188,** no. 3, 1425–1440.

Rhoades, D. A., and M. C. Gerstenberger (2009). Mixture models for improved short-term earthquake forecasting, *Bull. Seismol. Soc. Am.* **99,** no. 2A, 636–646.

Rhoades, D., M. Liukis, A. Christophersen, and M. Gerstenberger (2016). Retrospective tests of hybrid operational earthquake forecasting models for Canterbury, *Geophys. J. Int.* **204,** no. 1, 440–456.

Savran, W. H., M. J. Werner, W. Marzocchi, D. A. Rhoades, D. D. Jackson, K. Milner, E. Field, and A. Michael (2020). Pseudoprospective evaluation of UCERF3-etas forecasts during the 2019 Ridgecrest sequence, *Bull. Seismol. Soc. Am.* **110,** no. 4, 1799–1817.

Savran, W. H., M. J. Werner, D. Schorlemmer, and P. J. Maechling (2022). pycsep: A python toolkit for earthquake forecast developers, *Seismol. Soc. Am.* **93,** no. 5, 2858–2870.

Schoenberg, F. P. (2013). Facilitated estimation of etas, *Bull. Seismol. Soc. Am.* **103,** no. 1, 601–605.

Schorlemmer, D., A. Christophersen, A. Rovida, F. Mele, M. Stucchi, and W. Marzocchi (2010). Setting up an earthquake forecast experiment in Italy, *Ann. Geophys.* doi: 10.4401/ag-4844.

Schorlemmer, D., J. D. Zechar, M. J. Werner, E. H. Field, D. D. Jackson, T. H. Jordan, , and RELM Working Group (2010). First results of the regional earthquake likelihood models experiment, in *Seismogenesis and Earthquake Forecasting: The Frank*

Evison Volume II, Springer, Basel, Switzerland, 5–22, doi: 10.1007/978-3-0346-0500-7_2.

Seif, S., A. Mignan, J. D. Zechar, M. J. Werner, and S. Wiemer (2017). Estimating etas: The effects of truncation, missing data, and model assumptions, *J. Geophys. Res.* **122,** no. 1, 449–469.

Shebalin, P. N., C. Narteau, J. D. Zechar, and M. Holschneider (2014). Combining earthquake forecasts using differential probability gains, *Earth Planets Space* **66,** no. 1, 1–14.

Steacy, S., M. Gerstenberger, C. Williams, D. Rhoades, and A. Christophersen (2014). A new hybrid coulomb/statistical model for forecasting aftershock rates, *Geophys. J. Int.* **196,** no. 2, 918–923.

Strader, A., M. Schneider, and D. Schorlemmer (2017). Prospective and retrospective evaluation of five-year earthquake forecast models for California, *Geophys. J. Int.* **211,** no. 1, 239–251.

Taroni, M., W. Marzocchi, D. Schorlemmer, M. J. Werner, S. Wiemer, J. D. Zechar, L. Heiniger, and F. Euchner (2018). Prospective csep evaluation of 1-day, 3-month, and 5-yr earthquake forecasts for Italy, *Seismol. Res. Lett.* **89,** no. 4, 1251–1261.

Taroni, M., J. Zechar, and W. Marzocchi (2014). Assessing annual global m 6+ seismicity forecasts, *Geophys. J. Int.* **196,** no. 1, 422–431.

Tracton, M. S., and E. Kalnay (1993). Operational ensemble prediction at the national meteorological center: Practical aspects, *Weather Forecast.* **8,** no. 3, 379–398.

van der Elst, N. J., J. L. Hardebeck, A. J. Michael, S. K. McBride, and E. Vanacore (2022). Prospective and retrospective evaluation of the US geological survey public aftershock forecast for the 2019–2021 southwest Puerto Rico earthquake and aftershocks, *Seismol. Soc. Am.* **93,** no. 2A, 620–640.

Veen, A., and F. P. Schoenberg (2008). Estimation of space–time branching process models in seismology using an em–type algorithm, *J. Am. Stat. Assoc.* **103,** no. 482, 614–624.

