Reviewer #1

We thank the reviewer for the helpful comments, which have allowed us to clarify several key and complex aspects of the paper.

1. *line 90: The description of type the mechanisms for how uncertainty is incorporated into determinist physics-based models is oversimplified. With the use of surrogates or ROMs, many probabilistic aleatory scenario models can be posited and used with the surrogate/ROMS to calculate the probability of hazard levels or exceedances.*

We understand the reviewer's point. In that paragraph, our intention was to provide a brief overview of the different types of models that can be used to generate forecasts, without going into greater detail. To maintain this concise presentation while addressing the comment, we have revised the statement as follows:

" Depending on the hazard, forecasting models can be of different types, for example: i) Deterministic physics-based models, often accelerated through surrogate or reduced-order models, incorporate uncertainty into their point-forecasts either by introducing random perturbations to the initial and/or boundary conditions or by augmenting deterministic point predictions with probability distributions calibrated from historical discrepancies between model predictions and observations (Glahn, H. R. and Lowry, 1972; Gneiting et al., 2005; Koutsoyiannis and Montanari, 2021; Shabestanipour et al., 2023); ... "

2. *On the topic of physics-based models, there is no discussion in this paper of structural uncertainties. Even the best physics-based models (high-order, well-calibrated, etc.) are only an approximation to reality and often not reducible even with a concerted effort at minimizing epistemic uncertainties. This is perhaps the hardest kind of uncertainty to quantify, but it shouldn't go unmentioned. See Smith and Berger, 2019 10.1146/annurev-statistics-030718-105232*

In our framework, structural uncertainty is incorporated into the epistemic uncertainty, which represents our incomplete knowledge of the “true” underlying data-generating process and/or the limitations of its numerical representation. We have modified two paragraphs to explain better this point.

At the beginning of section 3: " Uncertainty arises from multiple sources, which can be grouped into two broad categories in forecasting: the natural variability of the data-generating process and our limited knowledge of the true data-generating process and/or the limitations of its numerical representation. These are commonly referred to as aleatory variability and epistemic uncertainty, respectively (see Table 1). Here, we implicitly assume the existence of a true data-generating process, an assumption to which we return later in the paper. "

In the first paragraph of subsection 3.2: " To illustrate this situation, Figure 3 shows two examples from long-term flood analysis and seismic hazard; in both cases, epistemic uncertainty is represented by alternative models reflecting different working hypotheses or numerical approximations, each producing a distinct forecast distribution (FD). In some disciplines, this source of uncertainty is referred to as structural uncertainty (Smith and Berger, 2019), which can be regarded as a component of epistemic uncertainty. "

3. *Line 264-271. These sentences seem to contradict themselves. The processes are random. For a tephra fall out event we don't know when, the volume of material ejected, the hight of the ash cloud, the distribution of grain sizes, the wind speed/direction, etc. We model this aleatory variability of such random scenarios with a pdf. So I don't know what "Attempting to attribute aleatory ... character directly to the process -- rather than the model is problematic, since they coincide only if the model is true" means. Of course we don't know the "true" model of the process, that's why we posit epistemic uncertainty about the aleatoric model.*

This is an important point that deserves further clarification. In the ideal case of having an infinite number of data, we can build the "true" unknown aleatory variability of the data-generating process numerically. This definition is model-independent, because it relies only on observations. Since we never have the possibility to collect a huge number of observations, we estimate this unknown distribution through a set of different forecasts obtained by different models. Each model has its own description of the aleatory variability which is a peculiarity of the model itself. For example, we can describe the aleatory variability of the eruption size adopting different probability distributions. The aleatory variability of the data-generating process and of the model coincides only if the model is the "true" one. We have deeply modified the text of section 4 "Modeling different uncertain ties". Regarding the specific issue raised in this comment, we state:

