Deliverable

D7.3 How to define the best OEF model to be used for societal purposes: ensemble modelling

Deliverable information	
Work package	WP7: Rigorous testing and validation of dynamic risk components
Lead	GFZ-Potsdam
Authors	Warner Marzocchi (UniNa), Marcus Herrmann (UniNa)
Reviewers	Maximilian J. Werner
Approval	Management Board
Status	Final
Dissemination level	Public
Will the data supporting this document be made open access? (Y/N)	Yes (upon request to the WP leader)
If No Open Access, provide reasons	
Delivery deadline	28.02.2022
Submission date	28.02.2022
Intranet path	[DOCUMENTS/DELIVERABLES/File Name]

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

Table of contents

Summary

Ensemble modeling is a general mathematical procedure to combine forecasts provided by different models (or different parametrizations of the same model) in one single ensemble forecast. Ensemble modeling is named in different ways and approached through different philosophies. Here we summarize and discuss the main generic concepts including their shortcomings and raise open issues, such as the possibility to validate the ensemble forecasting model and the importance of keeping uncertainties of different nature separate to obtain a complete description of what we know and what we do not know. We introduce the principles of ontological ensemble (OE) modeling, which is rooted in a unified probabilistic framework. To create the OE forecast, the individual models are weighted to maximize the skill of the OE itself. But instead of collapsing the individual forecast distributions into a single ensemble distribution, the OE forecast maintains various kinds of uncertainties to acknowledge the ignorance of the "true" model. Based on theoretical concepts and a practical application to the operational earthquake forecasting system in Italy, we show how the OE modeling allows us to overcome some of the shortcomings of classical ensemble methods.

1. Introduction

Uncertainties of different kinds are pervasive in natural systems. Accounting for these uncertainties implies that the evolution of such natural systems can be cast only in probabilistic terms. In this context, we use the term forecast to describe any probabilistic statement about the future occurrence of a natural threat (Jordan et al., 2011; see Gneiting and Katzfuss, 2014 for a full appraisal of the problem).

In the most advanced real applications, scientists use a set of forecasting models that may be rooted in different kinds of modeling, ranging from entirely empirical models (e.g., see Gerstenberger et al., 2021, for the case of seismic hazard) to deterministic models, which provide probabilities when sampling the unavoidable uncertainty in the initial and/or boundary conditions (e.g., Murphy and Palmer, 1996; Folch et al., 2022); a single deterministic model can also generate different forecasts given the uncertainty over its parametrization (e.g., Senior Seismic Hazard Analysis Committee, NRC 2017). Loosely speaking, the *i*-th forecast is expected to describe the intrinsic variability of the process (aleatory variability), which may be described by a probabilistic survival function $f_i(x) = P_i(X > x)$**,** where X is the hazard intensity of interest in one specific spatiotemporal window. Specifically, the function $f(x)$ provides the probability of exceedance (PoE) for each hazard intensity value x; it is often called the hazard curve and should not be confused with the hazard function commonly used to describe failure rate in survival analysis.

The post-processing integration of all forecasts $f_i(x)$ is meant to include also the uncertainty related to the lack of knowledge of the "true" model (epistemic uncertainty), with the goal of improving the forecasting skill (Palmer et al., 2004; Marzocchi et al., 2012; Krishnamurti et al., 2000; 2016). This post-processing operation is often named differently, such as multimodel ensemble or forecast (e.g., Khrishnamurti et al., 2000; Palmer et al., 2004; Tebaldi & Knutti, 2007) or superensemble (Khrishnamurti et al., 2016; Bottazzi et al., 2021). Here we use the general term ensemble modeling to describe any post-processing procedure that aims at providing a complete description of the hazard forecast, including all known uncertainties.

Ensemble modeling is currently approached through two different perspectives that we call point estimation and probabilistic (see Figure 1). In the point estimation perspective (Figure 1 bottom left), ensemble modeling is focused on getting the point forecast that is expected to be the closest to the real observation (e.g., Khrishnamurti et al., 2016). In the probabilistic perspective (Figure 1 bottom center), all forecasts $f_i(x)$ of one variable of interest (i.e., the intensity of the hazard) are collapsed into one single ensemble probabilistic forecast $\bar{f}(x)$ using different strategies (Gneiting and Ranjan, 2013).

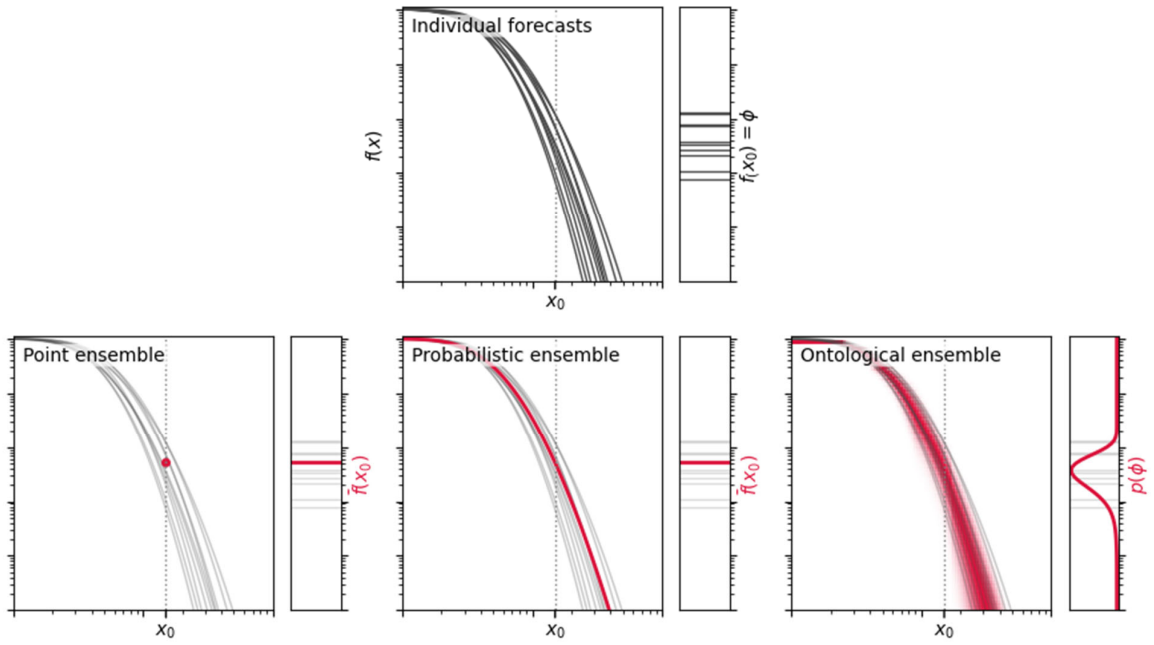

Figure 1. Schematic illustration of different ensemble approaches. Top: Probabilistic forecasts of ten individual models in terms of hazard curves (left) and a slice at hazard level x_0. Bottom: Ensembles (red) of the individual forecasts using different ensemble approaches. The ontological ensemble modelling (OEM) presented in this deliverable is shown on the right. For simplicity, the individual forecasts are weighted uniformly. Note that x and $f(x)$ is shown on a logarithmic scale.

Although the latter approach is widely used, it has some remarkable shortcomings and leaves open important questions that are worth being considered in detail. First, current methods of probabilistic ensemble modeling do not preserve the distinction between different kinds of uncertainty. De facto, these methods integrate all uncertainties into one single probability distribution (Figure 1 bottom center). We argue that this approach does not provide a full picture of what we know and what we do not know to the decision makers. Remarkably, this need is recognized in many fields. For example, the most recent IPCC reports (IPCC, 2021) implicitly call for the need to have a more complete way to describe uncertainties of different kinds; in these reports, each model produces a likelihood (as the ensemble model $\bar{f}(x)$) characterized by an additional heuristic "high, medium, and low confidence" (which is not the statistical confidence) to communicate the reliability of the model.

Figure 2 shows another example to illustrate the importance of keeping aleatory variability and epistemic uncertainty distinct for a complete information to the decision makers. It shows seismic hazard curves derived from different logic tree branches at two sites in the United States, in Memphis and on the San Andreas fault. In essence, if we just look at the mean hazard, $\bar{f}(x)$, as usually done in practice (the mean hazard could be seen as a sort of ensemble modeling among the different logic tree branches), the horizontal peak ground acceleration that has a PoE of 2% in 50 years is the same in Memphis (MEM) and in the San Andreas fault (SAF). However, the dispersion of all hazard curves $f_i(x)$ around the mean hazard is much larger in MEM than in SAF, which indicates a better constrained ensemble $\bar{f}(x)$ in SAF as opposed to MEM.

