

Reviews of Geophysics^{*}




















REVIEW ARTICLE

10.1029/2023RG000823

Developing, Testing, and Communicating Earthquake Forecasts: Current Practices and Future Directions

Key Points:

- We capture the state of earthquake forecasting systems in Italy, New Zealand, and the United States, and future plans in these and other countries
- Experts encourage benchmark comparison, prospective testing, reproducibility and transparency, but avoid endorsing specific models or tests
- Experts stress the need to co-design forecast communication products with end-users to ensure their societal relevance and usefulness

Leila Mizrahi¹ , Irina Dallo¹, Nicholas J. van der Elst² , Annemarie Christophersen³ , Ilaria Spassiani⁴ , Maximilian J. Werner⁵ , Pablo Iturrieta⁶ , José A. Bayona⁵ , Iunio Iervolino^{7,8}, Max Schneider⁹, Morgan T. Page² , Jiancang Zhuang¹⁰ , Marcus Herrmann⁷ , Andrew J. Michael⁹ , Giuseppe Falcone⁴ , Warner Marzocchi⁷ , David Rhoades³, Matt Gerstenberger³ , Laura Gulia¹¹, Danijel Schorlemmer⁶, Julia Becker¹², Marta Han¹ , Lorena Kuratle¹, Michèle Marti¹ , and Stefan Wiemer¹ 

¹Swiss Seismological Service at ETH Zurich, Zürich, Switzerland, ²US Geological Survey, Pasadena, CA, USA, ³GNS Science | Te Pū Ao, Lower Hutt, New Zealand, ⁴Istituto Nazionale di Geofisica e Vulcanologia (INGV), Rome, Italy, ⁵School of Earth Sciences, University of Bristol, Bristol, UK, ⁶GFZ German Research Centre for Geosciences, Potsdam, Germany, ⁷Università degli Studi di Napoli Federico II, Naples, Italy, ⁸IUSS – Scuola Universitaria Superiore di Pavia, Pavia, Italy, ⁹US Geological Survey, Moffett Field, CA, USA, ¹⁰The Institute of Statistical Mathematics, Tokyo, Japan, ¹¹Dipartimento di Fisica e Astronomia, Università di Bologna, Bologna, Italy, ¹²Joint Centre for Disaster Research, Massey University, Wellington, New Zealand

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

L. Mizrahi,
leila.mizrahi@sed.ethz.ch

Citation:

Mizrahi, L., Dallo, I., van der Elst, N. J., Christophersen, A., Spassiani, I., Werner, M. J., et al. (2024). Developing, testing, and communicating earthquake forecasts: Current practices and future directions. *Reviews of Geophysics*, 62, e2023RG000823. <https://doi.org/10.1029/2023RG000823>

Received 3 NOV 2023
Accepted 18 JUL 2024

Author Contributions:

Conceptualization: Leila Mizrahi, Irina Dallo, Stefan Wiemer
Funding acquisition: Stefan Wiemer
Methodology: Leila Mizrahi, Irina Dallo, Lorena Kuratle
Project administration: Leila Mizrahi, Stefan Wiemer
Supervision: Stefan Wiemer
Writing – original draft: Leila Mizrahi, Irina Dallo, Nicholas J. van der Elst, Annemarie Christophersen,

Abstract While deterministically predicting the time and location of earthquakes remains impossible, earthquake forecasting models can provide estimates of the probabilities of earthquakes occurring within some region over time. To enable informed decision-making of civil protection, governmental agencies, or the public, Operational Earthquake Forecasting (OEF) systems aim to provide authoritative earthquake forecasts based on current earthquake activity in near-real time. Establishing OEF systems involves several nontrivial choices. This review captures the current state of OEF worldwide and analyzes expert recommendations on the development, testing, and communication of earthquake forecasts. An introductory summary of OEF-related research is followed by a description of OEF systems in Italy, New Zealand, and the United States. Combined, these two parts provide an informative and transparent snapshot of today's OEF landscape. In Section 4, we analyze the results of an expert elicitation that was conducted to seek guidance for the establishment of OEF systems. The elicitation identifies consensus and dissent on OEF issues among a non-representative group of 20 international earthquake forecasting experts. While the experts agree that communication products should be developed in collaboration with the forecast user groups, they disagree on whether forecasting models and testing methods should be user-dependent. No recommendations of strict model requirements could be elicited, but benchmark comparisons, prospective testing, reproducibility, and transparency are encouraged. Section 5 gives an outlook on the future of OEF. Besides covering recent research on earthquake forecasting model development and testing, upcoming OEF initiatives are described in the context of the expert elicitation findings.

Plain Language Summary The exact location, time, and magnitude of future earthquakes cannot be predicted. However, based on past earthquake sequences, it is possible to assess probabilities for future earthquakes. This is called earthquake forecasting. Operational Earthquake Forecasting (OEF) systems are designed to provide near-real-time authoritative earthquake forecasts, based on current earthquake activity, to aid the decision-making of various societal stakeholders. Setting up these systems is complex, involving decisions about which model to use, how to best test the model, and how to turn earthquake probability estimates into practical information. This review captures the current state of OEF worldwide and analyzes expert recommendations on the development, testing, and communication of earthquake forecasts. Section 2 provides an overview of OEF-related research and the background knowledge required to understand the other parts. Section 3 describes existing OEF systems of Italy, New Zealand, and the United States in detail. Section 4 discusses an elicitation of expert views on modeling, testing, and communicating earthquake forecasts (Mizrahi, Dallo, & Kuratle, 2023, <https://doi.org/10.3929/ethz-b-000637239>). Data from the elicitation allow to identify consensus and dissent on OEF issues and provide guidance for future earthquake forecasting efforts. Finally, Section 5 gives an outlook on future OEF-related research and planned OEF efforts at various institutions.

© 2024 The Author(s). This article has been contributed to by U.S. Government employees and their work is in the public domain in the USA.

This is an open access article under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

Ilaria Spassiani, Pablo Iturrieta,
José A. Bayona, Iunio Iervolino,
Jiangang Zhuang, Marcus Herrmann,
David Rhoades, Laura Gulia
Writing – review & editing:
Leila Mizrahi, Irina Dallo, Nicholas J. van
der Elst, Annemarie Christophersen,
Ilaria Spassiani, Pablo Iturrieta,
José A. Bayona, Iunio Iervolino,
Max Schneider, Morgan T. Page,
Jiangang Zhuang, Marcus Herrmann,
Andrew J. Michael, Giuseppe Falcone,
Warner Marzocchi, David Rhoades,
Matt Gerstenberger, Danijel Schorlemmer,
Julia Becker, Marta Han, Michèle Marti,
Stefan Wiemer

1. Introduction

What is commonly referred to as *earthquake prediction*, namely the deterministic description of the location, time, and magnitude of a future earthquake “within narrow limits [...], so that a planned evacuation can take place” (Main, 1999), is currently not possible. The term *earthquake forecasting* refers to a probabilistic assessment of earthquake occurrence, over a range of magnitudes, and within some specified time frame and region (e.g., Kagan & Jackson, 2000). On the spectrum between phenomena of complete randomness, such as a fair coin toss, and entirely deterministic processes, the occurrence of earthquakes falls somewhere in-between, its most predictable components being the clustering behavior of groups of earthquakes in time and space (Kagan & Jackson, 1991; Omori, 1894; Zhuang et al., 2021), and the frequency distribution of earthquake magnitudes within a group (Gutenberg & Richter, 1944).

The understanding of occurrence patterns of earthquakes and their implications for the short-term increase in seismic hazard and risk can provide crucial information during an ongoing seismic crisis and enable informed decision-making of civil protection, governmental agencies, the public, or other user groups. As first defined by Jordan et al. (2011), “Operational earthquake forecasting comprises procedures for gathering and disseminating authoritative information about the time dependence of seismic hazards to help communities prepare for potentially destructive earthquakes.” The label *operational* has been borrowed from meteorology, where it describes the existence of automated and timely applications of forecasting models, which increases the actionability of the information provided. To be considered *authoritative*, information must originate from an official, trusted source that is legally responsible for providing this information. Thus, the authoritativeness of an information source is a quality that cannot be self-assigned but must be granted by a higher authority, such as the government. For the remainder of this review, we will use a broader definition of *operational earthquake forecasting (OEF)* as ***authoritative near-real-time applications of earthquake forecasting***. The concept of near-real-time here refers to the delivery of information as soon as possible when an interest in updated information arises, acknowledging the possibility of some noticeable, unavoidable processing time. This is in contrast to the concept of true real-time, which would refer to the delivery of information as soon as needed.

Few governmental agencies worldwide have systems in place that fall under this OEF definition, and one could argue that only Italy has fully deployed an OEF system under the definition of Jordan et al. (2011), even though their OEF information has not yet been open to the public. The Italian OEF system was established in 2013 and has produced weekly forecasts since then (Marzocchi et al., 2014), including the expected seismic losses based on such forecasts (Operational Earthquake Loss Forecasting, OELF; Iervolino et al., 2015). The forecasts can be accessed only by authorized personnel and, only recently, the OEF information has been released to regional governments to trigger a discussion on how to use and communicate this information in different contexts. In New Zealand, earthquake forecasts that satisfy our broader OEF definition have been provided to the public since the 2010 *M*7.1 Darfield earthquake, and the content of these forecasts has evolved over time (Christophersen, Rhoades, & Colella, 2017; Christophersen, Rhoades, Gerstenberger, et al., 2017). The forecasts disseminated by the U.S. Geological Survey (USGS) since 1989 for California (Reasenber & Jones, 1989) and since 2018 for the entire country (Michael et al., 2020), although fulfilling our OEF definition, are referred to as *aftershock forecasts*, to emphasize that forecasts are only issued after the occurrence of large events. Here, we consider (*operational*) *aftershock forecasting (OAF)* to be a differently named special case of (*operational*) *earthquake forecasting*.

There is currently no known difference in the physical mechanisms of what are often termed *foreshocks*, *aftershocks*, or *mainshocks*, that is, the earthquakes in a cluster, and their distinction can be considered a convention (Felzer et al., 2004). These labels can only be applied to earthquakes of a sequence in hindsight: the earthquake with the largest magnitude is generally referred to as the *mainshock*, while earthquakes that happened earlier than the mainshock are retrospectively defined to be *foreshocks*, and earthquakes that happened after the largest event are called *aftershocks*. However, because this distinction seems not to exist in nature, the classification cannot be done before the earthquake sequence has finished. During an ongoing earthquake sequence (the definition of which is itself ambiguous), it cannot be ruled out that a larger earthquake is yet to follow, rendering the mainshock-aftershock terminology often misleading. Between 1980 and 2019, more than 10 percent of worldwide *M* ≥ 6.0 earthquakes were followed by a larger one within a space-time window of 60 days and 100 km (Taroni, 2023). Aftershock forecasts published by the USGS clearly emphasize this possibility and it is a question of definition whether a larger earthquake following a first mainshock should be considered a second mainshock, a

large aftershock, or whether the previous mainshock is now a foreshock to the *only true mainshock*. In Japan, aftershock forecasts ignoring the possibility of a strong earthquake being followed by a stronger one were produced following 15 earthquakes between 1998 and 2016. After the 2016 Kumamoto sequence, in which a $M6.5$ event was followed by a $M7.3$ event 28 hr later, the Japan Meteorological Agency (JMA) stopped providing aftershock forecasts during the first week after a mainshock because of the “invalidity of the [previously applied] procedure” (Omi et al., 2019).

The notion of *foreshocks* can likewise be misleading. Many mainshocks occur without detectably anomalous prior seismic activity, and vice-versa, a temporal increase of seismic activity is not always followed by a large earthquake (Mogi, 1963). Series of earthquakes located close to each other in time and space that are not dominated by a clear mainshock are commonly referred to as *earthquake swarms* and are often associated with aseismic slip or slow-slip events, fluid intrusions or magmatic processes (Hainzl, 2004; Hill et al., 1975; Hirose et al., 2014; Kisslinger, 1975; Passarelli et al., 2021; Ross & Cochran, 2021; K. D. Smith et al., 2004; Vidale & Shearer, 2006). Although the differentiation of swarms from other, mainshock-dominated clusters is accompanied by proposed underlying physical mechanisms, the classification of earthquake clusters into mainshock-dominated clusters or swarms is, again, not free of ambiguity (Zaliapin & Ben-Zion, 2016).

Section 2 of this review provides an overview of the current state of research relevant for OEF. Besides an introduction to different model types currently applied in practice, it covers the history and necessary background information on forecast model testing as well as the state of the art of earthquake forecast communication. Thereafter, the current state of OEF systems in Italy, New Zealand, and the United States is described in Section 3 of this review.

Currently, most countries, even in regions considered of high seismic risk, do not have OEF systems in place and only estimate the time-invariant, long-term earthquake probability as part of their seismic hazard and risk assessment, which serves different user groups involved in long-term decision making (see Gerstenberger et al., 2020). In addition to long-term seismic hazard programs, so-called *earthquake early warning (EEW)* systems have been and are being established in various countries (e.g., Given et al., 2014; Nof & Kurzon, 2021; Orihuela et al., 2023; Suárez, 2022; see Cremen and Galasso (2020) for a review on EEW). These systems can disseminate warnings about earthquakes after they have already occurred, but, depending on the location and strength of the earthquake and the efficiency of the warning system, before the damaging seismic waves reach populated cities or critical infrastructure.

The scarcity of OEF systems could be attributed to factors such as insufficient data, knowledge gaps, resource constraints, or a lack of guidance required for their establishment. Amidst a wealth of forecasting models and model tests proposed and published every year, it may be difficult to choose the most appropriate ones. On top of the more technical choices involved in the development and testing of forecasting models comes the additional challenge of how to communicate earthquake forecasts to different user groups. In New Zealand, in particular, much effort has gone into testing and evaluating forecast communication products with end-users, providing initial guidance for other regions as well (e.g., Becker et al., 2020). In Europe and the United States, recent studies have similarly aimed to assess user needs, design preferences, and the impact of the communication format on people's risk perception (Dryhurst et al., 2022; Schneider, Wein, et al., 2023). It is crucial that earthquake forecasts serve a variety of stakeholders, are adapted over time as information needs evolve throughout the course of earthquake sequences, and are provided in different formats to consider stakeholders' knowledge, skills and requirements. In this process, two main challenges arise: (a) effectively communicating what can be considered low-probability, high-impact events and supporting stakeholders in interpreting probabilistic information and translating it into risk and risk mitigation actions, if needed; and (b) obtaining a thorough understanding of how earthquake forecasts influence people's emotions and, consequently, their behavior and well-being.

Section 4 of this review describes an elicitation of expert views on the three pillars *Model Development*, *Model Testing*, and *Forecast Communication* based on a Delphi study conducted among a group of experts in OEF (Mizrahi, Dallo, & Kuratle, 2023). The study involved two surveys and an online workshop, which facilitated meaningful discussions on the three pillars. The study aimed to establish areas of consensus among the experts and identify open questions within the research landscape to highlight future research directions. The results of the Delphi study provide a transparent and structured collection of expert opinions, offering valuable guidance for the future development of OEF systems.

Section 5 of this review outlines emerging forecasting techniques potentially relevant for OEF in the future, an outlook on forecast testing efforts, as well as forthcoming developments in earthquake forecasting at various institutions represented by the authors. These developments are presented in the context of the guiding framework derived from the Delphi study, demonstrating the practical implications of the study's findings and emphasizing how they contribute to shaping the future of OEF systems.

2. The Current State of Research Relevant for Operational Earthquake Forecasting

This Section 2 of the review introduces concepts, terms, and models that will be discussed in the subsequent sections: Examples of OEF Systems Worldwide, Elicitation of Expert Views, and Outlook. Readers who are familiar with the topic may wish to skip this section or parts of it.

2.1. Background on Earthquake Forecasting Models Used for OEF

2.1.1. Basics of Statistical Seismology

The study of earthquake occurrence patterns goes back a long time. In 1756, Immanuel Kant noted that the frequency of earthquakes occurring in a region is more related to its proximity to mountain ranges and “fire-spitting mountains” than to the degree of Christianness of its inhabitants (Kant, 1756). Since then, descriptions of such empirical observations have become more precise and quantitative; the field of *statistical seismology* has emerged. One of its most central (empirical) laws, the Omori-Utsu law (Omori, 1894; Utsu, 1961), describes a possible trend in the decrease of the expected number of aftershocks in time after a parent event as

$$N(t) = \frac{k}{(t + c)^p}, \quad (1)$$

with parameters k , c , and p . The number of aftershocks (or earthquakes) per unit time, here $N(t)$, is referred to as the *aftershock rate* (or *earthquake rate*). In short- or medium-term forecasting, it is often provided in the unit of earthquakes per day. The parameter k in Equation 1 describes the overall number of aftershocks of a certain parent event. Note that we use the term *parent event* (rather than *mainshock*) to emphasize the possibility that the parent event can trigger aftershocks that are larger than itself. The exponent p of the denominator in Equation 1 quantifies how fast the aftershock rate decays with time—large values of p indicate a faster decay, and small values indicate a slower decay. Typically, p takes values of around 1 (Page et al., 2016), which corresponds to an aftershock rate decay that behaves like t^{-1} . The parameter c avoids a singularity at time $t = 0$, and interpretations of the parameter range from it being an artifact of more noisy data recordings right after a large earthquake (Hainzl, 2016a, 2016b; Kagan, 2004; Lolli & Gasperini, 2006) to it being a consequence of the physics of earthquakes (Lippiello et al., 2007; Narteau et al., 2009; Peng et al., 2006). Typically observed values of c are of the order of magnitude of several seconds to several minutes (Page et al., 2016).

The operational provision of earthquake forecasts has practical value only if the earthquake probabilities provided are time-variant. And the variability with time of most models currently applied for OEF in Italy, New Zealand, and the United States is centered around the Omori-Utsu law. The EEPAS model (Every Earthquake a Precursor According to Scale; Rhoades & Evison, 2004; described later in this section) stands as an exception, although it, too, uses prior earthquakes as the sole precursory signals for impending earthquakes. Thus, one limitation common to all currently authoritatively applied forecasting models is their inability to predict (or forecast with probabilities considered high by decision-makers) earthquakes. Similarly common to these models is their main strength of capturing the occurrence patterns of aftershocks following a moderate to large event.

The possibility that an aftershock has a larger magnitude than the current mainshock can be captured in most short-term forecasting models through another central empirical law of statistical seismology—the Gutenberg-Richter relationship (Gutenberg & Richter, 1944). It describes the relative frequency of earthquakes in relation to their magnitude as

$$N(M) = 10^{a-b \cdot M}, \quad (2)$$

where $N(M)$ is the expected number of events with magnitudes greater than or equal to M , and a and b are parameters generally referred to as a -value and b -value. In this formulation, the a -value parametrizes the number of earthquakes with magnitude 0 or larger in the region and time horizon of interest. Note that using 0 as a threshold value is somewhat arbitrary and does not represent a lower limit on the magnitude. The a -value naturally tends to vary in space; places close to boundaries of tectonic plates for example, tend to have more earthquakes and thus higher a -values than places in the middle of tectonic plates. Spatial variations of the a -value are merely a more abstract formulation of Kant's (1756) observations mentioned at the beginning of this section: the expected number of earthquakes may differ from place to place. The b -value parametrizes the relative frequency of large magnitude events compared to small magnitude ones. In general, the b -value for moment magnitudes is found to be around 1 (Kagan, 1999; Kamer & Hiemer, 2015; Utsu, 1972) for regions that are considered large compared to the seismic faults they contain, which implies that 10 times as many $M \geq 5$ earthquakes are expected as $M \geq 6$ earthquakes, 10 times as many $M \geq 4$ earthquakes as $M \geq 5$ earthquakes, and so on. Variations of the b -value in time and space have been widely discussed and have been proposed to be related to stress, volcanic activity, rock heterogeneity, crack density, and other physical quantities (Main et al., 1992; Mori & Abercrombie, 1997; Murru et al., 2007; Scholz, 2015; Schorlemmer et al., 2004, 2005; Wiemer & Wyss, 1997, 2002).

Combining the expected decay of the number of aftershocks given by the Omori-Utsu law with the knowledge about the relative frequency of small and large magnitude events given by the Gutenberg-Richter relationship enables an estimation of the expected probability that an aftershock larger than the current mainshock (or larger than any magnitude of interest) is yet to follow. In this sense, such probabilistic models provide earthquake probabilities, rather than aftershock probabilities.

A crucial concept for the type of analysis of earthquake occurrence that has been so far described is that of an *earthquake catalog*, which is a list of earthquakes, where each earthquake is characterized through several attributes related to its source. Most catalogs contain each earthquake's epicenter (the point on the earth's surface above the point where the earthquake started), depth, origin time (the time when the earthquake started), and magnitude. Additional information such as details about the rupture (orientation of the fault plane, direction of slip, etc.), and other attributes or attribute uncertainties may be available too.

An earthquake catalog is usually compiled for a specific region or area based on the recordings of seismic waves captured by a network of seismometers covering that region or area. Earthquake catalogs can be heterogeneous for various reasons. The seismic network only records signals at specific points, hence, naturally, not all locations are observed equally well. Some subregions may be more densely covered by the network than others. The quality of catalogs can vary with time, for example, because more instruments are deployed over time, or because newer earthquake detection methods become available (e.g., Hutton et al., 2010). In the shorter term, the ability to detect earthquakes can fluctuate after the occurrence of a relatively large earthquake, when the waves of smaller earthquakes are hidden within the large amplitude waves of the bigger one (Kagan, 2004). When discussing the issue of varying detection capability, the concept of the *completeness* or *incompleteness* of a catalog is often used. In general, large earthquakes are more easily detectable than small ones. The *magnitude of completeness* (also: *completeness magnitude*), which we will here denote with M_c , is the magnitude above which all earthquakes, in the region and time span of interest, are assumed to be detected and thus to be present in a catalog. Because it is impossible to know about the earthquakes that have not been detected, *because* they have not been detected, M_c has to be estimated and various methods have been proposed to do this (e.g., Schorlemmer & Woessner, 2008; Wiemer & Wyss, 2000).

The magnitude itself is a quantity that deserves discussion. An earthquake catalog is a complex entity because the quantities it contains are inferred from seismograms through uncertain and non-unique analyses. For the purposes of earthquake forecasting, the magnitude deserves particular attention because it does not have a clear physical meaning and forecasts are particularly sensitive to it. In contrast to the intensity of shaking, which varies between different locations based on their distance to the earthquake and other relevant factors, the magnitude is meant to be a unique quantity describing the size of the earthquake, independently of where it was measured. Several different definitions of magnitudes (i.e., *magnitude scales*) have been developed in the literature, and the way magnitudes are calculated varies strongly between the different scales, but also between different study regions. Two of the most widely used magnitude scales are *local magnitudes* (M_L) and *moment magnitudes* (M_w). Local magnitudes are defined based on the maximum amplitudes of recorded transient displacement at a seismic station, and the effects of the location at which the recording was made on the recording itself are corrected for, to obtain a

unique value which ideally does not depend on the location of the recording. This magnitude scale was originally introduced by Richter (1935) and is often called the *Richter magnitude scale*. Moment magnitudes (Hanks & Kanamori, 1979) are more tightly linked to earthquake source features: they are based on the energy (the seismic *moment*) released by the earthquake. Obtaining the moment released by an earthquake requires more complex processing of the recorded data. It is usually only done for a subset of earthquakes for which local magnitudes can be calculated. Both local and moment magnitudes, as well as other magnitude types not discussed here, scale logarithmically with the physical quantity they intend to measure. Although different magnitude types are usually defined such that the values of these different magnitude types of individual earthquakes coincide as much as possible, an exact match is not generally possible and the presence of different magnitude types within a catalog is another source of catalog heterogeneity which may need to be addressed by analysts. For example, using local or moment magnitudes for the same earthquake catalog can result in substantially different *b*-value estimates (Deichmann, 2017) and some magnitude types adhere less closely to the Gutenberg-Richter relationship than others (Herrmann & Marzocchi, 2021). In the following, we will refer to magnitudes mostly without specifying magnitude type, assuming what is discussed holds for different magnitude types.

The macroseismic intensity scales, such as the Modified Mercalli Intensity scale (MMI, H. O. Wood & Neumann, 1931) on the other hand, are twelve-level scales of Roman (ordinal) numerals which qualitatively indicate an increasing level of shaking based on observed effects, in vast areas, of an earthquake on people, animals, objects, infrastructure and landscape. Ultimately, these effects are what one is interested in when assessing the risk posed by earthquakes. While an earthquake ideally has a unique magnitude (ignoring the just discussed different possible magnitude scales), it can cause shaking of different MMI levels at different locations. However, because of the qualitative nature of macroseismic intensities, and the fact that they do not relate to specific sites of interest, quantitative risk assessment preferentially expresses earthquake hazard in terms of ground motion intensity measures, which are coupled with vulnerability and consequence models to compute (probabilistic) seismic risk metrics.

2.1.2. Description of Model Types Used for OEF

2.1.2.1. The Reasenber and Jones Model

Combining the Omori-Utsu law with the Gutenberg-Richter relationship yields the arguably simplest model currently used for OEF: the Reasenber and Jones (R&J; 1989) model. It defines the rate $\lambda(t, M)$ of earthquakes of magnitude M or above at time t after a parent event of magnitude M_m as

$$\lambda(t, M) = 10^{a-b(M-M_m)} \cdot (t+c)^{-p}. \quad (3)$$

Here, the second factor is the denominator of the Omori-Utsu law, and the first factor describes the expected number of aftershocks above a certain magnitude according to the Gutenberg-Richter relationship. Note that the parameter a here has a different interpretation than in Equation 2; here it controls the number of aftershocks larger than the parent event. The R&J model is thus based on the assumption that the number of aftershocks larger than their parent event is independent of the size of the parent event. A b -value of 1 would imply that, counting all aftershocks above a fixed magnitude M , a magnitude 6 event has (on average) 10 times as many aftershocks as a magnitude 5 event. Using the same scaling for aftershock productivity and the Gutenberg-Richter relationship produces a self-similar process in which earthquakes in equal-size magnitude bins produce the same total number of aftershocks. It also implies that the aftershock and foreshock processes are similar and share the same scaling (e.g., Felzer et al., 2004).

Reading the previous paragraph, one may have noticed that it mentions parent events rather than mainshocks, in contrast to the originally used terminology of Reasenber and Jones (1989). As noted earlier, this is to emphasize that a parent event can trigger aftershocks of any magnitude. In fact, if a large aftershock occurs, it is expected to trigger aftershocks of its own. It is a main limitation of the R&J model that it does not specifically model such *higher-order aftershocks* (i.e., aftershocks of aftershocks, and so on), but considers all aftershocks as triggered by one event. This simplification allows the R&J model to be analytic and computationally efficient, advantages that were particularly important when it was introduced.

2.1.2.2. The Epidemic-Type Aftershock Sequence Model

Around the same time as when the R&J model was introduced, Ogata (1988) first described what is now widely known as the Epidemic-Type Aftershock Sequence (ETAS) model, and it was first used as a forecasting tool by Console et al. (2003). As suggested by its name, it models seismicity to resemble an epidemic, where earthquakes recursively trigger more earthquakes, similarly to how infected beings can recursively infect others around them. Numerous variants of ETAS formulations have been and keep being proposed (Asayesh et al., 2023; Console et al., 2003; Guo et al., 2015; Harte, 2013; Helmstetter & Sornette, 2002; Nandan et al., 2017; Ogata, 1993, 1998, 2011; Ogata & Zhuang, 2006; Veen & Schoenberg, 2008; Z. Xiong & Zhuang, 2023). Generally, ETAS describes the rate $\lambda(t,x,y)$ of earthquakes with magnitudes above a reference magnitude M_r at time t and location (x,y) as

$$\lambda(t, x, y) = \mu + \sum_{t_i < t} k(m_i) \cdot g(t - t_i) \cdot f(m_i, x - x_i, y - y_i). \quad (4)$$

The key idea of the model is that earthquakes are partitioned into background earthquakes and triggered earthquakes. All earthquakes not triggered by a previous one are considered background and are assumed to occur uniformly in time. The rate of background earthquakes with magnitudes above M_r is captured by the parameter μ .

In addition to the background earthquakes come the aftershocks with magnitudes above M_r of all prior earthquakes. The summation in Equation 4 is over all earthquakes which occurred at times t_i before time t and at locations (x_i, y_i) and had magnitudes m_i . Each prior earthquake contributes to the rate of triggered events, and the rate of its aftershocks is defined through three components. The total number of expected aftershocks or *aftershock productivity* $k(m_i)$ depends on the magnitude m_i of the parent earthquake; the temporal decrease of the aftershock rate is modeled through $g(t - t_i)$ and only depends on the elapsed time since the parent event. The spatial distribution of aftershocks, $f(m_i, x - x_i, y - y_i)$, can depend on the magnitude of the parent event and is often defined as isotropically decreasing with the distance to the parent event (Musmeci & Vere-Jones, 1992; Ogata, 1998). Figure 1 schematically illustrates the concept of aftershock triggering within the ETAS model.

The temporal aftershock decay is most often modeled according to the Omori-Utsu law given in Equation 1. The aftershock productivity $k(m_i)$ is often described to depend on the magnitude of the parent event through an exponential relationship of the form

$$k(m) = K_0 \cdot 10^{\alpha \cdot (m - M_r)}, \quad (5)$$

(Utsu, 1971). The parameter α parametrizes the increase of the aftershock number per magnitude unit of the parent event. If α equaled the b -value of the Gutenberg-Richter relationship, as is often suggested, observed, and imposed (Hainzl et al., 2008, 2013; van der Elst et al., 2022), the assumption made in the R&J model that the number of aftershocks larger than their parent event is independent of the parent event magnitude would be fulfilled.

A variety of spatial aftershock decay kernels has been proposed and will not be further discussed here; instead, we refer to Zhuang et al. (2011) for a detailed technical overview of spatio-temporal ETAS models. The description of the ETAS model given here is a simplistic one. Variants in which the background seismicity rate μ , but also other model parameters, vary in space and/or time are common. Generally, the estimation of ETAS parameters has been studied thoroughly, yielding improved parameter estimation techniques (e.g., Lippiello et al., 2014; Lombardi, 2015; Ross, 2021; Schneider & Guttorp, 2020; Veen & Schoenberg, 2008) and aiming to understand and address biases in parameter estimates due to data imperfections (e.g., Grimm et al., 2022; Harte, 2013; Seif et al., 2017).

The important quality of the ETAS model that it accounts for higher-order aftershocks is also the cause of a noteworthy drawback: even when the parameters describing background seismicity and aftershock triggering are known, the expected rate of future earthquakes cannot be calculated analytically. To obtain the expected rate of events at time t , the summation in Equation 4 is over all events that occurred prior to t . In practice, one may be interested in the expected rate of events hours, days, weeks or even years into the future. Thus, aftershocks of events that have not yet occurred when the forecast is calculated are not accounted for in Equation 4. Modelers address this by issuing many simulated *stochastic event sets*, containing simulated earthquakes and cascades of

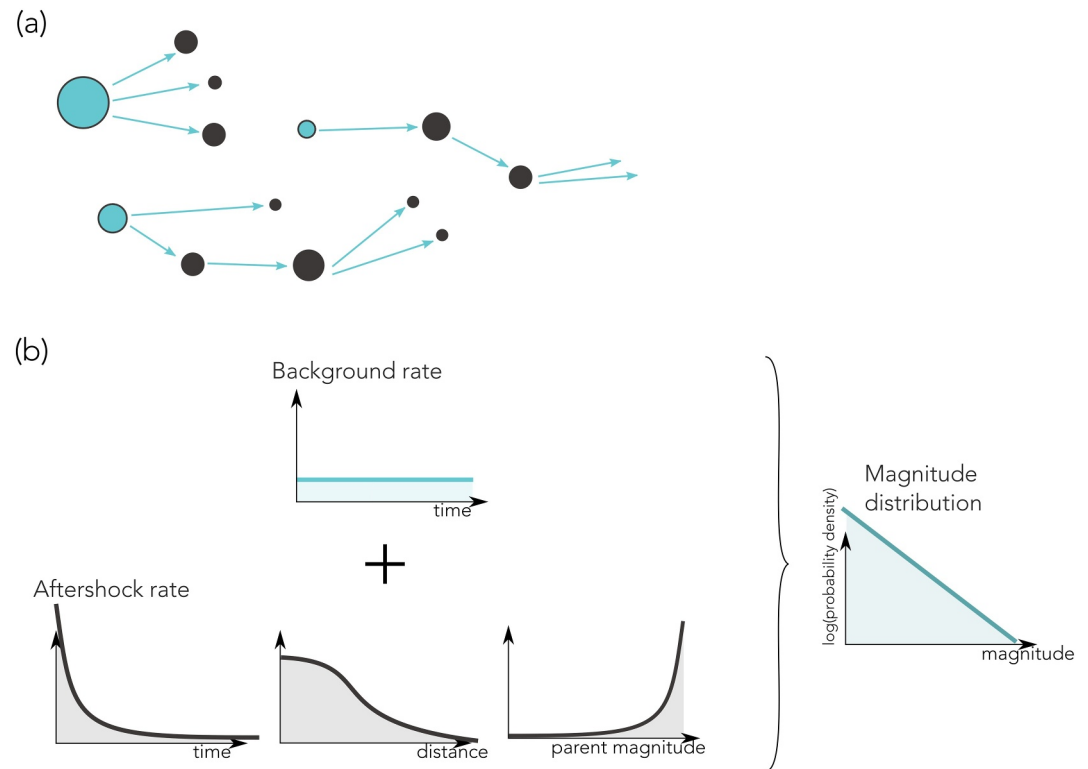


Figure 1. Schematic illustration of the ETAS model. (a) Background earthquakes (turquoise) and aftershock cascades (black). Arrows indicate triggering. (b) Composition of the earthquake rate into a temporally invariant background rate and aftershock rate. The aftershock rate itself has three independent components: it decays with time and distance from the parent event, and increases with the magnitude of the parent event. All earthquakes (background events and triggered events) follow the same magnitude distribution.

aftershocks, which yields more correct forecasts (Helmstetter & Sornette, 2003) but requires more computation time.

2.1.2.3. The Epidemic-Type Earthquake Sequence Model

Although some authors use the name Epidemic-Type Earthquake Sequence (ETES) model as a synonym for the ETAS model (e.g., Helmstetter et al., 2006), we maintain its original use as a special case of the ETAS model based on the Omori-Utsu law for the temporal aftershock decay, and a circular symmetry around the triggering event's epicenter for the spatial component of the rate (Console et al., 2003; Falcone et al., 2010). The main distinction between ETES and ETAS as defined in Equation 4 can be summarized in the following points.

1. The background component of the rate is still time-invariant, but it depends on a failure rate, defined as the ratio between the expected number of background events and the total number of events.
2. In ETES, the expected spatial density of triggered events does not depend on the magnitude of the triggering event and hence is of the form $f(x - x_p, y - y_i)$.
3. The parameter α in Equation 5 is set to 1.0. Physically, this means that the number of triggered events scales with the magnitude of the parent event in a similar way as its rupture area (Console et al., 2006; Hainzl et al., 2008; Murru et al., 2014). Note that assuming a b -value of 1.0, setting $\alpha = 1.0$ is equivalent to setting α equal to b , which was discussed earlier. The ETES model with $\alpha = 1$ was applied to the aftershock sequence following the 2011 Mw 9.0 Tohoku-Oki, Japan, earthquake and achieved a better performance at forecasting event numbers compared to other models submitted in the same test (Nanjo et al., 2012).

2.1.2.4. The Short-Term Earthquake Probability Model

The Short-Term Earthquake Probability (STEP) model (Gerstenberger et al., 2004; Woessner et al., 2010) is based on the same core principles as the R&J and ETAS models. It uses the R&J model to describe aftershock rate,

but differs from it in that it does, in addition, include a component that models the long-term background seismicity, does account for aftershocks of aftershocks, and does include a description of seismicity in space. It thus shares its main advantages over the R&J model with the ETAS model. The way in which these issues are addressed however differs between ETAS and STEP. While in ETAS, the background and the recursive space-time aftershock triggering are all self-consistently described within a single model, the STEP model is a combination of several components that could themselves be viewed as individual models.

Besides a time-invariant component, a STEP model can have several time-variant components: a generic one based on past seismicity, and sequence-specific ones which can be derived in different ways. In these time-variant components, recorded earthquakes are classified to be either primary events or secondary events. An event is considered secondary if it is an aftershock of a specific prior event, which can but does not need to be a primary event. Aftershocks are associated with their parent events through aftershock zones defined in space and time, and they can only be considered aftershocks of events of equal or larger magnitude. As in ETAS, all earthquakes are considered capable of triggering aftershocks. A distinction is made however between primary and secondary events, where a sequence-specific set of parameters can describe the expected aftershock rate of primary events. In contrast, secondary events always use generic parameters. Within one component, the forecasted rate of events at a given location is given as the maximum of forecasted aftershocks at that location of all previous primary or secondary events.

The different time-dependent components are combined by weighting them according to how well they fit the observed data, and the combined rate forecast is added to the time-independent rate forecast to obtain the final STEP forecast.

An advantage of the STEP model is that, similar to the R&J model, it is analytic and computationally efficient, while partially including secondary triggering (unlike the R&J model). Its process of combining time-dependent components with a long-term hazard model allows it to easily integrate, and be consistent with, authoritative national seismic hazard models. STEP can accommodate different spatial kernels and can either fit a two-segment fault model or take external finite fault input. A disadvantage of STEP, compared to ETAS, is that it does not fully model secondary triggering.

2.1.2.5. *The Every Earthquake a Precursor According to Scale Model*

The Every Earthquake a Precursor According to Scale (EEPAS) model was introduced by Evison and Rhoades (2004) and Rhoades and Evison (2004). Although it shares the key principle of the ETAS model of describing earthquake rate as the sum of a background term and the sum of rate contributions of prior events (see Equation 4), the specific way in which prior events contribute to the forecasted rate differs substantially from the ETAS formulation. There is no suggestion that each earthquake may trigger its own offspring. Rather, each earthquake is regarded as evidence of a medium-term seismogenic process taking place on the scale indicated by its magnitude (Rhoades, 2007). The probabilistic distribution of an earthquake's contribution is lognormal in time, isotropic bivariate normal in space and normal in magnitude (Rhoades & Evison, 2004).

The EEPAS model is underpinned by the observation that prior to most large earthquakes the rate and magnitude of smaller earthquakes increases in the source region of the large earthquake. This precursory scale increase (Ψ -phenomenon) takes place within a precursor time T_p and precursory area A_p . Evison and Rhoades (2004) showed that T_p , A_p , and the mainshock magnitude M_m all scale with the precursor magnitude M_p (defined as the average of the three largest precursory earthquake magnitudes). Thus, M_p is predictive of the time, magnitude and location of forthcoming mainshocks. The assumption behind EEPAS is that the three Ψ predictive relations are pervasive at all scales in the seismogenic process. The model regards every earthquake as a precursor, according to scale, of larger earthquakes expected to follow it in the coming months, years, or decades, depending on its magnitude and the regional earthquake rate.

2.1.2.6. *Ensemble Models*

Rather than relying on a single model to provide a forecast, one can combine multiple models into one ensemble model. Ensemble modeling proved beneficial in various fields and applications such as weather/climate forecasting (Hagedorn et al., 2005; Krishnamurti et al., 2000; Tebaldi & Knutti, 2007), flood forecasting (Cloke & Pappenberger, 2009), ecology (Buisson et al., 2010), computer security (Menahem et al., 2009), and long-term

seismic hazard assessment (Gerstenberger et al., 2020). Ensembles can emphasize the individual strengths of different types of models (e.g., physical, statistical, deterministic), aim to achieve better performance than the best single model, are more flexible than single models, and better capture (different kinds of) uncertainty.

In earthquake forecasting research, several ensembles have been created using various methods. On short-term (time-varying) scales, weighted averaging of rate-based statistical models is common (Helmstetter & Werner, 2014; Herrmann & Marzocchi, 2023; Llenos & Michael, 2019; Marzocchi, Zechar, & Jordan, 2012; Rhoades & Gerstenberger, 2009; Taroni et al., 2018), and some ensembles incorporate physical information (Király-Proag et al., 2018; Steacy et al., 2013) or make an existing model more flexible (Gerstenberger et al., 2004; Mizrahi, Nandan, et al., 2023). These studies have shown promising results, even though ensembles have not always performed much better than the best individual model. On long-term (time-invariant) scales, a multiplicative procedure has been used to incorporate models and data other than earthquake rates, like strain and fault slip rates (Bird et al., 2015; Rhoades et al., 2014, 2017). Prospective testing of some of those has shown mixed results: although they can outperform their candidate models on a global scale (Strader et al., 2018), they may not do so on a regional scale (Bayona et al., 2022). Despite a wide range of possible ensemble methods, only weighted averaging allows for considering the dispersion of the single forecasts to inform about the epistemic uncertainty (Marzocchi & Jordan, 2014, 2017), that is, our lack of knowledge. A reasonable estimation of the epistemic uncertainty should be based on weights that account for the correlation of the forecasts (Marzocchi, Zechar, & Jordan, 2012) or maximize the forecasting skill of the ensemble (Herrmann & Marzocchi, 2023).