Woessner, J., S. Hainzl, W. Marzocchi, M. Werner, A. Lombardi, F. Catalli, B. Enescu, M. Cocco, M. Gerstenberger, and S. Wiemer (2011). A retrospective comparative forecast test on the 1992 Landers sequence, *J. Geophys. Res.* **116,** no. B5, doi: 10.1029/2010JB007846.

Zechar, J. D., D. Schorlemmer, M. Liukis, J. Yu, F. Euchner, P. J. Maechling, and T. H. Jordan (2010). The collaboratory for the study of earthquake predictability perspective on computational earthquake science, *Concurrency Comput.* **22,** no. 12, 1836–1847.

Zhuang, J. (2012). Long-term earthquake forecasts based on the epidemic-type aftershock sequence (ETAS) model for short-term clustering, *Res. Geophys.* **2,** no. 1, e8–e8.

Appendix

Polygons

The polygons used in this study are defined via the lists of vertices defined in Tables A1 and A2.

Inverted parameters

Figure A1 shows the inverted parameters for the four ingredient models, with an increasing time horizon used for the calibration, for southern California and Italy. For the standard epidemic-type aftershock sequence (ETAS) model and flexible epidemic-type aftershock sequence (flETAS) in which only the background rate is free, the parameters a and k_0 are inverted directly during expectation maximization (EM), whereas for the flETAS models with free productivity, they are inferred afterward based on the κ_j values that result from the EM inversion.

Most parameters show remarkable changes in time in southern California, and generally, the parameters differ between Italy and southern California. The differences between parameters obtained for different ingredient models show similar trends in both regions.

For instance, the background rate μ is highest for the model that only allows the background rate to be free, followed by the model in which background and productivity are free, and is lowest when only the productivity is free. This is expected, because allowing the background to be free will allow the model to classify more events to be background events, whereas allowing the productivity to be free will allow it to classify more events to be aftershocks.

The exponent of the productivity law, a, is larger in the flETAS models that allow the background to be free, indicating a stronger magnitude dependency of the number of aftershocks an earthquake is expected to generate. Those models also have larger γ and much larger ρ values, which translates to a stronger magnitude dependency of the spatial region in which aftershocks occur, and a stronger spatial decay of the aftershock rate.

Interestingly, the flETAS model in which only productivity is free shows smaller k_0 values than standard ETAS in both regions, accompanied by values of a that are similar to standard ETAS. Both these effects would suggest lower overall productivity. However, the value of τ is larger in this model, indicating a slower long-term tapering off of aftershock rate in time, and ω is smaller in southern California (similar in Italy), further indicating a slower (similar) temporal decay of aftershock rate. Together with the observation that μ is smaller for this model, these results suggest that allowing productivity to be free leads to an overall slower decay of aftershock rate, and, thus, a large fraction of aftershocks is expected to occur later in an ongoing sequence.

The branching ratio η, which captures the average expected number of aftershocks of any event, is highest for the standard ETAS model, followed by flETAS with free productivity, flETAS with free background, and flETAS with free productivity and background with the lowest branching ratio. Thus, the degree of flexibility of a model is qualitatively opposite to the degree of criticality of the system that is inferred with that model.

TABLE A1
Southern California Polygon Boundary Vertices

Latitude	Longitude
32.7219	−116.3004
33.7424	−117.6512
33.7958	−117.966
33.9322	−118.0775
34.0984	−118.2611
34.1755	−118.9365
34.6027	−118.8775
34.8281	−119.343
36.525	−119.1988
36.4835	−115.6381
34.128	−115.5463
32.7219	−115.2578
32.6922	−115.448
32.7753	−115.7234
32.8109	−115.8545

TABLE A2
Italy Polygon Boundary Vertices

Latitude	Longitude
45.1	4.9
44.5	5.1
43.3	5.9
42.8	6.5
41.6	9.1
38.0	10.5
36.7	11.5
35.8	13.4
35.3	15.1
35.7	16.1
38.8	19.4
40.1	20.1
41.3	19.5
42.9	17.2
44.0	15.6
45.6	15.6
46.5	15.4
47.5	14.7
47.9	13.7
48.1	13.2
48.4	12.2
48.2	10.7
47.9	9.4
47.8	8.4
46.8	5.8
45.8	5.1
45.1	4.9