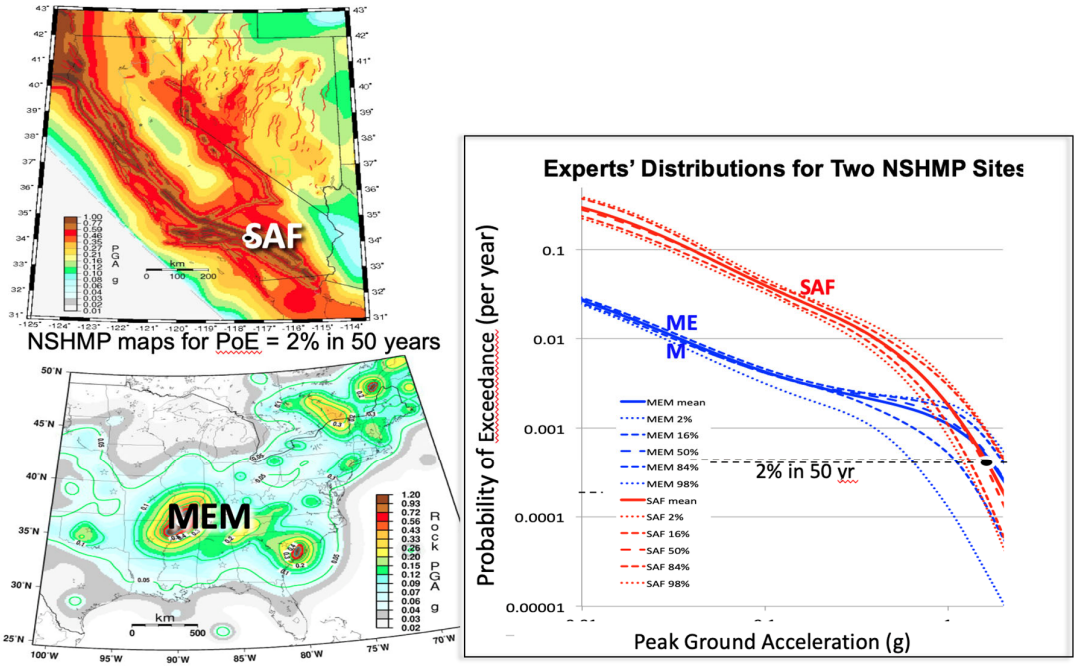

Figure 2. Example of the seismic hazard in United States according to the model of 2008. Looking only at the mean hazard Memphis (MEM) and the San Andreas Fault (SAF) in California have the same PGA relative to the PoE of 2% in 50 years which is an important parameter to define the building code. However, if we look at the dispersion around the mean hazard we may notice that the dispersion of the curves in MEM is much larger than in SAF. [The figure has been produced by USGS].

Second, current probabilistic ensemble methods aim to represent the best model given the present knowledge, but do not allow to be validated. This aspect is of fundamental importance for both practical applications and science. Regarding practical applications, the reliability of models is fundamental for their societal use (Jordan et al., 2011; Palmer, 2018). Regarding the more general scientific enterprise, science is rooted in the concept that a model can be tested against observations and rejected when necessary. The difficulty (or impossibility) to validate forecasting models of natural hazards is that threatening events happen in open systems. For example, Oreskes et al. (1994) state *"... it is impossible to demonstrate the truth of any proposition, except in a closed system"*, and *"Models can only be evaluated in relative terms, and their predictive value is always open to question. The primary value of models is heuristic".* This difficulty/impossibility of validation stands also at the root of a popular aphorism in statistics *"all models are wrong"* (Box, 1976), and so why waste time to validate them? We already know the answer. From a more technical point of view, Marzocchi and Jordan (2014, 2017) questioned that probabilistic models integrating all uncertainties in one single distribution cannot be meaningfully validated.

Third, current ensemble models do not explain if the reliability (or confidence in the IPCC language) of $\bar{f}(x)$ is expected to improve when increasing the number of available models; this should be expected under the assumption that errors tend to cancel if the models are independent, and thus decrease uncertainty as the number of models increases.

In this paper, we adopt a recently developed unified probabilistic framework (Marzocchi and Jordan, 2014) to build an ensemble model of earthquake forecasts that addresses all three points discussed above. In particular, this new method, which we call 'ontological ensemble modeling' (OEM, see Figure 1 bottom right), embeds a univocal hierarchy of uncertainties (aleatory variability, epistemic uncertainty, and ontological error) that have to be kept separated and allows model validation. We show the practical importance of these features by applying the procedure to the operational earthquake forecasting system in Italy.

2. Principles of ontological ensemble modeling (OEM)

Natural systems are pervaded by uncertainties of different kinds that are handled differently in different probabilistic frameworks. Marzocchi and Jordan (2014) introduced a unified probabilistic framework that is rooted in a univocal hierarchy of uncertainties. The term 'unified' comes from the fact that the framework is based on the interplay between subjective (Bayesian) and frequentist methods; it aligns well with the attitude of statistical unificationists (to our knowledge a first attempt in this direction is described by Rubin, 1984). In the unified framework discussed here, the probability is an unknown frequency treated as a random variable through the Bayesian mathematical apparatus.

The cornerstone and starting point of the framework is the definition of an experimental concept, external to the probabilistic model, that identifies collections of data (observed and not yet observed) judged to be stochastically exchangeable (i.e., with joint probability distributions invariant to data ordering) when conditioned on a set of explanatory variables (Draper et al. 1993). Pragmatically, the experimental concept is somehow related to the usefulness of the model, since the exchangeable sequence $\mathbf{e}_N = \{e_n : n = 1, 2, \dots, N\}$ represents what we want to describe with the model itself; at the same time, the definition of the experimental concept implicitly embeds the current state of knowledge about the process.

De Finetti's representation theorem states that an infinite exchangeable sequence has an unknown frequency $\hat{\phi}$, which can be interpreted as the aleatory variability of the experimental concept. This frequency is the target of a forecasting model, and the dispersion of the set of forecasts $\phi_i = f_i(x_0) = P_i(X > x_0)$ represents the epistemic uncertainty. In essence, given a set of forecasts ϕ_i we can build the extended experts' distribution (EED), $p(\phi)$. We can now define an ontological null hypothesis, which states that the aleatory representation of future occurrence of natural events—the data generating process—mimics a sample from the probability distribution of aleatory representations that describe the model's epistemic uncertainty. In other words, it states that the "true" unknown frequency $\hat{\phi}$ of the experimental concept is a sample of the EED, i.e., $\hat{\phi} \sim p(\phi)$. If the ontological null hypothesis of this relation is rejected, we found an ontological error, i.e., using the popular words of D. Rumsfeld, the existence of *"unknown unknowns".*

To summarize, the exchangeability judgment both on past data (needed to build the model) and future data (the object of forecasting; needed for model testing) implies that the future is stochastically predictable from past experience—a leap of faith that adds content and utility to the ontological null hypothesis. The difficulty in making proper exchangeability judgments underlies some problems of experimental reproducibility in the social sciences.

This probabilistic framework overcomes the drawback of the current ensemble modeling strategies that cannot be validated. For instance, in one of the most common applications (NRC 2017, Gneiting and Ranjani, 2011), all forecasts are collapsed into $\bar{f}(x)$ which is the linear combination (weighted average) of $f_i(x)$,

$$\bar{f}(x) = \sum_{i=1}^{m} f_i(x)\, \pi_i, \qquad (1)$$

where π_i is the weight of each forecast (discussed later on). It is worth noting that Eqn. 1 becomes the Bayesian model averaging (BMA) if we consider π_i as the probability of $f_i(x)$ to be the true forecast, or more pragmatically the one that should be used (Scherbaum and Kahn, 2011).

In this mathematical perspective, $\bar{f}(x)$ is not related to any physical process, but it is defined as conditional distribution relative to a (sub) sigma-algebra or, more informally, conditioned on the available information set of the physical process. Hence, $\bar{f}(x)$ is not expected to represent any distribution of data coming from a physical process, $\hat{f}(x)$, except when we have an infinite amount of information, the so-called "ideal" distribution for which $\bar{f}(x) = \hat{f}(x)$ (Gneiting and Katzfuss, 2014). In practice, the comparison of the distribution $\bar{f}(x)$ with real data is expected to lead to significant differences when $\bar{f}(x)$ contains epistemic uncertainty (i.e., when it has been built with a limited amount of information) and when we have enough independent data to detect it.

Besides keeping uncertainties of different kinds separated and allowing model validation, this probabilistic framework also offers an answer to the third point mentioned in the introduction. Since it is impossible to have fully independent models because they are all built using the same amount of information to some extent, the forecasts are meant to sample the current epistemic uncertainty that cannot be reduced by simply increasing the number of models. For example, let us consider the case of earthquake forecasting; even if models are built independently from different modelers, they all rely on the same earthquake catalog and try to describe it at best. However, the number of earthquakes occurred in the time period covered by the catalog could be on the left tail of the true distribution, i.e., be particularly low. This means that the forecasts coming from all models will be very likely affected by a common-mode error (here, an overall underestimation) that is unknown. In the unified probabilistic framework, we need more independent information, rather than more models, to reduce the epistemic uncertainty.