Even if some ensembles have not greatly outperformed the best-performing single model, they do not usually perform much worse. It remains to be seen whether the relatively limited skill improvements are due to sub-optimal ensemble methods, the limited skill of the single models, or because models did not complement each other well. A possible path for more skilled ensembles is to incorporate diverse models (e.g., physics-based approaches, for short-term earthquake forecasting like those proposed by Mancini et al. (2019), Cattania et al. (2018), Sharma et al. (2020); or Dahm and Hainzl (2022)), because diversity increases the predictive pool and enables exploiting well-performing forecasting models.

2.2. Background on Earthquake Forecast Model Testing

One decision-making process in OEF involves the selection of one or more *suitable* forecasting models. While many factors might affect the selection process (including less or non-scientific aspects such as required computational resources in real-time, ability to explain the model to non-experts, internal expertise to develop and maintain the model, maintaining some continuity with past model forecasts, etc., see also Marzocchi & Zechar, 2011), a key scientific question is how well a model can forecast future seismicity, especially in relation to other available models. OEF itself is motivated by scientific studies concluding that some time-dependent models forecast earthquake clustering better than time-independent models (e.g., Cattania et al., 2018; Helmstetter et al., 2006; Werner et al., 2011), most obviously during aftershock sequences. Thus, a major community effort has focussed on the performance evaluation of available earthquake and aftershock forecasting models, including OEF candidate models. This subsection reviews some of the basic concepts and methods that the community has used to assess model performance. We will focus primarily on the major community effort organized by the global Collaboratory for the Study of Earthquake Predictability (CSEP; Jordan, 2006; Field, 2007; Michael & Werner, 2018; Schorlemmer et al., 2018; Savran et al., 2022).

2.2.1. Basic Concepts

2.2.1.1. Testing

Model inference, evaluation, and selection are all well-established and still evolving fields, with many approaches and applications from across many disciplines. The philosophical foundations of (probabilistic) model evaluation approaches continue to be debated (Harte & Vere-Jones, 2005; Serafini et al., 2022), perhaps as a result of the different interpretations of probability (Marzocchi & Jordan, 2014). The seismological community has thus far largely focused on classical frequentist-type testing of probabilistic forecasts, some of which are described below (see Basic Methods). A recognition of the shortcomings associated with frequentist testing and likelihood-based model comparisons (e.g., Field, 2015; Nandan et al., 2019; Savran et al., 2022; Werner & Sornette, 2008) has recently led to more modern approaches, including simulation-based inference approaches (e.g., Page & van der Elst, 2018; Savran et al., 2020, 2022), Bayesian (e.g., Marzocchi, Zechar, & Jordan, 2012) and *unified* approaches

(Marzocchi & Jordan, 2014, 2017, 2018), the latter two of which are beyond our scope here but provide compelling frameworks. In the present context, we can define *testing* a forecasting model as formal procedures to identify discrepancies between forecasts and data, and interpreting these discrepancies in terms of their scientific and/or practical implications. Typically, models are also compared against each other in comparison tests.

2.2.1.2. Testability

A study of the “tumultuous history of earthquake prediction” (Hough, 2010) shows that many past claims of successful earthquake predictions and forecasts based on models were not reproducible (see, e.g., National Academies of Sciences, Engineering, and Medicine, 2019), including those in peer-reviewed publications. A range of reasons can explain this, but many can be summarized by a lack of testability of a model (e.g., Jackson, 1996). To be *testable*, there must be (a) a clearly defined region and depth over which the forecast or prediction applies, a magnitude range, and a time window, (b) a clear definition of which characteristic of an earthquake is being forecast (e.g., epicenter or centroid, moment magnitude or surface wave magnitude) and (c) a pre-specified authoritative, independent source of target/test data (e.g., Jackson & Kagan, 1999).

2.2.1.3. Authoritative, Independent Test Data Sources

An ambiguity in testing involves the choice of the target or test data set, at times for the same earthquake. Earthquake catalogs are data products derived from seismic waveform observations, quantitative models of the Earth, fixed or adjusted parameters, and sometimes ad hoc human intervention. The same earthquake might be assigned different parameters by different networks, or even multiple Psource parameter estimates (e.g., magnitude) by the same network. To avoid this ambiguity and any potential bias due to a post-hoc selection, the source and type of test data should be specified in advance, and originate from authoritative, independent networks and agencies.

2.2.1.4. Prospective Testing

The earthquake forecasting community places great emphasis on *prospective testing*, which comprises the evaluation of forecasts against future, yet-to-be-collected seismic data. For example, a forecast created and archived in real-time can be evaluated prospectively in the future against the earthquake data collected between when the forecast was issued and the end of the testing window (also called the *forecast horizon*). Prospective testing can also be conducted in a delayed manner, that is, with forecasts not created in real-time, when the model is entirely isolated from any data beyond the forecast issue time, and the test data set is fully specified before the forecast issue time. This requires fully automated and transparent procedures to ensure a zero-degrees-of-freedom environment. In its first phase, CSEP built and managed so-called testing centers (Schorlemmer & Gerstenberger, 2007; Zechar, Gerstenberger, & Rhoades, 2010; Zechar, Schorlemmer, et al., 2010) to implement delayed prospective testing by providing a controlled, secure and automated environment to test models with at least a 1-month delay. In this way, delayed prospective testing can remain entirely *out-of-sample* (i.e., model fitting and testing data are strictly separate) and, importantly, testing is against future data that neither model nor modeler have seen. Prospective testing provides the most objective view of the forecast skill of a model due to excluding even unconscious biases of modelers. It is often assumed to provide the most reliable estimate of future performance, although it is not yet clear for how long models must be tested to obtain stable and robust performance estimates. This is however not a specific problem of prospective testing, but one of testing in general.

2.2.1.5. Retrospective Testing

In contrast to prospective testing, *retrospective testing* is the testing of forecasts against past data. In retrospective testing, the test data were available to the modelers and thus consciously or unconsciously influenced modeling and/or testing choices (e.g., Schorlemmer et al., 2018). Test data may or may not have been explicitly used to fit or calibrate a model. To emphasize when test data were not known by the model, and when time-dependent causality was preserved when separating the test data from the data used for model fitting, the term pseudo-prospective testing, described below, is often used. Figure 2 visualizes the different testing modes in terms of the data split between test data, data known to the model, and data known to the modelers. An example of retrospective testing for which time-dependent causality is not preserved is shown at the bottom of the figure, where the test data are not known by the model, but cover a period before the period used for model calibration.

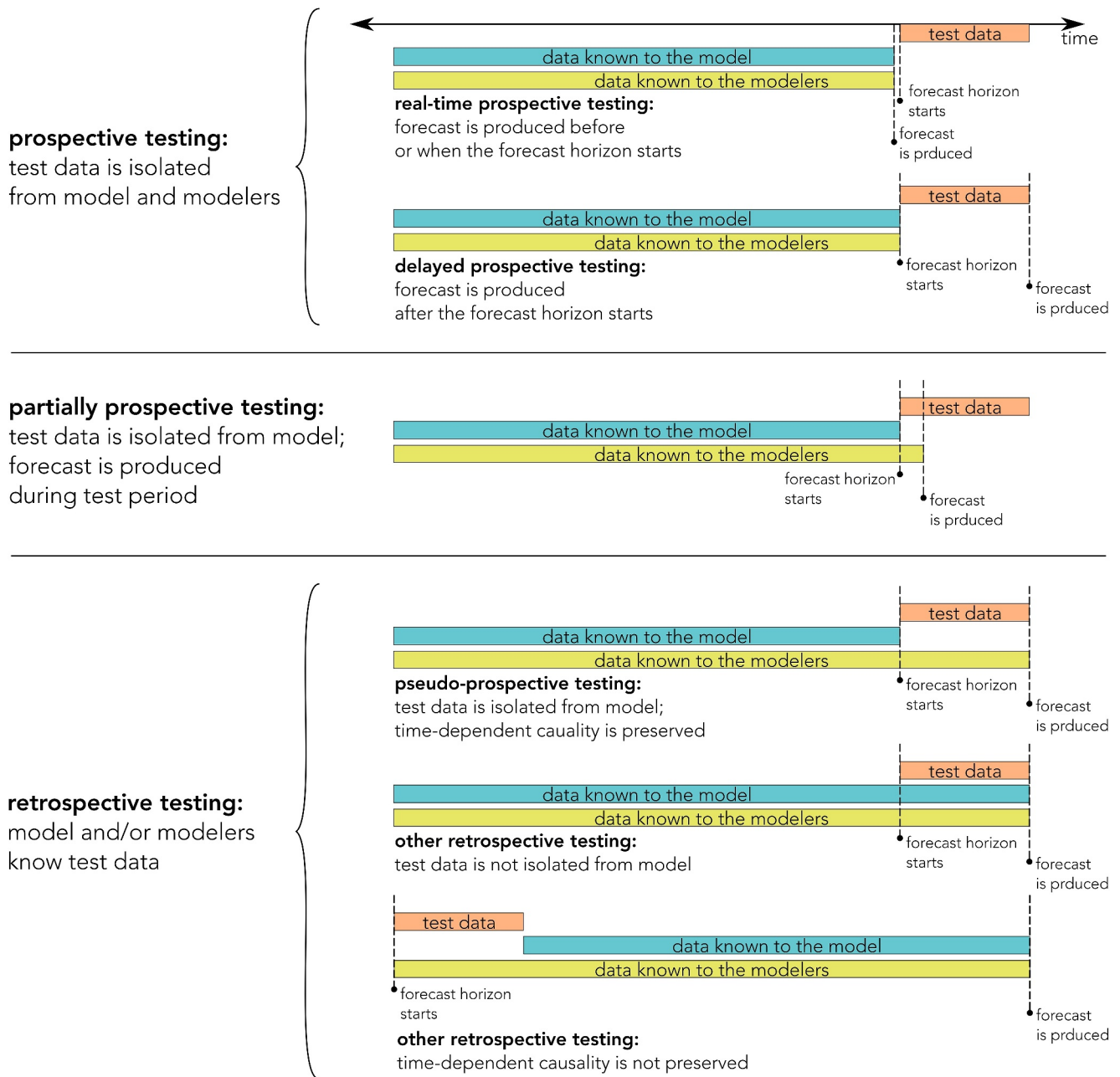


Figure 2. Different modes of testing, visualized in terms of the data split between test data, data known to the model, and data known to the modelers.

Because truly prospective testing is often associated with long waiting times to obtain the yet-to-be-collected data against which forecasts are evaluated, most scientific publications report model evaluations conducted retrospectively. The main drawback of retrospective testing is that modeling choices or testing choices (e.g., the choice of a particular data set for testing) may affect model performance, even when testing data were not explicitly used by the model. The extent of the resultant bias in performance is difficult to assess. However, a rule of thumb is that retrospective performance will provide an upper bound to truly prospective performance. Nevertheless, retrospective tests provide added value by acting as sanity checks, ensuring that the models function as intended by the modelers.

2.2.1.6. Pseudo-Prospective Testing

The subset of retrospective testing called *pseudo-prospective* testing aims to mimic the circumstances of prospective testing and preserve time-dependent causality. Models are being calibrated using data only until a certain time t_0 in the past, and are then used to produce forecasts for the time after t_0 . In this way, models and modelers pretend not to know what occurs after time t_0 , but the modelers might be biased due to their knowledge of the testing data or certain features thereof. Thus, while pseudo-prospective tests can formally be considered out-of-sample, modeling and/or testing choices may still be consciously or unconsciously influenced by the testing data, rendering pseudo-prospective tests less informative of a model's performance than truly prospective tests.

Thus, while pseudo-prospective tests can formally be considered out-of-sample, modeling and/or testing choices may still be consciously or unconsciously influenced by the testing data, rendering pseudo-prospective tests less informative of a model's prospective performance than truly prospective tests.

2.2.1.7. Partially Prospective Testing

In the literature, a testing approach that neither qualifies as prospective, nor can be considered fully retrospective and is thus not captured by the established terminology, can often be found. This approach, which we define as *partially prospective testing*, involves testing forecasts against partially future, partially past data. For instance, researchers might generate a forecast of aftershocks a few hours or days after a mainshock, for example, after quality-controlled information about the mainshock source parameters have become available, such as magnitude, faulting mechanism or the hosting fault, or researchers may have calibrated their aftershock model to a new region that experienced the mainshock (see e.g., Mancini et al., 2020; Milner et al., 2020, for forecasts issued during the 2019 Ridgecrest, California, sequence). During this time, aftershocks will already have occurred and been recorded, and indeed may influence the quality control of mainshock parameters, such as its rupture planes and spatial extent. The forecast horizon, however, might start in the past (at the time of the mainshock) and extend into the future, so as to be useful in some form. In addition, these forecasts might eventually be evaluated and published in a peer-reviewed publication, but no time-stamped archived record of the forecast might exist.

As illustrated in Figure 2, we make an overall distinction between prospective and retrospective testing, where prospective testing can be done in real time or in a delayed manner, and retrospective testing can be done pseudo-prospectively (with a data split that does preserve time-causality), or out-of-sample with a data split that does not preserve time causality, or in-sample. In-between these two categories lies partially prospective testing.

2.2.2. Basic Methods

2.2.2.1. Forecast Specification

During the first CSEP experiment, the Regional Earthquake Likelihood Models (RELM) experiment of 5-year forecast models in California (Field, 2007; Schorlemmer et al., 2007), modelers decided to define a forecast Λ as the expected number of events (mean rates) in a space-magnitude discretization (cells/bins of a grid), such that $\Lambda = \{\lambda_i\}$ for $i = 1, \dots, I$ bins. These long-term forecasts were time-independent, both in the sense of their mathematical formulations (statistical descriptions were not explicitly dependent on time) and the decision that models would not be updated with new data during the forecast time horizon. This implied that forecast models could be issued for testing as flat files of numerical data (Schorlemmer & Gerstenberger, 2007). Following this scheme, short-term 1-day RELM forecasts (relevant for OEF) were also specified as mean rates on a space-magnitude grid. However, it was required that the model source code be available for prospective testing, so the forecasts could be updated with new data after the 1-day time window, while preventing the modelers from modifying the original model formulation. In practice, the first generation of short-term forecasts handled by CSEP was tested in a similar fashion and with similar metrics as the long-term forecasts, with the only difference of providing a temporal evolution of the testing results (e.g., Cattania et al., 2018; Taroni et al., 2018).

2.2.2.2. Likelihood-Based Model Scoring, Comparison, and Ranking

The probabilistic nature of forecasts allows them to be evaluated using likelihood functions. That is, the joint probability $L(\Omega, \Lambda)$ of the observations Ω (i.e., the testing catalog), interpreted as a function of the issued probabilistic forecast Λ , measures how well a forecast explains the observed data. This concept is illustrated in Figure 3. Forecast scoring usually employs the log-likelihood score (or log-score) $LL(\Omega, \Lambda) = \ln L(\Omega, \Lambda)$, with

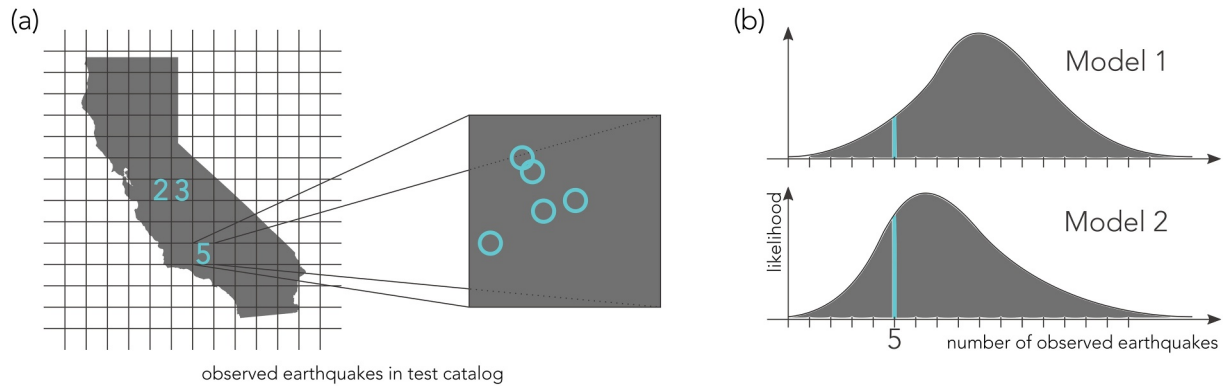


Figure 3. Schematic illustration of the concept of the log-likelihood score. (a) Observed numbers of earthquakes in individual cells of a spatial grid. (b) Comparison of likelihoods of the observed numbers under two alternative models for a particular cell. To obtain the log-likelihood score of a model, the logarithms of the likelihoods are summed over all grid cells (Figure adapted from Mizrahi et al. (2024)).

higher values implying a better performance. The original RELM testing formulation involved evaluating the log-likelihood function assuming a Poisson process, for multiple reasons: (a) it yields mathematical tractability for space-magnitude discretizations due to the independence of the distributions of event counts in the resulting bins/cells, (b) it can be derived from only one temporal parameter (the mean rate), and (c) its simple formulation enables confronting various types of models, even models that have no explicit likelihood functions. The multivariate Poisson log-likelihood is given as

$$LL(\Omega, \Lambda) = \sum_{i=1}^I \ln \frac{(\lambda_i)^{\omega_i}}{\omega_i!} e^{-\lambda_i}, \quad (6)$$

where ω_i is the number of events observed in the i th space-magnitude bin. The log-likelihood scores of different models can be compared to one another: a model with higher $LL(\Omega, \Lambda)$ than another provided a better forecast of the observations Ω . A ranking of the models' performance can then be built by ordering the forecasts according to their scores.

Differences in log-likelihood scores of different models are often referred to as information gain (IG), or, when divided by the number of earthquakes contained in the testing catalog, as information gain per earthquake (IGPE). Hereby, a positive IG (or IGPE) suggests a better performance of the alternative model, while negative IG suggests poorer performance and an IG of 0 suggests equal performance in terms of the log-likelihood score.

2.2.2.3. Frequentist (Consistency) Testing

Under the frequentist approach, predictive statistical distributions of a forecast model can be submitted to standard *hypothesis testing*, which informs us how (in)consistent a model is with new data. In traditional CSEP tests (Kagan & Jackson, 1995; Schorlemmer et al., 2007; Zechar, Gerstenberger, & Rhoades, 2010; Zechar, Schorlemmer, et al., 2010), referred to as consistency tests, the likelihood function evaluated on observed data is compared to likelihood values generated by the forecast thought of as a random process. Specifically, consistency tests determine whether $L(\Omega, \Lambda)$ falls within or outside a given confidence interval of a likelihood distribution $\hat{L} = \{L(\Lambda^j, \Lambda)\}$ obtained from a set of synthetic catalogs $\{\Lambda^j\}$ simulated using a Poisson process (e.g., Kagan & Jackson, 1995). A forecast is deemed *inconsistent* when a quantile score γ , defined as

$$\gamma = \frac{|\{L(\Lambda^j, \Lambda) \text{ such that } L(\Lambda^j, \Lambda) < L(\Omega, \Lambda)\}|}{|\{\Lambda^j\}|}, \quad (7)$$

satisfies $\gamma < p$ for a given critical value p , where $|A|$ denotes the number of elements contained in a set A . The quantile score can be interpreted as the probability that a likelihood score sampled from \hat{L} is lower than that of the observations Ω . This likelihood consistency test (L-test) is complemented with the Number (N-), magnitude (M-)

and Spatial (S-) tests, which follow the same principle, but the likelihoods are calculated from forecasts reduced to the corresponding N-, M-, or S-space (e.g., only the total number of events in the region; only the magnitude distribution; only the spatial distribution), by summing the mean rate of Poisson distributions across the bins of the remaining dimensions. An extensive description, reformulation and implementation of these tests can be found in Schorlemmer et al. (2007), Zechar, Gerstenberger, and Rhoades (2010), Zechar, Schorlemmer, et al. (2010), Werner et al. (2011), Rhoades et al. (2011), and Savran et al. (2022).

Due to the spatio-temporal clustering and correlations in earthquake distributions, limitations of assuming a Poisson process when estimating \hat{L} quickly became evident in short-term forecasting experiments (e.g., Lombardi, 2014; Lombardi & Marzocchi, 2010a, 2010b; Nandan et al., 2019; Werner & Sornette, 2008). But the Poisson and cell independence approximations can also break down in long-term forecast experiments at higher magnitude thresholds when clustering persists. To reduce the undesired strong influence of clustering within cells on testing results, Bayona et al. (2022) used a binary likelihood of observing either zero or $n \geq 1$ earthquakes in each cell. To avoid the unrealistically narrow Poisson distribution for the total number distribution, Werner et al. (2010) used an overdispersed negative-binomial distribution to describe and assess the total number forecasts of models (see also e.g. Kagan, 2010). Nandan et al. (2019) smoothed ETAS model simulations to derive ETAS-specific cell-wise likelihood functions, thus representing the marginals within each cell well, while maintaining the cell-independence assumption.

2.2.2.4. Simulation-Based Testing

From a modeling perspective, dependencies between earthquakes are already captured in simulation-based forecasting methods, such as ETAS, which provide synthetic catalogs directly through simulations, rather than from assumed statistical distributions of event counts in given cells. Savran et al. (2020) thus extended the traditional CSEP tests by relaxing the assumption that earthquakes follow independent Poisson distributions in discretized space-time-magnitude regions. Instead, CSEP catalog-based tests are based on pseudo-likelihoods (and other metrics) that are aggregated over the likelihood scores of earthquakes in the test set, rather than over the cells/bins of the forecast region (as in Equation 6), by using the log-likelihood definition for a continuous point process,

$$LL(\Omega, \Lambda) = \sum_{i=1}^N \ln \lambda(e_i|H_t) - \int_R \lambda(e|H_t) dR, \quad (8)$$

where $\lambda(H_t)$ is the conditional intensity function specifying the point process, H_t is the earthquake history until the time t of the forecast, and R is the (space-time-magnitude-) region for which the forecast is issued, and the summation is over all events $e_i \in \Omega$. We direct the interested reader to Daley and Vere-Jones (2003, 2008) for a thorough introduction to the theory of point processes. Because the analytical description of $\lambda(H_t)$ may not be given, it is approximated from simulated catalogs as $\lambda(H_t) = E[\lambda(e|H_t)|R_d]$ at a region discretization R_d . In this way, pseudo-likelihood distributions can be built (and tested for consistency) for models that fully represent overdispersion and spatio-temporal dependencies through simulations.

2.2.2.5. Other Methods

Additional tests have been proposed in the literature, which do not depend on (pseudo-) likelihood functions or have not been yet implemented in routine CSEP experiments. These include Turing-style tests that determine if a set of simulations can imitate the behavior of observed catalogs (Page & van der Elst, 2018). For this, they devise a series of statistical descriptors that summarize the rate and magnitude distributions of a catalog, as well as foreshock/aftershock productivity and spatio-temporal clustering. In the same fashion as consistency tests, Turing-style tests observe if the observed catalog's descriptors lie within a confidence interval from those generated by a set of simulations. Furthermore, earthquake forecasts have been submitted to tests based on second-order (spatial) statistics (e.g., Clements et al., 2011; Veen & Schoenberg, 2006), which are based on K-functions (Ripley, 2005) to determine the level of clustering or inhibition of a catalog. Clements et al. (2011) also describe a series of residual analysis for point-process methods, which consist of transforming the points of a simulated/observed catalog (e.g., by rescaling, thinning, superpositioning), such that the resulting transformed process should be homogeneous-Poisson if the original model were consistent with the observations.

2.3. State of Research in Earthquake Forecast Communication

“What will happen next?” is one of the most frequently asked questions after an earthquake felt by the public. Short-term earthquake forecasts allow us to reply to this question in a probabilistic way. Following the 2009 L’Aquila earthquake, a scientific discourse started on how (and if) earthquake forecasts should be made available to society (Jordan et al., 2011; Marzocchi, Iervolino, et al., 2015; Marzocchi, Jordan, & Woo, 2015). While the value of long-term forecasts was clear (i.e., building codes, risk management plans (Marti et al., 2019)), the effective use of short-term earthquake forecasts by societal stakeholders was still unexplored (Jordan et al., 2011). Through workshops with experts and stakeholders, key concepts and relevant issues to be considered when communicating forecasts to society were derived (Field et al., 2016; Marzocchi, Iervolino, et al., 2015; Marzocchi, Jordan, & Woo, 2015). Furthermore, New Zealand has evaluated earthquake forecast communication products (e.g., tables, maps) and provided advice for other regions aiming to communicate forecasts (Becker et al., 2018, 2019, 2020). Other research teams around the world have also been advancing the ways to best communicate earthquake forecasts (e.g., Dryhurst et al., 2022; Schneider, Wein, et al., 2023), which are summarized in the following paragraphs.

Earthquake forecasts are useful for various stakeholders ranging from first responders and critical infrastructure owners to the general public (Field et al., 2016). Depending on their professions and experiences as well as the severity of the event, stakeholders have different information needs that evolve throughout the sequence (Becker et al., 2019). Becker et al. (2020) have shown that agencies and the public in New Zealand make use of earthquake forecasts for, among others, timing infrastructure repair and rebuilding, land-use planning decisions, undertaking preparedness actions such as organizing essential supplies and emergency items, or deciding about safe access to buildings. This is in line with the purposes of earthquake forecasts in the United States (U.S.), where for example, used by the Federal Emergency Management Agency (FEMA) to set the temporal duration of disaster declarations (Michael et al., 2020; van der Elst et al., 2020). Furthermore, hospitals may use earthquake forecasts to postpone high-risk treatment and elective surgery until the chance of aftershocks has significantly declined (Field et al., 2016). However, making sense of the probabilities of forecasts in ways that translate into actions and measures is not trivial and needs further investigation (Becker et al., 2018), especially in countries with moderate seismic risk where people only rarely experience earthquakes.

Studies on stakeholders’ *information needs* have shown that a wide range of information is requested, from basic facts about aftershocks through to more technical information. The general public, for example, desires information about the likely future earthquake locations, the likely future magnitudes/intensities and locations of aftershocks, the expected duration of the aftershock sequence, and guidance on what to do during the aftershock sequence (Becker et al., 2019). Communicating forecasts can also fill information vacuums that set the stage for amateur earthquake predictions and misinformation (Marzocchi, Iervolino, et al., 2015; Marzocchi, Jordan, & Woo, 2015). Furthermore, recently conducted workshops in the United States, Mexico, and El Salvador by a multinational team revealed that for numerous stakeholders (from emergency managers to engineers to science communicators), the spatial distribution of forecasted shaking was of top priority, though other types of information could support profession-specific use cases (Schneider, Wein, et al., 2023). In general, it is crucial that the forecasts are put into a context relevant to the stakeholders, and the information is personalized (Becker et al., 2019). Moreover, one should make the purpose of the forecast clear, communicate the limitations, and explain the difference between a forecast and a warning (Freeman et al., 2023). Becker et al. (2020, p. 3343) thus recommend “developing a diversity of audience-relevant OEF information for communication purposes, alongside advice on how such information could be utilized.”

Forecasts should be provided in multiple formats (Becker et al., 2019) and with a *layered approach*, which means providing multiple pieces of information in a layered manner and, thus, allowing users to access the information they are interested in or need for their decision making (M. M. Wood et al., 2012, 2018). Studies have shown that maps were requested most frequently by users (e.g., Becker et al., 2018). If well-designed, they can be an effective way to display the spatial distribution of natural hazards and risks (Stieb et al., 2019; Thompson Clive et al., 2021). So far, only a few studies have analyzed short-term forecast maps (e.g., Becker et al., 2020; Schneider, McDowell, et al., 2022), but best practices from long-term hazard maps can be used as a reference point (e.g., Dallo et al., 2023; Marti et al., 2019; Schneider, Cotton, & Schweizer, 2023). When using tables, it is recommended to put the forecasted number of shocks of a certain magnitude alongside the actual observed number for periods that have already passed because it can reassure information recipients of the accuracy of the

forecast (Becker et al., 2019). Regarding textual information, Doyle et al. (2020) recommend to use the expression “within the next x ” instead of “in the next x ,” to avoid the propensity for people to think the event will happen at the end of this time window. Furthermore, Dryhurst et al. (2022) recommend presenting probabilities of events occurring as both percentages and expected frequencies. They conducted tests on diverse formulations of earthquake probabilities for online platforms, involving the general public across multiple countries. By considering best practices from various fields, they demonstrated that the following formulation enhances the public's comprehension of low percentages: “*With current levels of seismic activity the chance of an earthquake of [insert magnitude or intensity information here] happening in [insert location information here] between [insert dates here] is: $x\%$. Imagine 100,000 places with exactly the same chance of an earthquake as [insert location information here]. In the week of [insert dates here], with an $x\%$ chance, we would expect: An earthquake of [insert magnitude or intensity information here] to happen in y of these places. No earthquake of [insert magnitude or intensity information here] to happen in z of these places.*” Given the implications that such framings can have on people's interpretation and use of forecasts, it is important to continue exploring the best ways of presenting such information.

Studies have also identified various challenges. First, earthquake forecasts can cause anxiety, which can negatively affect people's psychosocial health and well-being (Becker et al., 2019; Schneider, Wein, et al., 2023; Wein, Becker, et al., 2016; Wein, Potter, et al., 2016). Second, earthquake forecasts may negatively influence sectors such as tourism, if people perceive the earthquake risk as too high and delay their activities (Becker et al., 2018). Third, professionals and the general public struggle to interpret probabilistic information and translate it into actions (Becker et al., 2020), though effective methods for visualizing uncertainty have been found (Schneider, Freeman, et al., 2022), and there lies potential in providing advice alongside probabilities to guide people in making decisions (e.g., “*check your preparedness*”; M. M. Wood et al., 2012). Thereby, accounting for individual variations in social background, knowledge, and the associated response capabilities remains a challenge (Mileti & Sorensen, 1990). Fourth, people tend to compare earthquake forecasts with weather forecasts, thus expecting that the precise location and time of a future earthquake can be estimated (Dryhurst et al., 2022). Fifth, the absence of significant damage caused by a large event can lead to a *normalization bias* among non-victims. This can limit their perception of the risk posed by subsequent damaging aftershocks and affect their responsiveness to forecasts (Mileti & O'Brien, 1992). Sixth, given that people have expressed willingness to know and understand the uncertainties associated with OEF (Becker et al., 2018), including stochastic (system variability) and epistemic (lack of knowledge) uncertainty, these warrant further study in terms of understanding how to communicate such concepts in ways that are meaningful to people and support action (Doyle et al., 2019; Hudson-Doyle et al., 2018).

Future research is needed to (a) test if people correctly interpret the forecasts and are able to apply them to practical actions (Doyle et al., 2020), (b) define thresholds, or impact-based contexts, for when it is recommended to take certain actions (Becker et al., 2018), (c) develop interactive systems where each user can set their own thresholds since they are profession- and use-case-specific (Schneider, Wein, et al., 2023), (d) assess users' needs in countries aiming to communicate earthquake forecasts, following the approach of Schneider, Wein, et al. (2023), also enabling cross-culture comparisons (Becker et al., 2018), and (e) evaluate how one could go from earthquake forecasts to earthquake loss forecasts (e.g., Iervolino et al., 2015).

3. Examples of OEF Systems Worldwide

We here provide a comprehensive review of three examples of OEF systems in Italy, New Zealand, and the United States. This encompasses an overview of the model types, tests applied, and communication tools employed by these countries to provide earthquake and aftershock forecasts. The aim is to foster transparency in the interest of the public as well as modelers worldwide who intend to develop OEF systems.

The countries covered here were selected because (a) they are among the few countries that do OEF according to the definition in this article, and (b) they are also represented by participants of the expert elicitation conducted by Mizrahi, Dallo, and Kuratle (2023) and presented in Section 4. For recent advancements in the development of OEF systems in China, we refer to Z. Liu et al. (2023), and to Omi et al. (2019) for the case of Japan. Furthermore, Kamer et al. (2021) and Nandan et al. (2021) describe the non-authoritative “earthquake prediction platform” RichterX, which has provided global earthquake forecasts since 2019 and allows users to submit earthquake predictions.

We divide the relevant issues into three pillars: the development, testing, and communication of earthquake forecasts. At the end of each pillar, a summary is provided in table format. After the description of the OEF systems of the three countries, the last part of this section of the review will describe the concept of Operational Earthquake Loss Forecasting (OELF): how to turn time-variant earthquake probabilities into time-variant estimates of seismic hazard and risk.

3.1. Model Development

3.1.1. Italy

The probabilistic forecasts delivered by the OEF-Italy system are obtained from an ensemble model, which is the weighted combination of three stochastic models typically used in statistical seismology: Epidemic-Type Aftershock Sequence (ETAS) (Lombardi & Marzocchi, 2010a, 2010b), Epidemic-Type Earthquake Sequence (ETES) (Falcone et al., 2010) and Short-Term Earthquake Probability (STEP) (Woessner et al., 2010). Following the score model averaging (SMA) procedure, the weights are assigned depending on the forecasting skill that the single models have shown considering the catalog of the last week, and are updated daily (Marzocchi, Zechar, & Jordan, 2012). The system also stores the forecasts computed by the single components of the ensemble. Nevertheless, they are not considered in the output. The ensemble model is expected to be the most reliable, or never much worse than any (unknown) best-performing model.

More precisely, the ensemble model in OEF-Italy derives from the merging model approach, which accounts for the uncertainties of the components (Vere-Jones, 1995). Necessarily, the assignment of the weight is susceptible to subjectivity, but it still gives a gain in reliability when the selection satisfies the following three properties (Garthwaite & Mubwandarikwa, 2012): dilution (assign smaller weights to the models highly correlated to others), strong dilution (the weight of a model has to be split with a newly added model that is equal to the first one), monotonicity (if a new model is added, the others' weights should not increase). This is compiled by OEF-Italy, where ETAS, ETES and STEP are combined through the score model averaging, according to which the weights are updated every week, on non-overlapping temporal windows, by considering the logarithm of the cumulative spatio-temporal likelihood scored by every model until the starting forecasting time (Marzocchi, Douglas Zechar, & Jordan, 2012; Marzocchi et al., 2014). This procedure guarantees the creation of ensemble forecasts never much worse than those produced by each component.

The models currently embedded in the OEF-Italy system are probabilistic models based on differently parameterized spatiotemporal kernels. Their forecasts are heavily dependent on past seismicity, as are all clustering models; they feed on progressively occurring events, and are particularly uncertain early in a sequence when limited observations poorly constrain the productivity. The magnitudes adopted to compute the probabilistic forecasts are all local magnitudes (M_L), this being the type estimated for the great majority of events recorded by the Italian Seismic Monitoring Room of the Istituto Nazionale di Geofisica e Vulcanologia (National Institute of Geophysics and Volcanology, INGV). Only for the few strongest events a moment magnitude (M_W) is also estimated, as the M_L may saturate. However, to keep the magnitudes homogeneous throughout the entire seismic catalog used to produce forecasts, only the M_L values are considered for all events.

As better specified in a later, the OEF-Italy system operates in real-time at a national scale by delivering weekly forecasts over a grid lattice of 10 km \times 10 km cells covering the entire Italian territory, with no site-specific parametrization. The models have been calibrated during a learning period from April 2005 to April 2009 for the whole national territory, without accounting for any spatio-temporal variation. Productivity is not updated during specific seismic sequences. Incompleteness entailed by strong events is also not automatically corrected in the current version of the system; so far, corrections for incompleteness are applied by hand only immediately after a large earthquake. A first attempt to make this correction applied automatically has been proposed by Stallone and Falcone (2021), but these authors' method has not been introduced in the OEF-Italy system yet.

Although the current version of the OEF-Italy model includes only clustering models, there are no limitations on the number of forecasting models to be added to the ensemble. Indeed, we argue that using models based on different ideas and thoughts is beneficial, and may allow the ensemble model to perform better. The only constraint imposed on the candidate forecasting models is that they must be submitted to one or more CSEP experiments to get an independent and transparent evaluation.

Table 1
Overview of the Time Scales, Observed Phenomena, and Models Used for Earthquake Forecasting in New Zealand

Time scale	Short-term	Medium-term	Long-term
Relevant time window	Hours, days and weeks	Months to years	Decades
Observed phenomenon	Temporal clustering following large earthquakes	Precursory seismicity increase prior to large earthquakes	Long-term spatial clustering
Models used	STEP, ETAS	EEPAS	PPE, NSHM background

3.1.2. New Zealand

In New Zealand, earthquake forecasting occurs on three different time scales: short-term, medium-term and long-term, which are each associated with different observed phenomena and are described by different models as outlined in Table 1. Researchers in New Zealand have long been engaged in the development of earthquake forecasting models, most notably, the Short-Term Earthquake Probability (STEP) model (Gerstenberger et al., 2004, 2005) and the medium-term Every Earthquake a Precursor According to Scale (EEPAS) model (e.g., Rhoades & Evison, 2004). The EEPAS model software package also includes an Epidemic-Type Aftershock Sequence (ETAS) model, and there is a stand-alone ETAS implementation for New Zealand, referred to here as ETAS-Harte (Harte, 2013). New Zealand has a suite of long-term models that contribute to earthquake forecasting (e.g., Gerstenberger et al., 2023) in hybrid forecast models.

The development, testing, and application of earthquake forecasting models is strongly dependent on the availability of earthquake catalog data. Prior to 2012, the earthquake location was not automated and there was often a delay in data processing, leading to delays and gaps in the earthquake catalog, particularly during aftershock sequences. Regional moment magnitudes were introduced in 2007 (Ristau, 2008) and highlighted that local magnitudes for moderate-sized earthquakes systematically overestimated the size of earthquakes compared to moment magnitude (Ristau, 2009).

In 2012, the earthquake location system changed to the automated system SeisComp3, which by default uses Californian attenuation relations for local magnitudes. Magnitudes of $M \geq 5.0$ were found to have a similar bias as previously (Rhoades et al., 2017), and are converted to the expected rate of moment magnitude for hazard estimates (Rhoades & Christophersen, 2017). In contrast, there are many fewer earthquakes in the magnitude range of 3–5, particularly for the central North Island (Harte, 2019). To overcome the differences between the magnitude scales in the different time periods of the catalog, a new magnitude was derived, M_{LNZ20} (Rhoades et al., 2021). M_{LNZ20} uses attenuation relations and a depth dependence to be on average as consistent with the regional moment magnitude as possible. For the revision of the NSHM, a catalog was derived with magnitudes as consistent as possible with moment magnitude for the instrumentally located since 1931 (Christophersen et al., 2022). Work is ongoing to implement M_{LNZ20} into the routine earthquake processing, and to integrate the revised magnitudes into the standard earthquake catalog.

At this stage, the models involved in public earthquake forecasting have their parameters derived prior to 2012. Therefore, there is hesitation to forecast earthquakes of magnitude $M < 5.0$ until the new magnitudes are fully implemented and available in real-time, as well as all model parameters refitted. Below we describe each model in more detail, as well as the hybrid forecasts that combine short-, medium and long-term models for public earthquake forecasting.

3.1.2.1. The STEP Model

The STEP model applied for public forecasting in New Zealand is an aftershock forecasting model purely based on seismicity. It applies the Reasenber and Jones model (Reasenber & Jones, 1989, 1990) with generic parameters for New Zealand (Pollock, 2007) to each earthquake above a chosen cut-off magnitude. Within a region, the earthquake with the highest forecast rate in a chosen time window contributes to the total forecast. Higher-order aftershocks are automatically included when updating the forecast during an ongoing sequence. Each rate contribution is initially circular in space. However, as the sequence evolves and rates from different regions dominate, the spatial pattern can become more complex. The model also has the option to fit a two-segment fault line, or have a simplified fault line supplied, and then spatially decay the rates with one over distance from the fault.

