Reponse to the comments of the reviewer #2, Dr. Francesco Serinaldi

By Nguyen Viet Dung

General comments

The paper under review introduces a probabilistic flood mapping strategy combining a quasi 2D hydrodynamic model fed by design hydrographs. These hydrographs are obtained by rescaling nondimensional typical hydrographs according to the flood peak and volume deduced from a bivariate distribution with dynamic (time varying) marginals. The paper is formally well written and the overall quality is good, however, as many other studies recently published in the literature, it suffers an original sin: the "mathematization" curse, just to use a Klemˇs' definition [Klemeˇs, 1986]. …

REPLY: First of all, We would like to thank the reviewer #2, Dr. Francesco Serinaldi for his insightful and contributive comments on the first version of the manuscript.

As mentioned already in the response to reviewer 1, we will completely rework the structure and focus of the manuscript. It will be more focussed on the statistical aspect with the overall aim to find a suitable method for flood hazard analysis of the Mekong Delta, taking the hydraulic and hydrological characteristics and the observed trends (Delgado et al, 2010) into account. We will discuss the pros and cons, limits and uncertainties of the stationary as well as non-stationary bi-variate approaches and finally give a recommendation for a suitable statistical procedure. The derivation of the hazard maps will be dropped and subject of a following publication.

Delgado, J. M., Apel, H., and Merz, B.: Flood trends and variability in the mekong river, Hydrology and Earth System Sciences, 14, 407-418, 10.5194/hess-14-407-2010, 2010.

SPECIFIC COMMENTS

I understand that the aim of the Authors is to propose a framework useful in nonstationary conditions; however, in principle, we are already able to introduce models even more complex than that proposed in this study. For example, we can incorporate dynamic copulas accounting for the time variation of the dependence structure, nonparametric link functions to allow for nonlinear variation of the copula and marginal parameters, exogenous covariates, and so forth. All these sophistications can be easily implemented and, in principle, we can push the degree of model complexity up to the exact (but uninformative) reproduction of the data. Therefore, the point is not how much we are clever in introducing complex models, but which is their correctness and usefulness when the available information (86 pairs of annual flood peaks and volumes) is just enough the reliably estimate summary statistics of central tendency. An example can help to clarify this point. Figure 1 reproduces the univariate and bivariate distributions fitted by Yue et al [1999]