3. OEM in practice; weighting the forecasts and building the ensemble

3.1 Building the ensemble forecast distribution

The essence of the OEM is to move from the set of forecasts $\{\phi_i = f_i(x_0), \pi_i\}$ to the ontological ensemble forecast $p(\phi)$. Here we assume that $p(\phi)$ has a Beta distribution, which is particularly suitable to describe random variables in the range [0, 1]. Although the choice of any distribution becomes a possible source of ontological error, we argue that this is unavoidable; for instance, choosing a mean forecast $\bar{f}(x)$ as made in the classical procedures is like assuming a Dirac distribution for $p(\phi)$, see Figure 1 bottom left and center.

The parameters of the Beta distribution, α and β, are related to the weighted average $\bar{\phi}$ and weighted variance σ_ϕ^2 of $\{\phi_i, \pi_i\}$ through

$$\alpha = \left(\frac{\bar{\phi}\,(1-\bar{\phi})}{\sigma_\phi^2} - 1 \right) \bar{\phi} \tag{2}$$

and

$$\beta = \left(\frac{\bar{\phi}\,(1-\bar{\phi})}{\sigma_\phi^2} - 1 \right) (1-\bar{\phi}). \tag{3}$$

The ontological null hypothesis distribution $p(\phi)$ is given by the distribution Beta(α, β), which represents where $\hat{\phi}$ of the exchangeable sequence is expected to be (see Figure 1 bottom right).

Here we briefly explain the importance of using a distribution instead of one single probability in the usual model calibration (Gneiting et al., 2007). Let's take the calibration curve as an example, which is widely used to check heuristically the frequency of observations with their uncertainty as a function of the probability of the forecasts; in essence a calibration curve is the plot of the observed frequencies as a function of the associated forecasts. An implicit assumption of this analysis is that the forecasts represent the true expected frequency. In the unified probabilistic framework this means to use a Dirac distribution for $p(\phi)$, which would mean that we do not have any epistemic uncertainty. Adding a horizontal bar to the points of the calibration curve could make this validation applicable also in the unified probabilistic framework. This issue will be addressed in future work devoted to the validation phase in the unified framework.

In the following section we introduce the new weighting scheme considering the data coming from the OEF system in Italy (OEF-Italy; Marzocchi et al., 2014), which will be further explored in the next chapter for a full application of the ontological ensemble modeling.

3.2 Weighting the forecasts

A fundamental step for OEM is to assign weights, π_i, to each forecast. The weights can be assigned depending on the problem at hand. For long-term forecasts, which are usually characterized by a few (if any) independent data for testing, weights are often assigned through qualitative experts' elicitation schemes (e.g., Cooke, 1991; see also Meletti et al., 2021). When independent data for testing are sufficiently available (like, for example, in short-term earthquake forecasting and weather forecasting), weights can be assigned more quantitatively.

As regards earthquake forecasting, Marzocchi et al. (2012) proposed a weighting procedure based on the scoring of each single model, i.e., the model that performs better receives a higher weight. Although the procedure sounds reasonable, Monteith et al. (2011) show that this is not the optimal procedure to weight the ensemble model. Following their suggestions, we weight the models to maximize the forecasting performance of the ensemble model, instead of weighting each model according to their individual performance.

Determining the weights using multivariate logistic regression

The goal is to find weights that maximize the skill of the ensemble, or in other words minimize the error between the ensemble forecast and the observation. Here we choose the logistic regression to fit the ensemble of forecasts to the observed data. This type of regression uses the logistic function (see Figure 3), here in particular its multivariate extension to incorporate more than one forecast, to model a binary observable:

$$p(\boldsymbol{\lambda}) = \frac{1}{1 + e^{-g(\boldsymbol{\lambda})}} \quad \text{with } g(\boldsymbol{\lambda}) = \beta_0 + \beta_1 \ln \lambda_1 + \cdots + \beta_m \ln \lambda_m, \tag{4}$$

in which p is the output probability between 0 and 1, β_0 is the intercept, $\beta_{1,\ldots,m}$ the coefficients, and λ_i the rates of the individual forecasts $f_i(x_0)$ in form of feature vectors at a specific hazard level x_0. The logistic regression has its name from being a linear combination on the logistic, or log-odds, scale $\ln \frac{p}{1-p} = g(\boldsymbol{\lambda})$, i.e., the logistic function converts log-odds to probability. Due to its binary form, the observable, y, needs to be discretized to 0 or 1 (e.g., inside each spatiotemporal bin, $N = 0$ or $N \geq 1$ target events above a threshold magnitude x_0, see Figure 3). The above formulation suggests that the logistic regression is limited to only provide ensemble forecasts for a single threshold of the observable (e.g., hazard level x_0) rather than a full probability distribution (Hamill et al. 2008), e.g., a hazard curve. However, this limitation has been alleviated by Wilks (2009), who extended the logistic regression by including the threshold x as an additional feature (and the corresponding forecast data for $f_i(x)$). In section 4, we will however not demonstrate this extension and instead remain at a specific hazard level for the sake of illustrating the principles of OEM.

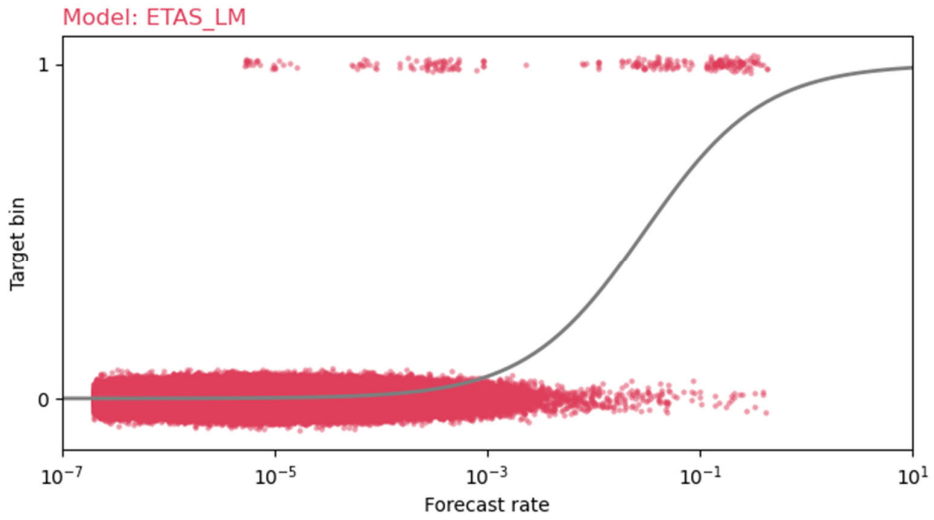

Figure 3. Example fit of the logistic function (gray curve) to one forecast model component (ETAS_LM) using one year of data. The observable (here: earthquakes with $M \geq 4$) is binned in spatiotemporal bins and becomes $y = 1$ if one or

more target events occur within it. For illustration purposes, the displayed data points are randomly jittered vertically (to not collapse them all on either 0 or 1) and shown with a transparency effect (to highlight denser parts).

The best fit to the observable, i.e., the β_0 and $\beta_{1,\dots,m}$ that maximize the skill of the ensemble, is obtained by maximizing the log-likelihood ℓ, the measure for the goodness of fit of the logistic regression:

$$\ell = \sum_{i:y_j=1} \ln(p_j) + \sum_{i:y_j=0} \ln(1 - p_j) = \sum_j [y_j \ln(p_j) + (1 - y_j) \ln(1 - p_j)], \quad (5)$$

in which y_j is the discretized observable and p_j the probability estimates in a particular spatiotemporal bin j.

To reduce the computational demand of fitting the logistic model (totaling 56 million samples, of which 99.996% are non-target bins), we only use 10% of the non-target bins (while keeping all target bins). This modification changes the prevalence of target bins and causes a bias in the estimated $\tilde{\beta}_0$ (King & Zeng, 2001), which can be corrected by subtracting $\beta_0^{\text{bias}} = -\ln 0.1$ to obtain β_0; the increased uncertainty in $\beta_{1,\dots,m}$ due to fewer data is negligible.