The start of the earthquake catalog is a model parameter, so the user can choose how much historical data to fit. To address incompleteness in an ongoing sequence, the model fits the magnitude of completeness in each run and then fits parameters to a particular earthquake sequence, once there are more than 100 earthquakes within the sequence above the completeness. For large sequences, the parameters can spatially vary between subsets of the sequences, if a minimum of 100 earthquakes per subset are available to fit parameters. As the parameters are refitted, the b -value can change for the whole sequence and for sub-sequences. For the 2016 Kaikōura earthquake sequence, the magnitude of completeness was fixed to 3.95 because there were problems with fitting the initial completeness. Consequently, only generic parameters are available and used for the Kaikōura sequence. The generic forecast overestimates the number of observed earthquakes.

3.1.2.2. ETAS-Harte

A spatially, temporally varying ETAS model is included in the hybrid forecast, based on the extensive work of David Harte (Harte, 2013, 2014, 2015, 2016, 2017, 2018, 2019). The currently used parameters were estimated for the period 1965 January 1 until 2010 December 31 (inclusive), with depths ≤ 40 km and magnitude $M \geq 4.0$. The estimated b -value for the data set is 1.11. Given the uncertainty around the magnitudes, the parameters were not refitted when the seismic processing was changed in 2012. However, to accommodate the significant drop in earthquake numbers, especially in the central North Island, the parameter for the background seismicity was halved (Harte, 2019).

In this version of the ETAS model, the parameter α describing the increase of the number of aftershocks with mainshock magnitude is small ($1.54/\ln(10) = 0.67$) compared to the b -value of 1.11. Therefore, the smaller earthquakes contribute more to the overall sequence than the larger events. Thus, the model is more affected by incompleteness and underestimates the overall seismicity in the early part of the sequence due to the missing small aftershocks in the earthquake catalog. For Kaikōura, the forecast was smaller than the observations for the first 2 days, and then tracked well with the observations (Harte, 2019).

3.1.2.3. The EEPAS Model

The EEPAS model is a well-established medium-term earthquake forecast model that is purely based on seismicity. When fitting the model, earthquake catalog data are considered from a time when the catalog is mostly complete above a minimum magnitude threshold. The model attempts to forecast earthquakes above a target magnitude threshold about two units higher than the minimum magnitude threshold. The contribution of each earthquake is only included 50 days after its occurrence to prevent aftershock clustering from skewing the fitting of the model. For forecasting, EEPAS uses fixed, previously optimized parameters. EEPAS is not a complete model of seismicity since it does not include aftershock clustering, and therefore does not forecast higher-order aftershocks. In applications, aftershocks can be down-weighted, or all earthquakes weighted equally.

The EEPAS model has been applied to a number of regional earthquake catalogs and consistently forecasts major earthquakes better than time-invariant models (Console et al., 2006; Rhoades, 2007; Rhoades & Evison, 2004, 2005, 2006). Development of the model is ongoing (Rastin et al., 2021; Rhoades et al., 2020, 2022). Several new features have yet to be included in the medium-term component of the hybrid used for operational forecasting.

3.1.2.4. Long-Term Models

Long-term earthquake models have been studied extensively in New Zealand in the context of developing the NSHM (Gerstenberger, Bora, et al., 2022; Gerstenberger, Van Dissen, 2022; Rastin, Rhoades, Rollins, Gerstenberger, Christophersen, & Thingbaijam, 2022; Stirling et al., 2012). The long-term component of the current hybrid forecast models is the average of three long-term models: NSHMBG, PPE-SSR and PPE1950, all described in more detail below.

NSHMBG is the background seismicity model from the 2010 update of the New Zealand NSHM (Stirling et al., 2012). It is a smoothed seismicity model with a 50-km Gaussian smoothing kernel and with the Gutenberg-Richter b -value varying between polygonal seismogenic zones. The rates in this model are based on a declustered catalog.

PPE-SSR is a multiplicative hybrid model constructed using the method of Rhoades et al. (2014). This method starts with a baseline earthquake likelihood and modifies it by applying a multiplier to the expected number of

earthquakes in each cell. The multiplier depends on the values of covariates in the same cell. PPE-SSR has a spatially uniform baseline model, a smoothed seismicity covariate based on Proximity to Past Earthquakes (PPE), and a shear strain rate (SSR) covariate computed from the GNSS observations over the period 1991–2011. The development of PPE-SSR is described by Rhoades et al. (2017). It was one of numerous multiplicative hybrids fitted to a period of the New Zealand catalog of earthquakes $M \geq 4.95$ from two smoothed seismicity models and six other gridded covariates. PPE-SSR was the best-performing hybrid in a retrospective test on an independent period of the same catalog. The PPE component of the PPE-SSR model was described by Rhoades and Evison (2004) and is based closely on a model proposed by Jackson and Kagan (1999). It incorporates spatial smoothing of the locations of past earthquakes with an inverse power-law kernel, weighted by earthquake magnitude, and includes a small, spatially uniform background term to allow for surprises. The PPE model rates are designed to forecast all earthquakes, including aftershocks and other clustered activity, above the minimum magnitude threshold of 4.95.

PPE1950 is a version of PPE based on the catalog from 1840 to 1950, comprising both historical earthquakes and the early instrumental catalog up to 1950. The minimum magnitude of completeness for this model is $M_c = 5.95$. For use at lower magnitudes, it is extrapolated down to magnitude 5.0 based on the Gutenberg-Richter frequency-magnitude relation.

3.1.2.5. Hybrid Forecast Models

The Canterbury earthquake sequence, which started with the $M7.1$ Darfield earthquake in September 2010 and included the devastating $M6.2$ Christchurch earthquake in February 2011, significantly changed the expected earthquake hazard for the Canterbury region for the coming decades (Gerstenberger et al., 2014). Therefore, a time-varying seismic hazard model was developed for the rebuilding of Christchurch (Gerstenberger et al., 2014, 2016). The seismicity model is a hybrid, consisting of time-varying models and time-invariant models. The time-varying component is a mixture of STEP, ETAS, EEPAS_OF and EEPAS_IF, the latter two models being versions of EEPAS with equal-weighting and aftershocks down-weighted, respectively. The time-invariant model is a mixture of several different smoothed seismicity models. The model is defined in annual steps up to 50 years on a 0.05-degree spatial grid with magnitude bins in steps of 0.1 from 5.0 to 8.0. For each magnitude bin within each spatial cell for a given year, the hybrid was defined as the maximum of the expected number of earthquakes in the time-invariant and time-varying components. The idea for this “Avmax” combination came from the STEP model (Gerstenberger et al., 2004, 2005). In the first stage of the model development, the models in each category were equally weighted. For the second stage, an expert panel was convened to discuss the model and weight the individual model contribution by structured expert judgment (Gerstenberger et al., 2014, 2016).

By the time of the Kaikōura earthquake response, further consideration had been given to the form of the hybrid and to the models to include in the long-term component. The decision was made to split the time-varying component into short-term and medium-term components, with the Avmax combination being retained, that is, the hybrid was defined as the maximum of the long-term, medium-term and short-term components within each magnitude bin, spatial cell and time step (Gerstenberger et al., 2023). The short-term component was defined as the average of STEP and EEPAS, the medium-term component as the average of EEPAS_OF and EEPAS_IF, and the long-term models as an average of NSHMBG, PPE-SSR and PPE1950, as described above.

A hybrid forecast tool (HFT) has been developed based on the hybrid model adopted for the Kaikōura response (Christophersen et al., 2018). The HFT combines all the software required to produce the various component models into a single system. A simple user interface allows any user to produce forecasts for any future time period (days, weeks, months, or years) for any specified subset of the New Zealand CSEP testing region. Work on further developing HFT is ongoing (Graham et al., 2022).

3.1.2.6. Structured Expert Judgment When Data Are Sparse

Structured expert elicitation is the process of using expert judgments like scientific data (Colson & Cooke, 2017; Cooke, 1991). There are two fundamentally different approaches to combining expert judgments—behavioral (e.g., O’Hagan et al., 2006) and mathematical (e.g., Cooke, 1991). Mathematical methods are generally more objective and auditable. In particular, the Classical Model of Cooke (1991) aims to capture the uncertainty across experts. At the heart of this method is the acknowledgment that it is not reasonable to expect scientific consensus for complex problems. The experts agree with the process, but not necessarily with the outcome of the elicitation.

Additionally, the overall aims of the process applied are to reduce bias in both the input from individual experts and from who is selected as an expert. An integral part of the method is the weighting of the experts' answers to the target questions according to their performance on so-called *seed* or *calibration questions*. These are questions that are similar in nature to the target questions and for which the answers are known to the analyst or will be known within the timeframe of the study (Aspinall, 2010). The experts provide their own uncertainty distribution for their answers to each question. The Classical Method was applied in the second stage of developing the Canterbury Seismic Hazard model (Gerstenberger et al., 2014, 2016) and most recently for weighting the logic tree branches in the revision of the NSHM (Gerstenberger, Bora, et al., 2022; Gerstenberger, Van Dissen, 2022). The Classical Model was also applied in response to the Kaikōura earthquake. Within 10 days of the $M7.8$ earthquake, continuous GPS observations indicated widespread slow slip events (SSE) on the Hikurangi subduction zone margin. Based on the timing of the onset of the SSE, it was assumed that they were probably dynamically triggered by the passing mainshock energy and locally increased stress. The seismological community was concerned about the impact of the SSE on future earthquake occurrence. On November 25th, the Ministry of Civil Defense and Emergency Management (MCDEM) was notified of the concerns and, on November 26th, information about the SSE and the potential for it to impact future earthquake occurrence was disseminated via the GeoNet website (<http://www.geonet.org.nz/news/3VICSAmmLuaYa2sYoue200>).

Following the briefing with MCDEM, it became apparent that there was an expectation of more formal advice from GNS Science on the likelihood of future $M \geq 7.8$ events in central New Zealand, including any potential impact of the ongoing SSE on this likelihood. At this point, GNS began informal discussions with overseas experts on this topic and on possible ways to model this impact. Concurrent with these discussions, GNS began compiling multiple streams of evidence to help inform the development of probabilities of future large earthquakes. Little existing research was available to guide the determination of a quantitative expectation of future events. Using this information as guidance and through an informal and unstructured elicitation process (Gerstenberger et al., 2016), each expert was asked to independently provide their best estimate, and 90% confidence bounds for the probability of a $M \geq 7.8$ within the next year (from 1 December 2016). The region was loosely defined as “the Lower half of North Island and Kaikōura Aftershock zone as presented on the GeoNet webpages today.” Results were collected from each expert within one day of the workshop and combined by averaging the individual results and confidence bounds. The final estimate was a 5% probability (2%–8%) of a $M \geq 7.8$ within the next year (from 1 December 2016). This was the first time such a modeling exercise for SSE had been undertaken anywhere in the world. This initial estimate was updated with another workshop discussion at the Southern California Earthquake Centre annual meeting in September of the following year and a structured expert elicitation workshop at GNS Science in November. In addition to updating the annual probability, the probabilities for the next 10 years of earthquakes of $M7$ and $M7.8$ were elicited. The region of the forecast and the results are still available on the GeoNet website (https://www.geonet.org.nz/earthquake/forecast/central_nz).

3.1.3. United States

The U.S. Geological Survey produces aftershock forecasts using several distinct methodologies: an automatic system used for domestic earthquakes (including earthquakes affecting American Samoa, Guam, the Northern Mariana Islands, Puerto Rico, and the U.S. Virgin Islands), and a manual system used for select international earthquakes in support of humanitarian response. The automated forecast system used for domestic earthquakes is based on the Reasenberg and Jones methodology (Reasenberg & Jones, 1989), and treats all aftershocks as originating from the mainshock. The forecast is a Bayesian solution using a generic model that describes the range of behavior seen in past sequences as a prior. This results in a forecast that is maximally broad at time zero, and that narrows in on the sequence-specific behavior as time elapses and data accumulate. Bayesian updates to the model occur according to a predetermined update schedule. At this time, the domestic forecasts only consider Bayesian updating of the productivity parameter a . The c and p parameters of Omori's law are kept fixed to generic point estimates until an analyst determines that the sequence is deviating sufficiently from the generic model, at which time one or both of the c and p parameters is freed to take on the sequence-specific value (Michael et al., 2020). The generic or prior models are specific to one of about a dozen unique tectonic environments (Page et al., 2016) or specific regions (Hardebeck et al., 2019). Epistemic model uncertainty is included by integrating over the Bayesian posterior distributions. Short-term aftershock incompleteness is modeled using a time-varying magnitude of completeness, $M_c(t)$, using an empirical function of the mainshock magnitude from Helmstetter

et al. (2006) for regional networks and with parameters trained for teleseismic data on the same global or regional data sets as the global prior/generic model (Page et al., 2016).

The manual system used for international response and particularly complex domestic sequences is based on the Epidemic-Type Aftershock Sequence (ETAS) model (Ogata, 1988), including higher-order aftershocks. The manual system uses a fully Bayesian model where all parameters of the generic model are characterized as distributions along with their covariance, allowing the model to smoothly transition from past behavior to the current sequence as the data require. The ETAS forecast is based on 10,000 or more stochastic event sets, each of which is generated using a different set of parameters sampled from the multivariate posterior distribution (van der Elst et al., 2022). Time-dependent catalog completeness is modeled using a fast approximation in which the Omori c -parameter is a function of mainshock magnitude (de Arcangelis et al., 2018; Hainzl, 2016b; Lippiello et al., 2019). Both the domestic and international forecast systems allow for specification of the b -value—for the international system, this is through a Bayesian framework—but time variations in the b -value are not considered. The Reasenberg and Jones and ETAS models are formulated such that the productivity term and the magnitude of the parent earthquake appear in summation. In updating the sequence specific productivity, the model thus takes into account both aleatory variability in the productivity of each earthquake as well as epistemic uncertainty in the estimate of its magnitude. The prior distribution on productivity models the full range of these effects as well, as it describes the full range of variability observed empirically in past data. In the ETAS model, potential uncertainty in the mainshock magnitude is addressed by adjusting the mainshock (primary) productivity as an additional parameter, independent of the secondary productivity, but constrained by the same prior. Mathematically, this is no different than a linear adjustment to the magnitude, thus allowing the model to adjust for potential real differences in primary and secondary productivity, as well as differences in how the mainshock and aftershock magnitudes may be calculated. Neither the domestic nor the international system include a forecast of the background component of seismicity. In very active regions, the aftershock forecasts may therefore give probabilities that underestimate the true hazard at later times in the sequence.

The international forecasting system also includes the possibility of generating forecast maps. The spatial distribution of future seismicity is modeled using an empirical spatial kernel developed for California earthquakes that is consistent with some physical expectations for rupture length and stress decay away from the rupture (van der Elst & Shaw, 2015). The spatial kernel is used to generate a gridded rate model that is then fed into the OpenSHA software platform (Field et al., 2003). Ground motion prediction equations translate the gridded rate into probabilistic estimates of shaking. The shaking can be expressed in terms of peak ground velocity (PGV), peak ground acceleration (PGA), peak spectral acceleration (PSA), or MMI. It can be mapped as either the shaking level with a given probability of exceedance or the probability of exceeding a given level of shaking. Based on previous discussions with a range of stakeholders (Schneider, Wein, et al., 2023), the default forecast map is set to show the probability of exceeding MMI level VI, which is the intensity associated with the onset of light damage in high-quality structures.

The USGS also has a third model that is available only in California, the UCERF3-ETAS model, which combines the ETAS model with a fault-based background rate model and adds elastic rebound effects (Field et al., 2017). This model can be run with or without modeled faults, and is consistent with the long-term National Seismic Hazard Map within California. UCERF3-ETAS is currently only run on-demand, for example, following large or notable earthquakes (e.g., Milner et al., 2020), as it requires high-performance computing and takes several hours to run. It can be used to create stochastic event sets, spatial maps of exceedance probabilities, as well as expected loss calculations. UCERF3-ETAS is unique in that it can also give fault-based probabilities, for example, the nucleation or participation probability of a particular fault segment, as well as probability gains relative to the long-term hazard. Also, in contrast to the domestic and international aftershock forecast products, the UCERF3-ETAS model includes background earthquakes, not just aftershocks, and can therefore be used to create realistic stochastic catalogs over very long time intervals (e.g., decades). So far, this model has primarily been used to inform additional discussion issued with forecasts calculated with other methods.

Finally, the USGS has used a temporal ETAS model with a time-varying background rate to forecast the behavior of swarms (e.g., McBride et al., 2020). While the aftershock forecasts cover time periods as long as a year, the swarm forecasts tend to focus on the immediate days and weeks, due to the difficulty in forecasting swarm duration. In certain regions, including the Salton Trough and San Jacinto Fault zones, the USGS uses an 'actuarial'

model in which the duration is treated as a random variable drawn from an empirical distribution for the region (Llenos & van der Elst, 2019).

3.1.4. Model Development Summary Table

Table 2 summarizes the information provided on the forecasting models used for OEF in Italy, New Zealand, and the United States.

3.2. Model Testing

3.2.1. Italy

The testing of models is an essential prerequisite for any candidate model to become part of the ensemble used in OEF-Italy. Taroni et al. (2018) show the results of prospective CSEP tests, for example, the N-test, for one-day earthquake forecasting models. Results show that for one-day experiments, the single models' forecasts and observations are consistent in number and magnitudes and, in the case of ETAS and ensemble forecasts, also spatially. Other tests carried out outside the CSEP umbrella, but using similar tests, show similar results (Marzocchi & Lombardi, 2009; Marzocchi, Murru, et al., 2012).

Besides the tests on each single model, the forecasts produced by the ensemble, that is, the released OEF-Italy forecasts, were also tested. The tests have been carried out both on single seismic sequences, such as the Central Italy earthquake sequence 2016–2017 (Marzocchi et al., 2017), and on all the forecasts made in the last decade (Spassiani et al., 2023). In this latter case, the authors used additional performance measures borrowed from other research fields, like meteorology, specific to validate alarm-based systems by a binary criterion (forecast: yes/no; occurrence: yes/no, Spassiani et al., 2023). Although, to date, the OEF-Italy system does not explicitly include alarm-based components, the probabilistic forecasts it delivers can be re-interpreted in terms of these measures, which are specific for extreme (rare) event probabilistic forecasting.

Specifically, Spassiani et al. (2023) evaluated the forecasts of $M_L \geq 4$ earthquakes from 2013 to 2020. The N-test is revised to account for the fact that OEF-Italy forecasts do overlap in time, since they refer to the next week, but are delivered at least every midnight; this test allows to evaluate the consistency between the number of forecasted events in all the space-time-magnitude bins of analysis, and the number of observed $M_L \geq 4$ events in a testing spatiotemporal window of interest (Figure 4). Performance diagrams (Molchan, Reliability) are also considered, together with the verification measures obtained from different combinations of the contingency variables True/False Negatives/Positives. Results highlight a good performance skill of the OEF-Italy system for almost all the analyses performed, except for a period during the Central Italy sequence (2016–2017), which entailed a strong incompleteness that induced an underestimation of the expected seismicity (see also Marzocchi et al., 2017).

The codes behind the OEF-Italy system are currently available only to authorized parties. All the forecasts produced from 2009 are stored and archived in a database internal to INGV, both for the ensemble model and its components.

3.2.2. New Zealand

Much effort has gone into testing both the individual models that are included in the operational hybrid and various hybrids of them, but not into testing the public forecasts themselves. The public forecasts cover a variety of time periods ranging from 24 hr to 100 years and the time periods of successive forecasts often overlap. These features make formal testing challenging. For shorter time periods, the number of observed and expected earthquakes above the minimum target magnitude threshold of 5.0 is rather small, so statistical tests tend to have low power. Aftershock models are better tested at lower magnitude thresholds, say, 4.0. Tests of the longer-term forecasts have greater power at the magnitude 5.0 threshold, but the time periods need to elapse before those tests can be carried out.

STEP and a simple ETAS model have been prospectively tested with a magnitude threshold of 4.0 and one-day updating by CSEP in the New Zealand Earthquake Forecast Testing Centre (Gerstenberger & Rhoades, 2010). Tests of the ETAS-Harte model during the Kaikoura aftershock sequence were presented by Harte (2019), also with a magnitude threshold of 4.0. The two versions of EEPAS used in the hybrid forecast tool and PPE have been

Table 2
Summary of the Model Specifications for the Models Used in Italy, New Zealand, and the United States

Base model(s)	Italy	New Zealand				United States domestic	United States international	United States California UCERF3-ETAS
	ETAS, ETES, STEP	STEP	ETAS	EEPAS	Long-term	Reasenber and Jones	ETAS	ETAS, long-term
Ensemble	Weighted average based on past forecasting skill (SMA); weights updated every week	Maximum of the three model classes short-, medium-, and long-term, which each are an average of different models (as described in the text)				No	No	No
Background seismicity?	Yes	No in HFT, but yes for standalone use	Yes	Yes	Yes	No	No	Yes
Higher-order aftershocks	Yes (aftershocks of observed aftershocks)	Yes (aftershocks of observed aftershocks)	Yes	No	No	No	Yes (aftershocks of observed and possible future aftershocks)	Yes (aftershocks of observed and possible future aftershocks)
Model updating	No	Yes, sequence-specific	No, parameters calibrated prior to 2012	No	No	Yes, Bayesian updating of productivity and sequence-specific Omori parameters	Yes, Bayesian updating of all parameters	No
Underlying data for model parameters	Past seismicity (2005–2009)	Sequence seismicity	Past seismicity	Past seismicity	past seismicity, shear strain rate	Global and local past seismicity	Global past seismicity, California seismicity for spatial kernel	Past seismicity, faults
Anisotropic aftershock triggering?	No	option to fit a two-segment fault line	No	N/A	N/A	N/A	Yes (primary aftershocks only)	Yes, along known faults
Is catalog incompleteness addressed? How?	Manual correction immediately after large earthquake	M_c estimated for ongoing sequence	No (this led to under-forecasting for Kaikōura in the first 1.5 days after the $M7.8$ event)	No	No	Empirical $M_c(t)$ depending on mainshock magnitude	Modeled with magnitude-dependent Omori c parameter	No
b -value variations	No	No, but b -value can be updated during ongoing sequence	No, set to 1.11	No	Yes	b -value can be specified as a parameter	continuous Bayesian updating with other parameters	No
Epistemic model uncertainty	No	No	No	No	No	Yes	Yes	Yes
Number of simulations	N/A	N/A	2000 in HFT, but can be specified by user	N/A	N/A	N/A	10'000+	10'000+
Other						This component is itself an ensemble		Elastic rebound effects

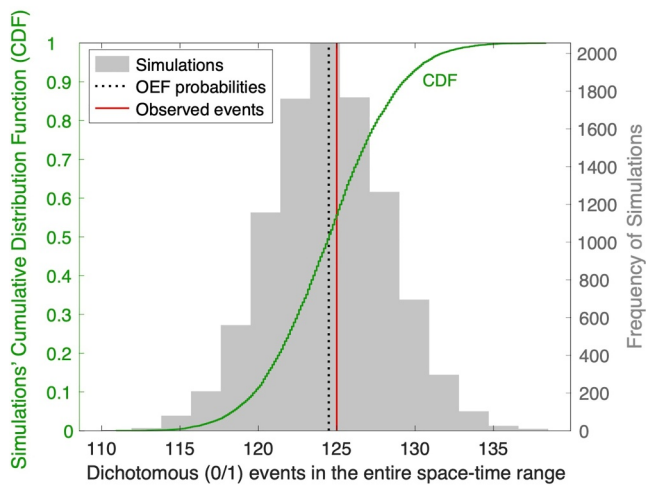


Figure 4. Cumulative Distribution Function (green step line) and frequency (gray bars) of the number of spatio-temporal OEF bins with $\geq 1 M_L \geq 4.0$ events, among 10,000 synthetic dichotomous (0/1) catalog observations, obtained from the ensemble OEF-Italy probabilities. The dotted vertical line represents the sum of the overall ensemble OEF-Italy probabilities. The continuous vertical red line represents the number of space-time bins with at least one $M_L \geq 4.0+$ event (Adapted from Spassiani et al. (2023)).

tested in the New Zealand Earthquake Forecast Testing Centre with a magnitude threshold of 5.0 and three-monthly updating. The results of 10 years of CSEP testing in New Zealand were given by Rhoades et al. (2018).

The Canterbury hybrid model and all its individual model components were tested in a special retrospective CSEP experiment using the whole New Zealand test region with 1-year updating and time lags ranging from zero to 25 years (Rhoades et al., 2016). That experiment was one of several that have demonstrated the good performance of hybrids of various forms relative to individual forecasting models (see also Rhoades & Gerstenberger, 2009; Rhoades & Stirling, 2012; Rhoades et al., 2014, 2015, 2017). The choice of contributing models and setting of mixing parameters in the hybrid model adopted for Kaikoura and the Hybrid Forecasting Tool (HFT) was informed by retrospective optimization of the mixing parameters in each of the model classes (short-term, medium-term and long-term) of the Avmax model (maximum of the three model classes short-, medium- and long-term, which each are an average of different models) with time-lags of up to 25 years (Gerstenberger et al., 2023).

The form of hybrid models, following both the Canterbury and Kaikōura earthquakes, was chosen subjectively by experts but informed by all available testing information on the individual components and hybrids of them. Future changes to the model are likely to follow the same path. In this way, it is expected that demonstrated model improvements in short-term, medium-term and long-term forecast modeling will be incorporated into the hybrid used for public forecasting.

For some model components, for example, STEP, the source code is freely available. For others, for example, EEPAS, the executable code can be freely licensed to other users. Making the source code more freely available is currently under consideration.

3.2.3. United States

The Reasenber and Jones forecasts have been compared to observations for the 2018 $M7.1$ Anchorage earthquake (Michael et al., 2020) but have not been comprehensively tested, and the USGS is unlikely to do so, because of the known shortcomings of the Reasenber and Jones approach. Retrospective tests show that the Reasenber and Jones model performs fairly well at capturing the probability of having at least one earthquake of a given magnitude, but fails to adequately predict the 95% confidence range on the aftershock numbers, because secondary aftershocks are ignored (van der Elst & Page, 2017; van der Elst et al., 2022). The ETAS model used for the international forecasts has been tested prospectively for the domestic 2020 SW Puerto Rico earthquake sequence, where the model has performed well over three years of operation (van der Elst et al., 2022). All of the USGS public aftershock forecasts are archived in the ANSS Comprehensive Earthquake Catalog (ComCat) (USGS, 2017) and are available for prospective testing. The code for both the automatic and manual systems is publicly available (<https://github.com/opensha/opensha-oaf>).

An R package (R Core Team, 2021) called OAFtools (Paris & Michael, 2022a) allows users to interactively view the forecasts stored in ComCat, and is used internally by the USGS for prospective testing and forecast evaluation. The functions that make up the interactive viewer can also be used to extract the forecasts from ComCat into an R dataframe for additional analysis and testing. The software can be downloaded from code.usgs.gov via <https://doi.org/10.5066/P9PZTYEN> (Paris & Michael, 2022b).

The UCERF-ETAS model, available only in California, has been tested retrospectively in several ways. “Turing tests” have compared the statistical features of the synthetic catalogs relative to the observed catalog (Page & van der Elst, 2018). The model performance in the 2019 Ridgecrest earthquake sequence was also evaluated using CSEP tests, although this was a pseudo-prospective evaluation (Savran et al., 2020).

3.2.4. Testing Summary Table

Table 3 summarizes the information provided on the testing of OEF models in Italy, New Zealand, and the United States.

3.3. Communication

3.3.1. Italy

Using OEF has been proposed in Italy by the International Commission for Earthquake Forecasting (ICEF) for Civil Protection, nominated by the Italian government after the M_W 6.3 earthquake occurred in L'Aquila (Italy) on 6 April 2009, which caused the loss of about 300 lives (Jordan et al., 2011; Marzocchi et al., 2014). Before operating the OEF-system in Italy, the INGV team had already produced forecasts immediately after the L'Aquila earthquake and subsequently for the seismic crisis in Emilia in 2012. The forecasts were released each day during the first period of the seismic activity and weekly when the seismicity began to decrease. The results were released for internal use and for the Civil Protection Department. At that stage, the forecasts were produced using two ETAS models, sometimes providing slightly different results. These differences were one of the reasons motivating the use of an ensemble model in which the weighted contribution is related to the daily performances of individual models.

Since 2013, the Seismic Hazard Center produced the first version of the OEF-Italy system that has been designed and developed to run in real-time 24/7 on a computer physically placed at the INGV in Rome (Marzocchi et al., 2014). At midnight of each day, and after the occurrence of every $M_L \geq 3.5$ event recorded in real-time by the Italian Seismic Network, the OEF-Italy system produces the next week's probabilistic forecast of *target* earthquakes, that is, events with local magnitude $M_L \geq 4.0$, $M_L \geq 5.5$, or with microseismic Modified Mercalli Intensity $MMI \geq VI$, $\geq VII$, $\geq VIII$, over a specific $0.1^\circ \times 0.1^\circ$ grid lattice covering the whole national territory. Specifically, the lattice is placed inside a polygon opportunely selected for Italy according to the standards of the Collaboratory for the Study of Earthquake Predictability (CSEP, https://scec.usc.edu/scecpedia/CSEP_Working_Group; Schorlemmer et al., 2018). Sardinia is not included in the analysis since instrumentally recorded seismicity on this island was insufficient to calibrate the model. Volcanic areas (such as around the Etna volcano, Sicily) are also excluded because the seismic activity beneath active volcanic regions is driven by different mechanisms (e.g., magma intrusions), which are not well-captured by the parameterization models behind OEF-Italy.

The choice of weekly forecasts has been agreed upon with the Italian Civil Protection for practical reasons, but the system can provide shorter-term forecasts (e.g., daily). The main outputs delivered by the system are earthquake rates in each grid cell, and the relative time-dependent probability maps.

The OEF-Italy system is equipped with a graphical interface, as shown in Figure 5 (Marzocchi et al., 2014). It consists of an embedded Leaflet Map reflecting the current weekly probability of the target events inside a spatial window (circle or rectangle), directly selected by the user from the interactive dashboard. The user can also fix the specific M_L or MMI threshold to identify the target events among those available. An additional interactive graph shows the temporal evolution of the weekly probabilities, which can be zoomed on, at any time interval of interest. The probability value of the last run produced by the system is also shown.

To date, the OEF-Italy system is not open to the public and is accessed only by authorized personnel. Since 2015, the forecasts have been released in a structured manner to the Major Risks Commission of the Italian Civil Protection in 4-month maps for the quarterly discussion of seismicity with the Major Risks Commission, and in daily reports during important seismic sequences, or upon specific request.

Recently, the OEF-Italy information has been officially released through the Department of Civil Protection to the Italian regional governments to handle seismic sequences in their territory (<https://www.protezionecivile.gov.it/en/notizia/short-term-seismic-hazard-meeting-ingv-and-regions>). The system is running in real-time, and the forecasts delivered by every run for each cell in the grid lattice are progressively stored in separate files for both the single and the ensemble models, thus enabling to continuously access the flow of information produced.

The communication of the OEF-Italy forecasts has been partially tested with the public (Savadori et al., 2022). The possibility of spreading results to the population and the best way to disseminate them is being discussed in periodic meetings with the Italian Civil Protection; presently the OEF information is released only to the Major

Table 3

Summary of the Testing of OEF Models in Italy, New Zealand, and the United States (Domestic, International, and for California)

	Italy	New Zealand	United States domestic	United States international	United States California UCERF3-ETAS
How was communication tested?	Periodic meetings with Italian Civil Protection, tests with public	Co-developed with user groups and support from social scientists	Elicitation of user preferences through formal and informal interactions For public user groups: media engagement evaluation, citizen surveys during Puerto Rico sequence		
How were the models tested?	Prospective CSEP experiment; tests of individual models and ensemble; additional performance measures from meteorology specific for alarm-based systems	Extensive prospective and retrospective testing of individual models and hybrids, also in official CSEP experiments	Retrospectively, prospectively for select sequences (2018 Anchorage)	Prospectively for select sequences (2020 SW Puerto Rico)	Retrospectively using “Turing tests,” CSEP tests
Are operationally issued forecasts tested?	Yes, systematically for all forecasts, and in addition for specific sequences	Not systematically	Not systematically	Yes, manually	Not systematically
Are forecasts archived?	Yes	Yes, but not in a well-structured archive	Yes	Yes	Not routinely
Are archived forecasts public?	No	News stories with forecast information continue to be online and are searchable; tables and figures are archived but not visible to the public at this stage	Yes (ANSS ComCat)	Yes (ANSS ComCat)	No
Are codes available?	Only to authorized personnel	STEP: within California and New Zealand CSEP testing centers, EEPAS: compiled code can be freely licensed	Yes: https://github.com/opensha/opensha-oaf (models), https://code.usgs.gov/esc/oaf/oaftools-R/-/tree/1.0.0 (forecast analysis)		

Risk Commission and to local governments. INGV officials have also given important food for thought and feedback about this delicate question. Major obstacles come out from a legal system that is unclear on the roles and responsibilities of scientists involved in delivering this information. This issue still has easy-to-predict major consequences in Italy after the infamous L'Aquila earthquake trial (Marzocchi, 2012), which indeed had a huge impact on seismologists and decision-makers, who now require legal protection before communicating any sort of public statement, especially of forecasting type. This highlights the need to recognize, in all sciences in which hazard is involved, that unlikely events can and will actually occur.

3.3.2. New Zealand

The current core products in an earthquake response in New Zealand are a table with expected rates, ranges and probabilities of earthquakes in different magnitude bands and time intervals (e.g., Table 4), a map of the forecast area (e.g., Figure 6), and maps of the probability of expected shaking (e.g., Figure 7). If possible, the rates are compared to the long-term rate from the NSHM. This quantitative material is usually accompanied by descriptive scenarios of what can be expected to happen (e.g., Table 5).

The forecast products have been developed in collaboration with user groups and with social scientists. For example, workshops with different user groups were held, including lifeline providers, engineers, emergency responders and the public, to test and further develop the products (Becker et al., 2018). Focus group discussions

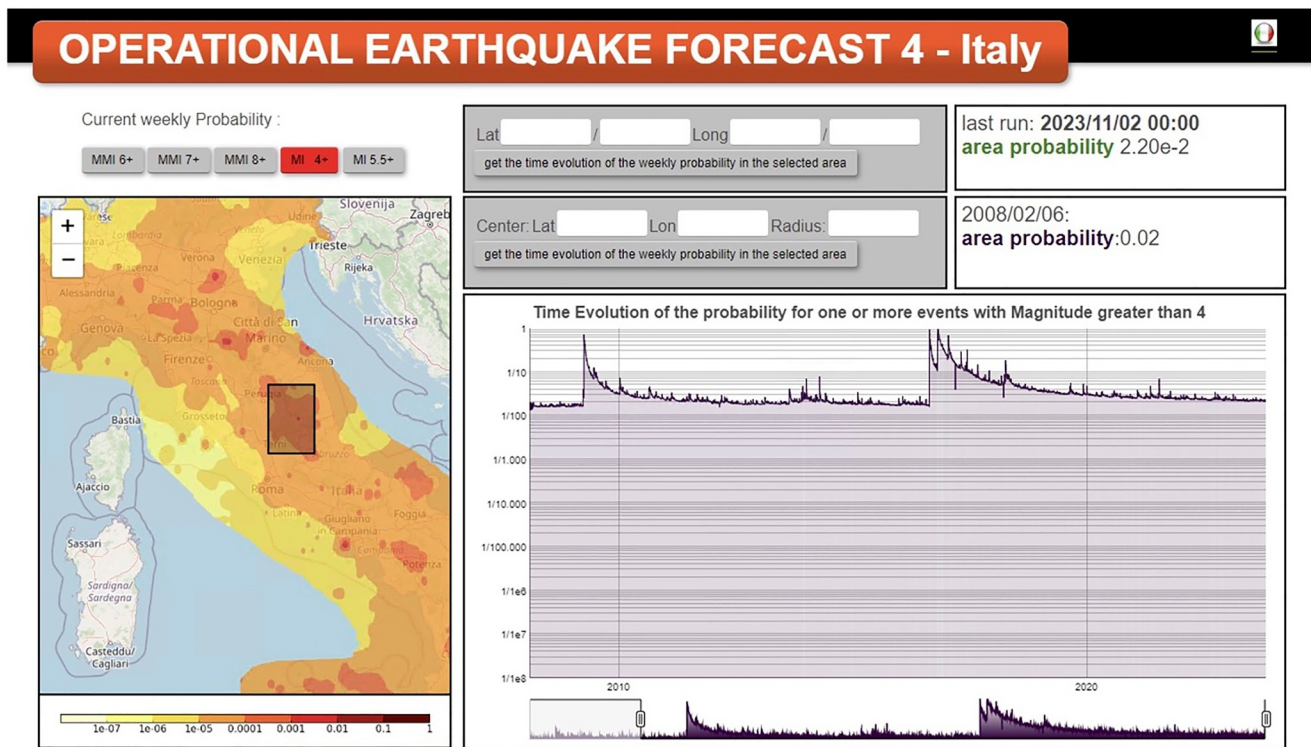


Figure 5. Graphical interface of the OEF-Italy system, example for a small rectangular area in Central Italy (black box in the map). On the left: embedded Leaflet Map of the current weekly probability for the selected area, where $M_L \geq 4.0$ events are selected as target. On the right: timeline of the probability history from January 2009 to January 2021 (bottom), and two boxes (top right) showing the probability of the last run and the probability computed at the date range selected along the timeline.

and surveys following earthquake responses also helped improve the forecast products (Becker et al., 2019, 2020; Wein, Becker, et al., 2016; Wein, Potter, et al., 2016). Additionally, forecasts and products have been developed collaboratively with users that targeted their specific and unique needs at that time. An example is a spatial forecast that highlighted shaking probability gains over building code requirements that informed decisions on mandatory retrofit requirements following the Kaikoura earthquake. An overview of different target audiences in New Zealand, along with potential applications of earthquake forecasts spanning different time frames, is provided in Figure 8.

The first recipients of the forecasts are employees within GNS Science, in particular duty officers and seismologists, for their situational awareness. The forecasts are then communicated to the National Emergency Management Agency (NEMA) before being published on the GeoNet website. GeoNet also uses social media, to communicate scientific information. Information about the forecasts themselves is provided, as well as empathetic messages and links to preparedness messages.

There is no set schedule for the release and the updating of the forecasts. The frequency of updating, as well as the time windows for the forecast, are adjusted as the sequence evolves. Initially, a forecast is provided for 24 hr, 7 and 30 days, and is updated at least daily. For smaller sequences, such as for example, the $M5.6$ Taupo earthquake of 30 November 2022, no daily forecasts were provided, and there were only two weekly updates, which were communicated as text in a news story rather than in the tables and with maps of expected shaking. The daily forecast is dropped as the expected rate of earthquakes diminishes. The forecast length is adjusted to the frequency of updating. At the time of monthly updating, a forecast for the next year is usually included.

Of crucial importance during a response is to connect directly with the users of the information. In response to a major earthquake, such as the 2016 Kaikoura earthquake, the National Crisis Management Centre (NCMC) is activated. A GNS seismologist is stationed at the NCMC to communicate what has happened and what is most likely to happen next, including the forecasts. GNS Science has good connections to the earthquake engineering community and also attends regular meetings to understand the information requirements and provide targeted

Table 4

Example of a Forecasting Table for the Kaikōura Earthquake as of 12:00 Noon NZDT 28 November 2016, Which Is About Two Weeks Following the M7.8 Earthquake

	Average number of M5.0–5.9	Range ^a of M5.0–5.9	Probability of 1 or more M5.0–5.9 (%)	Average number of M6.0–6.9	Range ^a of M6.0–6.9	Probability of 1 or more M6.0–6.9 (%)	Average number of M ≥ 7	Range ^a of M ≥ 7	Probability of 1 or more M ≥ 7 (%)
within 7 days	5.6	1–13	98	0.53	0–2	41	0.05	0–1	5
within 30 days	15.7	6–28	>99	1.5	0–4	77	0.15	0–1	14
within 365 days	44.2	27–64	>99	4.1	1–8	98	0.39	0–2	32

Note. The forecast was published on GeoNet with the following additional information: Forecast for rectangular box with the coordinates $-40.7, 171.7, -43.5, 171.7, -43.5, 175.5, -40.7, 175.5$ at 12 noon, Monday, 28 November. The table shows that, for example, there is a 41% chance of one or more M6.0–6.9 earthquakes occurring within the next week. Between 0 and 2 earthquakes in this magnitude range were estimated within the next week. The current rate of magnitude 6 and above for the next month is about 50 times larger than what would normally be expected for long-term seismicity represented in the National Seismic Hazard model. As the aftershock rates decrease, this difference will decrease as well. ^a95% confidence bounds.

information. GNS Science seismologists also engage in outreach activities such as talks to the public and school visits to communicate earthquake forecasts.

