Reviewer #2

We thank Dr. Francesca Pianosi for the helpful comments, which have allowed us to clarify several key and complex aspects of the paper, and to have brought to our attention some pertinent references that we did not include in our original version.

MAJOR POINTS

1. *The manuscript is the outcome of a national project that brought together scientists and practitioners from different natural hazard/risk sectors. While this may have been the first attempt of this kind in Italy (has it?) there certainly have been similar efforts in other countries. For example, I am aware of a similar project in the UK, which resulted in two NHESS papers very relevant to this manuscript: https://nhess.copernicus.org/articles/18/2741/2018/ - a review of the treatment of epistemic uncertainty in hazard assessment, covering floods, landslides, droughts, earthquakes, tsunamis, volcano eruptions, windstorms. https://nhess.copernicus.org/articles/18/2769/2018/ - a discussion of good practices in dealing with epistemic uncertainties across hazards. Some key messages from the two papers above seem to well align with those of this manuscript, for example on the importance of quantifying and communicating epistemic uncertainty per se, or the potential for defining common approaches across hazards. On other points, the authors' views may diverge. Either way, it would be good to see some acknowledgement and discussion of this past work (also beyond the two papers linked above, if others can be found) so to better place this manuscript's contribution in the context of previous literature.*

First, we apologize for overlooking several important references and thank the reviewer for the valuable suggestions. The rapid growth of the literature makes it increasingly challenging to keep track of all relevant publications, although this does not diminish the importance of the references that were brought to our attention. While our manuscript is intended as a perspective rather than a comprehensive review, we agree that the suggested references are highly relevant and improve the completeness and context of the discussion. We have therefore incorporated them into the revised manuscript where appropriate. We include these references in many parts of the revised version. Here, we report only the context in which they have been added:

* Beven, Almeida, et al., 2018 has been added when considering these two important points:

- The commonalities of uncertainties across different hazards.

- In practical applications, all forecasts are often collapsed into one single distribution that is expected to mimic aleatory variability.

* Beven, Aspinall, et al., 2018, has been added in our manuscript when considering these important points:

- aleatory variability is not irreducible: it can, be reduced when increasing our knowledge of the process.

- Emphasis on the challenges of testing a forecasting model against observations in the subjective framework (The Beven, Aspinall, et al., 2018's paper reports " Within the Bayesian paradigm, there are ways of avoiding the specification of a formal aleatory error model, such as in the use of expectations in Bayes linear methods (Goldstein and Wooff, 2007), in approximate Bayesian computation (Diggle and Gratton, 1984; Vrugt and Sadegh, 2013; Nott et al., 2014), or in the informal likelihood measures of the GLUE methodology (Beven and Binley, 1992, 2014; Smith et al., 2008). It is still possible to empirically control the performance of any such methodology in simulating past data, but, given the epistemic nature of uncertainties this is no guarantee of good predictive performance in future projections.").

- All models are wrong and they will fail test. The Beven, Aspinall, et al., 2018's paper reports " However, all natural hazards models are approximations, and if tested in too much detail (severely) are almost certain to fail."