One may argue that the logistic fit itself represents already the ensemble model, as it translates the individual forecasts into a probabilistic output. In fact, the logistic regression is typically used in probabilistic forecasting because the output is a probability rather than a discrete quantity, for instance in meteorology (Wilks 2009). However, unlike weather forecasting, we have to deal with the situation that the observable (here: spatiotemporal bins where moderate or large earthquakes occur) are very rare (~0.004%, without downsampling non-target bins). This low fraction has major implications for obtaining a reliable fit to the data (Hamill et al. 2008) and limits the calibration of the forecast models in the ensemble. However, we can use the logistic regression to assess the *relative* performance of the models, and use this information to build the ensemble. In this case, we have to account for the forecasting horizon $t_{\text{fh}} > 0$ days to avoid data leakage when fitting the ensemble: because the forecasts at time t_n refer to the future, we can only use forecast data up to $t_n - t_{\text{fh}}$ (i.e., the ensemble can only be assessed after t_{fh} has passed)—although we can incorporate the observable up to t_n. In other words, the training data is validated 'out-of-sample' to assure their independence from data that will be observed only in the future. Because our target events are so rare, the logistic fit strongly depends on them, in particular how much the distribution of $\ln \lambda_i$ in target bins ($y = 1$) differs from the one in non-target bins ($y = 0$). As a consequence, the delay due to t_{fh} leads to a reduced utility of the fitted logistic model to represent a calibrated ensemble at t_n, i.e., it may perform well retrospectively at $t_n - t_{\text{fh}}$ (provided $\ln \lambda_i\,(y = 0)$ and $\ln \lambda_i\,(y = 1)$ are well-separated), but not prospectively at t_n; the fit may not be exactly applicable anymore once the individual forecast models adjusted the forecast rate due to new events (especially after a mainshock, which considerably changes the mapping between $\ln \lambda_i$ and y). This misfit due to the delay strongly affects β_0, but also $\beta_{1,\dots,m}$ (i.e., the model importance weights) may not be exactly appropriate anymore at t_n.

To avoid the strong dependence on β_0, we take a different approach: we only make use of the fitted $\beta_{1,\dots,m}$ and use them to build a weighted average of λ_i. Since the coefficients can be negative, we map them to pseudo-weights as follows:

$$w_i = \begin{cases} e^{\beta_i} - 1, & \text{for } \beta_i > 0 \\ 0, & \text{otherwise} \end{cases} \quad \text{with } i = 1, \dots, m; \quad (6)$$

in this way, a forecast attributed with a $\beta_i \leq 0$ by the logistic regression receives zero weight. Choosing an exponential relationship is related to the fact that e^{β_i} represent the odds ratio of the i-th model, i.e., that the odds multiply by this value for every 1-unit increase in $\ln \lambda_i$. Note that the sum of pseudo-weights does not necessarily equal 1.0 (hence 'pseudo'), and they need to be normalized:

$$\pi_i = \frac{w_i}{\sum_j^m w_j}, \quad (7)$$

so that the weighted-average ensemble $\bar{f}(\boldsymbol{\lambda})$ can be calculated with Eqn. 1.

4. OEM applied to operational earthquake forecasting

Operational Earthquake Forecasting (OEF) comprises procedures for gathering and disseminating authoritative information about the time dependence of seismic hazards to help communities prepare for potentially destructive earthquakes (Jordan et al., 2011). Seismologists are not able to predict large earthquakes with high probability in small spatiotemporal windows, but they are able to model the spatiotemporal clustering, which is the most striking deviation of the earthquake occurrence process from complete randomness. Such clustering is particularly pronounced in time windows of days to weeks, but it may be still relevant for up to one decade or more (depending on the mainshock magnitude and fault loading rate, see Stein & Liu, 2009).

Here we consider the OEF system in Italy (OEF-Italy) which consists of a combination of three models through an ensemble approach. The ensemble is produced using Score Model Averaging (SMA), in which the weight of each model is proportional to the inverse of the logarithmic score of the same model that is calculated from prospective testing. A full description of the SMA ensemble and the individual models can be found in Marzocchi et al. (2014) and references therein. Since April 2005, these models provide weekly forecast rates for earthquakes with $M \geq 3.95$ in the testing region of the corresponding CSEP (Collaboratory for the Study of Earthquake Predictability) experiment (see orange area in Figure 4). To date the consistency of the OEF-Italy system has been tested prospectively during the 2012 Emilia earthquake sequence and more recently for the 2016 central Italy sequence (Marzocchi et al., 2017). However, the adequacy of the ensemble modeling strategy has never been tested. In this section, we aim to find if a novel weighting scheme can improve the forecasting performance of OEF-Italy, and to describe how we can validate the OEF system.

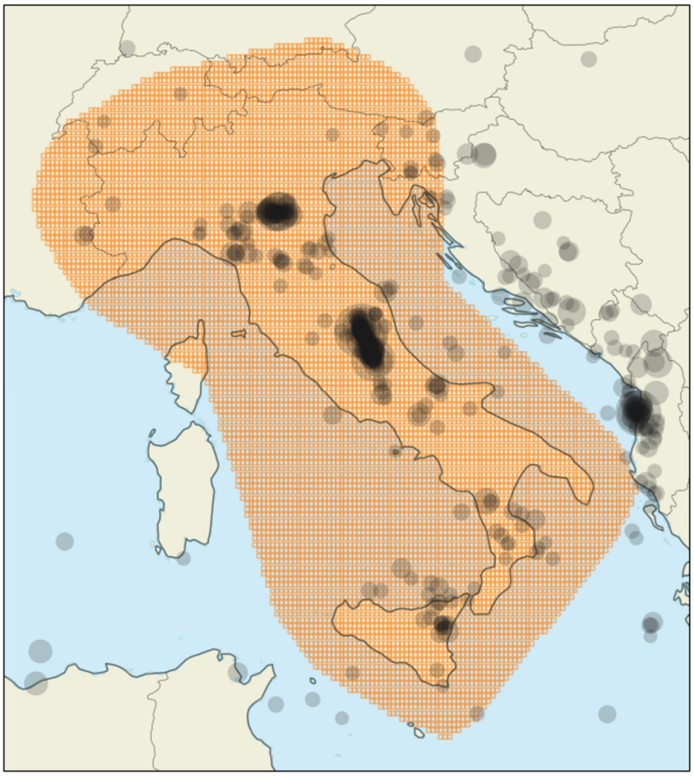

Figure 4. Map of the OEF system in Italy. Each model provides a rate forecast for each bin in the CSEP (Collaboratory for the Study of Earthquake Predictability) testing region of the Italian experiment (orange). $M \geq 3.95$ events between April 2005 and May 2020 are indicated by filled circles; only those inside the testing region are considered (as target events).

4.1 Building a new ensemble based on logistic regression

At each date (00:00 UTC) between April 2005 and May 2020, we fit the multivariate logistic model between the forecast rates of the individual models and the observation in the spatiotemporal bins (i.e., target earthquakes binned to the testing region's grid resulting in '0' [no targets] or '1' [one or more targets]) within the corresponding forecast period. To obtain model weights, we examined various temporal fitting schemes such as using all data since the beginning (Figure 5), or only data of the previous year (Figure 6). The first scheme lets the model weights π_i converge over time, whereas the latter scheme lets π_i reflect only the recent performance of the individual models. In both cases, the most notable weight change occurs during the L'Aquila sequence in April 2009. The other significant sequences in the recent past, i.e., Emilia in May 2012, and Central Italy in August—November 2016 cause only minor weight changes in the first scheme, likely because their relative contribution to the increasing amount of used data decreases over time. In the latter scheme, instead, these two sequences do not seem to play a role in readjusting the ensemble, or at least only for a brief period in time. Because the tends to "forget" model performance after one year, it provides some insights into the best performing model at a certain time. Accordingly, the ETAS_LM model appears to perform best during the L'Aquila and Central Italy sequences, but not during Emilia, in which ETES_FMC appears to perform best. STEP_LG generally appears to play a minor role.

Many more fitting schemes are imaginable, e.g., using exponentially decaying temporal window to pronounce the recent performance, or incorporating the spatial skill to account for the regionally varying seismicity and model performance.