" Before introducing the unified framework adopted here, it is useful to examine in greater depth the definitions of aleatory variability and epistemic uncertainty in the context of Earth sciences (Table 1). It is often stated that aleatory variability is an irreducible randomness of the natural process, whereas epistemic uncertainty can, at least in principle, be reduced (e.g., Oberkampf et al., 2004; Eldred et al., 2011; Hüllermeier and Waegeman, 2021). This definition encounters some obvious conundrum: what appears to be intrinsic randomness may instead reflect incomplete understanding of the underlying physical mechanisms. As knowledge improves, part of the apparent aleatory variability may be explained deterministically and therefore reclassified as epistemic uncertainty. Consequently, the true intrinsic variability of a process (if any) cannot be identified unless epistemic uncertainty is eliminated, a condition that is rarely, if ever, achievable in practice (Marzocchi and Jordan, 2014; Beven, Almeida, et al., 2018). A coherent and unambiguous definition of aleatory variability requires shifting the focus from the physical process to the data-generating process. In the idealized limit of an infinite amount of data, the FD of the data-generating process (its aleatory variability) could be estimated directly from the observations. This definition is model-independent because it relies solely on empirical evidence rather than on assumptions about the underlying physical model. Since we never have the possibility to collect a huge number of observations, we estimate this unknown FD through a FD-ensemble obtained by different models; each model has its own description of the aleatory variability which is a peculiarity of the model itself.

The irreducibility of the aleatory variability is strictly valid only at the model level (e.g., Der Kiureghian and Ditlevsen, 2009; Liou and Abrahamson, 2024): assuming to know the true model, the epistemic uncertainty arises solely from incomplete knowledge of its parameters; under these conditions, the acquisition of

additional data and information reduces epistemic uncertainty while leaving the aleatory variability of the process essentially unchanged. "

4. *Line 329-332. Related to 3, the sentence "Changing the experimental concept, i.e. changing the way we in which we collect the data that we want to describe, we change the aleatory variability" also does not make sense to me. Back to the tephra example, the randomness inherent. How do we have any power to change it? In both 3 and 4, perhaps you are trying to get at a more subtle point that needs further elucidation.*

Aleatory variability is defined relative to the chosen data-generating process (e.g., experimental concept). As our understanding improves, we may redefine the data-generating process by conditioning on additional variables, thereby changing the apparent aleatory variability without altering the underlying physical process. Marzocchi and Jordan (2014) illustrated the change of experimental concepts with one example showing that the same physical process may exhibit different apparent aleatory variability depending on how the observations are collected. Consider, for example, the annual occurrence of exceedances of a given hazard threshold. If the data are collected simply as a single annual time series, the aleatory variability 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. 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. The only change is our improved knowledge of the process, which leads us to define a different data-generating process through a different observational protocol, or experimental concept. The same principle applies in epidemiology. The mortality rate associated with a disease can be estimated from the entire population of infected individuals. However, if evidence shows that mortality depends strongly on sex, age, or other covariates, it is more appropriate to estimate separate mortality rates for each subgroup. Again, the disease itself has not changed; rather, our improved understanding of the factors controlling the outcome leads us to redefine the data-generating process and, consequently, the observed aleatory variability.

We have revised this paragraph to make it more specific:

" 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. 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 a different observational protocol, or experimental concept, which represents a different data-generating process."

5. *It would be useful to spend some more time discussing high-impact tail events for hazards that may have sufficient data on small scales, little-to-no-data at a large scale. How does this framework handle such extrapolation or can we only rely on "expert opinion"?*

This is a very important question. In general, the ability to extrapolate from the range of observed data to higher hazard intensities or different spatial scales is an inherent property of the model itself. From the perspective of defining a complete forecast, this does not introduce any new conceptual issues. The main challenge instead concerns model validation, since observations at the extreme tail of the distribution are typically too scarce to allow a robust empirical assessment.