2. The manuscript focuses on hazard, rather than risk assessment, but the authors claim that their work is “a crucial first step toward understanding uncertainties in risk forecasting as well” (L. 50). I think this is possibly oversimplifying a more complicated issue and may deserve a bit more nuance. I am familiar with a thread of literature that has attempted at formally quantifying the relative importance of hazard, vulnerability and exposure uncertainties on risk assessment. These studies tend to consistently find that the impact of hazard uncertainty is dwarfed by uncertainty in vulnerability. See for examples De Moel et al 2014, Metin et al 2018, Pianosi et al 2026 for floods or Dawkins et al 2023 for heatwaves. Another example: I recently reviewed a paper for NHESS (Ruf et al 2026) presenting a new computationally efficient approach to improve simulation of dyke failures. A key motivation for that work was to enable uncertainty-aware evaluation of the benefits of flood mitigation measures. However (and quite ironically!) the sensitivity analysis enabled by the new hazard model showed that, at least in the paper’s case study, risk metrics were completely controlled by uncertainty in flood damage estimation and essentially insensitive to uncertainty in hazard - including uncertain assumptions about dyke breaches! This is not to say that improving hazard assessment is not important – good quality hazard models are essential for a range of uses, from early-warning to emergency management to improving understanding of physical impacts of different infrastructure or land use decisions, etc. However, if hazard models are used to feed into larger risk assessment modelling chain, and decision-makers are only interested in summary output metrics (e.g. cost-benefit ratio of different mitigation measures; or annual average loss from a given hazard, etc.) then sensitivity analysis literature has shown us that aggregate outputs are typically controlled by a limited number of uncertainty sources (e.g. Wagener and Pianosi 2019), and in the field of natural risk assessment these dominant controls are likely to be in the vulnerability/exposure component rather than the hazard one. I do not expect this manuscript to go deep into this discussion as the focus here is hazard assessment, however I think the authors should at least mention how moving from hazard to risk opens up a range

of other issues due to even bigger uncertainties in vulnerability/exposure, and that the transition from analysing uncertainty in hazard to uncertainty in risk is more complex than somehow suggested by the manuscript in its present form.

Our original statement was misleading, and we appreciate the reviewer's comment. We did not intend to suggest that uncertainty in hazard is more important than uncertainty in vulnerability for risk assessment. Rather, our point was that the probabilistic framework proposed to construct "complete forecasts" can be applied not only to hazard forecasting but also to vulnerability assessment. This argument is independent of the relative magnitude of the uncertainties associated with hazard and vulnerability. We have revised the manuscript accordingly, as follows:

" Although this paper focuses on hazard forecasting, the proposed framework represents an important first step toward a comprehensive treatment of uncertainty in risk forecasting. In its simplest formulation, risk is a function of hazard, exposure, and vulnerability. While uncertainty in hazard has received considerable attention, vulnerability assessments are also affected by substantial uncertainty, which may in some cases exceed that associated with the hazard itself (e.g., Pianosi et al., 2026). We argue that the probabilistic framework proposed here for representing and propagating uncertainty in hazard forecasting can be extended naturally to the vulnerability domain. If this is the case, the resulting risk assessment becomes the combination of the uncertainties associated with both hazard and vulnerability, propagated through the risk model."

3. On the subjective view of probabilities, particularly Lines 260-264: “Although scientists often prefer the frequentist framework … testing against data becomes meaningless”: this is discussion very interesting although I am not sure I completely follow the argument here. I do not think that the ‘subjective approach’ necessarily implies that “testing models against data … becomes useless”. I tend to embrace the subjective view of probability, as described here: https://environment.blogs.bristol.ac.uk/2017/03/20/what-is-probability/ whereby “probabilities … represent a consensus of well-informed people” (and, of course, satisfying Kolmogorov axioms). I do not think this implies discarding the role of testing against data, indeed success (or better, lack of failure) in goodness-of-fit tests is one of the key pieces of ‘evidence’ by which well-informed people will build their consensus on probabilities. This point could be clarified? Also note that Lindley 2000 cited on line 264 does not that seem to be included in the reference list!

This is an important point that deserves a more careful discussion. We have modified section 4 to explain better our perspective on testing forecasts against data in the subjective Bayesian framework (see below). Subjective Bayesian statisticians hold different views regarding the role of model testing and validation. For example, Lindley (2000) argued that "the rejection of a model is not a reality except in comparison with an alternative that appears better." Similarly, Gelman (2007) wrote: "All models are wrong, and the purpose of model checking (as we see it) is not to reject a model but rather to understand the ways in which it does not fit the

data. From a subjective Bayesian point of view, the posterior distribution is what is being used to summarize inferences, so this is what we want to check." These views are consistent with Box's well-known statement that "all models are wrong, but some are useful" (Box, 1976) and also supported in Beven, Aspinall, et al., 2018.