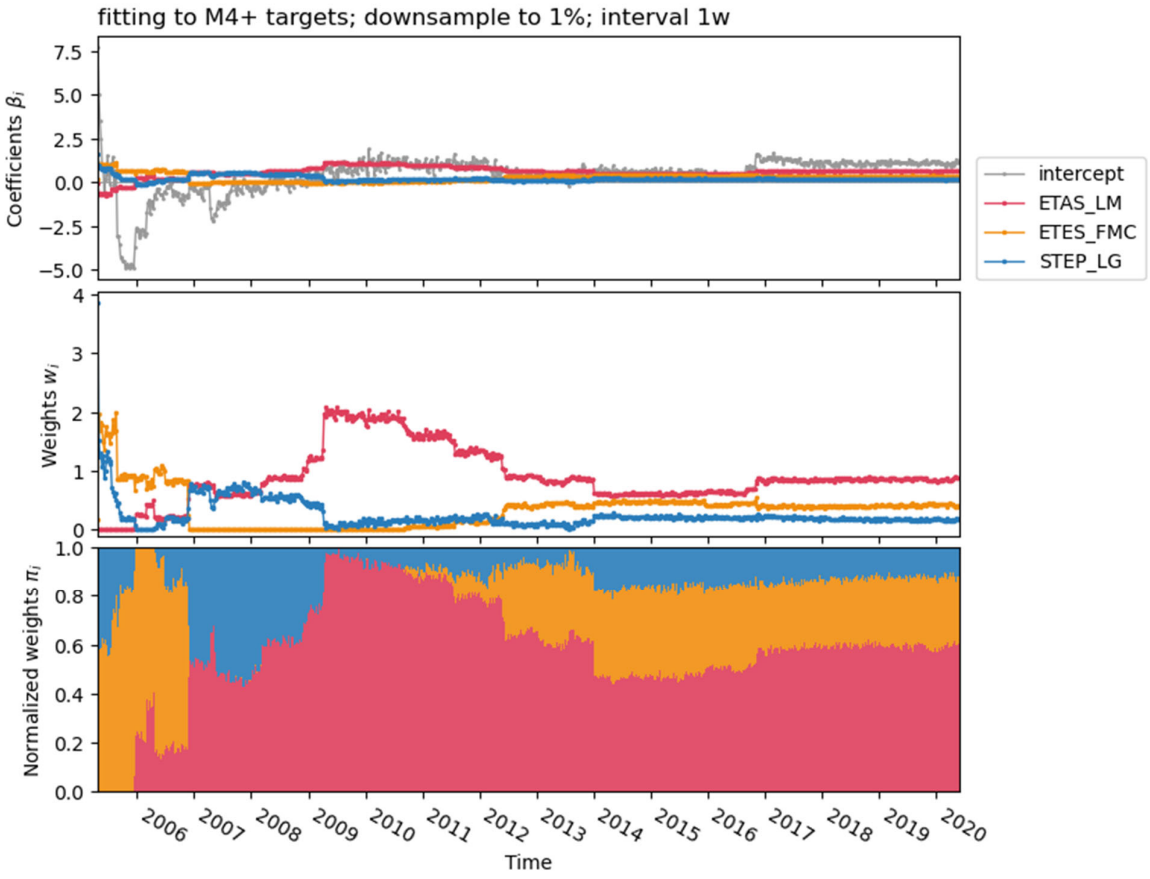

Figure 5. Applying the logistic regression over time using increasingly more data. Top: Logistic regression intercept β_0 after correcting for the bias due to downsampling, and coefficients $\beta_{1,2,3}$ for the three models; Middle: Regression coefficients mapped to pseudo-weights w_i; Bottom: Normalized weights π_i.

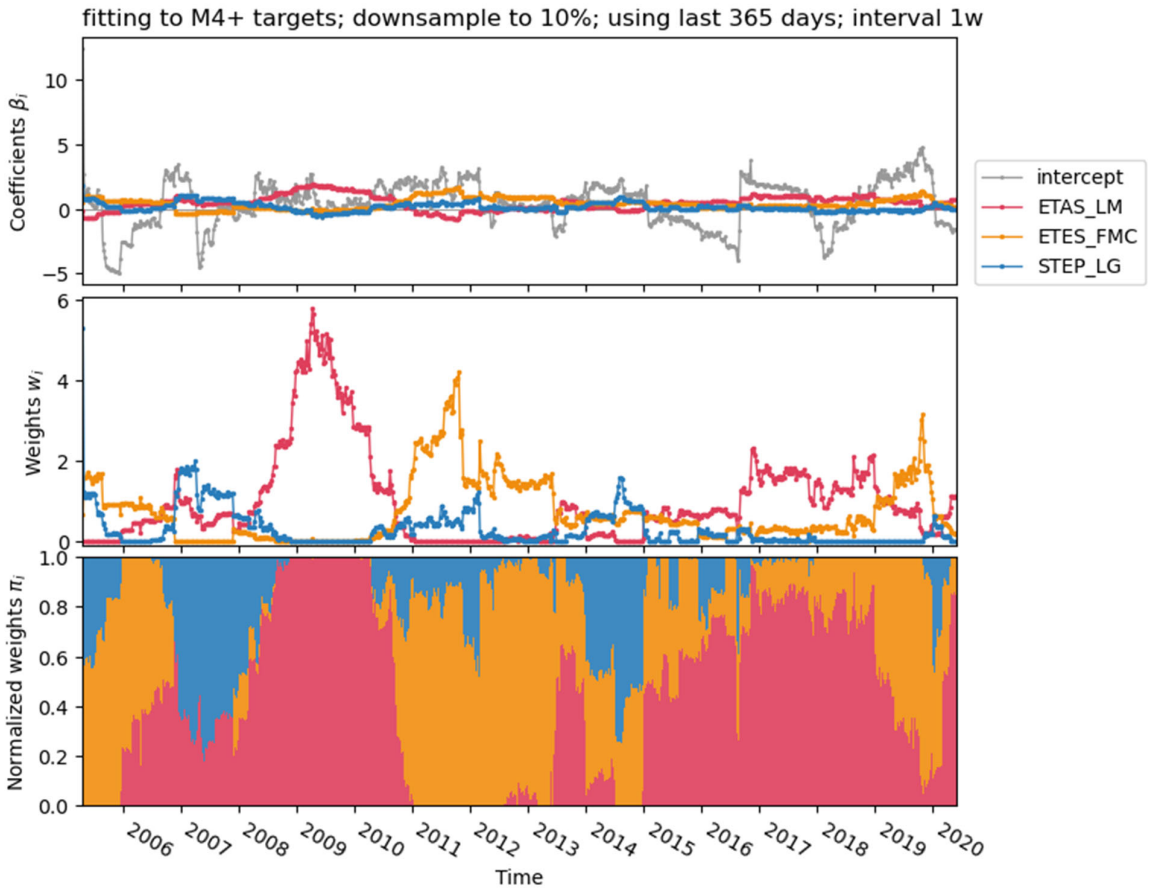

Figure 6. Like Figure 5, but applying the logistic regression only to data of the previous 365 days.

4.2 Assessing the performance

To evaluate the forecast performance quantitatively, we calculate the *information gain (IG) per event* (Rhoades et al. 2011) relative to the SMA ensemble as the reference model ():

$$I_{A,SMA} = \frac{1}{N}\sum_{k=1}^{N}\left(\log\left(\lambda_{j_k}^{A}\right) - \log\left(\lambda_{j_k}^{SMA}\right)\right) - \frac{\hat{N}^{A} - \hat{N}^{SMA}}{N}, \qquad (8)$$

in which model 'A' is synonymous for all the models for which we want to determine the (relative) performance, i.e., the logistic model $p(\boldsymbol{\lambda})$ itself, the weighted-average ensemble $\bar{f}(\boldsymbol{\lambda})$ based on the logistic fit, and each candidate model (ETAS_LM, ETES_FMC, and STEP_LG); λ_{j_k} is the forecast rate of any model in every bin j in which target event k occurs; $\hat{N} = \sum_j \lambda_j$ the total number of expected earthquakes of any model; and N the number of target events. As with model fitting, we only use forecasts issued on 00:00 UTC and skip all irregular forecast times.

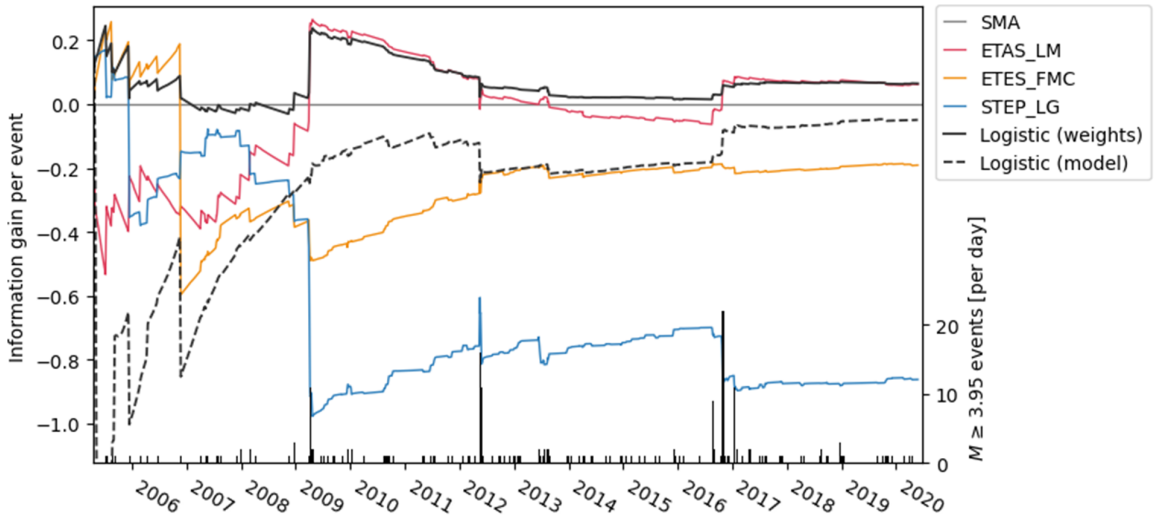

Figure 7. Information gain per event of each new ensemble and candidate model over the SMA ensemble using the first fitting scheme (increasingly more data, Figure 5). At the bottom, the daily rate of target events is displayed by vertical bars.