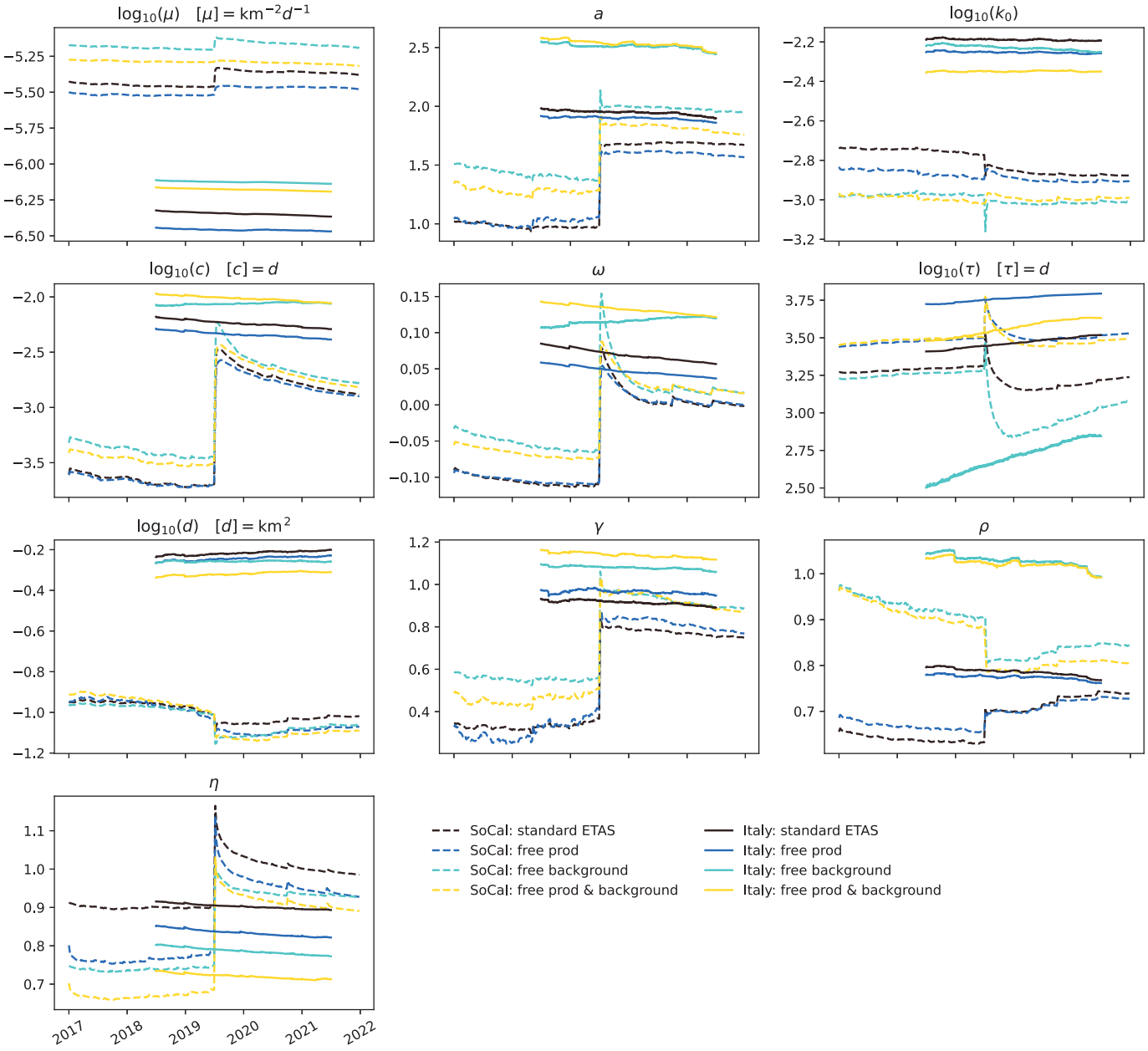

Figure A1. Evolution of inverted parameters with increasing length of the training catalog, for the four ingredient models. The branching ratio η is not individually inverted but is calculated from the other parameters. The dashed lines reflect southern California parameters; solid lines reflect Italian parameters.

Manuscript received 14 July 2022
Published online 18 January 2023