In all cases, we conclude that the subjective Bayesian framework is not designed to test a forecasting model against independent observations and reject it when necessary, which is a cornerstone of scientific practice. Bayesian coherence and internal consistency are important properties of a probabilistic forecast, but they cannot replace empirical validation whenever suitable observations are available (at the end of page 2777, Beven, Aspinall et al., 2018 wrote: " It is still possible to empirically control the performance of any such [Bayesian] methodology in simulating past data, but, given the epistemic nature of uncertainties this is no guarantee of good predictive performance in future projections.)

Our position differs more fundamentally. We fully recognize that all models are imperfect representations of reality (Box, 1976). However, we also assume that there exists a true forecast distribution (FD) associated with the data-generating process, even though it is unknown. Validation, in our framework, consists of testing, at least in principle, whether the available observations (assumed to be generated from this true FD) are statistically compatible with the proposed FD-ensemble. We regard this step as essential for maintaining natural-hazard forecasting within the scientific domain. In our view, empirical testability is not merely a useful diagnostic tool but a fundamental requirement of the scientific method (e.g., AAAS, 1989).

That said, it is important to distinguish between the subjective assessment (that can be obtained through experts' judgment) of the occurrence of a specific event and the subjective probability framework, in which the probability is a degree of belief or a bet quotient. For the former, we do not regard subjective assessments as an obstacle to any kind of testing, including validation. Marzocchi and Jordan (2014) showed that if the subjective assessments are expressed as probabilistic forecast distributions, they are, at least in principle, amenable to empirical evaluation (see, e.g., the discussion on the Galton's experiment: Galton F (1907) Vox populi. Nature 75(1949):450–451.). We also emphasize the essential role of expert judgment in selecting the models to be used, and in situations where meaningful validation is impossible because of the scarcity or absence of observations. As illustrated in Figure 7, expert elicitation is then required to assess the credibility of alternative models and forecast distributions until sufficient observational evidence becomes available.

To clarify this important point we modify the second paragraph of section 4 in this way:

"Although scientists often favor the frequentist interpretation because probabilities are defined in terms of observable frequencies (aleatory variability), applying this framework to natural hazards is challenging. In many cases, defining the frequency of a truly repeatable experiment is challenging because every hazardous event is unique to some extent. Moreover, the frequentist interpretation does not naturally accommodate epistemic uncertainty, which inevitably involves subjective choices, such as the selection and weighting of the models used to represent alternative models or numerical approximations. In the subjective Bayesian

interpretation, by contrast, probability quantifies a state of knowledge, and all uncertainty is therefore treated as epistemic. This framework is more flexible and broadly applicable (e.g., Beven, Aspinall, et al., 2018), but it is not designed to test a forecasting model against independent observations and reject it when warranted, which is a cornerstone of scientific practice (Box, 1976; AAAS 1989; Oreskes et al. 1994; Marzocchi & Jordan 2014).

Subjective Bayesian statisticians hold different views regarding the role of model testing (see also Beven, Aspinall, et al., 2018). For example, Lindley (2000) argued that "the rejection of a model is not a reality except in comparison with an alternative that appears better." Similarly, Gelman (2007) wrote: "All models are wrong, and the purpose of model checking (as we see it) is not to reject a model but rather to understand the ways in which it does not fit the data available. From a subjective Bayesian point of view, the posterior distribution is what is being used to summarize inferences, so this is what we want to check." These views are consistent with Box's well-known observation that "all models are wrong, but some are useful" (Box and Draper, 1987). In the subjective framework (e.g., Beven, Aspinall, et al., 2018), independent observations are incorporated through Bayesian updating to revise the state of knowledge, rather than being used to assess whether a probability distribution converges to an objectively existing data-generating process."