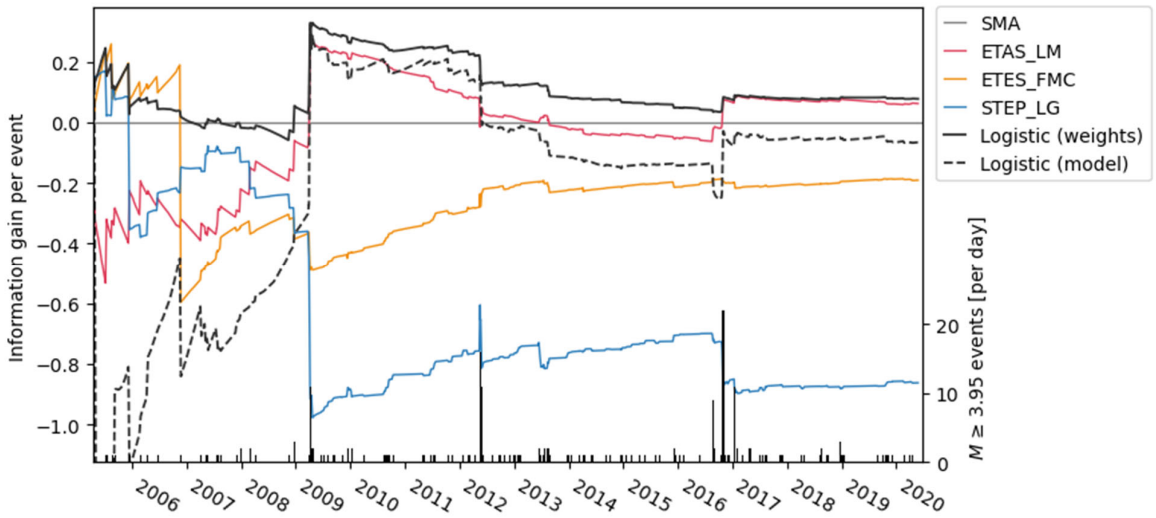

Figure 8. Like Figure 7, but using the second fitting scheme (only data of the last 365 days, Figure 6).

Table 1. Cumulative information gains of each model over the SMA ensemble for the two different temporal fitting schemes (see text), i.e., the sum of each curve shown in Figures 7 and 8, respectively.

Model	Individual	Fitting scheme #1	Fitting scheme #2
ETAS_LM	-132.3		
ETES_FMC	-1293.6		
STEP_LG	-3563.9		
Logistic ensemble (weights)		364.9	580.4
Logistic ensemble (model)		-1443.1	-1124.9

Figures 7 and 8 as well as Table 1 show the overall superior performance of the ensemble that uses the weights of the logistic fit, independently of the fitting scheme. The logistic model ensemble performs worse than simply using its coefficients as weights, although it performs better than STEP_LG and slightly better than ETES_FMC in the second fitting scheme. This misfit is likely caused by the problematics mentioned in section 3 (too few target bins), which led us to build the weight-based average in the first place. Figures 7 and 8 confirm what could be already inferred from the ensemble weights (Figures 5 and 6): ETAS_LM performs best during the 2009 L'Aquila and 2016 Central Italy sequence. The SMA ensemble apparently does not capture these sudden improvements in ETAS_LM's IG, likely because the badly performing STEP_LG model cancels those benefits. Our new weight-based ensemble does a better job at combining the models to obtain a well-performing ensemble, especially for the second fitting scheme (Figure 8), where our ensemble provides an even higher IG than any other model—at almost all times. In the following years after L'Aquila, STEP_LG and ETES_FMC slightly improve in IG, while the ETAS_LM slightly degrades. Our weight-based ensemble is able to capture this behavior, barely decreasing in IG. During the 2012 Emilia sequence, however, the ensemble dropped in IG, likely because it placed too much weight on STEP_LG (see Figure 6) due to its good performance during quiescent periods. Apparently, it was not the appropriate model to favor during the onset of this sequence; it will take the ensemble a week to adjust due to the delay caused by the forecasting horizon. In the following years, the IG slightly degrades, albeit less than ETAS_LM's IG. During the 2016 Central Italy sequence, ETAS_LM gains IG, which our weight-based ensemble, however, cannot fully exploit because it was briefly weighting all three models to an about equal amount. Overall, the second fitting scheme leads to a higher cumulative IG than the first, making this 365-day fitting scheme our preferred choice. Although the SMA is the second-best model, the IG of our weight-based ensemble is several times higher than the difference between SMA and ETAS_LM. This margin illustrates the advantage of *sound* ensemble modelling: exploiting the strength of each individual model to provide a significantly better model, even though ETES_FMC and STEP_LG are rarely more informative than ETAS_LM.

4.3 Building the ontological ensemble (OE) forecast distribution

Finally, we use our weight-based ensemble for the second fitting scheme to model an ontological ensemble (OE) forecast distribution at each date using: (i) the individual forecasts $\phi_i = f_i(x_0) = P_i(X > x_0)$ for the hazard level x_0 (here: M4); (ii) the ensemble mean forecast $\bar{\phi} = \overline{f_i}(x_0)$, (ii) the model weights π_i; and the beta distribution Beta(α, β), which describes $\mathrm{p}(\phi)$ within the range [0, 1]. We provide the OE forecast at three spatial locations over time: the spatial bins in which the mainshocks of the 2009 L'Aquila sequence (Figure 9), 2012 Emilia sequence (Figure 10), and 2016 Central Italy sequence (Figure 11, Norcia) occurred.

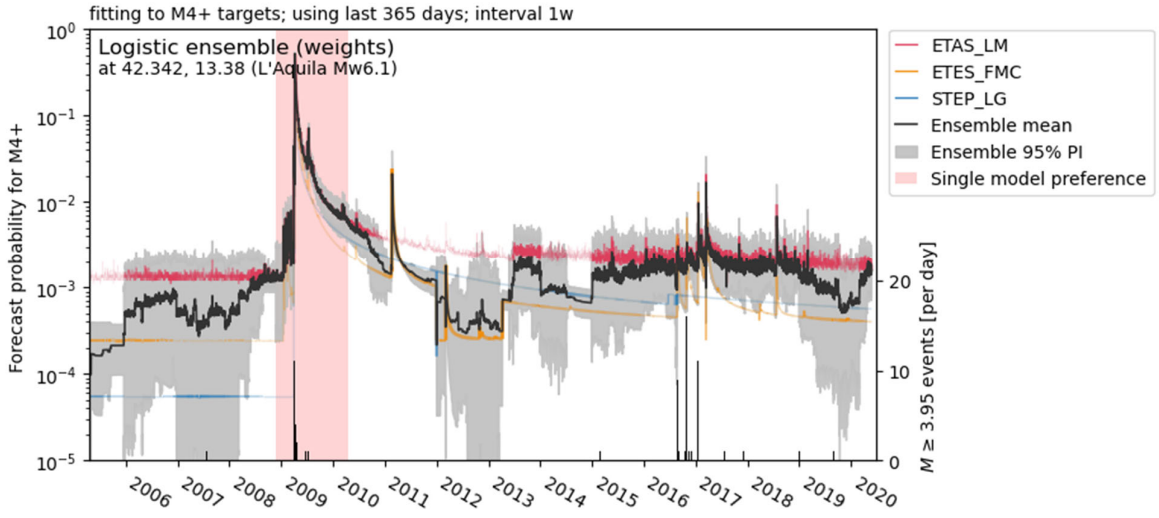

Figure 9. Ontological ensemble (OE) forecast for the spatial bin where the 2009 L'Aquila mainshock occurred. The forecast rate of the various models is indicated by the curves (see legend): each candidate model (colored curve),

our weighted-average ensemble (black curve), 95% prediction interval (PI) of the OE forecast (gray shaded band). The temporal evolution of the individual model weights in the ensemble is indicated by an opacity/transparency effect of the candidate forecast curves. The vertical red shaded band represents a time period in which a single model has all the weight (see text). At the bottom, the daily rate of target events within 50 km is displayed by vertical bars.