The communication of the forecast has not been as extensively tested as the individual models but was co-developed with the user groups and with the support of social scientists (Becker et al., 2018, 2019, 2020; Wein, Becker, et al., 2016; Wein, Potter, et al., 2016). The public forecasts are kept, but not in a well-structured archive. This is an area for improvement for transparency as well as ease of future retrospective testing.

3.3.3. United States

The United States (US) forecasts have been developed in partnership with stakeholders and social scientists, including communication specialists, through both informal meetings and formal workshops designed to elicit user needs (e.g., Field et al., 2016). Stakeholders have included emergency response organizations including FEMA, the California Office of Emergency Services, the California Department of Transportation, utility and lifeline operators, public transportation, civil engineers, and building inspectors, among others. The U.S. Geological Survey (USGS) forecast communication strategy developed largely out of the work done in New Zealand during and after the Canterbury sequence (Becker et al., 2019, 2020). This work emphasized the importance of communicating the forecast in both qualitative and quantitative terms, employing empathetic messaging, and including preparedness and action recommendations for the public.

The USGS issues forecasts automatically following magnitude 5 and larger domestic earthquakes, using a variation of the Reasenber and Jones (1989) methodology as described in Page et al. (2016) and Michael et al. (2020). Forecasts are not automatically issued in Hawaii due to concerns that volcanic processes undermine the validity of the methodology. In Hawaii, forecasts may be triggered manually after consultation with the USGS Hawaii Volcano Observatory to confirm that the sequence is tectonic. The domestic forecasts are publicly released on the mainshock's event page within the USGS website earthquake.usgs.gov 20 min after the mainshock. The 20-min delay provides time to obtain a stable estimate of the mainshock magnitude. The forecast is updated frequently at first, and then intermittently, following a predetermined schedule available on the website.

Forecast information is displayed graphically in several ways. An interactive forecast summary tab allows users to view the probability of one or more aftershocks and the uncertainty distribution for the expected number of aftershocks for magnitude thresholds and forecast durations that they choose (day, week, month, and year) and aftershock magnitude (magnitudes 3 through 7), as well as bar graphs and tables summarizing this information (see Figure 9).

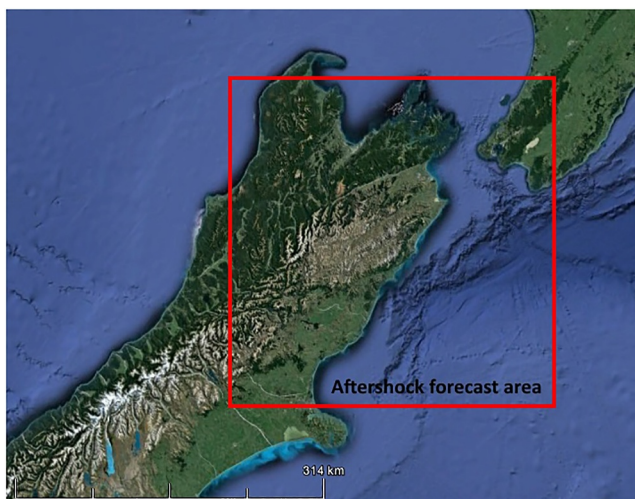
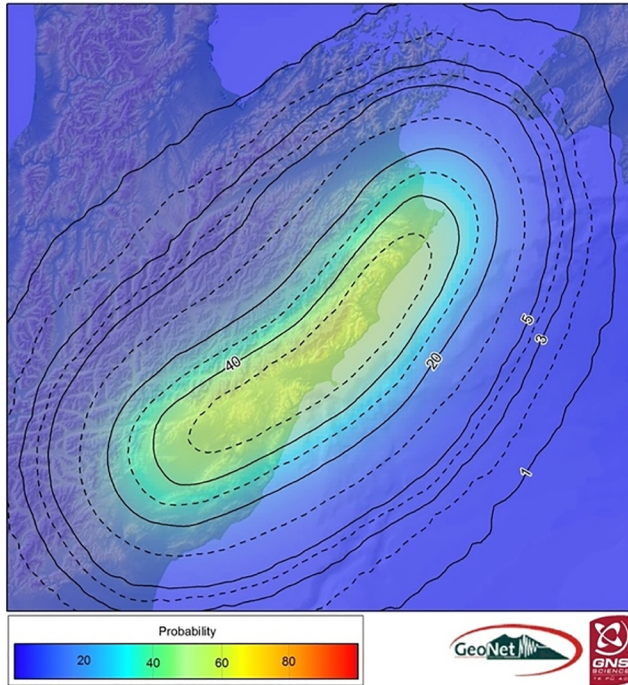


Figure 6. A simple representation of the aftershock forecast area for the Kaikōura earthquake as published on GeoNet in conjunction with the type of information shown in Table 4. The simplicity of the figure illustrates that in the wake of this major earthquake no further identifying features like longitude and latitude or place names were deemed necessary to inform the users of the map on its location.

Probability of damaging shaking (MM7) in the next 30 days

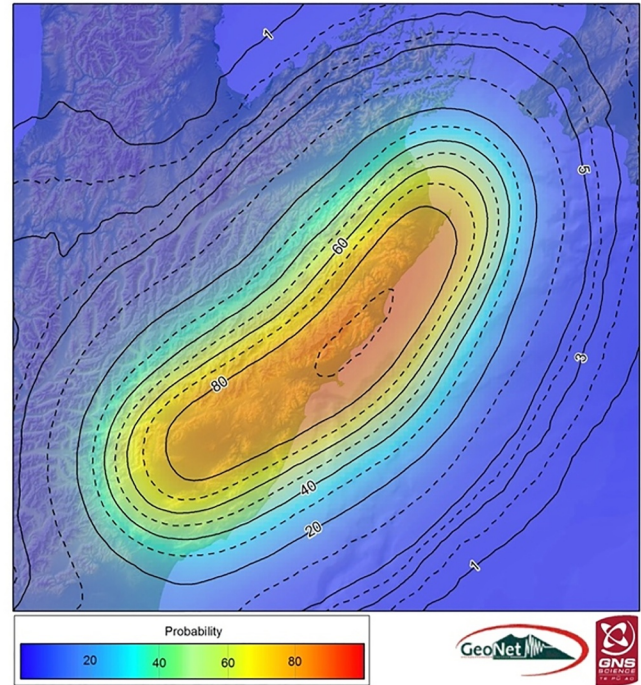
As at 28/11/2016



MM7 shaking corresponds with internal building damage, structural damage to a few weak buildings, and will be alarming to affected people

Probability of damaging shaking (MM7) in the next year

As at 28/11/2016



MM7 shaking corresponds with internal building damage, structural damage to a few weak buildings, and will be alarming to affected people

Figure 7. An example of maps of expected shaking for Modified Mercalli Intensity for MM7 for 30 days and one year from 28 November 2016, two weeks after the mainshock; the forecast is the full hybrid model described in the model subsection. The figures are an exact copy of what was published on GeoNet at the time (used with permission).

Forecast information is also given in text format in the commentary tab. The forecast commentary includes a narrative that employs empathetic messaging around preparedness and action recommendations, tailored for laypersons. The template is arranged in a tiered manner to present the simplest information first (e.g., “*be prepared for more earthquakes*”) and then proceeds to present more detailed information for users that read further.

Table 5

The Scenarios Produced and Published in a News Story on 26 January 2017, 2.5 Months Following the Kaikōura Earthquake to Cover One Year Rather Than 30 days as Initial Scenarios

Scenario One: Likely (approximately 70% within the next year)

The most likely scenario is that aftershocks will continue to decrease in frequency (and in line with forecasts) over the next year and no aftershocks of magnitude 7 or larger will occur. Felt aftershocks (e.g. over magnitude 5) will occur in the area from North Canterbury to Cape Palliser/Wellington. It's very likely (98% within the next year) that there will be at least one aftershock of magnitude 6.0–6.9 in the next year, regardless of there being a larger (magnitude 7.0+) earthquake

Scenario Two: Unlikely (approximately 25% within the next year)

An earthquake smaller than the mainshock and between magnitude 7.0 to magnitude 7.8 will occur. There are numerous mapped faults in the North Canterbury, Marlborough, Cook Strait and Southern North Island areas capable of such an earthquake. It may also occur on an unmapped fault. This earthquake may be onshore or offshore but close enough to cause severe shaking on land. This scenario includes the possibility of an earthquake in the Hikurangi Subduction Zone. Earthquakes originating from here or in the Cook Strait have the potential to generate localized tsunamis. The Hawke's Bay earthquake sequence in 1931 provides an analogy to scenario two, as a magnitude 7.3 aftershock occurred approximately 2 weeks after the initial magnitude 7.8 earthquake

Scenario Three: Very unlikely (5% within the next year)

A much less likely scenario than the previous two scenarios is that recent earthquake activity will trigger an earthquake larger than the magnitude 7.8 mainshock. This includes the possibility for an earthquake of greater than magnitude 8.0, which could be on the plate interface (where the Pacific Plate meets the Australian Plate). Although it is still very unlikely, the chances of this occurring have increased since before the magnitude 7.8 earthquake, and have also been slightly increased by the slow-slip events

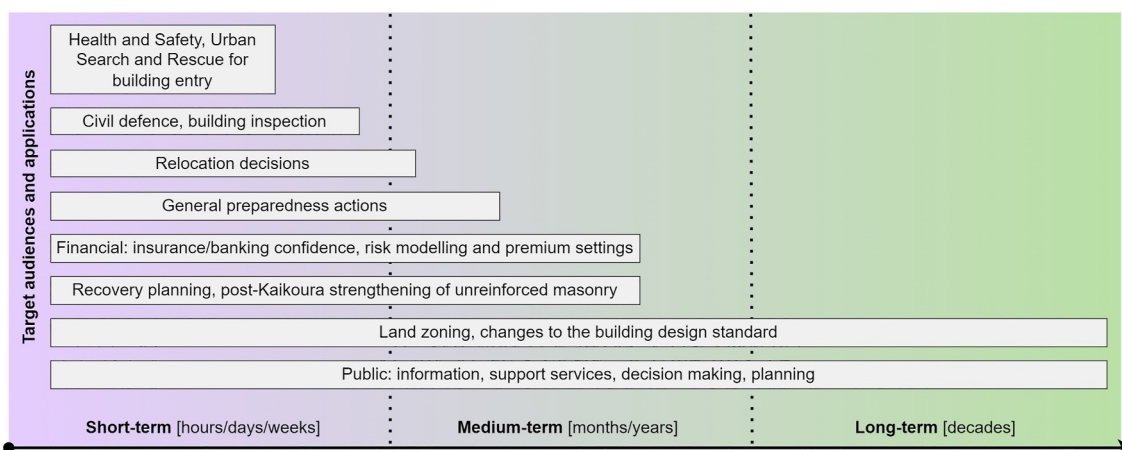


Figure 8. Schematic overview of applications of earthquake forecasting on different time scales in New Zealand.

Further information about the scientific background of the forecast models is available through links. As necessary, the USGS provides additional information about a sequence through the commentary section and/or within “USGS Top Stories” about significant sequences on the overall USGS website: www.usgs.gov.

For international earthquakes, or for sufficiently complex domestic earthquakes, the USGS deploys a manual forecast system based on the Epidemic-Type Aftershock Sequence model, operated through a Graphical User Interface (van der Elst et al., 2022). The international forecast methodology and communication products have been developed in partnership with the U.S. Agency for International Development—Bureau of Humanitarian Assistance (BHA). The software was developed to build forecasting capacity within foreign science agencies and allow for local ownership of the forecasts, and is used by the USGS to provide situational awareness during US humanitarian response in countries that have not yet been trained in the software. These forecasts are delivered through BHA to Urban Search and Rescue and Disaster Response Teams, U.S. embassies abroad, and other user groups identified during the response. Forecasts are also offered to local authorities and disaster management agencies through diplomatic channels. The international forecast product has been developed in conversation with the same set of stakeholders as the domestic forecasts, but with greater international representation through software training workshops and with an emphasis on the needs of emergency managers and international search and rescue teams (Schneider, Wein, et al., 2023).

The international forecasts are designed to be understood and used by emergency managers and other experts and include a range of information that is not available on the domestic forecast page. This information includes graphical representations of the forecast probabilities that allow an at a glance impression of the severity of the sequence (Figure 10); narrative forecast scenarios that describe in plain language three possible outcomes for the sequence along with associated probabilities (Figure 11); and a map of the probability of experiencing strong shaking (Figure 12). A complete example of a fictitious international forecast can be found in Text S1 in Supporting Information S1. The fictitious earthquake is placed in Corsica, France, to commemorate the 12th International Workshop on Statistical Seismology, which facilitated early conversations leading to this review.

Currently, The USGS does not make quantitative international forecasts public without an express request by the country in question. This can lead to fraught situations where a forecast exists in the country but is not being communicated to the people who live in the affected region. In these cases, the USGS strives to release at least basic information about the aftershock sequence and the chance of further damaging earthquakes through its communication offices, typically in the form of hand-crafted narrative forecast scenarios on the USGS public website earthquake.usgs.gov.

As mentioned previously, the communication products used for both domestic and international forecasting were developed through engagement and testing with stakeholders, using several approaches. User preferences and use cases for forecast products have been elicited in structured workshops, for example, in the HayWired project on Northern California aftershock scenarios (Detweiler & Wein, 2017), and at a workshop for California-wide stakeholders aimed at brainstorming new products (Field et al., 2016; Field & Milner, 2018), and other formal

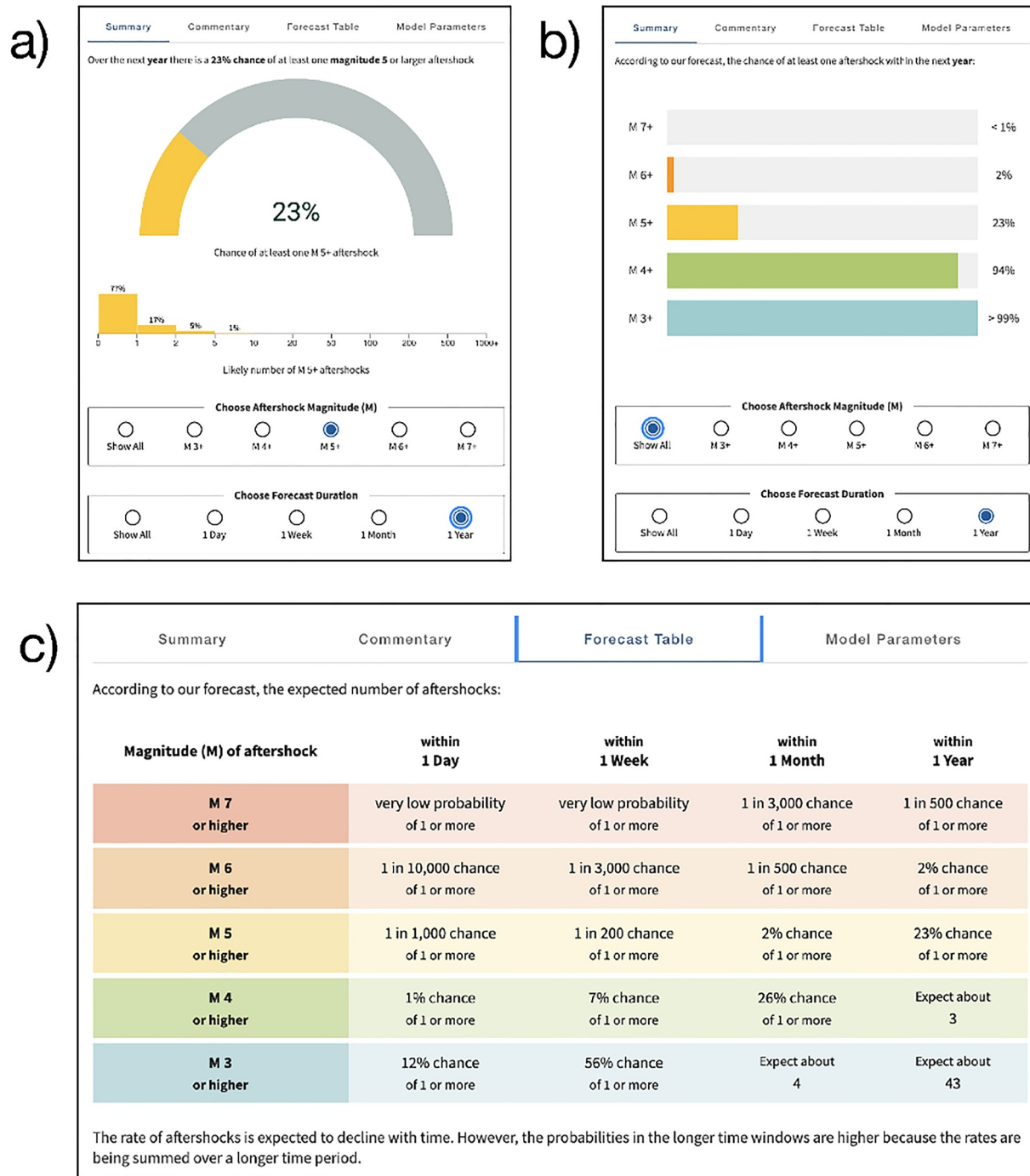


Figure 9. Interactive graphics examples from a U.S. Geological Survey aftershock forecast. (a) Probability of 1 or more aftershocks and number uncertainty visualization for a given magnitude and forecast duration. (b) Visualization of probability of 1 or more aftershocks for multiple magnitude levels and a given forecast duration. (c) Table summarizing probability of ≥ 1 event and median expected number of aftershocks for multiple magnitudes and forecast durations.

meetings with stakeholders, for example, from the BHA. Workshop participants have included emergency managers, critical infrastructure and lifeline operators, civil engineers, public information officials and geoscientists. Public reception of the forecast has further been evaluated through a review of media engagement (McBride et al., 2018; Michael et al., 2020). During the southwestern Puerto Rico sequence, citizens affected by aftershocks and aftershock forecasts were also surveyed to understand their response to different graphical products.



Example Forecast – Fictitious Earthquakes
Aftershock Advisory and Forecast



As of 20 Oct 2022, 10:34 (UTC) there is a
37% chance of an M5 or larger within the next week
in and around the area currently affected by aftershocks.

Mainshock Magnitude: M7.0
Mainshock Date: 19 Oct 2022, 10:34

ID: [Fictitious](#) Location: Fictitious
Forecast last updated: 20 Oct 2022, 10:34 UTC

- Expect more earthquakes in and around the area currently affected by aftershocks.
- Over the next week there may be 25 - 100 aftershocks of M3 or larger, which could be felt nearby.
- Aftershock rates will decrease over time, but may remain elevated over the following year or longer.
- This forecast will be updated as the sequences progresses and more information becomes available.

Aftershock Forecast starting 20 Oct 2022, 10:34 (UTC)

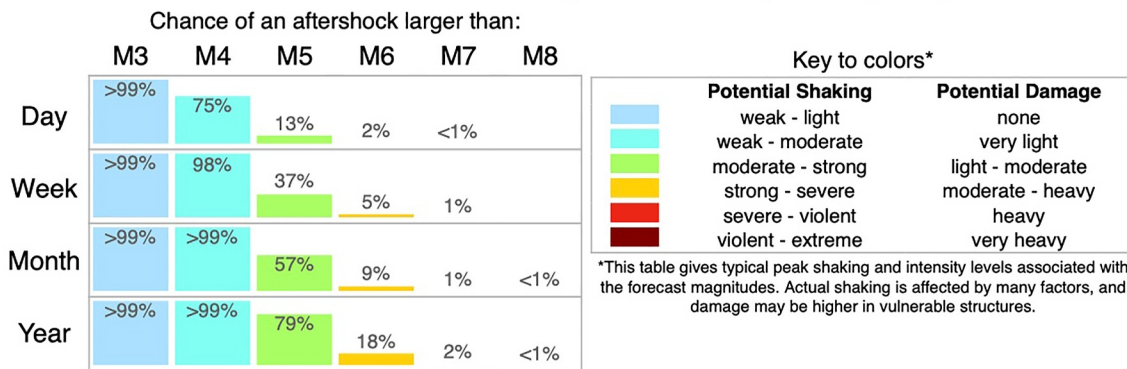


Figure 10. Headline forecast and graphical forecast summary for a fictitious magnitude 7 earthquake, following the U.S. Geological Survey template for international aftershock forecasts. This forecast template was developed with support of the U.S. Agency for International Development Bureau of Humanitarian Assistance (BHA), and would be delivered to BHA to support international response.

3.3.4. Communication Summary Table

Table 6 summarizes the information provided in this pillar on the forecast communication strategies of Italy, New Zealand, and the United States.

3.4. From OEF to Time-Dependent Hazard and Risk

Rational decision making, when there is uncertainty involved, requires assessment of the benefit of possible risk mitigation actions versus their costs, including the potential side effects (losses) they can cause. This also applies to the seismic case (e.g., Erto et al., 2016). Therefore, the earthquake forecasts outlined in this document must be used as an input to loss forecast, that is seismic risk assessment, to be used for risk management by user groups. Note that different users will make decisions in different manners and require different levels of information (Field et al., 2016).

The approach to seismic risk assessment around which there is most consensus is that framed by performance-based earthquake engineering (PBEE; Cornell & Krawinkler, 2000), where the loss metric is the rate of earthquakes exceeding a threshold of interest, and the computation of such a rate is carried out combining probabilistic descriptions of (a) the seismic hazard of the site, (b) the earthquake vulnerability of the system of interest (e.g., a building, an infrastructure, or - by extension - a community), and (c) the potential consequences (i.e., the losses) caused by the damages the seismic vulnerability can lead to (this term is also sometimes expressed as exposure).

Example Forecast – Fictitious Earthquakes

Aftershock Sequence Scenarios

These are three likely scenarios for how the aftershock sequence will evolve over the next **week** starting 20 Oct 2022, 10:34 (UTC)

Scenario One (Most likely)

95%

The most likely scenario is that aftershocks will continue to decrease in frequency with no aftershocks larger than M6 within the next week. Moderately sized aftershocks (M5 and larger) may still cause localized damage, particularly in weak structures. Smaller magnitude earthquakes (M3 and M4) may be felt by people close to the epicenters.

Scenario Two (Less likely)

5%

A less likely scenario would include one or more aftershocks larger than M6, but with none larger than the M7.0 mainshock. Aftershocks of this size could cause additional damage and temporarily re-energize the aftershock sequence. These aftershocks would most likely affect the area already impacted by the mainshock.

Scenario Three (Least likely)

1%

The least likely scenario is that the sequence could generate an aftershock of the same size or even larger than the M7.0 mainshock. Such an earthquake would likely affect communities both in and adjacent to the areas already impacted by the mainshock, and would trigger additional aftershocks.

Figure 11. Narrative forecast scenarios for a fictitious earthquake, generated as part of the U.S. Geological Survey international aftershock forecasts. The forecast scenarios would be delivered to U.S. Agency for International Development Bureau of Humanitarian Assistance to support international response operations.

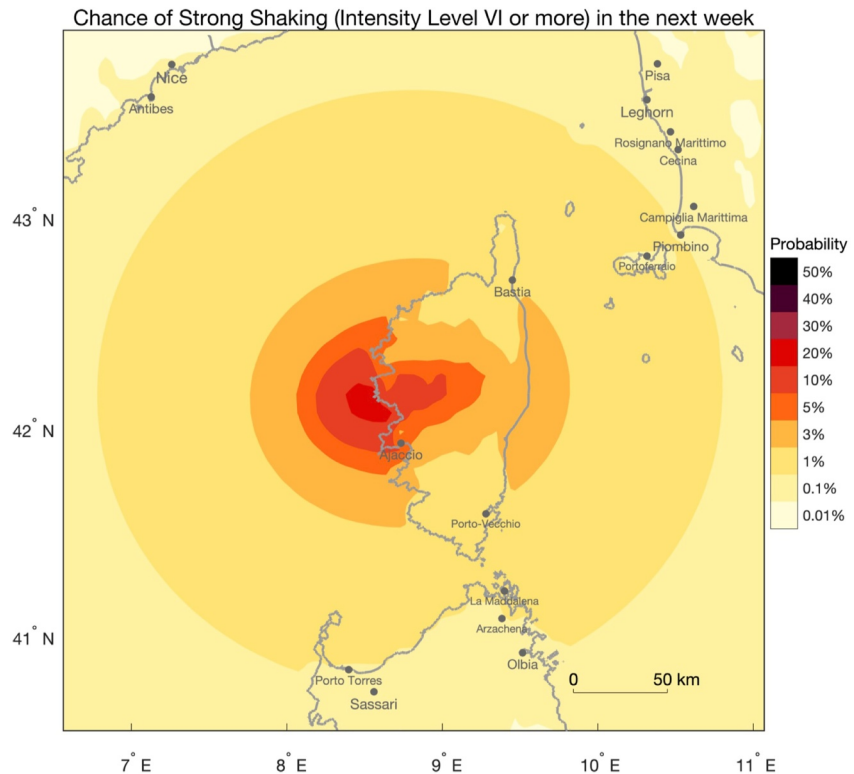


Figure 12. Map of potential shaking in the affected region for a fictitious earthquake. The map shows the probability of exceeding Intensity level VI on the Modified Mercalli Intensity scale. This forecast map would be delivered to the US-AID Bureau of Humanitarian Assistance to support international response operations.

Table 6

Summary of the Forecast Communication Strategies of Italy, New Zealand, and the United States (U.S.) (Domestic and International)

	Italy	New Zealand	United States domestic	United States international
To whom is the forecast communicated?	Authorized personnel: Major Risk Commission of the Italian Civil Protection, who can further distribute	First duty officers and seismologists, then NEMA, then public	Public	Urban Search & Rescue Disaster Response Teams, U.S. embassies abroad, other users identified during response
Distribution channel	Department of Civil Protection (to regional entities)	Personal communication to duty officers and seismologists, GeoNet website, Facebook, Twitter. Also professional channels such as Emergency Management or Lifeline Groups, or the NZ Earthquake Engineering Society	USGS website	US-AID Bureau of Humanitarian Assistance
Stakeholder interactions for forecast product development	Informal meetings with Italian Civil Protection and INGV officials	Workshops, focus group discussions and surveys with lifeline providers engineers, emergency managers, public	Formal & informal interactions with emergency response organizations, utility & lifeline operators, public transportation, civil engineers, building inspectors	Additional informal interactions with international stakeholders through training workshops
Testing of the communications	Partially tested with the public Periodic meetings with Italian Civil Protection	Not extensively tested Co-developed with user groups and with support of social scientists	Structured workshops and formal meetings Review of media engagement Public survey	Structured workshops and formal meetings
When is a forecast released?	At midnight of each day, and after the occurrence of every $M_L \geq 3.5$ event Continuously available to Italian Civil Protection, disseminated to the Major Risk Commission every 4 months or upon request	Manual after a significant event	20 min after a $M \geq 5$ event	Manual after a significant event (within 24 hr)
When is a forecast adjusted?	Does not apply, because forecasts are not specific to individual events	Frequently at first (at least daily), then intermittently. No pre-determined schedule	Frequently at first, then intermittently. Seventy-five automatic updates in the first year are made according to pre-determined schedule. Manual updates can also be made	Frequently at first, then intermittently on a schedule negotiated around user needs
What information is provided?				
Earthquake probabilities	Yes	Yes	Yes	Yes
Expected number of events/Rates	Yes, in dedicated additional reports available during a sequence	Yes, possible range	Yes	Yes
Maps	Yes, map of earthquake and ground shaking probabilities, users can select sub-region through GUI	Yes, expected ground shaking map, map of forecast area	Yes, map of forecast area and aftershocks	Yes, expected ground shaking map
Temporal evolution of probabilities/rates	Default is 1-week forecast	Yes, in table format	Yes, in table format	Yes, in table format
Comparison to long-term values	Yes, evolution of past forecasts is shown	Yes, with 2010 NSHM, still to be updated to 2020 revision	No	No
Ground motion	Yes	Yes	No	Yes
Losses and/or impacts	No (it is considered by a companion initiative named OELF)	No	No	No

Table 6
Continued

	Italy		New Zealand		United States domestic		United States international	
Scenarios	No	Yes			No	Yes		Yes
Explanations for lay people	A dedicated website is almost ready but not open to the public		Yes		Yes			Yes
Interactive forecast	Yes, but only for specific users. Static for lay people		No		Yes, users can select magnitude and duration of interest			No

PBEE was initially formulated for long-term risk assessment of single structures (e.g., buildings) in the *long term*, with this meaning that the seismic hazard is represented by classical probabilistic seismic hazard (Cornell, 1968), where earthquake occurrence is represented by a homogeneous Poisson process (HPP), and therefore only mainshocks are considered to contribute to the risk. Long-term also reflects on the vulnerability models used in the risk assessment, which typically neglect that losses can accumulate in multiple events, assuming, for example, that time between mainshocks is enough for repair/rebuild. This leads to risk metrics output of the risk assessment, for example, the mentioned loss exceedance rate or the unit-time expected loss, to be time-invariant, which greatly simplifies the probabilistic modeling of the problem at hand.

The adaptation of PBEE to short-term risk assessment is the natural way to translate earthquake forecasts into risk assessment. This requires replacing classical seismic hazard with its short-term counterparts. One worth-mentioning attempt is the aftershock hazard analysis of Yeo and Cornell (2009), to be used for risk assessment after a mainshock and in which the earthquake occurrence is described by a non-homogeneous Poisson process, the mean of which is provided by the Omori law, modified considering the advancements of Reasenberg and Jones mentioned above. This approach allows the estimation of the time-variant rate of earthquakes exceeding a ground motion intensity threshold at a site of interest. It can also be argued that during seismic sequences there is not enough time to repair the direct damage to the assets of interest and therefore seismic losses can accumulate in multiple events clustering in time and space. This requires developing vulnerability models able to probabilistically describe such a phenomenon (e.g., Iervolino et al., 2016, 2020). Similarly, exposure may also need appropriate models for short-term risk assessment. For example, human exposure can be time-variant during a sequence because of evacuation and relocation actions by civil protection. These issues, individually or all together, generally lead to time-variant risk metrics.

In this context, short-term risk assessment based on OEF has been developed (Herrmann et al., 2016; Iervolino et al., 2015), referred to as operational earthquake loss forecasting (OELF). For example, Iervolino et al. (2015) describe a prototype risk calculation engine set up for Italy, called Mantis-K, which receives each day or after a $M \geq 3.5$ event, the rate of earthquakes with $M \geq 4.0$ computed as discussed in Model Specifications, for the whole country. Based on a residential buildings inventory available at a national scale, and a vulnerability model applicable to this inventory, these rates are translated into risk metrics such as the expected number of unusable and collapsed buildings, and the expected number of injuries and fatalities for seismic causes. This risk assessment is valid for the same time horizon the earthquake OEF rates refer to and is updated at each earthquake rate release. For example, Figure 13 shows the forecasted earthquake rates in the week after 04/10/2024 (left) and the corresponding expected number of fatalities in the same week (right). In Figure 14, taken from Marzocchi, Iervolino, et al. (2015), the time-variation of the local personal (i.e., fatality) risk for someone exposed to building damage during the 2012 Pollino sequence in Italy is shown.

Adapting PBEE to include information available at small time scales can go as far as considering information about an earthquake after the rupture generating it has occurred already, but while its seismic waves are still traveling to the site where the exposed asset is located. This is referred to as earthquake early. In fact, seismology and earthquake engineering enable performing seismic hazard and risk assessment even in this situation (e.g., Iervolino, 2011; Iervolino et al., 2006), which is considered real-time, while the earthquake forecasting discussed in this review, which refers to forecasting rupture occurrence in the short-term, is typically referred to as near-real-time.

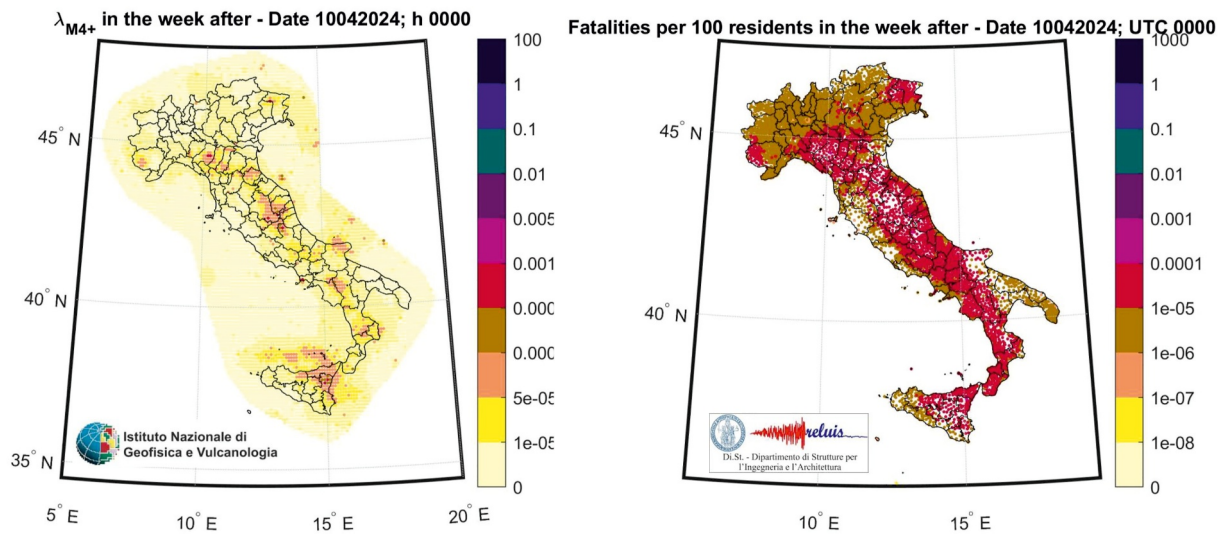


Figure 13. Forecasted rates of earthquakes with magnitude equal or larger than 4 in the week after 04/10/2024 in Italy and corresponding expected fatalities for the OELF system of the country named Mantis-K (as described in Iervolino et al. (2015)).

4. Elicitation of Expert Views

After the description of the forecasting systems currently in operation in Italy, New Zealand, and the United States, Section 4 of this review describes the results of and the process that was applied to elicit expert views on OEF good practice recommendations in the development, testing, and communication of earthquake forecasts (Mizrahi, Dallo, & Kuratle, 2023). In this study, experts are considered to be scientists who do research in this field and have contributed to the current state of OEF systems around the world.

This expert elicitation serves as a structured and transparent means of capturing expert views on a specific topic, complementing the current state of OEF systems described in Section 3. It should be noted that this

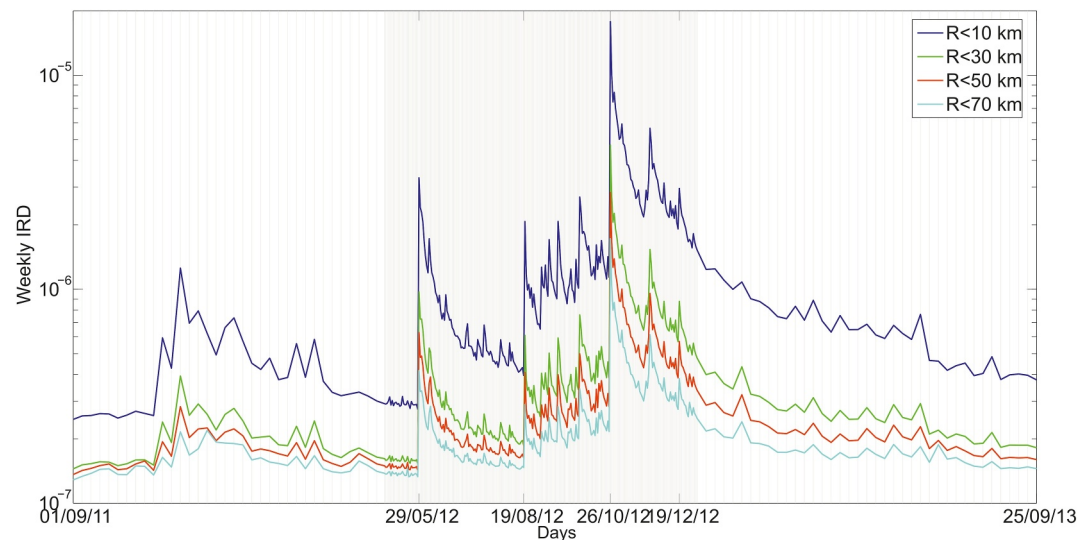


Figure 14. Time-variant local personal risk for residents in the Pollino region in Italy during the 2012 seismic sequence. The continuous lines of different colors show the weekly individual risk of death (IRD) caused by earthquakes for circular areas of different radius from the center of the seismic sequence located at latitude 39.85°N and longitude 16.05°E. The IRD is computed as the expected number of fatalities per radius bins divided by the total number of residents in the area (Figure taken from Marzocchi, Iervolino, et al. (2015)).



Figure 15. The three steps of the Delphi study, with N being the number of experts involved in each step.

elicitation shall not be considered as a definitive and authoritative source of truth regarding earthquake forecasting, but as a snapshot of opinions from a relatively large number of experts in the field, a group that is not representative of the entire seismological community. The aim of the elicitation was to capture international debates and discussions on OEF modeling, testing, and communication, seeking to inform future research endeavors, support nations interested in developing OEF systems, and identify research gaps and potential areas for collaboration.

4.1. Method and Limitations

4.1.1. Methodological Procedure

To elicit expert opinions and identify areas of consensus and dissent in a scientific field, Delphi studies are a common approach (Hsu & Sandford, 2007). A Delphi study is an iterative process in which selected experts are asked to anonymously rate given statements based on their level of agreement with these statements. If the expert group does not reach a consensus on certain statements (according to a pre-determined consensus definition), the reason for this dissent is discussed in a joint workshop among the group. After the discussion, the statements in question are adapted and re-evaluated by the experts in the next round of the anonymous survey. This process can be repeated until consensus is reached on all topics, or the process can be interrupted and the areas of dissent can be taken as indicators for current research gaps.

The Delphi method was applied to elicit expert views on issues related to the development, testing, and communication of earthquake or aftershock forecasts (Mizrahi, Dallo, & Kuratle, 2023). The survey will only mention earthquake forecasts, considering aftershock forecasts as a special case thereof. A survey was conducted first (15–29 March 2023), followed by a workshop (5 April 2023), and a second survey round (11–24 April 2023); see Figure 15. The workshop between the two surveys allowed a deeper investigation of certain comments raised in the first survey and a discussion of statements with dissent. The experts were encouraged to exchange ideas, discuss differing views, and find a common ground on which they could all agree. After the workshop, certain statements were adjusted, those where consensus was already reached were removed, and further statements brought up during the discussion were added. The first survey and the workshop were structured into three parts: *Model Development*, *Model Testing*, and *Forecast Communication*; the second survey additionally addressed interdependencies between these parts as a result of the workshop discussion.

In most survey questions, the study participants were asked to indicate their level of agreement with a statement on a 7-point Likert scale, with “1” representing strong disagreement and “7” representing strong agreement (Joshi et al., 2015). A 7-point Likert scale was chosen to allow a neutral/undecided response (level 4), and to have enough response options to capture nuanced expert opinions and reduce central tendency bias, compared to, for example, a 5-point Likert scale. In addition, two survey questions, which will specifically be pointed out, allowed a binary indication of agreement (yes/no), and other questions were designed to collect plain text answers by the expert group. Given that certain questions demanded technical understanding or reviewer experience, respondents were always given the option to select “I don’t know,” ensuring that only participants with a sufficient understanding addressed the various inquiries. To calculate consensus, we excluded the “I don’t know” answers. There were at least 15 responses to all questions, with the exception of three questions that are specifically marked in the text.