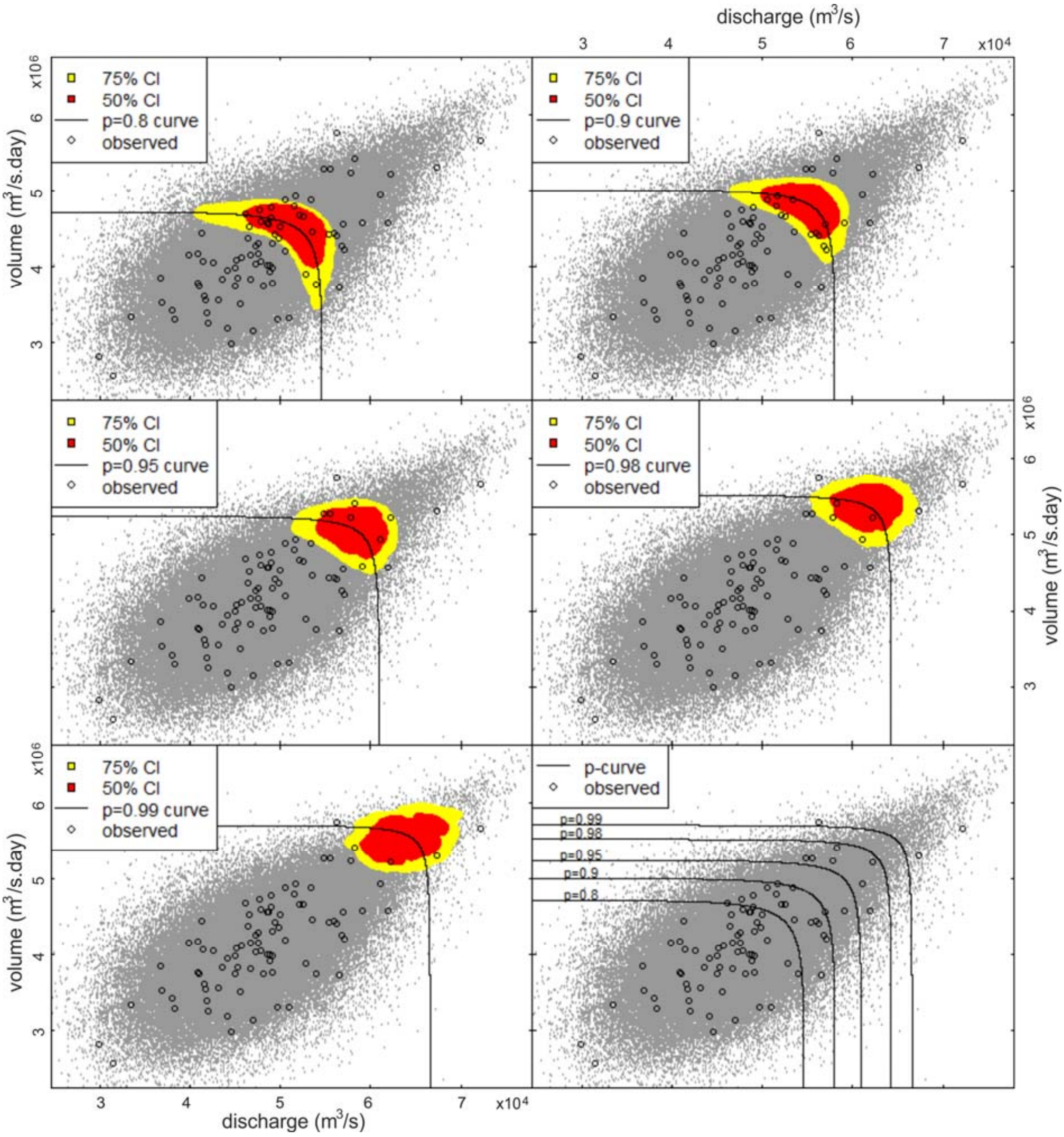
on 33 pairs of flood peaks and volumes. The figure also shows the confidence bands of the marginal distributions along with the sampling uncertainty areas related the 0.99 p-level curve. Since the uncertainty areas cover a large set of univariate quantiles and p-level curves, it is rather evident that the definition of the events with 0.99 "AND" and "OR" probabilities can just be an educated guess. Figure 2 shows that at least thousands iid points are required to reliably estimate the 0.99 p-level quantiles.

Now, suppose that we have 5000 years of real-world observations from an ideal (perfectly working) stream gauge: first, we can make inference directly from the observations and do not need any model, and second, it is reasonable that the observations refer to very different historical conditions of climate, drainage basin and river network, thus making questionable the basic hypotheses required by statistical inference procedures [Kleme˘s, 2000a,b; Koutsoyiannis, 2004]. In other words, refined statistical tools can be useful but cannot replace data. Introducing highly parameterized models to describe small samples does not add insight if we account for the increased uncertainty related to the additional parameters. The guidance should always be the Occam's razor (*multipla nonest ponenda praeter necessitatem*, which was stated before Einstein's sentence reported at the beginning of Ch. 3 of Dung [2011]), and a fair assessment of the true benefits resulting from the increased sophistication.

Moreover, in real-world design problems, statistical discrimination tools such as AIC and likelihood ratio tests have a limited value, as we are interested to improvements that are significant from a physical and economical point of view rather than purely statistical.

The overall meaning of the above remarks is that we need to reconcile statistics and engineering by using strictly necessary statistical tool based on the available information, paying attention to the fulfillment of the underlying hypotheses, trying to communicate the uncertainty, and avoiding ill-posed confidence in refined techniques that can be cool and fashionable but substantially no well devised for the problem at hand.

REPLY: We thank the Referee #2 for understanding the objective of our study and showing the constraint of the study using a limited data set. However, we have to bring to attention that 86 years of observation is very long in the context of hydrological observations. Most of the studies on flood risk have shorter time series than this. Defining a priori that this data series is not sufficient for an analysis more sophisticated that a standard uni-variate frequency analysis would eventually mean that a more complex flood hazard analysis is unlikely to be performed within this century world-wide. But many river systems actually require a multi-variate analysis in order to properly describe their hydrological and hydraulic characteristics in the statistical framework, as in presented study. However, we agree with the reviewer and Vit Klemes that the statistical analysis should not pretend certainty when there is none and acknowledge that is was too ambitious to publish such a complex study in a single publication. Thus we will shift the focus of the manuscript on this subject, i.e. discuss the prerequisites given by the characteristics of the river system and the different statistical approaches in much more detail. We will also introduce show the sampling uncertainty of the bi-variate stationary statistics as proposed. The following figure shows the sampling uncertainty for the Gumbel copula and the AND case joint probability of non-exceedance. As indicated by the reviewer, the uncertainty changes and increases with higher quantiles. A similar figure and its discussion will be presented in the revised manuscript.