And at the end of section 5:

" One final point concerns the role of experts' judgment and subjectivity assessment in science (for a detailed discussion see, e.g., Marzocchi and Jordan, 2014; Hanea et al., 2021). Since pure objectivity is unattainable, the traditional objectivity–subjectivity dichotomy is misleading. What ultimately matters is whether probabilistic models and their forecasts (FDs) can be subjected, at least in principle, to rigorous empirical tests of reliability (e.g., validation tests), regardless of the degree of subjectivity involved in their construction."

4. I found the description of the unified framework by Marzocchi and Jordan (pages 12-13) not always easy to follow. I think this section would benefit from some more practical examples of how different concepts (“experimental concept”, “explanatory variables”, etc.) would be applied/interpreted in different contexts, including those (floods & droughts, extreme weather) where the approach has not been applied so far. This would be a very valuable addition to help bridge gaps between communities and share best practice. Similarly on page 15, the discussion of reliability, calibration, and consistency (L. 366-379) is very abstract and would probably benefit from few concrete examples. For example, the difference between consistency and calibration test is not clear to me (maybe because I am from one of those communities where “calibration” is “the process of estimating parameters”!) and the description in the glossary and Figure 6 are not sufficient! Again, a few application examples across different hazards would probably be very useful here.

We agree that terminology is important, which is why we have included a glossary. We adopted the term calibration because it is well established in the statistical literature. We hope that the explanations provided

throughout the text, together with the glossary, will help readers understand the terminology and the concepts presented in the paper (see line 371 of the submitted version for the specific aspect of the term calibration). By the way, the distinction between calibration and consistency cannot be overstated. Calibration evaluates the forecast distribution (FD) against independent observations that were not used to construct the model, thereby avoiding overfitting. In contrast, consistency is assessed using observations that have contributed to model development, so the results may be substantially biased by overfitting and therefore provide much weaker evidence of predictive performance (see line 375).

We also add more explanations on the experimental concept and explanatory variables, highlighting why they are important to understand the aleatory variability and how it can vary also when we are considering the same physical process.

Specifically, we add some more explanation on the experimental concept at section 4:

" We emphasize once again that the true aleatory variability is not a property of the underlying "true" process (which, in practice, is never fully known), but of the data-generating process defined by the experimental concept. Consequently, the commonly stated irreducibility of aleatory variability should be interpreted with caution: as the experimental concept evolves and the data-generating process is redefined (e.g., by conditioning on newly identified explanatory variables), the estimated aleatory variability may also change, even though the underlying physical process remains the same. Consider, for example, the annual probability of exceedances of a given hazard intensity threshold, x_0, $\phi = f(x_0)$. If the experimental concept consists of data that are collected simply as a single annual time series, the aleatory variability, $\hat{\phi}^{(I)}$, is represented by the annual frequency of exceedances. Suppose, however, that we subsequently identify a binary variable A (A=0 or A=1) that significantly influences the probability of exceedance, with exceedances being more likely when A=1. Rather than treating all years identically, we would naturally stratify the observations into two separate time series, corresponding to A=0 and A=1; these two new exchangeable time series represent two distinct experimental concepts. Although the underlying physical process has not changed, the estimated aleatory variability differs between the two series and also differs from that obtained when the data are analyzed without conditioning on A, i.e., $\hat{\phi}^{(I)} \neq \hat{\phi}^{(II)}_{A=0} \neq \hat{\phi}^{(II)}_{A=1}$. The only change is our improved knowledge of the process, which leads us to define new different experimental concepts (i.e., new different observational protocols), which represent different data-generating process."