Consensus was defined to be reached when at least 70% of the experts agree with a statement (agreement level 6 or 7) or disagree with a statement (agreement level 1 or 2) or are undecided about a statement (agreement levels 3, 4,

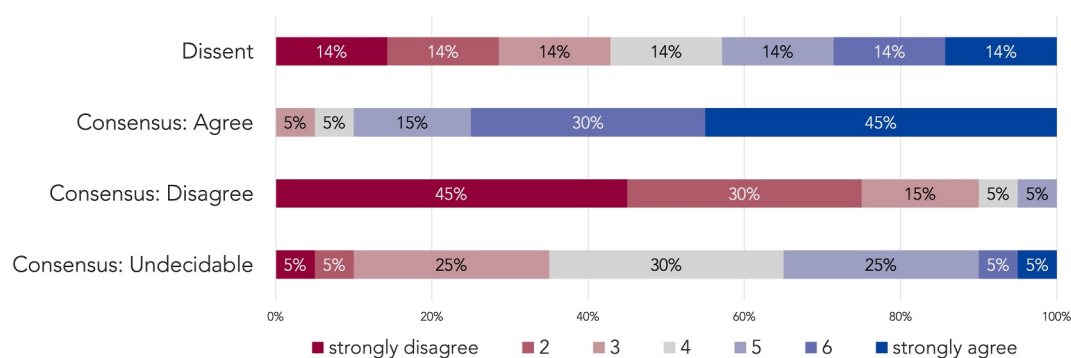


Figure 16. Examples of different types of consensus, and dissent (Mizrahi, Dallo, & Kuratle, 2023).

or 5), as has been proven appropriate in previous Delphi studies (e.g., Slade et al., 2014; Vogel et al., 2019). In the third case, the consensus is that this statement can currently not be decided. Dissent is defined as the absence of all three types of consensus. Figure 16 illustrates examples of possible outcomes. For the two questions where the indication of agreement was binary, only two types of consensus were distinguished: agreement ($\geq 70\%$ of experts indicating agreement) or disagreement ($\geq 70\%$ of experts indicating disagreement, which is equivalent to $\leq 30\%$ of experts indicating agreement). In case consensus is not reached, a tendency toward agreement, disagreement, or undecidedness is considered to be reached when at least 50% of experts agree, disagree, or are undecided about a statement.

The group of study participants was composed of individuals who met the eligibility criteria of holding or having held research positions and having published peer-reviewed articles in the fields of earthquake forecasting, forecast model testing, or forecast communication. An exception was made for PhD students who have presented their work in one of the relevant fields at a scientific conference. This inclusion of individuals with less mature viewpoints is vital for capturing diverse perspectives and considering the most recent advancements and developments in research.

The experts who ultimately participated in the study formed a group of 20 individuals working at different institutions in Germany, Italy, New Zealand, Switzerland, the United States, and the United Kingdom. The workshop and the second survey saw a reduction in group size to 17 participants, although it's worth noting that the composition of the group may have varied between these two phases. The expert group encompassed individuals at various career stages, ranging from doctoral students to emeritus professors with over 20 years of experience in the field (see Figure 17). Their self-perceived expertise was highest in developing (median 6 out of 7) and testing (median 6 out of 7) earthquake forecasting models. In comparison, they perceived their expertise in communicating earthquake forecasts as lower (median 4 out of 7). Their research primarily focused on regional and national geographical scopes, with a few experts concentrating on global and multinational scales, and only a couple on European and continental scales. 30% of the experts were female, and 70% were male. Their mean age was 46 years ($SD = 13.0$). In the invitation email, it was explicitly communicated that the elicitation centered on short-term forecasts to make the focus of the surveys and workshop clear.

A detailed description of all results, as well as the exact wording of the survey questions, can be found in Mizrahi, Dallo, and Kuratle (2023).

4.1.2. Limitations

While Mizrahi, Dallo, and Kuratle (2023) provides valuable insights into experts' views on how to best develop, test, and communicate earthquake forecasts, certain limitations should be acknowledged.

First, one main limitation of expert elicitations in general is that their results inherently depend on the composition of the group of participating experts. For instance, the experts' opinions could be biased because they all live in countries that have seismic hazard programs in place, and input from researchers living in more vulnerable regions is missing. This limitation is acknowledged and addressed by providing a transparent description of the expert group. Furthermore, much effort was invested in inviting a diverse and knowledgeable expert group to

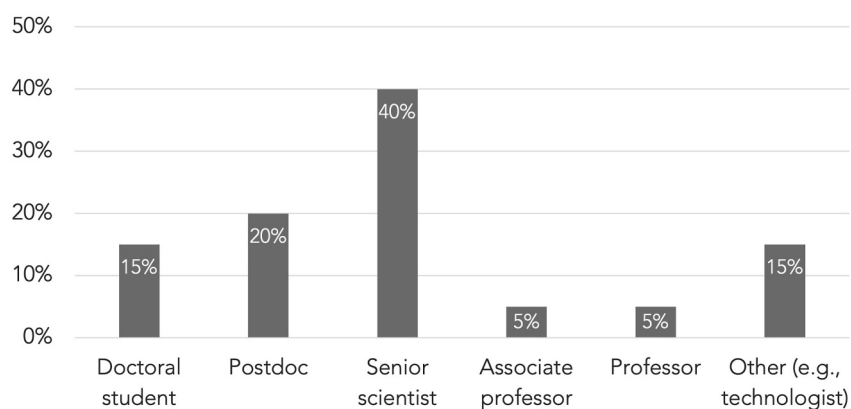


Figure 17. Distribution of scientific position of the study participants (as of the first survey) (Mizrahi, Dallo, & Kuratle, 2023).

participate in the study. The employment of eligibility criteria ensured a systematic approach to the aspect of knowledgeability. To broaden the pool of experts, recommendations were actively sought from those who agreed to participate. However, researchers in numerous regions are presently not actively working on OEF development and thus declined the invitation to participate. This underscores the necessity for future research and endeavors aimed at establishing OEF systems, especially in areas with elevated seismic hazard and risk.

Second, expert responses on a topic are a (possibly complex) function of their experience with the topic, such that the choice of response options and the definitions of agreement, undecidedness, disagreement, consensus, and dissent impact the study's findings. While the effect of the choice of response options on the results cannot be assessed in hindsight, the authors aimed to minimize the effect of chosen definitions of agreement, consensus, etc., by additionally discussing tendencies and always providing information on the percentage of agreement within the expert group. Furthermore, the detailed anonymized responses to all questions can be found in Mizrahi, Dallo, and Kuratle (2023).

A third limitation of the study in particular is the relatively small number of two iterations of surveys that were completed. Starting with a set of broadly formulated statements, subsequent iterations could have helped to identify additional, more specific areas of expert consensus.

Furthermore, almost by definition of the Delphi process, its results are not expected to be surprising. In contrast, the aim of the process is to identify, in a structured way, statements on which a large fraction of the expert group ($\geq 70\%$ in this case) already agrees. This means that many of the conclusions of the study may have been pointed out in previous publications on the topic in general, and by the 2011 report of the International Commission for Earthquake Forecasting (Jordan et al., 2011).

Overall, Mizrahi, Dallo, and Kuratle (2023) believe that the benefits of providing a structured collection of expert opinions on OEF-related issues outweigh the unavoidable drawbacks of expert elicitations.

4.2. Results

In this subsection, the findings of the Delphi study are described, separated into the three pillars *Model Development*, *Model Testing*, and *Forecast Communication* (Mizrahi, Dallo, & Kuratle, 2023). A summary of the key findings and a discussion of the results are given in the next subsection.

4.2.1. Model Development

Within the *Model Development* pillar, consensus among the panel of earthquake forecasting experts was reached only on a few statements in the first survey. These statements were designed to identify specific models and model features that experts agree are suited/necessary for earthquake forecasting. The workshop discussion however revealed that only a few explicit recommendations could be given, possibly because the development of

When developing a forecasting model for the following user groups [x-axis], the following pieces of information [y-axis] are important:

	General public	Emergency responders	Civil protection	Critical infrastructure providers	Decision makers	Insurance companies	Duty seismologists
Higher order aftershocks	71%	82%	82%	71%	82%	82%	77%
Catalog incompleteness	59%	59%	59%	59%	65%	71%	94%
Faults	59%	53%	59%	71%	59%	53%	77%
Available historical data	53%	47%	53%	59%	59%	59%	71%
More than just seismicity	35%	47%	47%	53%	59%	47%	59%
Spatial anisotropy	35%	41%	47%	47%	47%	59%	77%
Uncertainty of model parameters	35%	35%	35%	53%	47%	53%	77%
Geodesy	24%	24%	29%	29%	29%	29%	53%
(Spatial/temporal) b-value variations	18%	18%	24%	24%	29%	29%	59%

Figure 18. Important pieces of information per user group [in percentage]. Cells are highlighted according to the percentage of experts that indicated relevance, in shades between red ($\leq 30\%$) and blue ($\geq 70\%$). Cells with consensus ($\leq 30\%$ or $\geq 70\%$) are further highlighted with a black border (Mizrahi, Dallo, & Kuratle, 2023).

earthquake forecasting models may depend on a variety of factors including the user needs of the OEF product. These insights were used to formulate adjusted statements, which experts then evaluated in the second survey.

4.2.1.1. Model Ingredients

In the first survey, the experts were asked to indicate for different scientific model ingredients whether they should be considered in a forecasting model, where the agreement levels range from 1 (nice to have) to 7 (absolutely necessary). None of the items reached consensus ($\geq 70\%$ agreement of level 6 or 7). The experts rated the scientific ingredients in the following order (sorted by percentage of experts agreeing with level 6 or 7):

Earthquake forecasting models should:

- account for catalog incompleteness (66%)
- account for higher-order aftershocks (45%)
- account for spatial anisotropy (45%)
[note: anisotropic aftershock triggering is meant here]
- make use of available historical data (33%)
- account for magnitude heterogeneity in catalogs (30%)
- be based on more than purely on seismicity (23%)
- account for (spatial/temporal) *b*-value variations (18%)

After the workshop discussion highlighted that the way in which earthquake forecasts are developed may depend on the targeted user groups, an adjusted question was posed in the second survey to capture this dependency. The survey participants were given a matrix of user groups and information pieces and they were asked to indicate with a binary response (yes or no) whether “When developing a forecasting model for the following user group [x-axis], the following pieces of information [y-axis] are important.” Note that these pieces of information are to be considered by the model, not provided by the model. The user groups were selected based on the results of the first survey in which experts indicated which user groups might be interested in OEF products (see Target Audiences), and the information pieces were selected based on the responses to the first survey. Figure 18 shows the results of this question. Cells are colored based on the percentage of experts that indicated agreement. In addition to a consensus that something is important, there is the possibility for consensus that something is not important (i.e., at most 30% agreed to it being important). Note that the answers to this question can not be directly compared to the question posed in the first survey, since agreement corresponds to an ingredient being considered “important for a user group” versus “absolutely necessary to be considered in a forecasting model.”

There are clear trends in the importance of pieces of information seen relatively to each other: *b*-values are considered not to be important in the development of forecasting models for most user groups, while higher-order aftershocks should always be considered, independently of the user group. But comparing the importance of information pieces between different user groups, the trends are not clear. There is no piece of information on which there is expert consensus that it is important for one user group ($\geq 70\%$ agreement) but not important for another user group ($\leq 30\%$ agreement).

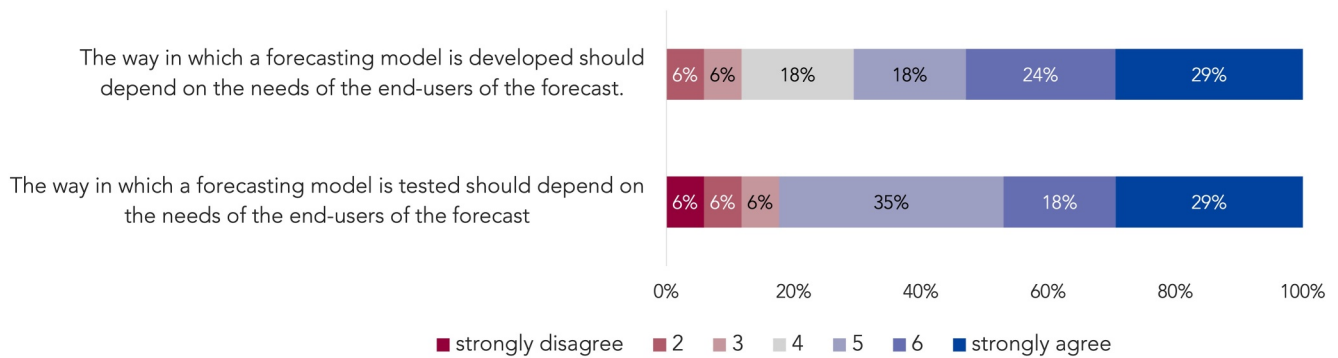


Figure 19. Evaluation of the two statements related to the role of user groups when developing and testing earthquake forecasting models (Mizrahi, Dallo, & Kuratle, 2023).

The assumption that a model feature that is considered important for one user group can therefore be considered important for other user groups too, can thus not be rejected. This is in line with the result of the second survey shown in Figure 19, which addresses whether experts agree on the idea that model development and testing *should* at all depend on the user group. No consensus was reached on these issues.

The question of whether experts believe that model development and testing should depend on user needs can therefore not be answered conclusively. This result is further highlighted by the diametral differences between comments made by experts at the end of the second survey. One survey participant was “confused by the idea that scientific ingredients of a model should depend on the users” and stated that “these ingredients should be selected to make a scientifically accurate model.” Another participant disagreed with the idea that “deployment of forecasts is primarily a scientific question” and claimed that “politics, culture, economics, and user preferences are just as important.”

Although consensus was still not reached for all items after distinguishing between user groups in the second survey, the general trend of the importance of each of these ingredients remained consistent. It is challenging to reach consensus among the expert panel both when distinguishing and when not distinguishing between user groups. Figure 18 highlights the model ingredients on which experts reach a consensus regarding their importance—and simultaneously reveals which items cause dissent.

4.2.1.2. Model Types

Besides investigating the importance of model ingredients, the experts were also asked to assess general model types: ETAS (Ogata, 1988), STEP (Woessner et al., 2010), EEPAS (Rhoades & Evison, 2004), Reasenber and Jones (Reasenber & Jones, 1989), or ensembles thereof (e.g., Herrmann & Marzocchi, 2023; Marzocchi, Zechar, & Jordan, 2012). After the first survey, there was consensus that:

- Ensembles thereof are suited for earthquake forecasting (70% agreement level 6 or 7).
- [almost consensus] The ETAS model is suited for earthquake forecasting (69% agreement level 6 or 7).
- It is unclear whether the EEPAS model and the Reasenber and Jones model are suited for earthquake forecasting (78% (EEPAS, 14 overall responses only) and 100% (R&J) of responses with agreement level 3–5).

There was dissent on the suitability of the STEP model for earthquake forecasting, with 40% of experts indicating agreement (level 6 or 7), and 60% indicating being undecided (agreement levels 3–5) about the statement.

Regarding the models for which it is not clear whether they are suited for earthquake forecasting, several points were raised in the comments. Some argued that models should be tested and studied by researchers outside the group that developed the model, others pointed out that some models are not well-defined in space, or that high model complexity might make a model less appealing from a communication perspective. It seems reasonable that the ETAS model, which has been widely used and tested since it was first introduced, was the one single model that could reach the highest level of acceptance. This was confirmed in the second survey, where more than 80% of the experts chose the ETAS model when asked “If you had to choose one simple base model to produce forecasts which are useful for a maximum number of end-users, which one would you choose?” (Mizrahi, Dallo,

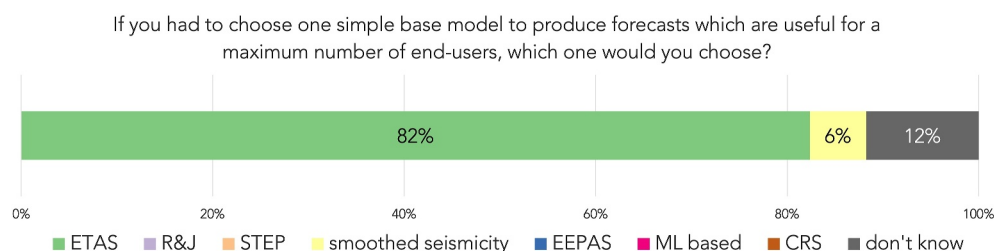


Figure 20. Visualization of the survey results regarding the experts' preferred default forecasting model. “CRS” and “ML based” refer to Coulomb-rate-state and machine learning based models, respectively, which will be discussed in the Outlook on Model Development (Mizrahi, Dallo, & Kuratle, 2023).

& Kuratle, 2023) and given the possibility to choose among all models that were discussed in the previous survey or mentioned in the comments or in the workshop (see Figure 20). The term “simple” was, among others, used to convey that the model should be suitable and adaptable for development and application in regions with limited data. Individuals or entities identified by the respondents in the initial survey as confirmed potential users were defined as “end-users.”

4.2.1.3. Model Updating and Transferability

There was no consensus (based on the first survey) on whether earthquake forecasting models should be (sorted by percentage of experts agreeing with level 6 or 7):

- updated during an ongoing sequence (63%)
- recalibrated regularly using the newest available data (45%)
- calibrated and tested once and not be updated without further testing (16%)

In the open comments, some experts stated that a forecasting model should be updated as often as practically feasible, when new scientific evidence arises, after large events, or when the catalog quality changes. In the second survey, the above statements were reformulated without using the imperative “should.” Due to the ambiguity in the strictness of the implied imperative, the use of the word was pointed out as a possible reason for dissent. Based on further inputs given in the workshop, the experts were tasked to quantify their agreement to statements regarding the transferability of models that have been previously tested in a specific setting. In the second survey, there was consensus on the following:

- If a model has been approved to be used for a given purpose, its parameters can be updated when new data becomes available (75%)
- If a model has been approved to be used for a given purpose, it **cannot** be applied for the same purpose in a different region without additional testing (75%)

No consensus was reached on the statement that “If a model has been approved to be used for a given purpose, it is expected to be useful for the same purpose in a different region,” though there is a slight tendency toward undecidability (56% with agreement levels 3–5).

4.2.2. Model Testing

A key result of the first survey was that there is dissent among the experts on whether there is a need to collectively define the minimum required tests a model should pass before it can be used for forecasting, as shown in Figure 21.

If there is no consensus on whether the community should define collectively how it recommends a forecasting model to be tested, it will be challenging to reach the key goal of the testing part of this Delphi study: to identify good practice recommendations for earthquake forecasting model testing. Nevertheless, the study could identify certain points—for example, regarding transparency and reproducibility of tests—on which the expert panel clearly agreed. For other points, especially such where the experts were asked for specific recommendations on which tests to use to test a forecasting model, no or little consensus was reached, as could be anticipated given the

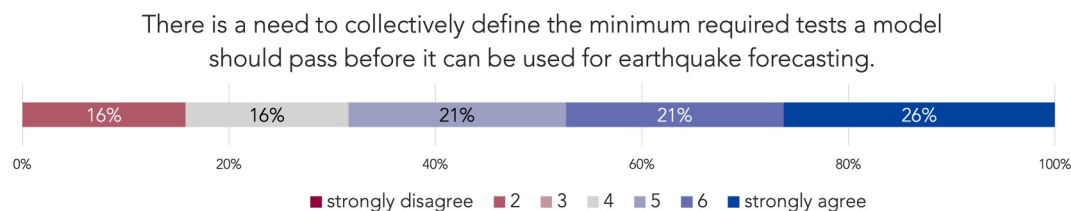


Figure 21. Evaluation of a statement regarding the need for collectively defined minimum required tests (Mizrahi, Dallo, & Kuratle, 2023).

result of Figure 21. In the remainder of this subsection, the results of the Delphi study for different aspects of earthquake forecast model testing (Mizrahi, Dallo, & Kuratle, 2023) are described.

4.2.2.1. Transparency and Reproducibility of Tests

Already in the first survey there was a strong consensus among the experts that (sorted by percentage of experts agreeing with level 6 or 7):

- Operationally issued forecasts should be archived for retrospective analysis (100%).
- Archived forecasts should be publicly available for retrospective analysis by the community (90%).
- Source code of forecasting models should be publicly available (80%).

4.2.2.2. Modes of Testing

The experts were asked to evaluate several statements related to ideal modes of testing, where some statements elicited a very clear consensus among the group, while others did not, as indicated by the percentage of experts rating the statement with agreement levels 6 or 7.

- A model that is already used for earthquake forecasting should continue to be tested (95%).
- For a forecasting model to be used, it is **necessary** to test the model pseudo-prospectively (i.e., excluding the testing data when training the model) (70%).
- For a forecasting model to be used, it is **necessary** to test the model truly prospectively (i.e., the testing data may not exist when the model is developed) (65%).
- For a forecasting model to be used, it is **sufficient** to test the model retrospectively (i.e., using the testing data when training the model) (6%).

For the latter statement, there was a tendency toward disagreement (53% of experts indicated agreement level 1 or 2). From a strictly logical point of view, the agreement that pseudo-prospective testing is necessary would imply disagreement with retrospective testing being sufficient. However, disagreement with the latter statement was not as strong as the acceptance of the former. Survey comments and workshop discussion revealed that some experts see important value in retrospective testing. Hence, by indicating medium instead of high disagreement levels for the statement of retrospective testing being sufficient, they may have wanted to indicate that although not sufficient, retrospective testing is still necessary or at least important to assess a forecasting model. In the second survey, the statements were thus reformulated to use less strict language and the following results were obtained.

- For a forecasting model to be used, it is **recommended** to test the model pseudo-prospectively (i.e., excluding the testing data when training the model) (93%).
- For a forecasting model to be used, it is **recommended** to test the model truly prospectively (i.e., the testing data may not exist when the model is developed) (73%).
- For a forecasting model to be used, it is **recommended** to test the model retrospectively (i.e., using the testing data when training the model) (53%).

By changing the formulation from “necessary” and “sufficient” to “recommended,” all statements reached higher acceptance. This suggests a hesitancy of the expert group to define strict testing requirements.

4.2.2.3. Testing Authority

Several questions in both surveys aimed to address the question of which entity should be entitled to test and confirm a model's readiness for application. The experts agreed on the following.

- A forecasting model is ready to be used if it has been tested by a third party (e.g., in a CSEP experiment) (74%).
- A forecasting model is **not** ready to be used if the model developers trust the model, but it has not been reviewed by anyone else (70%).
- It is **unclear** whether a forecasting model is ready to be used if the end-user of the forecast approves (80%).

Note that the wording of the latter two statements was modified compared to the wording used in the survey to reflect the consensus among the experts. The experts either disagreed (levels 1 or 2, second statement above) or were uncertain (levels 3–5, third statement above) about the original wording of those statements.

No consensus was reached on statements claiming that a model is ready to be used if it has been published in a peer-reviewed journal (57% unsure), or if it represents the best available science (56% unsure). These two statements were added to the second survey based on comments during the workshop.

4.2.2.4. Specific Tests

Consistent with the hesitancy of experts to define strict testing requirements is the fact that no consensus was reached on the statements declaring any specific CSEP test a “strict requirement for a forecasting model to be used” (sorted by percentage of experts with agreement level 6 or 7):

- CSEP Magnitude test (45%)
- CSEP Number test (45%)
- CSEP Spatial test (45%)
- CSEP (Pseudo-)Likelihood test (42%)

Also no consensus could be reached on what is required for an existing model to be replaced, with a majority (53%) of experts being unsure (levels 3–5) about whether it could be replaced by a model that has demonstrated positive information gain over the existing one, and 63% being unsure about whether it could be replaced by a model that passes the same consistency tests.

Based on inputs from the workshop, the second survey aimed to address this same question in the setting of a peer review process (although the second survey also revealed that there is no consensus on whether the peer review process is the adequate model testing mechanism). To the question “If you are peer-reviewing a paper describing a forecasting model, the results of which tests do you consider important for your decision,” experts evaluated different tests as follows (sorted by percentage of experts with agreement level 6 or 7, only 12–13 responses per option).

- Comparison to a benchmark model (79%)
- CSEP Spatial test (54%)
- CSEP Number test (54%)
- CSEP (Pseudo-)Likelihood test (46%)
- CSEP Magnitude test (39%)
- Specific metrics for the forecasting of rare events (34%)

The last point was added based on a user comment in the first survey. The only test on which there is a consensus among experts that its result would be useful for their decision during a peer review is the comparison to a benchmark model (e.g., Bayona et al., 2023; Mancini & Marzocchi, 2023; the definition of a benchmark model was left open-ended in the survey). On all other tests, there is dissent concerning their importance.

Several points were raised in the survey comments as well as the workshop as to what could be the reason for this dissent, ranging from a lack of specification of these tests (Should the N-test be passed every day? Over a long-term time horizon? Pass it 90% of the time?), a recommendation to test the tests (Show that the power of the tests is sufficiently high), to suggestions that no test is sufficient, or that no test should be a strict requirement. Others suggested that a model should never be replaced, but added to an ensemble, and the model weights would automatically be adjusted to give less power to the models that don't perform well.

4.2.3. Forecast Communication

4.2.3.1. Target Audiences

Operational earthquake forecasts can serve several different user groups—and the elicitation shows that experts agree that earthquake forecasts are relevant for (sorted by percentage of experts with agreement level 6 or 7):

- Civil protection (100%)
- Critical infrastructure providers (100%)
- Emergency managers and responders (95%)
- Search and rescue organizations (95%)
- National and cantonal authorities (90%)
- Communication experts (90%)
- Seismologists (85%)
- Policymakers (80%)
- Structural engineers (75%)
- Insurance companies (74%)
- General public (70%)

No consensus but a tendency toward agreement was reached regarding the relevance of earthquake forecasts for the following user groups.

- Geotechnical engineers (68%)
- Construction managers (64%)
- Health sector (64%)
- Media (64%)
- Business owners (61%)

4.2.3.2. Addressing User Needs

A point on which the experts largely agreed in the first survey is that the relevance of the information provided to users of OEF systems should be determined collaboratively with the user groups themselves. Specifically, there is consensus on the following statements (sorted by percentage of experts with agreement level 6 or 7).

- The way in which earthquake forecasts are communicated to society should be tested and co-designed with the end-users (e.g., civil protection, infrastructure owners, public) using surveys, workshops, or other activities (95%).
- The way in which earthquake forecasts are communicated to society should be regularly evaluated to check if the end-users' needs are still fulfilled (95%).
- The way in which earthquake forecasts are communicated to society should be discussed informally with the end-users (85%).
- The magnitude threshold above which earthquake/aftershock forecasts should provide a probabilistic assessment of the occurrence of earthquakes depends on building structure and vulnerability in the region (75%).
- The magnitude threshold above which earthquake/aftershock forecasts should provide a probabilistic assessment of the occurrence of earthquakes depends on end-user preference/needs. (71%).

No consensus but a tendency toward undecidability was reached on the idea that the way in which earthquake forecasts are communicated to society should be defined by the model developers (67% agreement levels 3–5) or follow best practices from other countries (55% agreement level 3–5). A tendency toward agreement (59% agreement level 6 or 7) was reached for the idea that it should be defined by the entity who provides the forecast.

4.2.3.3. Communication Guidelines

Given the strong consensus among experts that forecast communication products should be developed collaboratively with the end-users of forecasts, the results of the following subsection should be taken with caution. The experts were asked about their opinion on specific communication guidelines based on their experience. These results thus show nothing more than that: experience-based assessments of experts of what may be relevant to end-users. They can serve as a valuable starting point for the collaborative process between those designing forecast communication products and the end-users.

For the following user groups [x-axis] the following pieces of information [y-axis] are important.

	General public	Emergency responders	Civil protection	Critical infrastructure providers	Decision makers	Insurance companies	Duty seismologists
Earthquake probabilities	82%	82%	94%	100%	94%	100%	94%
Absolute earthquake rates	71%	82%	88%	82%	82%	88%	88%
Relative earthquake rates (relative to a normal day)	59%	47%	53%	47%	65%	82%	65%
Temporal evolution of the earthquake probabilities/rates	88%	94%	100%	100%	100%	100%	94%
Spatial distribution of the earthquake probabilities/rates	94%	88%	100%	100%	100%	94%	88%
Earthquake hazard/expected ground motion	82%	88%	100%	100%	94%	82%	65%
Earthquake risk	71%	65%	94%	82%	100%	100%	53%
Uncertainties in probabilities/rates	59%	65%	65%	88%	82%	100%	94%
Uncertainties in earthquake hazard/expected ground motion	59%	59%	88%	94%	88%	88%	65%
Uncertainties in earthquake risk	47%	53%	77%	77%	88%	94%	53%

Figure 22. Important pieces of information per user group [in percentage]. Cells are highlighted according to the percentage of experts that indicated relevance, in shades between red ($\leq 30\%$) and blue ($\geq 70\%$). Cells with consensus ($\leq 30\%$ or $\geq 70\%$) are further highlighted with a black border (Mizrahi, Dallo, & Kuratle, 2023).

Figure 22 shows the results of a question posed in the second survey regarding which pieces of information are considered important for which user groups. Analogously to the previously discussed question on the relevance of model ingredients for different user groups, the survey participants were given a matrix of user groups and information pieces and were asked to indicate with a binary response (yes or no) whether “For the following user groups [x-axis] the following pieces of information [y-axis] are important.” Information pieces include earthquake probabilities and rates, variations thereof in space or time, uncertainties, and other metrics such as hazard and risk, based on the responses to the first survey. Note that an important quantity, the magnitude, is missing in this list. It is assumed that a forecast is given for a magnitude range, thresholds for which were assessed through the statements in Addressing User Needs. The user groups are the same as in Figure 18.

Highlighted with a black border are the pieces of information that at least 70% of the survey participants agreed to be relevant for corresponding user groups, that is, there is a consensus among experts that this information item is important for this user group. For insurance companies, all provided pieces of information are considered important. Earthquake probabilities, absolute earthquake rates, and the spatial and temporal distribution of earthquake probabilities/rates are considered to be important for all user groups. Other information pieces did not reach the 70% consensus threshold for all user groups, though it is noteworthy that the percentage of agreement (level 6 or 7) lies above 47% for all matrix entries. No information piece is considered not important for any user group, which would have been indicated by at most 30% agreement of the expert group.

When asked about the ideal timing and regularity of forecast communication, there was consensus that “Ideally, earthquake forecasts should be permanently communicated to the society.” (82%). No consensus was reached on the following statements.

- Ideally, earthquake forecasts should only be communicated to society after earthquakes of a certain magnitude, that is, aftershock forecasting (64% disagreement).
- Ideally, earthquake loss forecasts would also be communicated to society. (65% agreement).

In the workshop, it was clarified that communicating forecasts regularly (not just after a significant event) is preferable. However, if not possible, providing forecasts only after the occurrence of significant events is a viable alternative. The idealizing formulations of these statements were chosen to replace imperative formulations of the first survey after the workshop discussions revealed that in some cases it may be desired by agencies but not possible due to missing resources to continuously disseminate earthquake forecasts.

Furthermore, the expert group was asked for their opinion on other specific communication guidelines based on their experience. Below are statements on which consensus was reached (sorted by percentage of experts with agreement level 6 or 7).

- Earthquake forecasts should be provided together with an explanation on how to interpret the numbers (95%).
- Scenarios should be used to communicate earthquake forecasts (e.g., most likely and least likely scenario) (79%).
- Earthquake forecast probabilities should be translated into recommended actions target audiences can/should take (70%).
- [almost consensus] Earthquake forecasts should be part of rapid impact assessment reports after an event (e.g., integrate it on rapid impact assessment leaflets such as PAGER) (69%).

In their comments, the experts clarified that the scenario-based forecast communication used in New Zealand and the United States (see Table 4 and Figure 11) could be tested in countries aiming at providing earthquake forecasts: (a) high probability sequence decays; (b) medium probability of similar level of shaking; and (c) low probability high-impact scenario. It was further mentioned that the time frame of the forecast might change with the time since the mainshock. One expert also stated that the scenarios should depend on user needs; the general public for example, may be more interested in a worst-case scenario, whereas emergency managers need to know what will realistically occur in the next week and month to organize the disaster response.

The experts furthermore had the chance to indicate how they would combine forecasts with existing communication products. Their answers revealed that (a) a one-page summary for key stakeholders might be useful; (b) the forecasts should come from the same source as the other communication products (e.g., notifications, rapid impact assessments); (c) a menu of available products on a website/app allows user groups to pick the products they need; (d) the forecast or a link to the forecast should be provided in the general earthquake notification.

Note that these were comments made by individual experts and do not necessarily represent the collective view of the expert group. Also, given the dissent regarding whether earthquake forecasting communication should follow best practices from other countries, these best practices can be a starting point for OEF communication product developers, but need to be tested with the specific user groups of the product in development.

4.2.3.4. Communication Challenges

According to the expert panel, a key challenge that could be relevant when communicating earthquake forecasts is the government/politicians not wanting earthquake forecasts to be publicly available (84% agreement level 6 or 7). Other challenges that were evaluated but on which no consensus was reached are the following (sorted by percentage of experts with agreement level 6 or 7).

- The legal basis to publish forecasts publicly does not exist (62% agree, 13 responses only).
- Civil protection does not want earthquake forecasts to be publicly available (59% agree).
- People struggle to interpret the forecasts (59% agree).
- The forecasts are misused by third parties (56% undecided).
- The public will not be able to correctly interpret the earthquake forecasts (53% undecided).
- The uncertainties of the forecasts are still too high to base any mitigation or recovery actions on (45% disagree).
- It is difficult to combine earthquake forecasts with other available communication products (e.g., earthquake notifications, rapid impact assessments) (60% undecided).

4.2.3.5. Communicating Institution

Experts were also asked who they think should be providing earthquake forecasts to society (Mizrahi, Dallo, & Kuratle, 2023). There was agreement that natural hazard institutions should do so (75%), but no consensus on other institutions such as national or regional authorities (65%), civil protection (62%), scientific institutions (59%), or the media (50%). Participants mentioned that it strongly depends on the country as to which institution is authorized to provide forecasts and is trusted by people. This possibly explains the lack of a clearer consensus on this question.

4.3. Discussion of the Expert Elicitation Results

This subsection summarizes and discusses the key insights gained from the expert elicitation. Note that all insights need to be interpreted considering the limitations of expert elicitations in general and of this expert elicitation in particular, as described in Limitations. Specifically, the insights depend on the composition of the expert group and on the definitions of agreement and consensus.

4.3.1. Model Development

Key insights:

- Considering higher-order aftershocks when developing an earthquake forecasting model is considered relevant independently of the user group for which the forecast is produced.
- Other scientific model ingredients were indicated to be relevant with a less uniform agreement by the experts.
- The inclusion of geodetic information or b -value variations is currently considered not to be important, except possibly for duty seismologists.
- Among the different existing model types, the ETAS model is recommended to be used as a default, and using ensemble models is encouraged.

The importance of different model ingredients in the development of forecasting models varies strongly both when accounting for and when not accounting for different user groups. The key insights drawn from Figure 18 are: (a) accounting for higher-order aftershocks is a model feature that the experts consider important regardless of the user group, and (b) accounting for spatial or temporal b -value variations or geodetic information is consensually considered not important for almost all user groups. The reasons for which a model feature is considered important or not were not assessed as part of the Delphi study. It is however crucial to remark that this assessment is based on the current state of research, and that it reflects the view of a specific group of experts that is not representative of the entire seismological community. Therefore, model features that are currently considered unimportant could become relevant once they are better understood. Specifically, the two topics rated least important by the experts, b -value variations and geodetic information, have recently been shown to be of potentially great interest for forecasting (Bletery & Nocquet, 2023; Gulia & Wiemer, 2019, see Outlook on Model Development). In contrast, extensive research exists on why earthquake forecasting models should, and how they can, account for higher-order aftershocks (e.g., Felzer et al., 2003; Nandan et al., 2019; Ogata, 1988) and catalog incompleteness (Hainzl, 2022; Mizrahi et al., 2021; Omi et al., 2014; Page et al., 2016; Seif et al., 2017), the two model ingredients rated most important. This suggests that experts are inclined to require exhaustive evidence before including additional ingredients in OEF models. This conservative behavior is further evidenced by the wide consensus of using the ETAS model as a default, most likely due to thorough research on its performance (e.g., Cattania et al., 2018; Savran et al., 2022; Strader et al., 2018; Zhuang, 2011). This observation, along with the lack of consensus on the importance of including certain ingredients, such as fault information or historical catalogs, indicates that further research is required to understand if and how they can be incorporated into existing (or yet-to-be-developed) models in a measurably useful way.

The topics rated more important tend to be topics that can be viewed as largely solved. For instance, as shown in Table 2, all three countries with OEF systems discussed in Section 3 have at least one model in place which does account for higher-order aftershocks. Yet, a distinction can be made between those who account “only” for aftershocks of aftershocks that have already been observed, and those who also consider aftershocks of possible future aftershocks that have not yet been observed and need to be simulated. Currently, the ETAS models used in New Zealand and the United States produce simulated catalogs, while the model applied in Italy neglects this latter type of higher-order aftershocks to save computation time.

A possible reason to explain the lack of consensus on the importance of many model ingredients could be the lack of their availability or harmonization both within and across regions; this issue is particularly notable in fault information, historical seismicity, and geodetic data.

Furthermore, experts encourage the use of ensemble models, possibly because they view it as a convenient way to exploratively include innovative or unconventional models into current OEF systems. If model weights are determined and regularly updated based on the recent performance of the ingredient models (e.g., Herrmann & Marzocchi, 2023), including such models can be beneficial to the forecast while poorly performing ingredient models will quickly be down-weighted and their negative impact thus minimized.

The lack of consensus on model updating strategies could be due to differences between the models. The R&J model depends only on the mainshock magnitude and the model parameters. With about an order of magnitude uncertainty in aftershock productivity between sequences within a single tectonic environment (Page et al., 2016) the forecasts are highly uncertain unless the model parameters are updated during a sequence. In contrast, the summation term in the ETAS and ETES models represents a degree of adaptation within a sequence without updating model parameters.

4.3.2. Model Testing

Key insights:

- There is dissent among experts on whether there is a need to collectively define the minimum required tests a forecasting model should pass before it can be used.
- As a consequence, clear requirements to an earthquake forecasting model or specific tests it should pass could not be identified.
- The expert panel agreed on the following recommended testing approaches:
 - Comparison to a benchmark model
 - Prospective and/or pseudo-prospective testing
 - Third-party involvement (such as a CSEP experiment)
- It is not considered sufficient if the model developers trust their model, without it having been reviewed by anyone else.
- If a model has been approved to be used for a given purpose, its parameters can be updated when new data become available, but it cannot be applied in a different region without additional testing.
- Transparency and reproducibility of forecasts is encouraged: sharing model source code and archived forecasts for use by the community is desirable.

A result worth some discussion is the dissent among experts on whether there is a need to collectively define a set of minimum required tests a model should pass before it can be used for forecasting. Several possible causes for this dissent were suggested by individual experts during the Delphi process, although their explanations varied significantly.

One possible cause is the fact that the experts have largely different kinds of experience in earthquake forecasting—some work at institutions that have specific approval procedures that need to be followed before a model can be used to issue forecasts. Others are more experienced in the theoretical aspects of model testing. These are two highly differing perspectives on the subject: On one side are those who develop tests to fairly compare models, agnostic of the model specifics; on the other side are those facing institutional approval procedures that precede operational work. Among such a diverse group, consensus on the topic is a difficult goal to reach.

Another possible distinction of perspectives is between a user-centric perspective and a science-centric perspective. Science-centric experts could argue that independently of how a model is used, there should be a way to definitively rule out models that perform too poorly, and are too far from describing reality. User-centric experts in comparison may argue that depending on the way a model is used, it should undergo different kinds of tests. Some may even be willing to use untested or imperfect models, based on the value of being able to provide a time-varying forecast at all. If the alternative is providing little or no information to users about an aftershock sequence then even a simple model, for example, R&J, may support improved decision-making.

The cause for dissent may also lie within the currently existing tests themselves: The passing of specific tests such as the CSEP N- or S-tests were not consensually considered strict requirements for forecasting models. The hesitancy of some experts to define such strict requirements might stem from the current lack of a test that is basic enough to be passed by all models considered sensible by these experts, without the test itself being meaningless. Sometimes, a model might fail an N-test and still provide information that can be useful to user groups. Furthermore, traditional CSEP tests relying on a Poissonian temporal occurrence model (Schorlemmer et al., 2010; Werner et al., 2010; Zechar, Gerstenberger, & Rhoades, 2010; Zechar, Schorlemmer, et al., 2010) have proven not to be adequate to evaluate forecasts that account for the over-dispersion of seismicity rates (e.g., Lombardi, 2014; Nandan et al., 2019), which may hinder the experts' confidence in using CSEP tests to rule out a model. This has been addressed by Savran et al. (2020), by designing catalog-based evaluations that account for

the inherent seismicity rate variability and how a forecast accounts for such. These tests have not yet been applied in truly prospective CSEP experiments, but are intended by the community to be deployed in the future.

But even more abstract tests such as “being published in a peer-reviewed journal” were not uniformly received by the expert group. When asked to rate statements about where the testing authority lies, there was consensus among the expert group that it is unclear whether a forecast is ready to be used if the users of the forecast approve. This, again, could be rooted in the different realities in which the experts operate: some are part of agencies with clearly defined pipelines for model approval, others are researchers whose day-to-day work is to develop and refine forecast model tests.