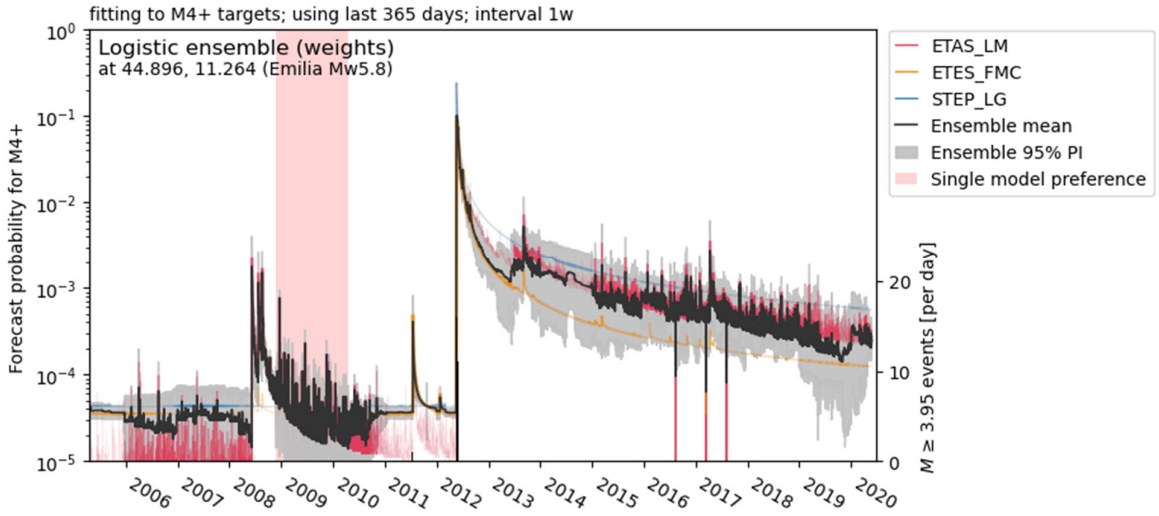

Figure 10. Like Figure 9, but for the spatial bin where the mainshock of the 2012 Emilia sequence occurred.

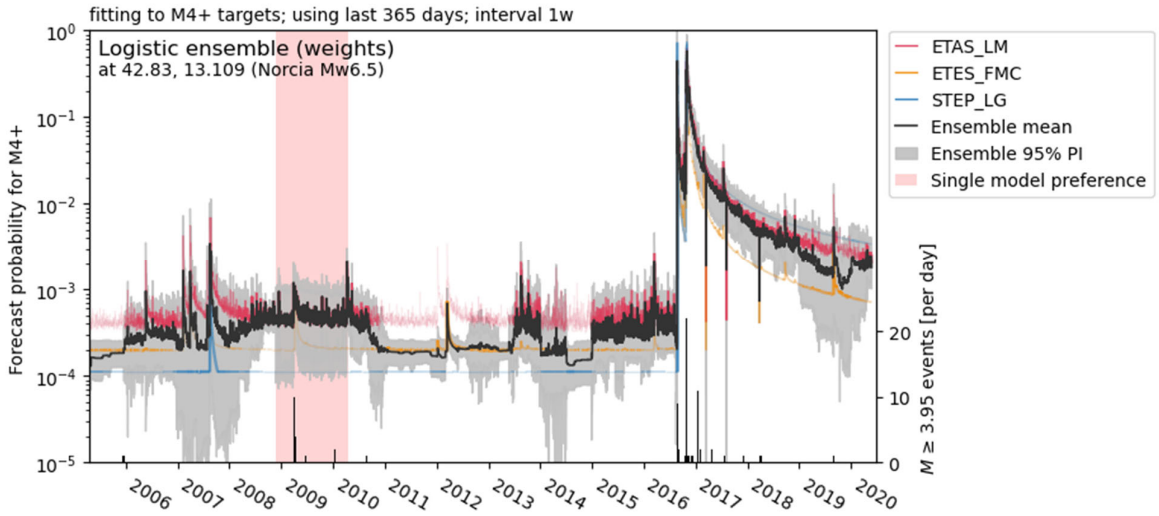

Figure 11. Like Figures 9 and 10, but for the spatial bin where the mainshock of the 2016 Central Italy sequence occurred (Norcia).

Note that when the ensemble fit prefers only a single model at a certain date (i.e., with 100% weight, shaded in red in Figures 9–11), the variance cannot be determined in our case (because our models do not provide a probability distribution by themselves), thus the ensemble forecast distribution becoming a Dirac distribution again. As a best guess for the variance at such a date, we determine the average relative variance, $\frac{\sigma_{\bar{\phi}}^2}{\bar{\phi}}$, over all prior forecast dates (to remain causal), and multiply it with $\bar{\phi}$.

In Figures 9–11, one can nicely see how $\bar{\phi}$ fluctuates between ϕ_i of individual models, and occasionally tracks the rate of single models if they have the major weight, e.g., ETAS_LM during the L'Aquila and Central Italy sequences, and ETES_FMC in the period of 2011–2013. It also becomes apparent that ETAS_LM provides the highest rate forecasts, and STEP_LG the lowest. This may

explain why ETAS_LM performs best during sequences, but not during times of quiescence, in which ETES_FMC typically receives most of the weight and gets tracked by the OE forecast (e.g., see 2011-2013 in L'Aquila or Central Italy).

But the novelty is that we now also model the uncertainty of the forecast, that is, given π_i and the dispersion of ϕ_i, we quantify the reliability of the ensemble. To illustrate this distribution from another perspective, Figure 12 provides slices of the OE forecast and the contributing forecasts at various times around the 2016 Central Italy sequence. For instance, before the sequence started, with a 95% reliability, the probability for a $M \geq 4$ was in the range of 0.02–0.1% per week; one week after the M_w6.0 Amatrice event, this probability was in the range of 2%–11% per week; before the Norcia event 8%–45% per week; and one day later 23%–77% per week.

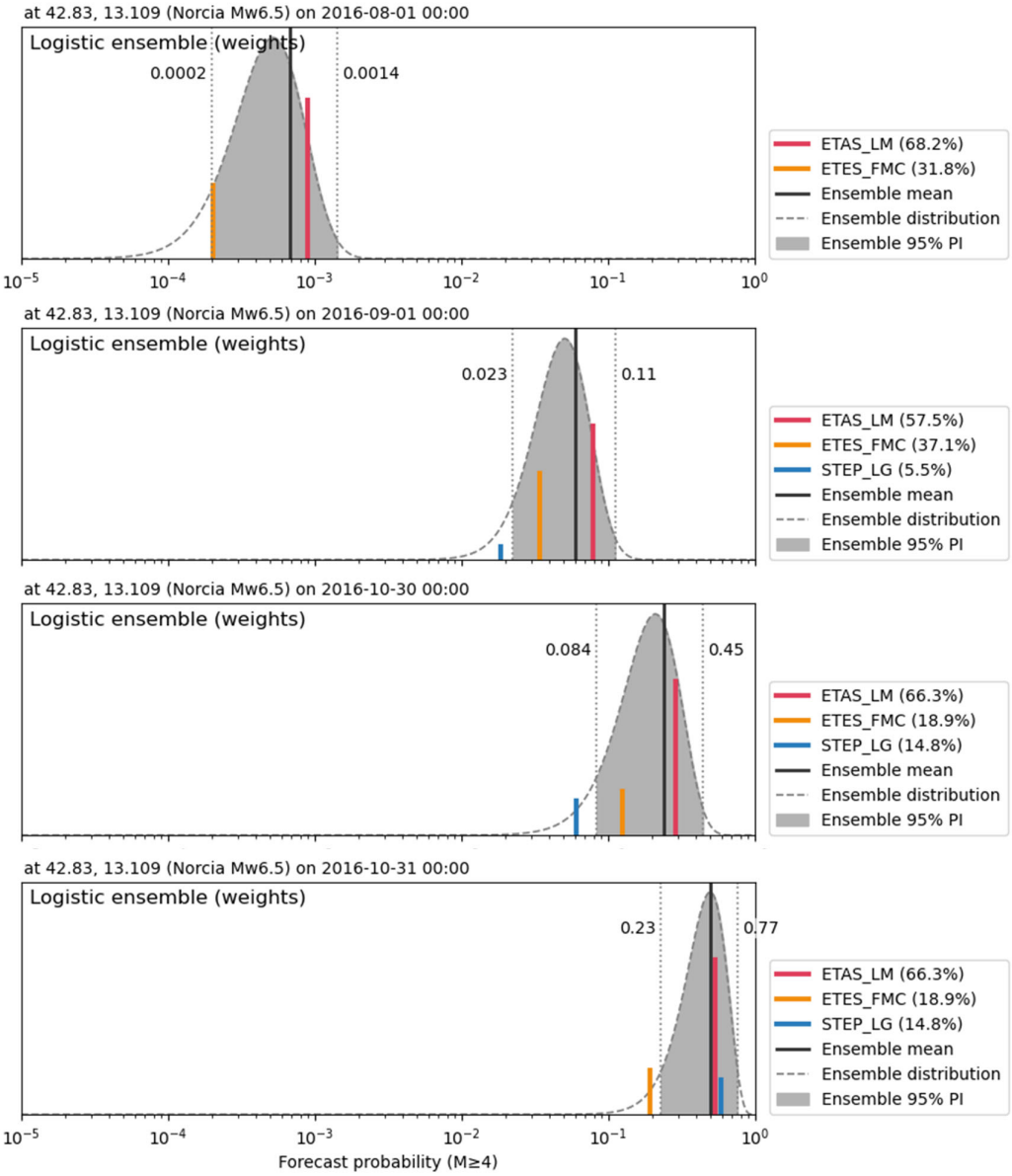

Figure 12. OE forecast distributions (dashed curves) corresponding to Figure 11 at various times during the Central Italy sequence. The forecasts of individual models are shown as bars at the appropriate forecast probability with their heights corresponding to the assigned weights $\boldsymbol{\pi_i}$. The 95% prediction intervals (vertical dotted lines) are annotated with their corresponding probabilities.