In principle, a complete forecast remains testable. In practice, however, testing its performance at very high hazard intensities is often infeasible because of the limited number of observations. We have attempted to illustrate this limitation in Figure 7. Complete forecasts can generally be evaluated when a sufficient amount of observational data is available. At the highest hazard intensities, however, direct validation may not be possible. In such cases, confidence in the forecast relies primarily on the model's ability to reproduce observations over the range of available data and on the assumption, e.g., supported by physical understanding and expert judgment, that the model remains valid when extrapolated beyond the observed range.

We have added this paragraph towards the end of section 5:

" Needless to say, the least desirable situation arises when few or no observational data are available. For example, a model may be testable only over a limited range of hazard intensities, while observations of the highest, and often societally most relevant, hazard intensities are unavailable (Beven, Almeida, et al., 2018). In such cases, confidence in the forecast relies primarily on the model's ability to reproduce observations within the available range and on the assumption, supported, for example, by physical understanding and expert judgment, that the model remains valid when extrapolated beyond the observed domain. In even more challenging situations, the available observations may be insufficient to perform any formal empirical testing. In these cases, expert judgment plays a fundamental role in assessing whether a given FD or FD-ensemble adequately represents the range of scientifically plausible interpretations relevant to the societal application under consideration (see the lower panel of Fig. 7). Expert elicitation may also be used to assign

relative weights to alternative FDs when there is no justification for considering all of them equally plausible (Beven, Almeida, et al., 2018; Selva et al., 2024). "

6. *It would be nice if you discussed non-stationarity, particularly in the context of exceedance probabilities (over some time frame). If aleatory models are built on existing data sets and stationary assumptions, hazard forecasts could go wrong quickly. This is particularly true in the context of a warming planet that is changing the underlying physical processes that drive many oceanic and weather related hazards. Again, I'm not expecting you to "solve" this problem in your framework, but it is certainly worth discussing. This could perhaps be expanded into the Communication section as well -- what does a 100-year flood mean under such non-stationarity? How do we get a handle on that kind of epistemic uncertainty?*

This is another very important point. Non-stationarity does not pose any conceptual difficulty for the probabilistic framework we propose, but it can considerably complicate the testing of complete forecasts. To clarify this issue, we have added two paragraphs to the revised manuscript, explaining how non-stationarity can be accommodated within the proposed framework and discussing its implications for forecast testing.

We have added this paragraph at section 4:

" 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".

And this paragraph at section 5:

"Forecast validation becomes more challenging in the presence of non-stationarity. In this case, it is generally impossible to estimate an empirical exceedance frequency for a given hazard threshold x_0 at a specific time because only one (or very few) observations are available for each time instant. However, if the non-stationary model is correctly specified, the exceedance probabilities associated with the observations should be uniformly distributed, according to the probability integral transform (PIT) (Gneiting and Katzfuss, 2014). Although this problem has not yet been fully addressed in the literature, the PIT framework can, in principle, be generalized to account for epistemic uncertainty. Let $f_m(\cdot)$ denote the FD associated with the m-th model and let $x(t_i)$ be the observation at time t_i. For each model, we compute the PIT values $\phi_m(t_i) = f_m(x(t_i))$. If a generic model m correctly represents the data-generating process, the corresponding set of PIT values $\{\phi_m(t_i)\}$ should follow a uniform distribution. When epistemic uncertainty is represented by an ensemble of M alternative FDs ($m=1,\ldots,M$), this requirement can be generalized by requiring that the ensemble of PIT distributions adequately encompasses the uniform distribution. "

7. *Line 324: One last small point, you should reference a news article that contains this quote. Junior researchers may not be familiar with it and it would be good to point them toward that cultural reference.*

We have added the reference to Logan (2009) and to Smith and Berger (2019), which claim that not all uncertainties are amenable to probabilistic quantification (our "unknown unknowns").

Logan, D. C.: Known knowns, known unknowns and unknown unknowns and the propagation of scientific enquiry, J. Exp. Bot., 60, 712–714, 2009.