Furthermore, while CSEP experiments entail many of the ideas on which the experts reached consensus, such as the desirability of prospective testing, comparative testing between models, and third-party involvement, there has not been true replicability of model results across regions so far, despite near-universal consensus on the advisability thereof. That is, a model's formulation and computational implementation have not been simultaneously used by different OEF systems. Although benchmark comparisons are common, different studies use different benchmark models (e.g., Bayona et al., 2023; Cattania et al., 2018; Dascher-Cousineau et al., 2023; Nandan et al., 2021; van der Elst & Page, 2017). Based on the result that the expert group considers the ETAS model the preferred default model, it would make a good candidate benchmark model. However, various ETAS implementations exist and our study did not aim to identify a specific benchmark model or model implementation. The expert-encouraged sharing of model source codes with the community may facilitate the establishment of a community-approved benchmark. At the same time, along with making test results accessible, source code sharing may encourage OEF teams to expand their model pools and cross-validate their models' reliability.

4.3.3. Forecast Communication

Key insights:

- Forecast communication products should be developed in collaboration with the user groups of the product.
- Earthquake forecasts...
 - are relevant for a wide range of user groups.
 - ideally should be permanently communicated.
 - should be communicated together with an explanation for how to interpret them. This includes the use of scenarios and, possibly, translation of probabilities into potential impacts and recommended action (e.g., “*There is a 30% probability of a M7 event, which could cause severe shaking and falling masonry. Ensure you know how to take protective action, and have your preparedness items in order.*”).
- A key challenge in OEF communication is that the government or politicians may not want earthquake forecasts to be publicly available.

The first point in the above list underscores an inherent limitation of this study: An expert elicitation cannot assess user needs. Earthquake forecasting experts are not the ideal group to decide on specific communication good practices for earthquake forecasting. To truly understand the information that is relevant for users of OEF products, one should ask the users themselves, as has been done in previously mentioned studies (e.g., Becker et al., 2020; Dryhurst et al., 2022; Schneider, Wein, et al., 2023). Possibly, the experts' acknowledgment of the need for interactions with users to understand their needs can explain the responses to the question in Figure 22, where no piece of information was deemed irrelevant for any user group.

The lack of consensus on whether earthquake impact and/or loss forecasts should ideally also be communicated to society could possibly be explained by experts' lack of experience in earthquake loss forecasting communication and the limited research on user needs related to impact or loss forecasts. Experts might make a distinction between the entities responsible for communicating forecasts of geophysical metrics (e.g., earthquake rate, ground motion) and for communicating loss forecasts (e.g., damage, casualties, monetary losses). Civil protection entities might handle the latter, while the former may fall under the purview of scientific institutions. Literature exists on impact-based warnings, and some of this could be used to inform the forecast space, to consider whether impact-based forecasts are a useful aspect of earthquake forecasting.

From a purely logical point of view, one would expect agreement with permanent communication of earthquake forecasts as an ideal to imply disagreement with only aftershock forecasting as an ideal. However, experts agreed to the permanent communication ideal and reached dissent on aftershock forecasting as an ideal. From the

anonymous survey responses, it can be inferred that at least three participants who strongly agreed with the aftershock forecasting ideal must have also agreed to a minimum level of 4 with the permanent communication ideal. This suggests that these individuals might have a preference for aftershock communication but no strong views against the permanent communication ideal.

After this Delphi study, a main challenge of earthquake forecast communication, namely finding the most effective means of communicating low-probability, high-impact events, remains. It is crucial to gain a deeper understanding of how individuals perceive low probabilities and the resulting behavioral actions they take. Several factors contribute to this, including people's inclination to underestimate events that are less immediate or familiar to them, the constraints on their attention and resources, and their sense of psychological distance from these events, which may make them feel less personally affected (Slovic, 2016). Furthermore, there is a need to conduct comprehensive analyses of the current means used to convey information about uncertain futures, that is, most likely and unlikely scenarios (e.g., Schneider, McDowell, et al., 2022). These analyses should focus on assessing their effectiveness in enhancing society's resilience toward low-probability, high-impact events.

4.3.4. Cross-Pillar Reflections

4.3.4.1. Consensus Due To Lack of Expertise?

An interesting observation is that consensus was more often reached among the expert group regarding statements on the communication of forecasts compared to statements on the development and testing of forecasting models. This agreement perhaps reflects the lack of experience of the experts in this field suggested by their self-perceived expertise levels. The dissent in model development and testing strategies could conversely be explained by the higher experience of the experts: they each have their particular viewpoint on which they insist.

4.3.4.2. User-Specificity

The results of the Delphi study provide evidence for two fundamentally different perspectives of expert group members on certain topics covered. Some experts have a more user-centric perspective on earthquake forecasting while others have a science-centric perspective, especially on questions related to the development and testing of forecasting models. Some experts may also have both perspectives depending on the pillar (e.g., user-centric for communication and science-centric for development/testing). The science-centric perspective views earthquake forecasting models as tools to approximate and understand reality, where the perfect model would perfectly describe future seismicity. Under this perspective, any model feature that allows a more precise description of future seismicity is a model feature worthwhile developing. It could not possibly cause any harm, regardless of who the user groups of the forecast will be. The user-centric perspective on the other hand views earthquake forecasting models as tools to serve the needs of specific user groups. If the user group is interested in understanding the risk of a specific building or structure, spatial details of the forecast are crucial, while for another user group, a spatially more aggregated forecast may be enough.

Possibly, the dissent among the expert group on the questions of whether earthquake forecasting models and tests thereof should be user-dependent can be simply explained through different perceptions of the word “should.” The science-centric experts with the ideal model in mind do not think that the ideal model should depend on the user group; the user-centric experts, who know that resources are limited and that a non-perfect but hopefully useful model can be developed and tested with one focus or another, think that this focus should definitely depend on the forecast's user groups.

5. Outlook

5.1. Future Developments of OEF Systems Worldwide

5.1.1. Italy

The Italian group (INGV, University of Naples) is currently working on two major fronts. First, improving the performance of the current OEF-Italy system by including automated algorithms that address (a) the incompleteness issues, and (b) the uncertainty related to the estimation of the models' parameters. More precisely, as regards the former point, the group is working on including in the system the RESTORE algorithm by Stallone and Falcone (2021), which accounts for Short-Term Aftershock Incompleteness (STAI). This algorithm

implements a stochastic gap-filling method that detects STAI gaps and reconstructs the missing events in a space-time-magnitude domain, thus extending the work by Zhuang et al. (2017, 2020), which replenishes the portions of an incomplete seismic catalog through empirical functions describing only the time–magnitude range of missing data. In this way, the problem related to the underestimation of the expected seismicity due to the high incompleteness is overcome. As regards point (b), the possibility in OEF-Italy of estimating the model's parameters by means of a Bayesian procedure, as proposed by Omi et al. (2014), is being discussed to reduce uncertainty in the forecasts (Michael et al., 2020; van der Elst et al., 2022). During the sequences of L'Aquila 2009 and Pianura Padana Emiliana 2012, a first attempt at a daily calibration of the OEF-Italy models was made but, in both cases, an overestimation of the events' number in the tails of the sequences was observed. This is likely due to the very large amount of data used to calibrate the models, thus freezing the estimation of the parameters for a certain long time such that the modeled temporal decay did not reflect the effective course of the sequence. It is also worth mentioning that the Bayesian estimation of the models' parameters would imply a higher computational cost. On foot of these reasons, before any adjustments to the estimation techniques, a discussion with experts in the field will be opened. This will also allow addressing the need, raised by the expert elicitation, of updating the parameters when new data become available, once the model has been approved to be used for the specific purpose.

The second front of currently ongoing work is related to communication. This is a delicate point, as the Italian legal system does not clearly define scientists' roles and responsibilities on the information delivered. The possibility of releasing the forecasts to the public and the best way to disseminate and explain them is being discussed in periodic meetings with the Italian Civil Protection. This meets the experts' agreement, from the Delphi study, that the way in which earthquake forecasts are communicated to society should be tested and co-designed with the end-users, provided that Civil Protection and the general public represent indeed two of the main Italian user groups for which forecasts are relevant. The key statements on which the Delphi expert group reached consensus, such as the need to communicate the forecasts together with an explanation for how to interpret them, will be proposed as food for thought. The international exchange on how to communicate earthquake forecasts will be a valuable tool on which future decisions will be based.

Concurrently with these two fronts, the group is working to overcome some technical problems related to computational time. To date, every 15 min the system checks the earthquake catalog, recorded in the INGV Seismic Monitoring Room in Rome, for the occurrence of any $M_L \geq 3.5$ event, after which a forecast is to be delivered (besides the programmed midnight ones). It is necessary to reduce this time, because it introduces a temporal delay that may entail underestimation (e.g., the check of the catalog is at 00:00, and an $M_L \geq 3.5$ event occurs at 00:01).

Future work is to include additional models in the ensemble of the OEF-Italy system, such to improve its performance skill. In line with the recommendation of the expert elicitation, the plan is first to include ETAS models based on simulated catalogs, such as to account for higher-order aftershocks.

Models that explicitly account for the new ingredients that were positively welcomed by the expert group of the Delphi study, like the catalog incompleteness as proposed in Mizrahi et al. (2021), will also be considered. Finally, the aim is to include physics-based models (e.g., Mancini et al., 2019) in OEF-Italy, thus allowing the system to calibrate the forecasts according to a different perspective.

On the engineering side, a version of Mantis-K, the OELF system currently running in Italy, that is able to account for damage accumulation in the vulnerability and exposure models it uses, has been developed and is now in the process of testing and calibration.

5.1.2. New Zealand

Consistent earthquake magnitude reports are important for earthquake forecasting and are currently not available in New Zealand. GeoNet has plans to introduce a local magnitude M_{LNZ20} that has been derived to be as consistent as possible with moment magnitude (Rhoades et al., 2021). For the revision of the NSHM, this new magnitude was calculated for earthquakes since the early 2000s with sufficient digital waveform recordings, and regression relations were derived for earlier magnitude estimates. The model parameters of the individual models need to be re-estimated to account for the new magnitudes. Also, the New Zealand forecast testing center is in the process of being revived. The composition of the hybrid model components will be kept continually under review. Since the construction of the present model more than 5 years ago, there have been several new insights that have the

potential to affect the construction of the hybrid. Recent innovations in EEPAS modeling include compensation for the time lag (Rhoades & Christophersen, 2019) and the lead time (Rhoades et al., 2020), as well as improved understanding of the space-time trade-off of precursory seismicity (Rastin et al., 2021) and the important effect of the regional strain rate or long-term earthquake rate on time scaling in the EEPAS model (Christophersen, Rhoades, & Colella, 2017; Christophersen, Rhoades, Gerstenberger, et al., 2017; Rhoades et al., 2022). When these new insights have been optimally incorporated into the EEPAS model, it will be necessary to review the medium-term component of the hybrid model and perhaps the Avmax form of the hybrid model. Recent studies of distributed seismicity in the revision of the NSHM have drawn to attention the question of whether strain-rate models, which contribute to the long-term component of the present hybrid, are well-correlated with earthquake occurrence in the long-term, that is, for more than a decade or two around the time of the geodetic observations on which they are based (Rastin, Rhoades, Rollins, & Gerstenberger, 2022), and have also called into question the previous selection of smoothed seismicity models contributing to the long-term component. All these issues point to the necessity for a thorough re-evaluation of the hybrid model components in the future.

For the future, it is planned to shift the emphasis from responding to major events to issuing regular national forecast updates for a range of time periods. Such forecasts can be produced using the present HFT. In quiet times, the medium-term and long-term components of the hybrid model will dominate the forecasts.

5.1.3. United States

The U.S. Geological Survey (USGS) is currently developing the ETAS model for use in the automated forecast system and is developing regionalized parameters for that system. This step will put USGS efforts more in line with results from the Delphi study that higher-order aftershocks are important and ETAS is the preferred single model. The first step will be a switch to temporal ETAS. Unlike some of the existing ETAS systems, the USGS system includes Bayesian updating of parameters during a sequence and includes the triggering from future, simulated, events. Both steps require attention to computational efficiency. ETAS parameter estimation can be difficult and must be reliable in an automatic system, and including the impacts of future, simulated events requires attention to avoid super-critical parameters. Thus, both approaches complicate the implementation and need careful consideration. Whether these, and other operational choices, improve the forecasts will be determined by pseudo-prospective testing. Later, spatio-temporal ETAS and map-based forecast information may be added, as is currently delivered for the international forecast product. The USGS also plans to eventually provide automatic aftershock forecasts following (possibly, a subset of) large and/or damaging international earthquakes on its website, but these plans have not yet been finalized.

Ongoing work regarding forecast communication includes a cross-country study that seeks to understand the user needs of stakeholders from multiple user groups (e.g., emergency managers, civil engineers, geoscientists, public information officials) in three countries (the United States, Mexico and El Salvador) to design novel forecast graphics and maps, for future testing in a user experiment (Schneider, Cotton, & Schweizer, 2023). This work starts by explaining to users what aftershock forecasting can provide and then eliciting from the users what decisions they make during aftershock sequences, what information would support those decisions, and how they want to receive that information. This, along with previous user research, will inform any updates to USGS forecast communication products.

5.1.4. Switzerland and Europe

The Swiss Seismological Service (Schweizerischer Erdbebendienst, SED) at ETH Zurich is developing and testing earthquake forecasting models for Switzerland and Europe, and developing the IT infrastructure and communication products required to disseminate the forecasts produced by these models regularly. In the case of Switzerland, the development of OEF capabilities is part of a larger ongoing initiative aimed at creating a “dynamic, harmonized and user-centered earthquake risk framework” (see Böse et al., 2023). The SED's ongoing efforts align closely with the recommendations of the expert panel derived from the Delphi study. The aim is to develop simple models that are consistent with the existing long-term forecasting models which underlie the respective hazard models (i.e., Danciu et al., 2021; Wiemer et al., 2016). For the European region, it is important to note that the planned harmonized forecast is not meant to overrule or replace local forecasts where these are available. Rather, they enable cross-national or cross-regional comparisons to be made.

In two studies of Han et al. (2024) and Mizrahi et al. (2024), different variants of the ETAS model are considered, which is one of the recommendations of the expert group. ETAS models naturally account for higher-order aftershocks, which was the model ingredient rated most important in the expert elicitation. The second most important information to consider during model development, catalog incompleteness, is also addressed by allowing temporal or spatio-temporal variations of the completeness magnitude of the catalogs used for calibrating the models, using the approach of Mizrahi et al. (2021).

To test the suitability of the models, retrospective and pseudo-prospective tests are conducted. For small magnitude events for which more data are available, pseudo-prospective tests comparing different model variants to time-independent benchmarks are performed, revealing the usefulness of the time-dependent nature of these models as well as advantages and drawbacks of individual model variants compared to each other. The relatively low number of large magnitude events in Switzerland makes it challenging to obtain statistically meaningful results from pseudo-prospective tests about the occurrence frequency of large magnitude, hence relevant, events. Thus, to complement the pseudo-prospective tests, retrospective tests are conducted for further insight. Once the forecasting models are agreed upon and forecasts are produced automatically, these forecasts will be stored and archived to enable prospective testing, in agreement with the recommendations of the expert group. The code base used for calibrating these models and for calculating forecasts is available online (<https://github.com/lmizrahi/etas>; Mizrahi, Schmid, & Han, 2023).

The main target group of the earthquake forecasts are professional stakeholders often familiar with risk and emergency management but not specifically trained to handle seismic crises, as well as the public. Communication products must be designed accordingly. While the detailed communication products are yet to be determined, the SED is planning to follow the recommendations provided through the Delphi study presented in this article. This means that the communication products will be designed and tested with the target groups to ensure a correct interpretation of the forecasts. To this end, public surveys and workshops with professionals will be conducted, for instance with members of the Swiss Federal Nuclear Safety Inspectorate (ENSI) and SED-internal duty seismologists, who are interested in receiving information tailored to their specific needs. Different forecast communication formats will be tested, including best practices from other countries (e.g., using scenarios). This will enable an evaluation of the effectiveness of these products for Swiss and European stakeholders, as well as their inclusivity, considering factors such as accessibility and information processing skills.

The expert group agreed that regular forecast release would be ideal. The main reasoning behind this recommendation is that users can familiarize themselves with the forecasts and the type of information provided and get an idea of what a forecast looks like on a normal day. This would provide them with a mental benchmark to compare to on a day with increased earthquake probabilities. The anticipated usual update frequency of one day will be temporarily increased when a significant event occurs. However, user tests must show whether these recommendations prove sensible and lead to an improved understanding of OEF.

After the establishment of basic OEF systems for Switzerland and Europe, the SED is considering the development of several refinements to these systems. These refinements include, but are not limited to, (a) adding a mechanism for sequence-specific model updating, (b) estimating and communicating the implications for hazard and risk that result from temporally elevated earthquake probabilities, (c) continued research on the development of superior forecasting models by including more information on earthquake physics, or by exploring alternative or complementary models for earthquake forecasting using machine learning (ML) techniques.

5.2. Outlook on OEF-Relevant Research

5.2.1. Outlook on Model Development

The earthquake forecasting models presented so far are those currently applied for OEF in practice. In this subsection, we provide a brief overview of emerging forecasting techniques, and direct the interested reader to a comprehensive discussion by Hardebeck et al. (2024) in their recent review on aftershock forecasting. As pointed out in previous sections, emerging models must undergo thorough testing before they can be considered suitable in an OEF context.

Earthquake forecasting models currently used for OEF are primarily empirical, meaning that they employ observed relationships from past earthquake sequences without considering the physical mechanisms involved in aftershock triggering. A widely accepted possible explanation for earthquake triggering is provided by the static

stress transfer hypothesis (Harris & Simpson, 1992). Its idea is that a fault can be brought closer to failure due to the redistribution of stresses in the crust caused by an earthquake. Coupled with the concept of rate-state-friction, which can be used to describe the temporal evolution of seismicity rate (Dieterich, 1994), Coulomb-rate-state (CRS) seismicity models have been established and refined recently, showing comparable skill to ETAS models and promising potential for improvement (Cattania et al., 2015, 2018; Mancini et al., 2019, 2020).

However, CRS models suffer from the same main limitation as empirical models: they are clustering models, most successful at modeling the evolution of seismicity without encoding information on the magnitude of the next earthquake, which is assumed to be a random sample from a magnitude-frequency distribution. Another approach is to look for precursory signals that only occur prior to large events or scale with the size of the impending earthquake. The search for precursory signals has, so far, not yielded any findings applicable for forecasting, but is a field of ongoing progress. One type of precursor, which is not controversial (but may not be predictive of future earthquake size), is foreshocks, that is, smaller earthquakes that occur prior to a larger one. There are currently two conceptual models that explain the occurrence of foreshocks: the *cascade model* and the *pre-slip model* (e.g., Dresen et al., 2020; Ellsworth & Beroza, 1995; McLaskey & Lockner, 2014; Mignan, 2014). In the former, foreshocks are regular earthquakes that happen to trigger a more significant event. In the latter, they do not trigger each other, but are a byproduct of the nucleation process of a mainshock. A recent study by Bletery and Nocquet (2023) appears to support the pre-slip model. The authors claim to have found evidence in GPS time series for a two-hour-long exponential acceleration of slip leading up to large earthquakes. As pointed out by the authors, this observation might be the very end of a much longer process of precursory slip. Their result is, however, controversial, and ongoing follow-up analyses suggest that common-mode errors among the GPS stations may cause these observations. Other precursors such as ionospheric disturbances prior to large earthquakes (Heki, 2011) have been claimed to be detected, though the robustness of this result is widely debated (Kamogawa & Kakinami, 2013; Masci et al., 2015). A general problem with the analysis of precursory signals is that they are mostly done in retrospect after a large earthquake has already occurred. Many parameters of a precursor analysis depend on the knowledge of information about the impending earthquake such as its epicenter or direction of slip. Moreover, phenomena such as ionospheric disturbances happen very often. Even if they did systematically occur prior to large earthquakes, their occurrence does not guarantee a large earthquake to be in preparation.

The topic of false or missed alarms was also relevant in the debate (Dascher-Cousineau et al., 2020, 2021; Gulia et al., 2020; Gulia & Wiemer, 2021) following the proposition of the foreshock traffic-light system (FTLS) by Gulia and Wiemer (2019). After a large earthquake, this system analyzes the temporal variation of the *b*-value to decide whether the earthquake is likely to be followed by a larger one (making the first earthquake a *foreshock*) or not. Not only after the occurrence of a large event, a change in *b*-value has been proposed to be a candidate precursor of impending large earthquakes, supported by theoretical considerations and observations in the laboratory and the field (El-Isa & Eaton, 2014; Imoto, 1991; Main et al., 1989; Mogi, 1962; Scholz, 1968; Schorlemmer et al., 2004, 2005; W. D. Smith, 1981). Laboratory experiments highlight a *b*-value decrease before the main-slip events, followed by a post-rupture increase (Bolton et al., 2020; Goebel et al., 2013, 2015; Jiang et al., 2017; Johnson et al., 2013; Lei & Ma, 2014; Main et al., 1989; McLaskey & Lockner, 2014; Rivière et al., 2018). Gulia et al. (2018), inspired by observations from both laboratory and single case studies (Ogata & Katsura, 2014; Tamaribuchi et al., 2018; Tormann et al., 2012, 2014, 2015; Wiemer et al., 2002; Wiemer & Katsumata, 1999), conducted a systematic analysis that confirmed this behavior in several worldwide aftershock sequences. This led to the insight that is at the basis of the FTLS: the *b*-value evolution, analyzed as a proxy for the average stress condition of a fault, can act as a first-order discriminator between normal aftershocks and likely precursory sequences. In cases where the *b*-value does not increase, but decreases, a second larger event is likely to happen. While in principle this is a promising approach, estimates of the *b*-value vary naturally and as a consequence of several sources of bias (Marzocchi & Sandri, 2003; Marzocchi et al., 2020; Shi & Bolt, 1982), and further testing is required to turn color alerts into probabilistic forecasts that can be compared to existing models.

The field of earthquake forecasting based on ML techniques has been approached with hesitation after a controversial model proposed by DeVries et al. (2018), which forecasted aftershock locations using deep learning, was revealed to be as informative as a far simpler parameterization (Mignan & Broccardo, 2019). More recently, however, ML techniques are gaining traction as multiple research teams have developed simple ML models that can match or exceed the performance of traditional forecasting approaches. Researchers at Google proposed the Forecasting Earthquake Rates with Neural networks (FERN) spatio-temporal model and found that it performed slightly better than the ETAS model (Zlydenko et al., 2023). Similar approaches proposed by

Stockman et al. (2023) and Dascher-Cousineau et al. (2023) likewise found improved performance of temporal neural point process models relative to the ETAS model, particularly when more earthquakes were used in training. For large data sets, these models are faster to train than ETAS. Another noteworthy advantage of neural point process models is that they are extremely adaptive to nonstationarities in earthquake catalogs. In particular, these models can make use of the small, numerous earthquakes available in recently developed ML-based earthquake catalogs (e.g., M. Liu et al., 2020; Ross et al., 2018), even when these cataloged earthquakes are not complete (e.g., Herrmann & Marzocchi, 2021; Mancini et al., 2022).

Although earthquake catalogs are improving continuously, the large earthquakes that drive hazard and risk remain rare. To compensate for the paucity of well-recorded large earthquakes, small earthquakes, which are more abundant, are studied and the relationships identified for these events are extrapolated to forecast larger earthquakes. The simplest example for this is to use the Gutenberg-Richter relationship to estimate the frequency of large (rare) earthquakes based on the relative frequency of smaller (more abundant) ones. Besides high-resolution catalogs of natural earthquakes, human-made laboratory earthquake catalogs can nowadays be produced to study the earthquake processes and the evolution of seismicity in a controlled environment (e.g., Gischig et al., 2020; Lei & Ma, 2014; Selvadurai et al., 2023). The level at which knowledge about laboratory earthquakes can be applied to natural earthquakes is not fully understood and is another topic of ongoing research (e.g., Q. Xiong et al., 2023). Rouet-Leduc et al. (2017) proposed a ML model to estimate the remaining time to failure in a laboratory fault mimicking the earth's faulting and found promising results relevant for the advancement of natural earthquake forecasting.

5.2.2. Outlook on Model Testing

There is a general consensus among researchers on the availability of seismicity models for pseudo-prospective and retrospective analyses that can help institutions identify candidate models for OEF. However, given the strongly stochastic nature of earthquakes, only prospective evaluations can be considered rigorous enough to assess the predictive skills of these models and, consequently, to build confidence around them (Jordan, 2006). This long-term task involves experimenting with earthquake forecasting models over reasonably long periods (e.g., decades), across multiple tectonic regions, and through a transparent and reproducible framework. Experimentation time scales require the testing frameworks to be persistent during the constant evolution of programming practices and changes of the researchers involved in forecasting and testing. Therefore, CSEP is currently devoting efforts to improve earthquake predictability research through open, long-lasting science practices. The first of these, pyCSEP (Savran et al., 2022), is an open-access software toolkit designed for earthquake forecast users to represent, visualize, and evaluate seismicity forecasts. pyCSEP's software development is community-oriented, meaning that most of its utilities are continually contributed by CSEP members and reviewed by the broader earthquake forecasting community.

Second is the implementation of reproducibility packages, which are sets of code, data and other resources needed to reproduce the computations, results and figures described in CSEP-related publications (e.g., Bayona et al., 2022, 2023; Khawaja et al., 2023). These packages can act as practical guides for students and non-experts to build on their original work, as they are fully documented and easy to run on local computers, provided sufficient computational power. Third is the development of a "software ecosystem," within which the source codes of experiments, forecasts, and tests become interoperable. This would allow the seamless deployment of prospective earthquake forecast experiments and support the evaluation of OEF systems. Using these platforms, agencies could compare OEF models against each other or with other experimental/benchmark models. At present, CSEP is working on a novel experiment format, managed by an application called floatCSEP (Iturrieta et al., 2023), which decentralizes the testing process carried out at CSEP servers and applies best-practice principles in open science software development and data management. This new format ensures the reproducibility of any forecast experiment and can be the basis for future experiments developed by both CSEP and independent researchers.

Finally, CSEP's prospects for future prospective testing involve diversifying statistical methods, forecast formats, and input data. With regard to testing methods, it is important to recognize that each method or metric is designed to assess only a particular aspect of a forecast, with inherent limitations arising from such reductionism (Iturrieta et al., 2024; Serafini et al., 2022). Hence, the testing procedure should include a battery of tests that are as diverse as possible, each of which should be clearly defined theoretically and computationally. CSEP recently introduced,

for example, new likelihood tests that consider synthetic catalogs simulated from seismicity models instead of synoptic earthquake rate/probability maps, thus relaxing the Poisson assumption in traditional CSEP tests (see Savran et al., 2020). Additionally, Bayona et al. (2022) introduced a set of new consistency tests that depend on a binary (or Bernoulli) likelihood function, which is less sensitive than the Poisson distribution to the spatiotemporal clustering behavior of earthquakes. When it comes to forecast formats, Asim et al. (2023) introduced a new data-driven approach that maps single-resolution earthquake forecasts to multi-resolution, Quadtree forecasts, thus reducing computational burdens and data storage issues.

In the near future, CSEP envisions to evaluate a more diverse range of earthquake forecasts, including earthquake-source forecasts; much of which are relevant to time-independent and time-dependent seismic hazard assessments. Other action plans that could be particularly useful for OEF could be to prospectively evaluate RichterX models (i.e., Kamer et al., 2021; Nandan et al., 2017) using CSEP metrics, strengthen collaborations between multiple earthquake forecasting communities (e.g., those that make use of the RESTORE (Stallone & Falcone, 2021) and NESTOREv1 (Gentili et al., 2023) software toolboxes), and diversify the pool of data/models that describe the occurrence of earthquakes. This will only be possible if data are openly accessible and if model source codes are freely available and fully documented. Releasing data and model source codes provides substantial benefits since, for example, the “reliability” of a model could be replicated across regions and the underlying hypotheses, if encoded as parameters in the algorithm, could be subjected to significance testing by modifying the parameter space. In addition, given that the code would be open to scrutiny, potential errors could be routinely identified and corrected by externals. Finally, making models available in a rigorous and standardized manner would also provide benchmarks against which new models can be compared during their development phase. Thus, we expect open-science approaches to dramatically increase the improvement of OEF models and, hence, our knowledge of the earthquake generation process.

5.3. Conclusions

In summary, this review captures the current state of and future perspectives on operational earthquake forecasting (OEF). The overview of existing OEF systems in conjunction with expert views represent a useful resource for agencies with OEF systems in place, who can use this as a tool for comparison and inspiration. Countries aiming to newly establish an OEF system may use it for guidance on a complex journey of developing, testing, and communicating earthquake forecasts.

In Section 2, we provide an overview of the research relevant for OEF. We cover the aspects of model development, where different model types and variants are described, model testing, where common means of forecast testing are discussed, and forecast communication, where relevant scientific work that ensures effectiveness of the provided forecasts is summarized.

In Section 3, our overview of the OEF systems of Italy, New Zealand, and the United States, also structured according to the three pillars of development, testing, and communication of earthquake forecasts, shows a large heterogeneity between the three countries. Although Epidemic-Type Aftershock Sequence (ETAS) models are used in some way by each country, the specifically used models differ substantially (i.e., none of the model specification questions are answered uniformly by all countries in the model development summary table (Table 2)). Similarly, their models were tested in different ways, comprising a mix of prospective and retrospective testing, with experiments run by the Collaboratory for the Study of Earthquake Predictability (CSEP) using CSEP test metrics being a common theme. Although their core common purpose is to increase societies' resilience against earthquake hazard, these three OEF systems evolved in countries with different seismicity, with different target user groups in mind, and that lived through different earthquake sequences which shaped forecasting needs. Naturally, this manifests in heterogeneous forecast communication products used by the different countries.

Section 4 of this review aimed to find points of consensus within the heterogeneity of views elicited by the Delphi study (Mizrahi, Dallo, & Kuratle, 2023). Results that emerged from the expert elicitation are not surprising new insights, but a collection of statements to which at least 70% of experts who participated in the study agree. A main result of the expert elicitation is that prospective testing (as opposed to retrospective testing) and benchmark comparisons are encouraged, as are transparency and reproducibility. The widely used ETAS model is identified as the preferred simple default model and therefore makes a good benchmark. Because no two implementations of the model are identical, the earthquake forecasting community faces the challenge that a truly standardized benchmark version of the ETAS model is currently lacking. Community-driven efforts to share source codes and

prospectively issued forecasts could facilitate the establishment of a standard benchmark model and thus substantially improve OEF research. The most central result regarding the communication of earthquake forecasts is that communication products must be developed in close collaboration with end-users. Thus, while experts working on the development and testing of forecasting models may provide relevant inputs in designing prototype products, the dialog with end-users is indispensable.

Planned future developments of the different countries outlined in Section 5 are again shaped by the individual challenges faced by the three countries and encompass technical improvements to the models as well as revisions of the way in which forecasts are communicated. The future developments of OEF systems worldwide are described out of the perspective of the groups involved in running these systems; a broader picture of the future of OEF-relevant research is given in Outlook on OEF-Relevant Research. Models that are currently used for OEF have a common limitation: they are clustering models and fail to forecast future large earthquakes with high probability. Recent efforts to advance forecasting capabilities include incorporating knowledge about physical processes, identifying precursory signals, utilizing insights from laboratory seismicity or applying ML techniques and novel high-resolution data sets. The Collaboratory for the Study of Earthquake Predictability is devoting efforts to facilitate open, long-lasting earthquake predictability research. This includes providing open-access software toolkits and reproducibility packages, as well as research that supports the diversification of statistical methods and forecast formats used in third-party prospective tests of existing models. Regarding forecast communication, future efforts are needed to systematically test how low-probability high-impact events can be best communicated to societal stakeholders, thus supporting policymakers' decision-making processes and individuals' perception of and reaction to forecasts. In this regard, cross-cultural comparisons can shed light on personal and social factors that influence people's perception and reaction.

Much like future earthquakes are not known today, the future of earthquake forecasting is not known today; perhaps advancements in earthquake forecasting have occurrence patterns similar to those of earthquakes. One day, a big discovery may occur unexpectedly and trigger numerous subsequent discoveries that advance the field. Until then, smaller discoveries keep being made, and possibly, a smaller one can trigger a bigger one.

Acknowledgments

The authors thank the experts who participated in the expert elicitation (Mizrahi, Dallo, & Kuratle, 2023) for their time and effort. The expert elicitation was approved by the Ethics Committee of ETH Zurich (EK 2023-N-38). We also thank the editors and the three reviewers for their valuable comments, which strongly improved the clarity of our article, and Athanasios Papadopoulos for reviewing an earlier version of this article and for the insightful discussions held throughout the writing process. This work was supported by 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), and Grant Agreement Number 101021746, sScience and human factOr for Resilient sociEty (CORE); by the Swiss Federal Nuclear Safety Inspectorate (ENSI); by the European Union project "A Digital Twin for Geophysical Extremes" (DT-GEO) (No: 101058129). OEF-Italy is funded by the Seismic Hazard Center - WP3 Short-Term Probabilistic Seismic Hazard (CPS, INGV). The New Zealand contribution was supported by the New Zealand Ministry of Business, Innovation and Employment (MBIE) through the Hazards and Risks Management programme (Strategic Science Investment Fund, contract C05X1702). Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government. Open access funding provided by Eidgenössische Technische Hochschule Zurich.

Data Availability Statement

The data collected and analyzed in the expert elicitation is available through the ETH Research Collection through <https://doi.org/10.3929/ethz-b-000637239> (Mizrahi, Dallo, & Kuratle, 2023).

References

- Asayesh, B. M., Hainzl, S., & Zöller, G. (2023). Depth-dependent aftershock trigger potential revealed by 3D-ETAS modeling. *Journal of Geophysical Research: Solid Earth*, 128(6), e2023JB026377. <https://doi.org/10.1029/2023jb026377>
- Asim, K. M., Schorlemmer, D., Hainzl, S., Iturrieta, P., Savran, W. H., Bayona, J. A., & Werner, M. J. (2023). Multi-resolution grids in earthquake forecasting: The Quadtree approach. *Bulletin of the Seismological Society of America*, 113(1), 333–347. <https://doi.org/10.1785/0120220028>
- Aspinall, W. (2010). A route to more tractable expert advice. *Nature*, 463(7279), 294–295. <https://doi.org/10.1038/463294a>
- Bayona, J. A., Savran, W. H., Iturrieta, P., Gerstenberger, M. C., Graham, K. M., Marzocchi, W., et al. (2023). Are regionally calibrated seismicity models more informative than global models? Insights from California, New Zealand, and Italy. *The Seismic Record*, 3(2), 86–95. <https://doi.org/10.1785/0320230006>
- Bayona, J. A., Savran, W. H., Rhoades, D. A., & Werner, M. J. (2022). Prospective evaluation of multiplicative hybrid earthquake forecasting models in California. *Geophysical Journal International*, 229(3), 1736–1753. <https://doi.org/10.1093/gji/ggac018>
- Becker, J. S., Potter, S. H., McBride, S. K., Gerstenberger, M., & Christophersen, A. (2018). Effective communication of Operational Earthquake Forecasts (OEFs): Findings from a New Zealand workshop. *GNS Science*. <https://doi.org/10.21420/10.21420/G2DH00>
- Becker, J. S., Potter, S. H., McBride, S. K., Doyle, E. E. H., Gerstenberger, M. C., & Christophersen, A. (2020). Forecasting for a fractured land: A case study of the communication and use of aftershock forecasts from the 2016 Mw 7.8 Kaikōura earthquake in Aotearoa New Zealand. *Seismological Research Letters*, 91(6), 3343–3357. <https://doi.org/10.1785/0220190354>
- Becker, J. S., Potter, S. H., McBride, S. K., Wein, A., Doyle, E. E. H., & Paton, D. (2019). When the earth doesn't stop shaking: How experiences over time influenced information needs, communication, and interpretation of aftershock information during the Canterbury Earthquake Sequence, New Zealand. *International Journal of Disaster Risk Reduction*, 34, 397–411. <https://doi.org/10.1016/j.ijdr.2018.12.009>
- Bird, P., Jackson, D. D., Kagan, Y. Y., Kreemer, C., & Stein, R. S. (2015). GEAR1: A global earthquake activity rate model constructed from geodetic strain rates and smoothed seismicity. *Bulletin of the Seismological Society of America*, 105(5), 2538–2554. <https://doi.org/10.1785/0120150058>
- Bletery, Q., & Nocquet, J. M. (2023). The precursory phase of large earthquakes. *Science*, 381(6655), 297–301. <https://doi.org/10.1126/science.adg2565>
- Bolton, D. C., Shreedharan, S., Rivière, J., & Marone, C. (2020). Acoustic energy release during the laboratory seismic cycle: Insights on laboratory earthquake precursors and prediction. *Journal of Geophysical Research: Solid Earth*, 125(8), e2019JB018975. <https://doi.org/10.1029/2019jb018975>
- Böse, M., Danciu, L., Papadopoulos, A., Clinton, J., Cauzzi, C., Dallo, I., et al. (2023). Towards a dynamic earthquake risk framework for Switzerland. *EGU Sphere*, 2023, 1–33.