As mentioned above, I understand the general idea to describe both copulas and marginals with (linearly) time varying parameters for the sake of generality and transferability (P285L10), but the final model actually reads as a bivariate Gaussian distribution applied to preliminarily log-transformed data, that is, probably the most classical bivariate approach that have been applied in hydrology for at least 50 years. Therefore, cannot section 3.1 be synthesized in

a few sentences saying that the log-transformed peak-volume pairs are described by a Gaussian bivariate distribution, which can be replaced by alternative families, if this is required based on the empirical evidence? In passing, the main advantage of copulas in not the possibility to incorporate different marginals (this is already allowed in the meta-Gaussian framework and is the basis of the hydrological applications of Box-Jenkins ARMA modeling framework, for instance), but the possibility of using dependence structures different from the meta-Gaussian (under "suitable conditions", the Sklar's theorem guarantees that plugging whatever marginals in a copula the resulting model is always a valid joint distribution).

The use of marginal distributions with time-varying parameters is an element of distinction with respect to a classical bivariate Gaussian distribution. However, this is exactly one of the above-mentioned practical cases in which statistics must meet engineering and common sense. I agree with the Koutsoyiannis [2006] statement that "*stationarity is closely associated with a model (rather than a real world system) either deterministic or stochastic, which implies an ensemble of infinite realizations*.". The Authors justify the use of dynamic marginals based on the results reported in Section 6.3.2.1 of Dung [2011] that reads

There exist several methods which can be used for detecting trends. This study uses a simple but robust non-parametric test, the Mann-Kendall test (Mann, 1945), to test for trends in the peak and volume series. The null hypothesis is that there is no trend in the peak, resp. volume series at the significance level of 10%. The Mann-Kendall test shows that trends are present in both peak and volume series. It means that it is reasonable to apply the non-stationary flood frequency analysis for both series.

In my opinion, this analysis is not enough for two reasons: (1) based on my experience, and looking at the time series shown in Fig. 4 of the manuscript, I am rather confident that the null hypothesis is not rejected at the 5% or 1% significance levels, and (2), more important, it is often overlooked that MK test checks for monotic trends, meaning that the alternative hypothisis is not unique (e.g. linear trend) but multiple, thus emcompassing whatever linear or nonlinear, abrupt or slowly varying monotonic patterns (this is why MK is less pawerful then its parametric competitors when the alternative matches the specific alternative embedded in the parametric techniques). Therefore, assuming a linear pattern for the functional relationships of the distribution parameters is arbitrary and not justified neither theoretically or empirically. Even though MK indicates rejection, it does not provide any information about the shape of the possible trend: for instance, it can be related to a regime shift, or an S-shaped pattern converging to some asymptotic level.

The key point is however that the data are not enough to infer about the shape of the possible trends and their evolution in the future [e.g., Guerreiro et al, 2013]. Moreover, without a physical (deterministic) justification for possible

trends (e.g. antropic intereventions on the basin), it is more likely that we are just dealing with fluctuations of natural processes that evolve (in an unknown manner) on time scales that go beyond the period of observation. From a modeling point view, the large uncertainty of the model parameter estimates discussed above makes the difference between stationary and nonstationary p-level curves likely insignificant. Indeed, this is confirmed by the small difference in the final flood maps. Accounting for the sampling and parameter uncertainty, the difference is probably even less evident.