And on the time as explanatory variable in the same section:

"The definition of the experimental concept can be naturally extended to account for non-stationarity, with time acting as the conditioning explanatory variable. In this case, a correctly specified non-stationary data-generating model provides the appropriate FD at each point in time, while the interpretation of exceedance probabilities remains unchanged. The main consequence of non-stationarity is therefore not conceptual but practical: it complicates forecast validation because observations collected at different times are generated

from different forecast distributions. We will be back to this point in section 5 "Evaluating the reliability and skill of forecasts"

5. Lines 412: “Needless to say, the least desirable situation is having few or no data” and Lines 423: “… consensus could also be understood as acceptance of procedures and models, not necessarily as agreement on their results (i.e. similarity)”. For hazards such as floods and droughts, this situation where observations are scarce, completely missing, or of very poor quality (i.e. potentially affected by errors of the same magnitude as the observations themselves) is extremely common! This is why this community has developed very comprehensive approaches to “model validation” including many that focus on “acceptance of procedures and models” rather than fit-to-data. There is a very good discussion of this topic for example in Mertz et al 2024, including the distinction between “outcome-based” and “process-based” validation/evaluation approaches. Again, a more comprehensive discussion and linking to previous literature would be good here!

   References:

   De Moel et al 2014: https://doi.org/10.5194/nhess-18- 3089-2018

   Metin et al 2018: https://doi.org/10.1016/j.scitotenv.2013.12.015

   Wagener and Pianosi 2019: https://doi.org/10.1016/j.earscirev.2019.04.006

   Pianosi et al 2026: https://doi.org/10.5194/nhess-26-1727-2026

   Dawkins et al 2023: https://doi.org/10.1016/j.crm.2023.100511

   Ruf et al 2026: https://doi.org/10.5194/egusphere-2025-4875

   Mertz et al 2024: https://doi.org/10.5194/nhess-24-4015-2024

In this paper, we have considered a range of natural hazards while highlighting their important differences. It is therefore not correct to state that all hazards lack sufficient data to build a complete forecast or for model validation. Figure 7 is intended to encompass this diversity by illustrating that the appropriate validation strategy depends critically on data availability. For floods, for example, we present a case (Figs. 3 and 4) in which sufficient data are available to construct complete forecasts.

Some of the suggested references focus primarily on sensitivity analysis. Although sensitivity analysis is valuable for identifying the factors that influence risk estimates, it is not directly relevant to the issue addressed here, namely the scientific credibility of probabilistic forecasts. Instead, in this context, the reference by Merz et al. is particularly relevant, as it explicitly addresses the scientific validation of flood models for societal applications, and we have therefore included it at the beginning of section 5.

MINOR POINTS

6. L. 15: “These challenges” unclear what this refers to (which challenges?)

Replaced by "common challenges".

7. L. 49: “across a range of natural hazards – including floods, earthquakes, volcanic phenomena, landslides, climate projections”. I am not sure I would include climate projections in this list. The ‘key’ source of uncertainty in climate projections is uncertainty in the evolution of input forcings (carbon emissions, population growth/decline, technology development) over long-term future. This type of uncertainty is fundamentally irreducible (we do not know how the world will look like in 50 or even 100 years from now and will not be able to test our guesses against observations!) and so should probably be kept distinct ‘epistemic’ uncertainty which, at least in principle, could be reduced with more investigation, experiments, etc.

That's correct. We have explained better that climate projections have to be evaluated under the specific emission scenario that will represent the future. At line 52 we write:
" climate projections (which can be interpreted as forecasts conditioned on a specific emissions scenario), ..."

8. L. 72: “its limited knowledge” maybe should be “our limited knowledge”

Done.

9. Figure 1: I wonder if the “context” box in this Figure should include a larger set of “model uses”. For example, where would insurers (such as Assicurazioni Generali) sit? Is their way of using models captured by the three points currently listed?

Our goal is not to describe the current state of practice, but rather to propose what we believe should constitute best practice. Different stakeholders may use forecasts in different ways; however, the existence of multiple approaches does not imply that they are all equally appropriate.

10. L. 173: “In many other forecasts” maybe should be “in many other sectors”

Done.