- Buisson, L., Thuiller, W., Casajus, N., Lek, S., & Grenouillet, G. (2010). Uncertainty in ensemble forecasting of species distribution. *Global Change Biology*, *16*(4), 1145–1157. <https://doi.org/10.1111/j.1365-2486.2009.02000.x>
- Cattania, C., Hainzl, S., Wang, L., Enescu, B., & Roth, F. (2015). Aftershock triggering by postseismic stresses: A study based on Coulomb rate-and-state models. *Journal of Geophysical Research: Solid Earth*, *120*(4), 2388–2407. <https://doi.org/10.1002/2014jb011500>
- Cattania, C., Werner, M. J., Marzocchi, W., Hainzl, S., Rhoades, D., Gerstenberger, M., et al. (2018). The forecasting skill of physics-based seismicity models during the 2010–2012 Canterbury, New Zealand, earthquake sequence. *Seismological Research Letters*, *89*(4), 1238–1250. <https://doi.org/10.1785/0220180033>
- Christophersen, A., Bourguignon, S., Rhoades, D. A., Allen, T. I., Salichon, J., Ristau, J., et al. (2022). Consistent magnitudes over time for the revision of the New Zealand National Seismic Hazard Model. In *GNS Science Report 2021/56*. GNS Science.
- Christophersen, A., Canessa, S., Huso, R., Gerstenberger, M. C., Harte, D. S., & Rhoades, D. A. (2018). An automated computational-system to support operational earthquake forecasting. In *GNS Science Consultancy Report 2018/163*. GNS Science.
- Christophersen, A., Rhoades, D. A., & Colella, H. V. (2017). Precursory seismicity in regions of low strain rate: Insights from a physics-based earthquake simulator. *Geophysical Journal International*, *209*(3), 1513–1525. <https://doi.org/10.1093/gji/ggx104>
- Christophersen, A., Rhoades, D. A., Gerstenberger, M. C., Bannister, S., Becker, J., Potter, S. H., & McBride, S. (2017). Progress and challenges in operational earthquake forecasting in New Zealand. In *New Zealand Society for Earthquake Engineering Annual Technical Conference*.
- Clements, R. A., Paik Schoenberg, F., & Schorlemmer, D. (2011). Residual analysis methods for space–time point processes with applications to earthquake forecast models in California (pp. 2549–2571).
- Cloke, H. L., & Pappenberger, F. (2009). Ensemble flood forecasting: A review. *Journal of Hydrology*, *375*(3–4), 613–626. <https://doi.org/10.1016/j.jhydrol.2009.06.005>
- Colson, A. R., & Cooke, R. M. (2017). Cross validation for the classical model of structured expert judgment. *Reliability Engineering & System Safety*, *163*(Supplement C), 109–120. <https://doi.org/10.1016/j.res.2017.02.003>
- Console, R., Murru, M., & Lombardi, A. M. (2003). Refining earthquake clustering models. *Journal of Geophysical Research*, *108*(B10), 2468. <https://doi.org/10.1029/2002jb002130>
- Console, R., Rhoades, D. A., Murru, M., Evison, F. F., Papadimitriou, E. E., & Karakostas, V. G. (2006). Comparative performance of time-variant, long-range and short-range forecasting models on the earthquake catalogue of Greece. *Journal of Geophysical Research*, *111*(B9), B09304. <https://doi.org/10.1029/2005JB004113>
- Cooke, R. M. (1991). *Experts in uncertainty: Opinion and subjective probability in science*. Oxford University Press.
- Cornell, C. A. (1968). Engineering seismic risk analysis. *Bulletin of the Seismological Society of America*, *58*(5), 1583–1606. <https://doi.org/10.1785/bssa0580051583>
- Cornell, C. A., & Krawinkler, H. (2000). *Progress and challenges in seismic performance assessment*. PEER Center News. Spring 2000.
- Cremen, G., & Galasso, C. (2020). Earthquake early warning: Recent advances and perspectives. *Earth-Science Reviews*, *205*, 103184. <https://doi.org/10.1016/j.earscirev.2020.103184>
- Dahm, T., & Hainzl, S. (2022). A Coulomb stress response model for time-dependent earthquake forecasts. *Journal of Geophysical Research: Solid Earth*, *127*(9), e2022JB024443. <https://doi.org/10.1029/2022JB024443>
- Daley, D. J., & Vere-Jones, D. (2003). *An introduction to the theory of point processes: Volume I: Elementary theory and methods*. Springer.
- Daley, D. J., & Vere-Jones, D. (2008). *An introduction to the theory of point processes: Volume II: General theory and structure* (pp. 13–14). Springer Science & Business Media.
- Dallo, I., Marti, M., Valenzuela, N., Crowley, H., Dabbeek, J., Danciu, L., et al. (2023). The communication strategy for the release of the first European Seismic Risk Model and the updated European Seismic Hazard Model. *Natural Hazards and Earth System Sciences Discussions*, 1–24. <https://doi.org/10.5194/nhess-2023-107>
- Danciu, L., Nandan, S., Reyes, C. G., Basili, R., Weatherill, G., Beauval, C., et al. (2021). The 2020 update of the European Seismic Hazard Model-ESHM20: Model overview. *EFEHR Technical Report* (Vol. 1).
- Dascher-Cousineau, K., Lay, T., & Brodsky, E. E. (2020). Two foreshock sequences post Gulia and Wiemer (2019). *Seismological Society of America*, *91*(5), 2843–2850. <https://doi.org/10.1785/0220200082>
- Dascher-Cousineau, K., Lay, T., & Brodsky, E. E. (2021). Reply to “Comment on ‘Two Foreshock Sequences Post Gulia and Wiemer (2019)’ by Kelian Dascher-Cousineau, Thorne Lay, and Emily E. Brodsky” by Laura Gulia and Stefan Wiemer. *Seismological Society of America*, *92*(5), 3259–3264. <https://doi.org/10.1785/0220210059>
- Dascher-Cousineau, K., Shchur, O., Brodsky, E. E., & Günemann, S. (2023). Using deep learning for flexible and scalable earthquake forecasting. *Geophysical Research Letters*, *50*(17), e2023GL103909. <https://doi.org/10.1029/2023gl103909>
- de Angelis, L., Godano, C., & Lippiello, E. (2018). The overlap of aftershock coda waves and short-term postseismic forecasting. *Journal of Geophysical Research: Solid Earth*, *123*(7), 5661–5674. <https://doi.org/10.1029/2018jb015518>
- Deichmann, N. (2017). Theoretical basis for the observed break in M_L/M_w scaling between small and large earthquakes. *Bulletin of the Seismological Society of America*, *107*(2), 505–520. <https://doi.org/10.1785/0120160318>
- Detweiler, S. T., & Wein, A. M. (2017). In *The HayWired earthquake scenario—Societal consequences (ver. 1.1, October 2021)*. U.S. Geological Survey Scientific Investigations Report 2017–5013–R–W. <https://doi.org/10.3133/sir20175013v3>
- DeVries, P. M., Viégas, F., Wattenberg, M., & Meade, B. J. (2018). Deep learning of aftershock patterns following large earthquakes. *Nature*, *560*(7720), 632–634. <https://doi.org/10.1038/s41586-018-0438-y>
- Dieterich, J. (1994). A constitutive law for rate of earthquake production and its application to earthquake clustering. *Journal of Geophysical Research*, *99*(B2), 2601–2618. <https://doi.org/10.1029/93jb02581>
- Doyle, E. E. H., Johnston, D. M., Smith, R., & Paton, D. (2019). Communicating model uncertainty for natural hazards: A qualitative systematic thematic review. *International Journal of Disaster Risk Reduction*, *33*, 449–476. <https://doi.org/10.1016/j.ijdrr.2018.10.023>
- Doyle, E. E. H., McClure, J., Potter, S. H., Lindell, M. K., Becker, J. S., Fraser, S. A., & Johnston, D. M. (2020). Interpretations of aftershock advice and probabilities after the 2013 Cook Strait earthquakes, Aotearoa New Zealand. *International Journal of Disaster Risk Reduction*, *49*, 101653. <https://doi.org/10.1016/j.ijdrr.2020.101653>
- Dresen, G., Kwiatek, G., Goebel, T., & Ben-Zion, Y. (2020). Seismic and aseismic preparatory processes before large stick–slip failure. *Pure and Applied Geophysics*, *177*(12), 5741–5760. <https://doi.org/10.1007/s00024-020-02605-x>
- Dryhurst, S., Dallo, I., Luoni, G., Marti, M., & Freeman, A. L. J. (2022). *Field evaluation of OEF communications (Society: Data Gathering and Information Sharing with the Public and Policy-Makers)* [Deliverable]. European Horizon-2020 project RISE.
- El-Isa, Z. H., & Eaton, D. W. (2014). Spatiotemporal variations in the b -value of earthquake magnitude–frequency distributions: Classification and causes. *Tectonophysics*, *615*, 1–11. <https://doi.org/10.1016/j.tecto.2013.12.001>
- Ellsworth, W. L., & Beroza, G. C. (1995). Seismic evidence for an earthquake nucleation phase. *Science*, *268*(5212), 851–855. <https://doi.org/10.1126/science.268.5212.851>

- Erto, P., Giorgio, M., & Iervolino, I. (2016). About knowledge and responsibility in probabilistic seismic risk management. *Seismological Research Letters*, 87(5), 1161–1166. <https://doi.org/10.1785/0220160001>
- Evison, F. F., & Rhoades, D. A. (2004). Demarcation and scaling of long-term seismogenesis. *Pure and Applied Geophysics*, 161(1), 21–45. <https://doi.org/10.1007/s00024-003-2435-8>
- Falcone, G., Console, R., & Murru, M. (2010). Short-term and long-term earthquake occurrence models for Italy: ETES, ERS and LTST. *Annals of Geophysics*, 53(3), 41–50. <https://doi.org/10.4401/ag-4760>
- Felzer, K. R., Abercrombie, R. E., & Ekstrom, G. (2004). A common origin for aftershocks, foreshocks, and multiplets. *Bulletin of the Seismological Society of America*, 94(1), 88–98. <https://doi.org/10.1785/0120030069>
- Felzer, K. R., Abercrombie, R. E., & Ekström, G. (2003). Secondary aftershocks and their importance for aftershock forecasting. *Bulletin of the Seismological Society of America*, 93(4), 1433–1448. <https://doi.org/10.1785/0120020229>
- Field, E. H. (2007). Overview of the working group for the development of regional earthquake likelihood models (RELM). *Seismological Research Letters*, 78(1), 7–16. <https://doi.org/10.1785/gssrl.78.1.7>
- Field, E. H. (2015). All models are wrong, but some are useful. *Seismological Research Letters*, 86(2A), 291–293. <https://doi.org/10.1785/02201401213>
- Field, E. H., Jordan, T. H., & Cornell, C. A. (2003). OpenSHA: A developing community-modeling environment for seismic hazard analysis. *Seismological Research Letters*, 74(4), 406–419. <https://doi.org/10.1785/gssrl.74.4.406>
- Field, E. H., Jordan, T. H., Jones, L. M., Michael, A. J., Blanpied, M. L., & workshop Participants. (2016). The potential uses of operational earthquake forecasting. *Seismological Research Letters*, 87(2A), 313–322. <https://doi.org/10.1785/0220150174>
- Field, E. H., & Milner, K. R. (2018). Candidate products for operational earthquake forecasting illustrated using the HayWired planning scenario, including one very quick (and not-so-dirty) hazard-map option. *Seismological Research Letters*, 89(4), 1420–1434. <https://doi.org/10.1785/0220170241>
- Field, E. H., Milner, K. R., Hardebeck, J. L., Page, M. T., van der Elst, N., Jordan, T. H., et al. (2017). A spatiotemporal clustering model for the third Uniform California Earthquake Rupture Forecast (UCERF3-ETAS): Toward an operational earthquake forecast. *Bulletin of the Seismological Society of America*, 107(3), 1049–1081. <https://doi.org/10.1785/0120160173>
- Freeman, A. L., Dryhurst, S., & Luoni, G. (2023). *Good practice recommendations report on OEF (OELF) communication (Society: Data Gathering and Information Sharing with the Public and Policy-Makers)*. [Deliverable]. European Horizon-2020 project RISE. Retrieved from http://www.rise-eu.org/export/sites/rise/galleries/Deliverables/Deliverable_5.5.pdf
- Garthwaite, P. H., & Mubwandarikwa, E. (2012). Selection of weights for weighted model averaging. *Australian & New Zealand Journal of Statistics*, 52(4), 363–382. <https://doi.org/10.1111/j.1467-842x.2010.00589.x>
- Gentili, S., Brondi, P., & Di Giovambattista, R. (2023). NESTOREv1.0: A MATLAB Package for Strong Forthcoming Earthquake Forecasting. *Seismological Society of America*, 94(4), 2003–2013. <https://doi.org/10.1785/0220220327>
- Gerstenberger, M., Bora, S., Bradley, B. A., DiCaprio, C., Van Dissen, R. J., Chamberlain, C., et al. (2022). New Zealand National Seismic Hazard Model 2022 revision: Model, hazard and process overview. In *GNS Science Report 2022/57*. G. Science.
- Gerstenberger, M. C., Marzocchi, W., Allen, T., Pagani, M., Adams, J., Danciu, L., et al. (2020). Probabilistic seismic hazard analysis at regional and national scales: State of the art and future challenges. *Reviews of Geophysics*, 58(2), e2019RG000653. <https://doi.org/10.1029/2019RG000653>
- Gerstenberger, M. C., McVerry, G. H., Rhoades, D. A., & Stirling, M. (2014). Seismic hazard modelling for the recovery of Christchurch, New Zealand. *Earthquake Spectra*, 30(1), 17–29. <https://doi.org/10.1193/021913EQS037M>
- Gerstenberger, M. C., & Rhoades, D. A. (2010). New Zealand earthquake forecast testing centre. *Pure and Applied Geophysics*, 167(8–9), 877–892. <https://doi.org/10.1007/s00024-010-0082-4>
- Gerstenberger, M. C., Rhoades, D. A., Litchfield, N., Abbott, E., Gode, T., Christophersen, A., et al. (2023). A time-dependent seismic hazard model following the Kaikōura M7.8 earthquake. *New Zealand Journal of Geology and Geophysics*, 66(2), 1–25. <https://doi.org/10.1080/00288306.2022.2158881>
- Gerstenberger, M. C., Rhoades, D. A., & McVerry, G. H. (2016). A hybrid time-dependent probabilistic seismic-hazard model for Canterbury, New Zealand. *Seismological Research Letters*, 87(6), 1311–1318. <https://doi.org/10.1785/0220160084>
- Gerstenberger, M. C., Van Dissen, R., Rollins, C., DiCaprio, C., Chamberlain, C., Christophersen, A., et al. (2022). The Seismicity Rate Model for the 2022 New Zealand National Seismic Hazard Model. In *GNS Science Report; 2022/46*. GNS Science.
- Gerstenberger, M. C., Wiemer, S., & Jones, L. (2004). Real-time forecasts of tomorrow's earthquakes in California: A new mapping tool. *United States Geological Survey Open-File Report 2004-1390*.
- Gerstenberger, M. C., Wiemer, S., Jones, L. M., & Reasenber, P. A. (2005). Real-time forecasts of tomorrow's earthquakes in California. *Nature*, 435(7040), 328–331. <https://doi.org/10.1038/nature03622>
- Gischig, V. S., Giardini, D., Amann, F., Hertrich, M., Krietsch, H., Loew, S., et al. (2020). Hydraulic stimulation and fluid circulation experiments in underground laboratories: Stepping up the scale towards engineered geothermal systems. *Geomechanics for Energy and the Environment*, 24, 100175. <https://doi.org/10.1016/j.gete.2019.100175>
- Given, D. D., Cochran, E. S., Heaton, T., Hauksson, E., Allen, R., Hellweg, P., et al. (2014). *Technical implementation plan for the ShakeAlert production system: An earthquake early warning system for the west coast of the United States* (p. 25). US Department of the Interior, US Geological Survey.
- Goebel, T. H. W., Sammis, C. G., Becker, T. W., Dresen, G., & Schorlemmer, D. (2015). A comparison of seismicity characteristics and fault structure between stick-slip experiments and nature. *Pure and Applied Geophysics*, 172, 2247–2264.
- Goebel, T. H. W., Schorlemmer, D., Becker, T. W., Dresen, G., & Sammis, C. G. (2013). Acoustic emissions document stress changes over many seismic cycles in stick-slip experiments. *Geophysical Research Letters*, 40(10), 2049–2054. <https://doi.org/10.1002/grl.50507>
- Graham, K., Christophersen, A., Gerstenberger, M. C., & Rhoades, D. A. (2022). Current state of New Zealand's Operational Earthquake Forecasting (OEF). In *Poster Presentation at 2022 SCEC Annual Meeting (Palms Spring, California, USA)*.
- Grimm, C., Hainzl, S., Käser, M., & Küchenhoff, H. (2022). Solving three major biases of the ETAS model to improve forecasts of the 2019 Ridgecrest sequence. *Stochastic Environmental Research and Risk Assessment*, 36(8), 2133–2152. <https://doi.org/10.1007/s00477-022-02221-2>
- Gulia, L., Rinaldi, A. P., Tormann, T., Vannucci, G., Enescu, B., & Wiemer, S. (2018). The effect of a mainshock on the size distribution of the aftershocks. *Geophysical Research Letters*, 45(24), 13–277. <https://doi.org/10.1029/2018gl080619>
- Gulia, L., & Wiemer, S. (2019). Real-time discrimination of earthquake foreshocks and aftershocks. *Nature*, 574(7777), 193–199. <https://doi.org/10.1038/s41586-019-1606-4>
- Gulia, L., & Wiemer, S. (2021). Comment on “Two Foreshock Sequences Post Gulia and Wiemer (2019)” by Kelian Dascher-Cousineau, Thorne Lay, and Emily E. Brodsky. *Seismological Research Letters*, 92(5), 3251–3258. <https://doi.org/10.1785/0220200428>

- Gulia, L., Wiemer, S., & Vannucci, G. (2020). Pseudoprospective evaluation of the foreshock traffic-light system in Ridgecrest and implications for aftershock hazard assessment. *Seismological Research Letters*, 91(5), 2828–2842. <https://doi.org/10.1785/0220190307>
- Guo, Y., Zhuang, J., & Zhou, S. (2015). An improved space-time ETAS model for inverting the rupture geometry from seismicity triggering. *Journal of Geophysical Research: Solid Earth*, 120(5), 3309–3323. <https://doi.org/10.1002/2015jb011979>
- Gutenberg, B., & Richter, C. F. (1944). Frequency of earthquakes in California. *Bulletin of the Seismological Society of America*, 34(4), 185–188. <https://doi.org/10.1785/bssa0340040185>
- Hagedorn, R., Doblas-Reyes, F. J., & Palmer, T. N. (2005). The rationale behind the success of multi-model ensembles in seasonal forecasting - I. Basic concept. *Tellus, Series A: Dynamic Meteorology and Oceanography*, 57(3), 219–233. <https://doi.org/10.1111/j.1600-0870.2005.00103.x>
- Hainzl, S. (2004). Seismicity patterns of earthquake swarms due to fluid intrusion and stress triggering. *Geophysical Journal International*, 159(3), 1090–1096. <https://doi.org/10.1111/j.1365-246x.2004.02463.x>
- Hainzl, S. (2016a). Apparent triggering function of aftershocks resulting from rate-dependent incompleteness of earthquake catalogs. *Journal of Geophysical Research: Solid Earth*, 121(9), 6499–6509. <https://doi.org/10.1002/2016jb013319>
- Hainzl, S. (2016b). Rate-dependent incompleteness of earthquake catalogs. *Seismological Research Letters*, 87(2A), 337–344. <https://doi.org/10.1785/0220150211>
- Hainzl, S. (2022). ETAS-approach accounting for short-term incompleteness of earthquake catalogs. *Bulletin of the Seismological Society of America*, 112(1), 494–507. <https://doi.org/10.1785/0120210146>
- Hainzl, S., Christophersen, A., & Enescu, B. (2008). Impact of earthquake rupture extensions on parameter estimations of point-process models. *Bulletin of the Seismological Society of America*, 98(4), 2066–2072. <https://doi.org/10.1785/0120070256>
- Hainzl, S., Zakharova, O., & Marsan, D. (2013). Impact of aseismic transients on the estimation of aftershock productivity parameters. *Bulletin of the Seismological Society of America*, 103(3), 1723–1732. <https://doi.org/10.1785/0120120247>
- Han, M., Mizrahi, L., & Wiemer, S. (2024). Towards a harmonized operational earthquake forecasting model for Europe. *EGU sphere*, 2024, 1–26.
- Hanks, T. C., & Kanamori, H. (1979). A moment magnitude scale. *Journal of Geophysical Research*, 84(B5), 2348–2350. <https://doi.org/10.1029/jb084i05p02348>
- Hardebeck, J. L., Llenos, A. L., Michael, A. J., Page, M. T., Schneider, M., & van der Elst, N. J. (2024). Aftershock forecasting. *Annual Review of Earth and Planetary Sciences*, 52(1), 61–84. <https://doi.org/10.1146/annurev-earth-040522-102129>
- Hardebeck, J. L., Llenos, A. L., Michael, A. J., Page, M. T., & Van Der Elst, N. (2019). Updated California aftershock parameters. *Seismological Research Letters*, 90(1), 262–270. <https://doi.org/10.1785/0220180240>
- Harris, R. A., & Simpson, R. W. (1992). Changes in static stress on southern California faults after the 1992 Landers earthquake. *Nature*, 360(6401), 251–254. <https://doi.org/10.1038/360251a0>
- Harte, D., & Vere-Jones, D. (2005). The entropy score and its uses in earthquake forecasting. *Pure and Applied Geophysics*, 162(6–7), 1229–1253. <https://doi.org/10.1007/s00024-004-2667-2>
- Harte, D. S. (2013). Bias in fitting the ETAS model: A case study based on New Zealand seismicity. *Geophysical Journal International*, 192(1), 390–412. <https://doi.org/10.1093/gji/ggs026>
- Harte, D. S. (2014). An ETAS model with varying productivity rates. *Geophysical Journal International*, 198(1), 270–284. <https://doi.org/10.1093/gji/ggu129>
- Harte, D. S. (2015). Log-likelihood of earthquake models: Evaluation of models and forecasts. *Geophysical Journal International*, 201(2), 711–723. <https://doi.org/10.1093/gji/ggu442>
- Harte, D. S. (2016). Model parameter estimation bias induced by earthquake magnitude cut-off. *Geophysical Journal International*, 204(2), 1266–1287. <https://doi.org/10.1093/gji/ggv524>
- Harte, D. S. (2017). Probability distribution of forecasts based on the ETAS model. *Geophysical Journal International*, 210(1), 90–104. <https://doi.org/10.1093/gji/ggx146>
- Harte, D. S. (2018). Effect of sample size on parameter estimates and earthquake forecasts. *Geophysical Journal International*, 214(2), 759–772. <https://doi.org/10.1093/gji/ggy150>
- Harte, D. S. (2019). Evaluation of earthquake stochastic models based on their real-time forecasts: A case study of Kaikoura 2016. *Geophysical Journal International*, 217(3), 1894–1914. <https://doi.org/10.1093/gji/ggz088>
- Heki, K. (2011). Ionospheric electron enhancement preceding the 2011 Tohoku-Oki earthquake. *Geophysical Research Letters*, 38(17), L17312. <https://doi.org/10.1029/2011gl047908>
- Helmstetter, A., Kagan, Y. Y., & Jackson, D. D. (2006). Comparison of short-term and time-independent earthquake forecast models for southern California. *Bulletin of the Seismological Society of America*, 96(1), 90–106. <https://doi.org/10.1785/0120050067>
- Helmstetter, A., & Sornette, D. (2002). Subcritical and supercritical regimes in epidemic models of earthquake aftershocks. *Journal of Geophysical Research*, 107(B10), ESE 10-1–ESE 10-21. <https://doi.org/10.1029/2001jb001580>
- Helmstetter, A., & Sornette, D. (2003). Predictability in the epidemic-type aftershock sequence model of interacting triggered seismicity. *Journal of Geophysical Research*, 108(B10), ESE 8-1–ESE 8-18. <https://doi.org/10.1029/2003jb002485>
- Helmstetter, A., & Werner, M. J. (2014). Adaptive smoothing of seismicity in time, space, and magnitude for time-dependent earthquake forecasts for California. *Bulletin of the Seismological Society of America*, 104(2), 809–822. <https://doi.org/10.1785/0120130105>
- Herrmann, M., & Marzocchi, W. (2021). Inconsistencies and lurking pitfalls in the magnitude–frequency distribution of high-resolution earthquake catalogs. *Seismological Research Letters*, 92(2A), 909–922. <https://doi.org/10.1785/0220200337>
- Herrmann, M., & Marzocchi, W. (2023). Maximizing the forecasting skill of an ensemble model. *Geophysical Journal International*, 234(1), 73–87. <https://doi.org/10.1093/gji/ggad020>
- Herrmann, M., Zechar, J. D., & Wiemer, S. (2016). Communicating time-varying seismic risk during an earthquake sequence. *Seismological Research Letters*, 87(2A), 301–312. <https://doi.org/10.1785/0220150168>
- Hill, D. P., Mowinkel, P., & Peake, L. G. (1975). Earthquakes, active faults, and geothermal areas in the Imperial Valley, California. *Science*, 188(4195), 1306–1308. <https://doi.org/10.1126/science.188.4195.1306>
- Hirose, H., Matsuzawa, T., Kimura, T., & Kimura, H. (2014). The Boso slow slip events in 2007 and 2011 as a driving process for the accompanying earthquake swarm. *Geophysical Research Letters*, 41(8), 2778–2785. <https://doi.org/10.1002/2014gl059791>
- Hough, S. E. (2010). *Predicting the unpredictable: The tumultuous science of earthquake prediction*. Princeton University Press.
- Hsu, C. C., & Sandford, B. A. (2007). The Delphi technique: Making sense of consensus. *Practical Assessment, Research and Evaluation*, 12(1), 10.
- Hudson-Doyle, E. E., Paton, D., & Johnston, D. (2018). Reflections on the communication of uncertainty: Developing decision-relevant information. In *Proceedings of the ISCRAM Asia Pacific Conference*. *ISCRAM2015_Conference_PaperTemplate.docx*.
- Hutton, K., Woessner, J., & Hauksson, E. (2010). Earthquake monitoring in southern California for seventy-seven years (1932–2008). *Bulletin of the Seismological Society of America*, 100(2), 423–446. <https://doi.org/10.1785/0120090130>

- Iervolino, I. (2011). Performance-based earthquake early warning. *Soil Dynamics and Earthquake Engineering*, 31(2), 209–222. <https://doi.org/10.1016/j.soildyn.2010.07.010>
- Iervolino, I., Chioccarelli, E., Giorgio, M., Marzocchi, W., Zuccaro, G., Dolce, M., & Manfredi, G. (2015). Operational (short-term) earthquake loss forecasting in Italy. *Bulletin of the Seismological Society of America*, 105(4), 2286–2298. <https://doi.org/10.1785/0120140344>
- Iervolino, I., Chioccarelli, E., & Suzuki, A. (2020). Seismic damage accumulation in multiple mainshock–aftershock sequences. *Earthquake Engineering & Structural Dynamics*, 49(10), 1007–1027. <https://doi.org/10.1002/eqe.3275>
- Iervolino, I., Convertito, V., Giorgio, M., Manfredi, G., & Zollo, A. (2006). Real time risk analysis for hybrid earthquake early warning systems. *Journal of Earthquake Engineering*, 10(6), 867–885. <https://doi.org/10.1080/13632460609350621>
- Iervolino, I., Giorgio, M., & Chioccarelli, E. (2016). Markovian modeling of seismic damage accumulation. *Earthquake Engineering & Structural Dynamics*, 45(3), 441–461. <https://doi.org/10.1002/eqe.2668>
- Imoto, M. (1991). Changes in the magnitude–Frequency b -value prior to large ($M \geq 6.0$) earthquakes in Japan. *Tectonophysics*, 193(4), 311–325. [https://doi.org/10.1016/0040-1951\(91\)90340-x](https://doi.org/10.1016/0040-1951(91)90340-x)
- Iturrieta, P., Bayona, J. A., Werner, M. J., Schorlemmer, D., Taroni, M., Falcone, G., et al. (2024). Evaluation of a decade-long prospective earthquake forecasting experiment in Italy. *Seismological Research Letters*. <https://doi.org/10.1785/0220230247>
- Iturrieta, P., Savran, W. H., Khawaja, M. A. M., Bayona, J., Maechling, P. J., Silva, F., et al. (2023). Modernizing earthquake forecasting experiments: The CSEP floating experiments. In *AGU Fall Meeting Abstracts* (Vol. 2023).
- Jackson, D., & Kagan, Y. Y. (1999). Testable earthquake forecasts for 1999. *Seismological Research Letters*, 70(4), 393–403. <https://doi.org/10.1785/gssrl.70.4.393>
- Jackson, D. D. (1996). Hypothesis testing and earthquake prediction. *Proceedings of the National Academy of Sciences*, 93(9), 3772–3775. <https://doi.org/10.1073/pnas.93.9.3772>
- Jiang, X., Liu, H., Main, I. G., & Salje, E. K. (2017). Predicting mining collapse: Superjerks and the appearance of record-breaking events in coal as collapse precursors. *Physical Review E*, 96(2), 023004. <https://doi.org/10.1103/physreve.96.023004>
- Johnson, P. A., Ferdowsi, B., Kaproth, B. M., Scuderi, M., Griffa, M., Carmeliet, J., et al. (2013). Acoustic emission and microslip precursors to stick-slip failure in sheared granular material. *Geophysical Research Letters*, 40(21), 5627–5631. <https://doi.org/10.1002/2013gl057848>
- Jordan, T., Chen, Y.-T., Gasparini, P., Madariaga, R., Main, I., Marzocchi, W., et al. (2011). Operational earthquake forecasting: State of knowledge and guidelines for implementation. *Annals of Geophysics*, 54(4), 315–391. <https://doi.org/10.4401/ag-5350>
- Jordan, T. H. (2006). Earthquake predictability, brick by brick. *Seismological Research Letters*, 77(1), 3–6. <https://doi.org/10.1785/gssrl.77.1.3>
- Joshi, A., Kale, S., Chandel, S., & Pal, D. K. (2015). Likert scale: Explored and explained. *British Journal of Applied Science & Technology*, 7(4), 396–403. <https://doi.org/10.9734/bjast/2015/14975>
- Kagan, Y. Y. (1999). Universality of the seismic moment-frequency relation. *Seismicity patterns, their statistical significance and physical meaning* (pp. 537–573).
- Kagan, Y. Y. (2004). Short-term properties of earthquake catalogs and models of earthquake source. *Bulletin of the Seismological Society of America*, 94(4), 1207–1228. <https://doi.org/10.1785/012003098>
- Kagan, Y. Y. (2010). Statistical distributions of earthquake numbers: Consequence of branching process. *Geophysical Journal International*, 180(3), 1313–1328. <https://doi.org/10.1111/j.1365-246x.2009.04487.x>
- Kagan, Y. Y., & Jackson, D. D. (1991). Long-term earthquake clustering. *Geophysical Journal International*, 104(1), 117–133. <https://doi.org/10.1111/j.1365-246x.1991.tb02498.x>
- Kagan, Y. Y., & Jackson, D. D. (1995). New seismic gap hypothesis: Five years after. *Journal of Geophysical Research*, 100(B3), 3943–3959. <https://doi.org/10.1029/94jb03014>
- Kagan, Y. Y., & Jackson, D. D. (2000). Probabilistic forecasting of earthquakes. *Geophysical Journal International*, 143(2), 438–453. <https://doi.org/10.1046/j.1365-246x.2000.01267.x>
- Kamer, Y., & Hiemer, S. (2015). Data-driven spatial b value estimation with applications to California seismicity: To b or not to b . *Journal of Geophysical Research: Solid Earth*, 120(7), 5191–5214. <https://doi.org/10.1002/2014jb011510>
- Kamer, Y., Nandan, S., Ouillon, G., Hiemer, S., & Sornette, D. (2021). Democratizing earthquake predictability research: Introducing the RichterX platform. *The European Physical Journal Special Topics*, 230(1), 451–471. <https://doi.org/10.1140/epjst/e2020-000260-2>
- Kamogawa, M., & Kakinami, Y. (2013). Is an ionospheric electron enhancement preceding the 2011 Tohoku-Oki earthquake a precursor? *Journal of Geophysical Research: Space Physics*, 118(4), 1751–1754. <https://doi.org/10.1002/jgra.50118>
- Kant, I. (1756). *Geschichte und Naturbeschreibung der merkwürdigsten Vorfälle des Erdbens welches an dem Ende des 1755ten Jahres einen grossen Theil der Erde erschüttert hat*. Gedruckt und Verlegt von Johann Heinrich Hartung.
- Khawaja, A. M., Hainzl, S., Schorlemmer, D., Iturrieta, P., Bayona, J. A., Savran, W. H., et al. (2023). Statistical power of spatial earthquake forecast tests. *Geophysical Journal International*, 233(3), 2053–2066. <https://doi.org/10.1093/gji/ggad030>
- Király-Proag, E., Gischig, V., Zechar, J. D., & Wiemer, S. (2018). Multicomponent ensemble models to forecast induced seismicity. *Geophysical Journal International*, 212(1), 476–490. <https://doi.org/10.1093/gji/ggx393>
- Kisslinger, C. (1975). Processes during the Matsushiro, Japan, earthquake swarm as revealed by leveling, gravity, and spring-flow observations. *Geology*, 3(2), 57–62. [https://doi.org/10.1130/0091-7613\(1975\)3<57:pdtmje>2.0.co;2](https://doi.org/10.1130/0091-7613(1975)3<57:pdtmje>2.0.co;2)
- Krishnamurti, T. N., Kishtawal, C. M., Zhang, Z., LaRow, T., Bachiochi, D., Williford, E., et al. (2000). Multimodel ensemble forecasts for weather and seasonal climate. *Journal of Climate*, 13(23), 4196–4216. [https://doi.org/10.1175/1520-0442\(2000\)013<4196:MEFFWA>2.0.CO;2](https://doi.org/10.1175/1520-0442(2000)013<4196:MEFFWA>2.0.CO;2)
- Lei, X., & Ma, S. (2014). Laboratory acoustic emission study for earthquake generation process. *Earthquake Science*, 27(6), 627–646. <https://doi.org/10.1007/s11589-014-0103-y>
- Lippiello, E., Bottiglieri, M., Godano, C., & de Arcangelis, L. (2007). Dynamical scaling and generalized Omori law. *Geophysical Research Letters*, 34(23), L23301. <https://doi.org/10.1029/2007gl030963>
- Lippiello, E., Cirillo, A., Godano, C., Papadimitriou, E., & Karakostas, V. (2019). Post seismic catalog incompleteness and aftershock forecasting. *Geosciences*, 9(8), 355. <https://doi.org/10.3390/geosciences9080355>
- Lippiello, E., Giacco, F., Arcangelis, L. D., Marzocchi, W., & Godano, C. (2014). Parameter estimation in the ETAS model: Approximations and novel methods. *Bulletin of the Seismological Society of America*, 104(2), 985–994. <https://doi.org/10.1785/0120130148>
- Liu, M., Zhang, M., Zhu, W., Ellsworth, W. L., & Li, H. (2020). Rapid characterization of the July 2019 Ridgecrest, California, earthquake sequence from raw seismic data using machine-learning phase picker. *Geophysical Research Letters*, 47(4), e2019GL086189. <https://doi.org/10.1029/2019gl086189>
- Liu, Z., Jiang, H., Li, S., Li, M., Liu, J., & Zhang, J. (2023). Implementation and verification of a real-time system for automatic aftershock forecasting in China. *Earth Science Informatics*, 16(2), 1891–1907. <https://doi.org/10.1007/s12145-023-00960-6>

- Llenos, A. L., & Michael, A. J. (2019). Ensembles of ETAS models provide optimal operational earthquake forecasting during swarms: Insights from the 2015 San Ramon, California Swarm. *Bulletin of the Seismological Society of America*, *109*(6), 2145–2158. <https://doi.org/10.1785/0120190020>
- Llenos, A. L., & van der Elst, N. J. (2019). Improving earthquake forecasts during swarms with a duration model. *Bulletin of the Seismological Society of America*, *109*(3), 1148–1155.
- Lolli, B., & Gasperini, P. (2006). Comparing different models of aftershock rate decay: The role of catalog incompleteness in the first times after main shock. *Tectonophysics*, *423*(1–4), 43–59. <https://doi.org/10.1016/j.tecto.2006.03.025>
- Lombardi, A. M. (2014). Some reasoning on the RELM-CSEP likelihood-based tests. *Earth, Planets and Space*, *66*, 1–6. <https://doi.org/10.1186/1880-5981-66-4>
- Lombardi, A. M. (2015). Estimation of the parameters of ETAS models by Simulated Annealing. *Scientific Reports*, *5*(1), 8417. <https://doi.org/10.1038/srep08417>
- Lombardi, A. M., & Marzocchi, W. (2010a). The assumption of Poisson seismic-rate variability in CSEP/RELM experiments. *Bulletin of the Seismological Society of America*, *100*(5A), 2293–2300. <https://doi.org/10.1785/0120100012>
- Lombardi, A. M., & Marzocchi, W. (2010b). The ETAS model for daily forecasting of Italian seismicity in the CSEP experiment. *Annals of Geophysics*, *53*(3), 155–164. <https://doi.org/10.4401/ag-4848>
- Main, I. (1999). Is the reliable prediction of individual earthquakes a realistic scientific goal. *Nature*, *397*(1). <https://doi.org/10.1038/nature28107>
- Main, I. G., Meredith, P. G., & Jones, C. (1989). A reinterpretation of the precursory seismic b-value anomaly from fracture mechanics. *Geophysical Journal International*, *96*(1), 131–138. <https://doi.org/10.1111/j.1365-246x.1989.tb05255.x>
- Main, I. G., Meredith, P. G., & Sammonds, P. R. (1992). Temporal variations in seismic event rate and b-values from stress corrosion constitutive laws. *Tectonophysics*, *211*(1–4), 233–246. [https://doi.org/10.1016/0040-1951\(92\)90061-a](https://doi.org/10.1016/0040-1951(92)90061-a)
- Mancini, S., & Marzocchi, W. (2023). SimpleTAS: A benchmark earthquake forecasting model suitable for operational purposes and seismic hazard analysis. *Seismological Research Letters*, *95*(1), 38–49. <https://doi.org/10.1785/0220230199>
- Mancini, S., Segou, M., Werner, M. J., & Cattania, C. (2019). Improving physics-based aftershock forecasts during the 2016–2017 Central Italy Earthquake Cascade. *Journal of Geophysical Research: Solid Earth*, *124*(8), 8626–8643. <https://doi.org/10.1029/2019jb017874>
- Mancini, S., Segou, M., Werner, M. J., & Parsons, T. (2020). The predictive skills of elastic Coulomb rate-and-state aftershock forecasts during the 2019 Ridgecrest, California, earthquake sequence. *Bulletin of the Seismological Society of America*, *110*(4), 1736–1751. <https://doi.org/10.1785/0120200028>
- Mancini, S., Segou, M., Werner, M. J., Parsons, T., Beroza, G., & Chiaraluce, L. (2022). On the use of high-resolution and deep-learning seismic catalogs for short-term earthquake forecasts: Potential benefits and current limitations. *Journal of Geophysical Research: Solid Earth*, *127*(11), e2022JB025202. <https://doi.org/10.1029/2022jb025202>
- Marti, M., Stauffacher, M., & Wiemer, S. (2019). Difficulties in explaining complex issues with maps. Evaluating seismic hazard communication – The Swiss case. *Natural Hazards and Earth System Sciences*, *19*(12), 2677–2700. Article 12. <https://doi.org/10.5194/nhess-19-2677-2019>
- Marzocchi, W. (2012). Putting science on trial. *Physics World*, *25*(12), 17–18. <https://doi.org/10.1088/2058-7058/25/12/27>
- Marzocchi, W., Zechar, J. D., & Jordan, T. H. (2012). Bayesian forecast evaluation and ensemble earthquake forecasting. *Bulletin of the Seismological Society of America*, *102*(6), 2574–2584. <https://doi.org/10.1785/0120110327>
- Marzocchi, W., Iervolino, I., Giorgio, M., & Falcone, G. (2015). When is the probability of a large earthquake too small? *Seismological Research Letters*, *86*(6), 1674–1678. <https://doi.org/10.1785/0220150129>
- Marzocchi, W., & Jordan, T. H. (2014). Testing for ontological errors in probabilistic forecasting models of natural systems. *Proceedings of the National Academy of Sciences*, *111*(33), 11973–11978. <https://doi.org/10.1073/pnas.1410183111>
- Marzocchi, W., & Jordan, T. H. (2017). A unified probabilistic framework for seismic hazard analysis. *Bulletin of the Seismological Society of America*, *107*(6), 2738–2744. <https://doi.org/10.1785/0120170008>
- Marzocchi, W., & Jordan, T. H. (2018). Experimental concepts for testing probabilistic earthquake forecasting and seismic hazard models. *Geophysical Journal International*, *215*(2), 780–798. <https://doi.org/10.1093/gji/gyg276>
- Marzocchi, W., Jordan, T. H., & Woo, G. (2015). Varena workshop report. Operational earthquake forecasting and decision making. *Annals of Geophysics*, *58*(4), 4. <https://doi.org/10.4401/ag-6756>
- Marzocchi, W., & Lombardi, A. M. (2009). Real-time forecasting following a damaging earthquake. *Geophysical Research Letters*, *36*(21), L21302. <https://doi.org/10.1029/2009gl040233>
- Marzocchi, W., Lombardi, A. M., & Casarotti, E. (2014). The establishment of an operational earthquake forecasting system in Italy. *Seismological Research Letters*, *85*(5), 961–969. <https://doi.org/10.1785/0220130219>
- Marzocchi, W., Murru, M., Lombardi, A. M., Falcone, G., & Console, R. (2012). Daily earthquake forecast during the May–June 2012 earthquake sequence (Northern Italy). *Annales Geophysicae*, *55*, 561–567.
- Marzocchi, W., & Sandri, L. (2003). A review and new insights on the estimation of the b-value and its uncertainty. *Annals of Geophysics*.
- Marzocchi, W., Spassiani, I., Stallone, A., & Taroni, M. (2020). How to be fooled searching for significant variations of the b-value. *Geophysical Journal International*, *220*(3), 1845–1856. <https://doi.org/10.1093/gji/egz541>
- Marzocchi, W., Taroni, M., & Falcone, G. (2017). Earthquake forecasting during the complex Amatrice-Norcia seismic sequence. *Science Advances*, *3*(9), e1701239. <https://doi.org/10.1126/sciadv.1701239>
- Marzocchi, W., & Zechar, J. D. (2011). Earthquake forecasting and earthquake prediction: Different approaches for obtaining the best model. *Seismological Research Letters*, *82*(3), 442–448. <https://doi.org/10.1785/gssrl.82.3.442>
- Masci, F., Thomas, J. N., Villani, F., Secan, J. A., & Rivera, N. (2015). On the onset of ionospheric precursors 40 min before strong earthquakes. *Journal of Geophysical Research: Space Physics*, *120*(2), 1383–1393. <https://doi.org/10.1002/2014ja020822>
- McBride, S. K., Llenos, A. L., Page, M. T., & van der Elst, N. (2020). #EarthquakeAdvisory: Exploring discourse between government officials, news media, and social media during the 2016 Bombay Beach Swarm. *Seismological Research Letters*, *91*(1), 438–451. <https://doi.org/10.1785/0220190082>
- McBride, S. K., Wein, A., Becker, J. S., Potter, S. H., & Doyle, E. E. H. (2018). An evidence-based approach for supporting scientists communicating earthquake forecasts. In *National Conference of Earthquake Engineering, Los Angeles, CA*.
- McLaskey, G. C., & Lockner, D. A. (2014). Preslip and cascade processes initiating laboratory stick slip. *Journal of Geophysical Research: Solid Earth*, *119*(8), 6323–6336. <https://doi.org/10.1002/2014jb011220>
- Menahem, E., Shabtai, A., Rokach, L., & Elovici, Y. (2009). Improving malware detection by applying multi-inducer ensemble. *Computational Statistics and Data Analysis*, *53*(4), 1483–1494. <https://doi.org/10.1016/j.csda.2008.10.015>
- Michael, A. J., McBride, S. K., Hardebeck, J. L., Barall, M., Martinez, E., Page, M. T., et al. (2020). Statistical seismology and communication of the USGS operational aftershock forecasts for the 30 November 2018 Mw 7.1 Anchorage, Alaska, earthquake. *Seismological Research Letters*, *91*(1), 153–173. <https://doi.org/10.1785/0220190196>