REPLY: As mentioned earlier trend have been detected and proven to be significant by several test and also for lower significance levels in Delgado et al. (2010). This is the original motivation to include also non-stationary models in the study. If studies show that the iid framework is actually no longer valid, this should be accounted for by selecting appropriate methods. Thus in the revised manuscript we will still include the non-stationary approaches, but we will discuss it in much more depth comparing it with the stationary approach. As the results won't change much, the conclusion will be that given the rather similar results of the two approaches, the stationary model can be used for flood hazard analysis of the Mekong Delta although – theoretically – not valid, but a viable solution for the engineering practice. We will also discuss the uncertainties of bi-variate approaches on the background of limited length of time series. However, it has to be noted that deriving sampling uncertainty in a non-stationary environment is not straight forward. We will also drop the extrapolation of the trends in the revised manuscript acknowledging the inherent uncertainties of extrapolation of trends, as the reviewer correctly states.

THE "MULTIVARIATE-NONSTATIONARY" BUSINESS

Moving from simple techniques to slightly more complicated statistical tools, some concepts are not so easy to extend. In the present context, the discussion about the bivariate return periods raised in the text and review process seems to reveal some confusion about the meaning and consequences of work ing in a nonstationary multivariate framework. Unfortunately, it seems that this misunderstanding is more and more spread in the recent literature. The Authors correctly acknowledge that the choice between "AND" and "OR" bivariate return periods concerns the purpose of the study; however, the subsequent selection (P290L20) is based on a widespread misunderstanding which is generated by the use of the apparently friendly joint return periods instead of the joint probabilities. In more detail, the underlying joint probability of TOR is pOR = Pr[Q _ q [V _ v]. Referring to the orthogonal blue lines in Fig. 9, they define four quadrants while pOR defines a probability measure on a subset of the bi-dimensional domain corresponding the first, second and fourth quadrants (counted from the top right counter-clockwise). pOR describes the probability that a realization of (Q, V) exceeds (q, v) in terms of q or v or both. The statement "...*the OR definition most of the ob served data fall below the 10-yr return period, even the event of 2000 with the*

***historically largest damage. This is not plausible, and thus the AND definition is selected*.” is therefore incorrect because even though the event of 2000 is not exceeded in terms of volume, it is exceeded by seven events in terms of peak (to visualize this, it is sufficient to trace orthogonal lines crossing in the 2000 point and focus on the points falling in the three mentioned quadrants). This means that we observed in average eight events exceeding or equaling the 2000 event in 86 years, i.e. one event every 10 years in average, which is exactly the information coherently provided by pOR and thus TOR. On the other hand, pAND defines a probability measure on the first quadrant and describes the probability that a realization of (Q, V) simultaneously exceeds (q, v) in terms of both q and v. Focusing on the first quadrant defined by the orthogonal lines crossing in the 2000 point, it is clear that only this event occurred in 86 year, thus leading to pAND _ 0.99 and TAND _ 100.**

The inequalities in Eq. 8 are also a natural consequence of the above definitions: without resorting to probabilistic reasoning, it is intuitive and evident that observing an event falling in a wider domain (three quadrants) is more probable than observing an event in a smaller domain (the top right quadrant).

Actually, as it does not make sense to say that pAND is better than pOR and vice versa, the same holds for the corresponding joint return periods. Joint return periods give values (in years) which appear falsely friendly and easy to be interpreted; however, they simply hide the true meaning of the underlying joint (or conditional) distributions, leading to misunderstandings and wrong statements. Unfortunately, the literature does not help shedding light on this concepts and proposes incoherent comparisons of concepts that are essentially incomparable, thus increasing the confusion. This is rather dangerous, especially if we plan to deliver results to unskilled policy-makers and administrative authorities.

Based on the above discussion, it is evident that that the joint return period cannot be chosen from a statistical reasoning but selecting the joint probability that describes the scenarios that are critical for the (hydraulic) system: if the failure of the system or device occurs when q or v are exceeded, we must assess the risk by pOR; if the system collapses only when both q and v are exceeded but is not damaged when only one quantity exceeds the critical value, pAND is therefore required, whereas pOR does not apply at all, as it describes scenarios that do not match the system/device operation rules.

Talking about return periods, it should be also noted that further sourcesof confusion raise when we move from stationary to nonstationary framework. Namely, the definition of return period as the reciprocal of the probability of exceedance (not “of occurrence”) holds only under iid conditions. Unfortunately, the derivation of this relationship seems to be forgot by too many people, thus allowing for extensions in the nonstationary framework that are actually incorrect. Some informal remarks and references on these aspects

(as well as on the definition of joint return periods) can be found in Serinaldi [2012]. Such comments apply to the present manuscript as well (especially Sections 3, 4 and 5 therein).