5. Conclusions

The aim of this work is to put forward a novel procedure that we call ontological ensemble modeling (OEM) to combine forecasts coming from different models. After having introduced the main theoretical background, which is rooted in a unified probabilistic framework recently introduced in the field of natural hazards, we apply OEM to the real case of operational earthquake forecasting in Italy to show how the method can be applied in real cases.

Here we summarize the main features and findings:

- OEM uses a univocal hierarchy of uncertainties keeping them separated; this allows scientists to validate a probabilistic model.
- The separation of the different kinds of uncertainty completely describes the ensemble forecast, highlighting clearly what we know and what we do not know about the future evolution of the process.
- The method clarifies that by increasing the number of models/forecasts, the epistemic uncertainty cannot be reduced because it can only be reduced by introducing new information. In other words, the models, even if they have been built independently, can never be fully independent because they all rely on the same information to some extent. Technically speaking, the lack of complete independence does not allow canceling out epistemic uncertainty by increasing the number of models, that can only sample more finely the epistemic uncertainty. This calls for creating and incorporating more diverse models into the ensemble, e.g., physics-based forecast models.
- The application to OEF-Italy demonstrates the benefits of OEM in two different ways: it provides a superior ensemble model than the current SMA ensemble, and it represents the reliability of the ensemble forecast by providing a probability distribution. A *sound* ensemble like our weighted-average based on the logistic fit is less likely to fail dramatically than a single model or a simple average of all candidate models. Moreover, we can modify the fitting scheme of the ensemble to address the multipurpose & authoritative character of OEF for an end user (i.e., with a focus on recent seismicity, overall rate, spatial skill, etc.). The additional quantification of the ensemble forecast's reliability allows scientists a more honest and versatile communication of forecast probabilities.

References

Bottazzi, M., Scipione, G., Marras, G. F., Trotta, G., D'Antonio, M., Chiavarini, B., Caroli, C., Montanari, M., Bassini, S., Gascón, E., Hewson, T., Montani, A., Cesari, D., Minguzzi, E., Paccagnella, T., Pelosini, R., Bertolotto, P., Monaco, L., Forconi, M., … Pieralice, A. (2021). The Italian open data meteorological portal: MISTRAL. *Meteorological Applications*, *28*(4). doi: 10.1002/met.2004

Box, G. E. P. (1976). Science and statistics. *Journal of the American Statistical Association*, 71(356), 791–799. doi: 10.1080/01621459.1976.10480949

Cooke RM (1991) *Experts in Uncertainty: Opinion and Subjective Probability in Science*. Oxford Univ Press, New York.

Gerstenberger, M. C., Marzocchi, W., Allen, T., Pagani, M., Adams, J., Danciu, L., Field, E. H., Fujiwara, H., Luco, N., Ma, K. -F., Meletti, C., & Petersen, M. D. (2020). Probabilistic seismic hazard analysis at regional and national scales: State of the art and future challenges. *Reviews of Geophysics*, *58*(2). doi: 10.1029/2019RG000653

Gneiting, T., Balabdaoui, F., & Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *69*(2), 243–268. doi: 10.1111/j.1467-9868.2007.00587.x

Gneiting, T. and Katzfuss, M. (2014). Probabilistic forecasting. *Annual Review of Statistics and Its Application, 1*(1), 125–151. doi: 10.1146/annurev-statistics-062713-085831

Hamill, T. M., Hagedorn, R., & Whitaker, J. S. (2008). Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part II: Precipitation. *Monthly Weather Review, 136(7*), 2620–2632. doi: 10.1175/2007MWR2411.1

Evans, R. E., Harrison, M. S. J., Graham, R. J., & Mylne, K. R. (2000). Joint medium-range ensembles from the met. Office and ecmwf systems. *Monthly Weather Review*, *128*(9), 3104–3127. doi: 10.1175/1520-0493(2000)128<3104:JMREFT>2.0.CO;2

King, G., & Zeng, L. (2001). Logistic Regression in Rare Events Data. Political Analysis, 9(2), 137–163. doi: 10.1093/oxfordjournals.pan.a004868

Krishnamurti, T. N., Kishtawal, C. M., Zhang, Z., LaRow, T., Bachiochi, D., Williford, E., Gadgil, S., & Surendran, S. (2000). Multimodel ensemble forecasts for weather and seasonal climate. *Journal of Climate*, *13*(23), 4196–4216. doi: 10.1175/1520-0442(2000)013<4196:MEFFWA>2.0.CO;2

Krishnamurti, T. N., Kumar, V., Simon, A., Bhardwaj, A., Ghosh, T., & Ross, R. (2016). A review of multimodel superensemble forecasting for weather, seasonal climate, and hurricanes. *Reviews of Geophysics*, *54*(2), 336–377. doi: 10.1002/2015RG000513

IPCC (2021). *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change.* Cambridge University Press. url: https://www.ipcc.ch/report/ar6/wg1

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. doi: 10.1785/0120110327

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. doi: 10.1073/pnas.1410183111

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. doi: 10.1785/0220130219

Marzocchi, W., Taroni, M., & Falcone, G. (2017). Earthquake forecasting during the complex Amatrice-Norcia seismic sequence. *Science Advances*, *3*(9), e1701239. doi: 10.1126/sciadv.1701239

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. doi: 10.1785/0120170008

Meletti, C., Marzocchi, W., D'Amico, V., Lanzano, G., Luzi, L., Martinelli, F., Pace, B., Rovida, A., Taroni, M., Visini, F., & Group, M. W. (2021). The new Italian seismic hazard model (MPS19). *Annals of Geophysics*, *64*(1), 6. doi: 10.4401/ag-8579

Monteith, K., J. L. Carroll, K. Seppi, and T. Martinez (2011). Turning Bayesian model averaging into Bayesian model combination. *Proceedings of International Joint Conference on Neural Networks, San Jose, California, 31 July–5 August 2011*, 2657–2663. doi: 10.1109/IJCNN.2011.6033566

Murphy, J. and Palmer, T.N. (1986). Experimental monthly long-range forecasts for the United Kingdom. II: A real time extended-range forecast by an ensemble of numerical integrations. *Meteorological Magazine, 115(1372)*, 337–348.

NRC (2018). *Updated Implementation Guidelines for SSHAC Hazard Studies (NUREG-2213)*. Washington, DC, Office of Nuclear Regulatory Research. url: https://www.nrc.gov/reading-rm/doc-collections/nuregs/staff/sr2213/index.html

Oreskes, N., Shrader-Frechette, K., & Belitz, K. (1994). Verification, validation, and confirmation of numerical models in the earth sciences. *Science*, *263*(5147), 641–646. doi: 10.1126/science.263.5147.641

Palmer, T. N., Alessandri, A., Andersen, U., Cantelaube, P., Davey, M., Délécluse, P., Déqué, M., Díez, E., Doblas-Reyes, F. J., Feddersen, H., Graham, R., Gualdi, S., Guérémy, J.-F., Hagedorn, R., Hoshen, M., Keenlyside, N., Latif, M., Lazar, A., Maisonnave, E., … Thomson, M. C. (2004). Development of a European multi-model ensemble system for seasonal to inter-annual prediction (DEMETER). *Bulletin of the American Meteorological Society, 85*(6), 853–872. doi: 10.1175/BAMS-85-6-853

Palmer, T. (2019). The ECMWF ensemble prediction system: Looking back (more than) 25 years and projecting forward 25 years. *Quarterly Journal of the Royal Meteorological Society*, *145*(S1), 12–24. doi: 10.1002/qj.3383

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. doi: 10.2478/s11600-011-0013-5

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, *12*(4). doi: 10.1214/aos/1176346785

Scherbaum, F., & Kuehn, N. M. (2011). Logic tree branch weights and probabilities: Summing up to one is not enough. *Earthquake Spectra*, *27*(4), 1237–1251. doi: 10.1193/1.3652744

Stein, S., & Liu, M. (2009). Long aftershock sequences within continents and implications for earthquake hazard assessment. *Nature*, *462*(7269), 87–89. doi: 10.1038/nature08502

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 and Engineering Sciences*, *365*(1857), 2053–2075. doi: 10.1098/rsta.2007.2076

Wilks, D. S. (2009). Extending logistic regression to provide full-probability-distribution MOS forecasts. *Meteorological Applications, 16*(3), 361–368. doi: 10.1002/met.134

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. doi: 10.1785/0120090192