- Michael, A. J., & Werner, M. J. (2018). Preface to the focus section on the Collaboratory for the Study of Earthquake Predictability (CSEP): New results and future directions. *Seismological Research Letters*, 89(4), 1226–1228. <https://doi.org/10.1785/0220180161>
- Mignan, A. (2014). The debate on the prognostic value of earthquake foreshocks: A meta-analysis. *Scientific Reports*, 4(1), 4099. <https://doi.org/10.1038/srep04099>
- Mignan, A., & Broccardo, M. (2019). One neuron versus deep learning in aftershock prediction. *Nature*, 574(7776), E1–E3. <https://doi.org/10.1038/s41586-019-1582-8>
- Mileti, D. S., & O'Brien, P. W. (1992). Warnings during disaster: Normalizing communicated risk. *Social Problems*, 39(1), 40–57. <https://doi.org/10.2307/3096912>
- Mileti, D. S., & Sorensen, J. H. (1990). *Communication of emergency public warnings: A social science perspective and state-of-the-art assessment*. Oak Ridge National Laboratory.
- Milner, K. R., Field, E. H., Savran, W. H., Page, M. T., & Jordan, T. H. (2020). Operational earthquake forecasting during the 2019 Ridgecrest, California, earthquake sequence with the UCERF3-ETAS model. *Seismological Research Letters*, 91(3), 1567–1578. <https://doi.org/10.1785/0220190294>
- Mizrahi, L., Dallo, I., & Kuratle, L. D. (2023). Supplement of “Developing, testing, and communicating earthquake forecasts: Current practices and an elicitation of expert recommendations” [Dataset]. *ETH Zurich*. <https://doi.org/10.3929/ethz-b-000637239>
- Mizrahi, L., Nandan, S., Mena Cabrera, B., & Wiemer, S. (2024). suiETAS: Developing and testing ETAS-based earthquake forecasting models for Switzerland. *Bulletin of the Seismological Society of America*, 2024. <https://doi.org/10.1785/0120240007>
- Mizrahi, L., Nandan, S., Savran, W., Wiemer, S., & Ben-Zion, Y. (2023). Question-driven ensembles of flexible ETAS models. *Seismological Research Letters*, 94(2 A), 829–843. <https://doi.org/10.1785/0220220230>
- Mizrahi, L., Nandan, S., & Wiemer, S. (2021). Embracing data incompleteness for better earthquake forecasting. *Journal of Geophysical Research: Solid Earth*, 126(12), e2021JB022379. <https://doi.org/10.1029/2021jb022379>
- Mizrahi, L., Schmid, N., & Han, M. (2023). Imzrahi/etas: ETAS with fit visualization (3.2) [Software]. *Zenodo*. <https://doi.org/10.5281/zenodo.7584575>
- Mogi, K. (1962). Magnitude-frequency relation for elastic shocks accompanying fractures of various materials and some related problems in earthquakes. *Bulletin of the Earthquake Research Institute, University of Tokyo*, 40, 831–853.
- Mogi, K. (1963). Experimental study on the mechanism of the earthquake occurrences of volcanic origin. *Bulletin of Volcanology*, 26(1), 197–208. <https://doi.org/10.1007/bf02597286>
- Mori, J., & Abercrombie, R. E. (1997). Depth dependence of earthquake frequency-magnitude distributions in California: Implications for rupture initiation. *Journal of Geophysical Research*, 102(B7), 15081–15090. <https://doi.org/10.1029/97jb01356>
- Murru, M., Console, R., Falcone, G., Montuori, C., & Sgroi, T. (2007). Spatial mapping of the *b* value at Mount Etna, Italy, using earthquake data recorded from 1999 to 2005. *Journal of Geophysical Research*, 112(B12), B12303. <https://doi.org/10.1029/2006jb004791>
- Murru, M., Zhuang, J., Rodolfo, C., & Giuseppe, F. (2014). Short-term earthquake forecasting experiment before and during the L' Aquila (central Italy) seismic sequence of April 2009. *Annals of Geophysics*.
- Musmeci, F., & Vere-Jones, D. (1992). A space-time clustering model for historical earthquakes. *Annals of the Institute of Statistical Mathematics*, 44, 1–11. <https://doi.org/10.1007/bf00048666>
- Nandan, S., Kamer, Y., Ouillon, G., Hiemer, S., & Sornette, D. (2021). Global models for short-term earthquake forecasting and predictive skill assessment. *The European Physical Journal Special Topics*, 230(1), 425–449. <https://doi.org/10.1140/epjst/e2020-000259-3>
- Nandan, S., Ouillon, G., Sornette, D., & Wiemer, S. (2019). Forecasting the full distribution of earthquake numbers is fair, robust, and better. *Seismological Research Letters*, 90(4), 1650–1659. <https://doi.org/10.1785/0220180374>
- Nandan, S., Ouillon, G., Wiemer, S., & Sornette, D. (2017). Objective estimation of spatially variable parameters of epidemic type aftershock sequence model: Application to California. *Journal of Geophysical Research: Solid Earth*, 122(7), 5118–5143. <https://doi.org/10.1002/2016jb013266>
- Nanjo, K. Z., Tsuruoka, H., Yokoi, S., Ogata, Y., Falcone, G., Hirata, N., et al. (2012). Predictability study on the aftershock sequence following the 2011 Tohoku-Oki, Japan, earthquake: First results. *Geophysical Journal International*, 191(2), 653–658. <https://doi.org/10.1111/j.1365-246x.2012.05626.x>
- Narteau, C., Byrdina, S., Shebalin, P., & Schorlemmer, D. (2009). Common dependence on stress for the two fundamental laws of statistical seismology. *Nature*, 462(7273), 642–645. <https://doi.org/10.1038/nature08553>
- National Academies of Sciences, Engineering, and Medicine. (2019). Reproducibility and replicability in science.
- Nof, R. N., & Kurzon, I. (2021). TRUAA—Earthquake early warning system for Israel: Implementation and current status. *Seismological Research Letters*, 92(1), 325–341. <https://doi.org/10.1785/0220200176>
- Ogata, Y. (1988). Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical Association*, 83(401), 9–27. <https://doi.org/10.1080/01621459.1988.10478560>
- Ogata, Y. (1993). Space-time modeling of earthquake occurrences. *Bulletin of the International Statistical Institute*, 55(2), 249–250.
- Ogata, Y. (1998). Space-time point-process models for earthquake occurrences. *Annals of the Institute of Statistical Mathematics*, 50(2), 379–402. <https://doi.org/10.1023/a:1003403601725>
- Ogata, Y. (2011). Significant improvements of the space-time ETAS model for forecasting of accurate baseline seismicity. *Earth, Planets and Space*, 63(3), 217–229. <https://doi.org/10.5047/eps.2010.09.001>
- Ogata, Y., & Katsura, K. (2014). Comparing foreshock characteristics and foreshock forecasting in observed and simulated earthquake catalogs. *Journal of Geophysical Research: Solid Earth*, 119(11), 8457–8477. <https://doi.org/10.1002/2014jb011250>
- Ogata, Y., & Zhuang, J. (2006). Space-time ETAS models and an improved extension. *Tectonophysics*, 413(1–2), 13–23. <https://doi.org/10.1016/j.tecto.2005.10.016>
- O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., et al. (2006). *Uncertain Judgements: Eliciting experts' probabilities*. Wiley.
- Omi, T., Ogata, Y., Hirata, Y., & Aihara, K. (2014). Estimating the ETAS model from an early aftershock sequence. *Geophysical Research Letters*, 41(3), 850–857. <https://doi.org/10.1002/2013gl058958>
- Omi, T., Ogata, Y., Shiomi, K., Enescu, B., Sawazaki, K., & Aihara, K. (2019). Implementation of a real-time system for automatic aftershock forecasting in Japan. *Seismological Research Letters*, 90(1), 242–250. <https://doi.org/10.1785/0220180213>
- Omorii, F. (1894). On after-shocks. *Seismological Journal of Japan*, 19, 71–80.
- Orihuela, B., Dallo, I., Clinton, J., Strauch, W., Protti, M., Yani, R., et al. (2023). Earthquake early warning in Central America: The societal perspective. *International Journal of Disaster Risk Reduction*, 97, 103982. <https://doi.org/10.2139/ssrn.4348227>

- Page, M. T., van der Elst, N., Hardebeck, J., Felzer, K., & Michael, A. J. (2016). Three ingredients for improved global aftershock forecasts: Tectonic region, time-dependent catalog incompleteness, and intersequence variability. *Bulletin of the Seismological Society of America*, 106(5), 2290–2301. <https://doi.org/10.1785/0120160073>
- Page, M. T., & van der Elst, N. J. (2018). Turing-style tests for UCERF3 synthetic catalogs. *Bulletin of the Seismological Society of America*, 108(2), 729–741. <https://doi.org/10.1785/0120170223>
- Paris, G. M., & Michael, A. J. (2022a). An interactive viewer to improve operational aftershock forecasts. *Seismological Research Letters*, 94(1), 473–484. <https://doi.org/10.1785/0220220108>
- Paris, G. M., & Michael, A. J. (2022b). *OAF Tools - R package, Version 1.0.0*. U.S. Geological Survey software release. <https://doi.org/10.5066/P9PZTYEN>
- Passarelli, L., Selvadurai, P. A., Rivalta, E., & Jónsson, S. (2021). The source scaling and seismic productivity of slow slip transients. *Science Advances*, 7(32), eabg9718. <https://doi.org/10.1126/sciadv.abg9718>
- Peng, Z., Vidale, J. E., & Houston, H. (2006). Anomalous early aftershock decay rate of the 2004 Mw6. 0 Parkfield, California, earthquake. *Geophysical Research Letters*, 33(17), L17307. <https://doi.org/10.1029/2006gl026744>
- Pollock, D. (2007). *Aspects of short-term and long-term seismic hazard assessment in New Zealand* [Master Thesis]. ETH Zurich.
- Rastin, S. J., Rhoades, D. A., & Christophersen, A. (2021). Space–time trade-off of precursory seismicity in New Zealand and California revealed by a medium-term earthquake forecasting model [Online]. *Applied Sciences*, 11(21), 10215. <https://doi.org/10.3390/app112110215>
- Rastin, S. J., Rhoades, D. A., Rollins, C., & Gerstenberger, M. C. (2022). How useful are strain rates for estimating the long-term spatial distribution of earthquakes? *Applied Science*, 12(13), 6804. <https://doi.org/10.3390/app12136804>
- Rastin, S. J., Rhoades, D. A., Rollins, C., Gerstenberger, M. C., Christophersen, A., & Thingbaijam, K. (2022). Spatial distribution of earthquake occurrence for the New Zealand National Seismic Hazard Model revision. In *GNS Science Report; 2021/51*. GNS Science.
- R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Retrieved from <https://www.R-project.org/>
- Reasenber, P. A., & Jones, L. M. (1989). Earthquake hazard after a mainshock in California. *Science*, 243(4895), 1173–1176. <https://doi.org/10.1126/science.243.4895.1173>
- Reasenber, P. A., & Jones, L. M. (1990). California aftershock hazard forecast. *Science*, 247(4940), 345–346. <https://doi.org/10.1126/science.247.4940.345>
- Rhoades, D. A. (2007). Application of the EEPAS model to forecasting earthquakes of moderate magnitude in southern California. *Seismological Research Letters*, 78(1), 110–115. <https://doi.org/10.1785/gssrl.78.1.110>
- Rhoades, D. A., & Christophersen, A. (2017). Magnitude conversion of earthquake rate forecasts. *Bulletin of the Seismological Society of America*, 107(6), 3037–3043. <https://doi.org/10.1785/0120170225>
- Rhoades, D. A., & Christophersen, A. (2019). Time-varying probabilities of earthquake occurrence in central New Zealand based on the EEPAS model compensated for time-lag. *Geophysical Journal International*, 219(1), 417–429. <https://doi.org/10.1093/gji/ggz301>
- Rhoades, D. A., Christophersen, A., Bourguignon, S., Ristau, J., & Salichon, J. (2021). A depth-dependent local magnitude scale for New Zealand earthquakes consistent with moment magnitude. *Bulletin of the Seismological Society of America*, 111(2), 1056–1066. <https://doi.org/10.1785/0120200252>
- Rhoades, D. A., Christophersen, A., & Gerstenberger, M. C. (2015). Multiplicative earthquake likelihood models based on fault and earthquake data. *Bulletin of the Seismological Society of America*, 105(6), 2955–2968. <https://doi.org/10.1785/0120150080>
- Rhoades, D. A., Christophersen, A., & Gerstenberger, M. C. (2017). Multiplicative earthquake likelihood models incorporating strain rates. *Geophysical Journal International*, 208(3), 1764–1774. <https://doi.org/10.1093/gji/ggw486>
- Rhoades, D. A., Christophersen, A., Gerstenberger, M. C., Liukis, M., Silva, F., Marzocchi, W., et al. (2018). Highlights from the first ten years of the New Zealand earthquake forecast testing center. *Seismological Research Letters*, 89(4), 1229–1237. <https://doi.org/10.1785/0220180032>
- Rhoades, D. A., & Evison, F. F. (2004). Long-range earthquake forecasting with every earthquake a precursor according to scale. *Pure and Applied Geophysics*, 161(1), 47–72. <https://doi.org/10.1007/s00024-003-2434-9>
- Rhoades, D. A., & Evison, F. F. (2005). Test of the EEPAS forecasting model on the Japan earthquake catalogue. *Pure and Applied Geophysics*, 162(6–7), 1271–1290. <https://doi.org/10.1007/s00024-004-2669-0>
- Rhoades, D. A., & Evison, F. F. (2006). The EEPAS forecasting model and the probability of moderate-to-large earthquakes in central Japan. *Tectonophysics*, 417(1/2), 119–130. <https://doi.org/10.1016/j.tecto.2005.05.051>
- Rhoades, D. A., & Gerstenberger, M. C. (2009). Mixture models for improved short-term earthquake forecasting. *Bulletin of the Seismological Society of America*, 99(2 A), 636–646. <https://doi.org/10.1785/0120080063>
- Rhoades, D. A., Gerstenberger, M. C., Christophersen, A., Zechar, J. D., Schorlemmer, D., Werner, M. J., & Jordan, T. H. (2014). Regional earthquake likelihood models II: Information gains of multiplicative hybrids. *Bulletin of the Seismological Society of America*, 104(6), 3072–3083. <https://doi.org/10.1785/0120140035>
- Rhoades, D. A., Liukis, M., Christophersen, A., & Gerstenberger, M. C. (2016). Retrospective tests of hybrid operational earthquake forecasting models for Canterbury. *Geophysical Journal International*, 204(1), 440–456. <https://doi.org/10.1093/gji/ggv447>
- Rhoades, D. A., Rastin, S. J., & Christophersen, A. (2020). The effect of catalogue lead time on medium-term earthquake forecasting with application to New Zealand data [Online]. *Entropy*, 22(11), 1264. <https://doi.org/10.3390/e22111264>
- Rhoades, D. A., Rastin, S. J., & Christophersen, A. (2022). A 20-year journey of forecasting with the “every earthquake a precursor according to scale” model [Online]. *Geosciences*, 12(9), 349. <https://doi.org/10.3390/geosciences12090349>
- Rhoades, D. A., Schorlemmer, D., Gerstenberger, M. C., Christophersen, A., Zechar, J. D., & Imoto, M. (2011). Efficient testing of earthquake forecasting models. *Acta Geophysica*, 59(4), 728–747. <https://doi.org/10.2478/s11600-011-0013-5>
- Rhoades, D. A., & Stirling, M. W. (2012). An earthquake likelihood model based on proximity to mapped faults and cataloged earthquakes. *Bulletin of the Seismological Society of America*, 102(4), 1593–1599. <https://doi.org/10.1785/0120110326>
- Richter, C. F. (1935). An instrumental earthquake magnitude scale. *Bulletin of the Seismological Society of America*, 25(1), 1–32. <https://doi.org/10.1785/bssa0250010001>
- Ripley, B. D. (2005). *Spatial statistics*. John Wiley & Sons.
- Ristau, J. (2008). Implementation of routine regional moment tensor analysis of New Zealand. *Seismological Research Letters*, 79(3), 400–415. <https://doi.org/10.1785/gssrl.79.3.400>
- Ristau, J. (2009). Comparison of magnitude estimates for New Zealand earthquakes: Moment magnitude; local magnitude and teleseismic body-wave magnitude. *Bulletin of the Seismological Society of America*, 99(3), 1841–1852. <https://doi.org/10.1785/0120080237>
- Rivière, J., Lv, Z., Johnson, P. A., & Marone, C. (2018). Evolution of b-value during the seismic cycle: Insights from laboratory experiments on simulated faults. *Earth and Planetary Science Letters*, 482, 407–413. <https://doi.org/10.1016/j.epsl.2017.11.036>

- Ross, G. J. (2021). Bayesian estimation of the ETAS model for earthquake occurrences. *Bulletin of the Seismological Society of America*, 111(3), 1473–1480. <https://doi.org/10.1785/0120200198>
- Ross, Z. E., & Cochran, E. S. (2021). Evidence for latent crustal fluid injection transients in Southern California from long-duration earthquake swarms. *Geophysical Research Letters*, 48(12), e2021GL092465. <https://doi.org/10.1029/2021gl092465>
- Ross, Z. E., Meier, M. A., & Hauksson, E. (2018). P wave arrival picking and first-motion polarity determination with deep learning. *Journal of Geophysical Research: Solid Earth*, 123(6), 5120–5129. <https://doi.org/10.1029/2017jb015251>
- Rouet-Leduc, B., Hulbert, C., Lubbers, N., Barros, K., Humphreys, C. J., & Johnson, P. A. (2017). Machine learning predicts laboratory earthquakes. *Geophysical Research Letters*, 44(18), 9276–9282. <https://doi.org/10.1002/2017gl074677>
- Savadori, L., Ronzani, P., Sillari, G., Di Bucci, D., & Dolce, M. (2022). Communicating seismic risk information: The effect of risk comparisons on risk perception sensitivity. *Frontiers in Communication*, 7, 743172. <https://doi.org/10.3389/fcomm.2022.743172>
- Savran, W. H., Bayona, J. A., Iturrieta, P., Asim, K. M., Bao, H., Bayliss, K., et al. (2022). pycsep: A python toolkit for earthquake forecast developers. *Seismological Society of America*, 93(5), 2858–2870. <https://doi.org/10.1785/0220220033>
- Savran, W. H., Werner, M. J., Marzocchi, W., Rhoades, D. A., Jackson, D. D., Milner, K., et al. (2020). Pseudopropective evaluation of UCERF3-ETAS forecasts during the 2019 Ridgecrest sequence. *Bulletin of the Seismological Society of America*, 110(4), 1799–1817. <https://doi.org/10.1785/0120200026>
- Schneider, C. R., Freeman, A. L. J., Spiegelhalter, D., & van der Linden, S. (2022). The effects of communicating scientific uncertainty on trust and decision making in a public health context. *Judgment and Decision Making*, 17(4), 849–882. <https://doi.org/10.1017/S1930297500008962>
- Schneider, M., Cotton, F., & Schweizer, P.-J. (2023). Criteria-based visualization design for hazard maps. *Natural Hazards and Earth System Sciences*, 23(7), 2505–2521. <https://doi.org/10.5194/nhess-23-2505-2023>
- Schneider, M., & Guttorp, P. (2020). Bayesian ETAS: Towards improved earthquake rate models in the Pacific Northwest. In *AGU Fall Meeting Abstracts* (Vol. 2020, p. NH003-0010).
- Schneider, M., McDowell, M., Guttorp, P., Steel, E. A., & Fleischhut, N. (2022). Effective uncertainty visualization for aftershock forecast maps. *Natural Hazards and Earth System Sciences*, 22(4), 1499–1518. <https://doi.org/10.5194/nhess-22-1499-2022>
- Schneider, M., Wein, A., van der Elst, N., McBride, S. K., Becker, J., Castro, R., et al. (2023). Visual communication of aftershock forecasts based on user needs: A case study of the United States. *Mexico and El Salvador*. <https://doi.org/10.31219/osf.io/5qam4>
- Scholz, C. H. (1968). The frequency-magnitude relation of microfracturing in rock and its relation to earthquakes. *Bulletin of the Seismological Society of America*, 58(1), 399–415. <https://doi.org/10.1785/bssa0580010399>
- Scholz, C. H. (2015). On the stress dependence of the earthquake b value. *Geophysical Research Letters*, 42(5), 1399–1402. <https://doi.org/10.1002/2014gl062863>
- Schorlemmer, D., & Gerstenberger, M. C. (2007). RELM testing center. *Seismological Research Letters*, 78(1), 30–36. <https://doi.org/10.1785/gssrl.78.1.30>
- Schorlemmer, D., Gerstenberger, M. C., Wiemer, S., Jackson, D. D., & Rhoades, D. A. (2007). Earthquake likelihood model testing. *Seismological Research Letters*, 78(1), 17–29. <https://doi.org/10.1785/gssrl.78.1.17>
- Schorlemmer, D., Werner, M., Marzocchi, W., Jordan, T., Ogata, Y., Jackson, D., et al. (2018). The collaboratory for the study of earthquake predictability: Achievements and priorities. *Seismological Research Letters*, 89(4), 1305–1313. <https://doi.org/10.1785/0220180053>
- Schorlemmer, D., Wiemer, S., & Wyss, M. (2004). Earthquake statistics at Parkfield: 1. Stationarity of b values. *Journal of Geophysical Research*, 109(B12), B12307. <https://doi.org/10.1029/2004jb003234>
- Schorlemmer, D., Wiemer, S., & Wyss, M. (2005). Variations in earthquake-size distribution across different stress regimes. *Nature*, 437(7058), 539–542. <https://doi.org/10.1038/nature04094>
- Schorlemmer, D., & Woessner, J. (2008). Probability of detecting an earthquake. *Bulletin of the Seismological Society of America*, 98(5), 2103–2117. <https://doi.org/10.1785/0120070105>
- Schorlemmer, D., Zechar, J. D., Werner, M. J., Field, E. H., Jackson, D. D., Jordan, T. H., & RELM Working Group. (2010). First results of the regional earthquake likelihood models experiment. *Seismogenesis and Earthquake Forecasting: The Frank Evison Volume II*, 167(8–9), 5–22. <https://doi.org/10.1007/s00024-010-0081-5>
- Seif, S., Mignan, A., Zechar, J. D., Werner, M. J., & Wiemer, S. (2017). Estimating ETAS: The effects of truncation, missing data, and model assumptions. *Journal of Geophysical Research: Solid Earth*, 122(1), 449–469. <https://doi.org/10.1002/2016jb012809>
- Selvadurai, P. A., Galvez, P., Mai, P. M., & Glaser, S. D. (2023). Modeling frictional precursory phenomena using a wear-based rate-and state-dependent friction model in the laboratory. *Tectonophysics*, 847, 229689. <https://doi.org/10.1016/j.tecto.2022.229689>
- Serafini, F., Naylor, M., Lindgren, F., Werner, M. J., & Main, I. (2022). Ranking earthquake forecasts using proper scoring rules: Binary events in a low probability environment. *Geophysical Journal International*, 230(2), 1419–1440. <https://doi.org/10.1093/gji/ggac124>
- Sharma, S., Hainzl, S., Zöeller, G., & Holschneider, M. (2020). Is Coulomb stress the best choice for aftershock forecasting? *Journal of Geophysical Research: Solid Earth*, 125(9), e2020JB019553. <https://doi.org/10.1029/2020JB019553>
- Shi, Y., & Bolt, B. A. (1982). The standard error of the magnitude-frequency b value. *Bulletin of the Seismological Society of America*, 72(5), 1677–1687. <https://doi.org/10.1785/bssa0720051677>
- Slade, S. C., Dionne, C. E., Underwood, M., & Buchbinder, R. (2014). Standardised method for reporting exercise programmes: Protocol for a modified Delphi study. *BMJ Open*, 4(12), e006682. <https://doi.org/10.1136/bmjopen-2014-006682>
- Slovic, P. (2016). *The perception of risk*. Routledge.
- Smith, K. D., von Seggern, D., Blewitt, G., Preston, L., Anderson, J. G., Wernicke, B. P., & Davis, J. L. (2004). Evidence for deep magma injection beneath Lake Tahoe, Nevada-California. *Science*, 305(5688), 1277–1280. <https://doi.org/10.1126/science.1101304>
- Smith, W. D. (1981). The b -value as an earthquake precursor. *Nature*, 289(5794), 136–139. <https://doi.org/10.1038/289136a0>
- Spassiani, I., Falcone, G., Murru, M., & Marzocchi, W. (2023). Operational earthquake forecasting in Italy: Validation after 10 years of operativity. *Geophysical Journal International*, 234(3), 2501–2518. <https://doi.org/10.1093/gji/ggac256>
- Stallone, A., & Falcone, G. (2021). Missing earthquake data reconstruction in the space-time-magnitude domain. *Earth and Space Science*, 8, e2020EA001481. <https://doi.org/10.1029/2020ea001481>
- Stacy, S., Gerstenberger, M., Williams, C., Rhoades, D., & Christophersen, A. (2013). A new hybrid Coulomb/statistical model for forecasting aftershock rates. *Geophysical Journal International*, 196(2), 918–923. <https://doi.org/10.1093/gji/ggt404>
- Stieb, D. M., Huang, A., Hocking, R., Crouse, D. L., Osornio-Vargas, A. R., & Villeneuve, P. J. (2019). Using maps to communicate environmental exposures and health risks: Review and best-practice recommendations. *Environmental Research*, 176, 108518. <https://doi.org/10.1016/j.envres.2019.05.049>
- Stirling, M., McVerry, G., Gerstenberger, M., Litchfield, N., Van Dissen, R., Berryman, K., et al. (2012). National seismic hazard model for New Zealand: 2010 update. *Bulletin of the Seismological Society of America*, 102(4), 1514–1542. <https://doi.org/10.1785/0120110170>

- Stockman, S., Lawson, D. J., & Werner, M. J. (2023). Forecasting the 2016–2017 Central Apennines earthquake sequence with a neural point process. *Earth's Future*, 11(9), e2023EF003777. <https://doi.org/10.1029/2023EF003777>
- Strader, A., Werner, M., Bayona, J., Maechling, P., Silva, F., Liukis, M., & Schorlemmer, D. (2018). Prospective evaluation of global earthquake forecast models: 2 yrs of observations provide preliminary support for merging smoothed seismicity with geodetic strain rates. *Seismological Research Letters*, 89(4), 1262–1271. <https://doi.org/10.1785/0220180051>
- Suárez, G. (2022). The seismic early warning system of Mexico (SASMEX): A retrospective view and future challenges. *Frontiers in Earth Science*, 10, 827236. <https://doi.org/10.3389/feart.2022.827236>
- Tamaribuchi, K., Yagi, Y., Enescu, B., & Hirano, S. (2018). Characteristics of foreshock activity inferred from the JMA earthquake catalog. *Earth, Planets and Space*, 70(1), 1–13. <https://doi.org/10.1186/s40623-018-0866-9>
- Taroni, M. (2023). Against Bath's law: When aftershocks became mainshocks—Implications for earthquake forecasting communication. *Seismological Research Letters*, 94(6), 2565–2568. <https://doi.org/10.1785/0220230080>
- Taroni, M., Marzocchi, W., Schorlemmer, D., Werner, M. J., Wiemer, S., Zechar, J. D., et al. (2018). Prospective CSEP evaluation of 1-day, 3-month, and 5-yr earthquake forecasts for Italy. *Seismological Research Letters*, 89(4), 1251–1261. <https://doi.org/10.1785/0220180031>
- Tebaldi, C., & Knutti, R. (2007). The use of the multi-model ensemble in probabilistic climate projections. *Philosophical Transactions of the Royal Society A: Mathematical, Physical & Engineering Sciences*, 365(1857), 2053–2075. <https://doi.org/10.1098/rsta.2007.2076>
- Thompson Clive, M. A., Lindsay, J. M., Leonard, G. S., Lutteroth, C., Bostrom, A., & Corballis, P. (2021). Volcanic hazard map visualisation affects cognition and crisis decision-making. *International Journal of Disaster Risk Reduction*, 55, 102102. <https://doi.org/10.1016/j.ijdrr.2021.102102>
- Tormann, T., Enescu, B., Woessner, J., & Wiemer, S. (2015). Randomness of megathrust earthquakes implied by rapid stress recovery after the Japan earthquake. *Nature Geoscience*, 8(2), 152–158. <https://doi.org/10.1038/ngeo2343>
- Tormann, T., Wiemer, S., & Hardebeck, J. L. (2012). Earthquake recurrence models fail when earthquakes fail to reset the stress field. *Geophysical Research Letters*, 39(18), L18310. <https://doi.org/10.1029/2012gl052913>
- Tormann, T., Wiemer, S., & Mignan, A. (2014). Systematic survey of high-resolution *b* value imaging along Californian faults: Inference on asperities. *Journal of Geophysical Research: Solid Earth*, 119(3), 2029–2054. <https://doi.org/10.1002/2013jb010867>
- U.S. Geological Survey (USGS), Earthquake Hazards Program. (2017). Advanced National Seismic System (ANSS) Comprehensive Catalog of Earthquake Events and Products: Various. <https://doi.org/10.5066/F7M33QZH>
- Utsu, T. (1961). A statistical study on the occurrence of aftershocks. *Geophysical Magazine*, 30, 521–605.
- Utsu, T. (1971). Aftershocks and earthquake statistics (2): Further investigation of aftershocks and other earthquake sequences based on a new classification of earthquake sequences. *Journal of the Faculty of Science, Hokkaido University - Series 7: Geophysics*, 3(4), 197–266.
- Utsu, T. (1972). Aftershocks and earthquake statistics (3): Analyses of the distribution of earthquakes in magnitude, time and space with special consideration to clustering characteristics of earthquake occurrence (1). *Journal of the Faculty of Science, Hokkaido University - Series 7: Geophysics*, 3(5), 379–441.
- van der Elst, N., Hardebeck, J. L., & Michael, A. J. (2020). *Potential duration of aftershocks of the 2020 southwestern Puerto Rico earthquake* (No. 2020-1009). US Geological Survey.
- van der Elst, N. J., Hardebeck, J. L., Michael, A. J., McBride, S. K., & Vanacore, E. (2022). Prospective and retrospective evaluation of the U.S. Geological Survey Public Aftershock Forecast for the 2019–2021 Southwest Puerto Rico Earthquake and Aftershocks. *Seismological Research Letters*, 93(2A), 620–640. <https://doi.org/10.1785/0220210222>
- van der Elst, N. J., & Page, M. T. (2017). Nonparametric aftershock forecasts based on similar sequences in the past. *Seismological Research Letters*, 89(1), 145–152. <https://doi.org/10.1785/0220170155>
- van der Elst, N. J., & Shaw, B. E. (2015). Larger aftershocks happen farther away: Nonseparability of magnitude and spatial distributions of aftershocks. *Geophysical Research Letters*, 42(14), 5771–5778. <https://doi.org/10.1002/2015gl064734>
- Veen, A., & Schoenberg, F. P. (2006). *Assessing spatial point process models using weighted K-functions: Analysis of California earthquakes*. Springer.
- Veen, A., & Schoenberg, F. P. (2008). Estimation of space–time branching process models in seismology using an em–type algorithm. *Journal of the American Statistical Association*, 103(482), 614–624. <https://doi.org/10.1198/01621450800000148>
- Vere-Jones, D. (1995). Forecasting earthquakes and earthquake risk. *International Journal of Forecasting*, 11(4), 503–538. [https://doi.org/10.1016/0169-2070\(95\)00621-4](https://doi.org/10.1016/0169-2070(95)00621-4)
- Vidale, J. E., & Shearer, P. M. (2006). A survey of 71 earthquake bursts across southern California: Exploring the role of pore fluid pressure fluctuations and aseismic slip as drivers. *Journal of Geophysical Research*, 111(B5). <https://doi.org/10.1029/2005jb004034>
- Vogel, C., Zwolinsky, S., Griffiths, C., Hobbs, M., Henderson, E., & Wilkins, E. (2019). A Delphi study to build consensus on the definition and use of big data in obesity research. *International Journal of Obesity*, 43(12), 2573–2586. <https://doi.org/10.1038/s41366-018-0313-9>
- Wein, A., Becker, J. S., McBride, S. K., Potter, S. H., Doyle, E. E. H., Detweiler, S. T., & Pollitz, F. (2016). Constructing better communication for Operational Earthquake Forecasting of aftershocks. In *Proceedings of the 11th United States-Japan Natural Resources Panel for Earthquake Research* (pp. 87–90).
- Wein, A., Potter, S., Johal, S., Doyle, E., & Becker, J. (2016). Communicating with the public during an earthquake sequence: Improving communication of geoscience by coordinating roles. *Seismological Research Letters*, 87(1), 112–118. <https://doi.org/10.1785/0220150113>
- Werner, M. J., Helmstetter, A., Jackson, D. D., & Kagan, Y. Y. (2011). High-resolution long-term and short-term earthquake forecasts for California. *Bulletin of the Seismological Society of America*, 101(4), 1630–1648. <https://doi.org/10.1785/0120090340>
- Werner, M. J., & Sornette, D. (2008). Magnitude uncertainties impact seismic rate estimates, forecasts, and predictability experiments. *Journal of Geophysical Research*, 113(B8). <https://doi.org/10.1029/2007jb005427>
- Werner, M. J., Zechar, J. D., Marzocchi, W., & Wiemer, S. (2010). Retrospective evaluation of the five-year and ten-year CSEP-Italy earthquake forecasts. arXiv preprint arXiv:1003.1092.
- Wiemer, S., Danciu, L., Edwards, B., Marti, M., Fäh, D., Hiemer, S., et al. (2016). Seismic Hazard Model 2015 for Switzerland (SUIhaz2015). <https://doi.org/10.12686/A2>
- Wiemer, S., Gerstenberger, M., & Hauksson, E. (2002). Properties of the aftershock sequence of the 1999 M_w 7.1 Hector Mine earthquake: Implications for aftershock hazard. *Bulletin of the Seismological Society of America*, 92(4), 1227–1240. <https://doi.org/10.1785/0120000914>
- Wiemer, S., & Katsumata, K. (1999). Spatial variability of seismicity parameters in aftershock zones. *Journal of Geophysical Research*, 104(B6), 13135–13151. <https://doi.org/10.1029/1999jb900032>
- Wiemer, S., & Wyss, M. (1997). Mapping the frequency–magnitude distribution in asperities: An improved technique to calculate recurrence times? *Journal of Geophysical Research*, 102(B7), 15115–15128. <https://doi.org/10.1029/97jb00726>
- Wiemer, S., & Wyss, M. (2000). Minimum magnitude of completeness in earthquake catalogs: Examples from Alaska, the western United States, and Japan. *Bulletin of the Seismological Society of America*, 90(4), 859–869. <https://doi.org/10.1785/0119990114>

- Wiemer, S., & Wyss, M. (2002). Mapping spatial variability of the frequency-magnitude distribution of earthquakes. In *Advances in geophysics* (Vol. 45, p. 259-V). Elsevier. [https://doi.org/10.1016/s0065-2687\(02\)80007-3](https://doi.org/10.1016/s0065-2687(02)80007-3)
- Woessner, J., Christophersen, A., Zechar, J. D., & Monelli, D. (2010). Building self-consistent, short-term earthquake probability (step) models: Improved strategies and calibration procedures. *Annals of Geophysics*, 53(3), 141–154. <https://doi.org/10.4401/ag-4812>
- Wood, H. O., & Neumann, F. (1931). Modified Mercalli intensity scale of 1931. *Bulletin of the Seismological Society of America*, 21(4), 277–283. <https://doi.org/10.1785/bssa0210040277>
- Wood, M. M., Mileti, D. S., Bean, H., Liu, B. F., Sutton, J., & Madden, S. (2018). Milling and public warnings. *Environment and Behavior*, 50(5), 535–566. <https://doi.org/10.1177/0013916517709561>
- Wood, M. M., Mileti, D. S., Kano, M., Kelley, M. M., Regan, R., & Bourque, L. B. (2012). Communicating actionable risk for terrorism and other hazards. *Risk Analysis*, 32(4), 601–615. <https://doi.org/10.1111/j.1539-6924.2011.01645.x>
- Xiong, Q., Brudzinski, M. R., Gossett, D., Lin, Q., & Hampton, J. C. (2023). Seismic magnitude clustering is prevalent in field and laboratory catalogs. *Nature Communications*, 14(1), 2056. <https://doi.org/10.1038/s41467-023-37782-5>
- Xiong, Z., & Zhuang, J. (2023). SETAS: A Spherical Version of the Space–Time ETAS Model. *Seismological Society of America*, 94(3), 1676–1688. <https://doi.org/10.1785/0220220198>
- Yeo, G. L., & Cornell, C. A. (2009). A probabilistic framework for quantification of aftershock ground-motion hazard in California: Methodology and parametric study. *Earthquake Engineering & Structural Dynamics*, 38(1), 45–60. <https://doi.org/10.1002/eqe.840>
- Zaliapin, I., & Ben-Zion, Y. (2016). A global classification and characterization of earthquake clusters. *Geophysical Journal International*, 207(1), 608–634. <https://doi.org/10.1093/gji/ggw300>
- Zechar, J. D., Gerstenberger, M. C., & Rhoades, D. A. (2010). Likelihood-based tests for evaluating space–rate–magnitude earthquake forecasts. *Bulletin of the Seismological Society of America*, 100(3), 1184–1195. <https://doi.org/10.1785/0120090192>
- Zechar, J. D., Schorlemmer, D., Liukis, M., Yu, J., Euchner, F., Maechling, P. J., & Jordan, T. H. (2010). The Collaboratory for the Study of Earthquake Predictability perspective on computational earthquake science. *Concurrency and Computation: Practice and Experience*, 22(12), 1836–1847. <https://doi.org/10.1002/cpe.1519>
- Zhuang, J. (2011). Next-day earthquake forecasts for the Japan region generated by the ETAS model. *Earth, Planets and Space*, 63(3), 207–216. <https://doi.org/10.5047/eps.2010.12.010>
- Zhuang, J., Matsu'ura, M., & Han, P. (2021). Critical zone of the branching crack model for earthquakes: Inherent randomness, earthquake predictability, and precursor modelling. *The European Physical Journal Special Topics*, 230(1), 409–424. <https://doi.org/10.1140/epjst/e2020-000272-7>
- Zhuang, J., Ogata, Y., & Wang, T. (2017). Data completeness of the Kumamoto earthquake sequence in the JMA catalog and its influence on the estimation of the ETAS parameters. *Earth, Planets and Space*, 69, 1–12. <https://doi.org/10.1186/s40623-017-0614-6>
- Zhuang, J., Wang, T., & Kiyosugi, K. (2020). Detection and replenishment of missing data in marked point processes. *Statistica Sinica*, 30(4), 2105–2130. <https://doi.org/10.5705/ss.202017.0403>
- Zhuang, J., Werner, M. J., Hainzl, S., Harte, D., & Zhou, S. (2011). Basic models of seismicity: Spatiotemporal models. *Community Online Resource for Statistical Seismicity Analysis*. <https://doi.org/10.5078/corssa-07487583>
- Zlydenko, O., Elidan, G., Hassidim, A., Kukliansky, D., Matias, Y., Meade, B., et al. (2023). A neural encoder for earthquake rate forecasting. *Scientific Reports*, 13(1), 12350. <https://doi.org/10.1038/s41598-023-38033-9>