REPLY: We completely agree with and thank the Referee #2 for this comment. As a consequence the notion of "return period" will be removed in the revised manuscript. Instead, we will use the joint probability to define the scenarios related to risk for our study. We will also justify the use of the AND joint probability by the characteristics of the Vietnamese Mekong Delta. This is characterized by a large number of channels, dikes, and control structures such as sluice gates. The presence of a wide spread dike system requires certain water level, i.e. discharges to be exceeded to cause inundations at all. But as socio-economical and agricultural systems are well adapted to the annual floods, this does not automatically means that a flood exceeding the dike levels is a disaster. For a flood event to become a disaster it needs also a high flood volume, which means that given a certain water level is exceeded, larger areas, and also those that normally are not flooded, can be inundated. This is typically causing the reported flood damages and are thus define the events that pose a high risk. This can be seen in Figure 9, where flood with a similar flood peak discharge can be disastrous or not, depending on the flood volume. The flood in 2000 is the most significant example for this. In the revised manuscript we will highlight this point and discuss the most disastrous flood events in terms of their Q and V characteristics.

Minor and editing remarks

P277L18: "People" perhaps is a typo.

REPLY: corrected

P278L10-15: Based on the above discussion, I would avoid statements introducing nonstationarity as something taken for granted both in general and especially for the data at hand.

REPLY: We base this statement on the detected and published trends in Delgado et al (2010). We will make this clear in the revised manuscript.

P280L24: "...the combination of the natural hydraulic peculiarities in combination with the large anthropogenic influence". Maybe it is better "Thus, the combination of the natural hydraulic peculiarities with the large anthropogenic influence..." or "Thus, the natural hydraulic peculiarities in combination with the large anthropogenic influence..."

REPLY: added

P282L22: "adapted"

REPLY: corrected

P284L15-25: Please, specify how the typical nondimensional hydrographs are rescaled. The text specifies that the peak is multiplied by the peak values simulated from the bivariate distributions and then the volume is adjusted to match the simulated volumes. Is this done by removing the peak value from the hydrograph? Please, add a few technical details.

REPLY: As we drop this part in the revised manuscript, we don't extend the discussion on this. However, we will consider this comment in the foreseen additional manuscript on the hazard map derivation.

P286L20-26: Tail dependence is defined as the limit of the conditional probability Pr[Q _ t|V _ t] as t ! 1. It is defined for every copula but the value of the limit is zero for some families. Therefore it is better to say that copulas can have tail dependence equal to zero. Nonetheless, the discussion about tail dependence can be avoided as it is not applied in the analysis and is not used to discriminate between the candidates. On the other hand, the sample size is certainly insufficient to assess whatever asymptotic property.

REPLY: The author thank the Referee #2 for this comment. We agree that the given data is not long enough to have a proper analysis on the (upper) tail dependence although this can be obtained using a simple Chi-plot (Abberger, 2005) or non-parametric estimators of the upper tail dependence coefficient (Serinaldi, 2008).

Abberger, K.: A simple graphical method to explore tail- dependence in stock-return pairs, Appl. Financ. Econom., 15, 43–51, 2005.

Serinaldi, F.: Analysis of inter-gauge dependence by Kendall's τK, upper tail dependence coefficient, and 2-copulas with applica- tion to rainfall fields, Stoch. Environ. Res. Risk. A, 22, 671–688, 2008

Sections 3.2 and 3.3: AIC and its modifications are performance measures and not goodness-of-fit tests. I suggest to apply at least a likelihood ratio test or better some ECDF based test. These tests are implemented in the R package copula. Moreover the scatter plots do not provide a good graphical diagnostic. The diagram of the Kendall distribution, i.e. the distribution function of the copula is a better tool. A description can be found in the paper by Genest and Favre (2007) cited in the manuscript.

REPLY: The author thank the Referee #2 for this comment. A formal goodness fit test based on Cramer von Mises statistic (Genest and Favre, 2007) was additionally run. The result of the test reveals that both Gaussian copula and Gumbel perform well. The statistic of the test as below will be added in the manuscript.

Genest, C., and Favre, A.-C.: Everything you always wanted to know about copula modeling but were afraid to ask, Journal of Hydrologic Engineering, 12, 347-347, 10.1061/(asce)1084-0699(2007)12:4(347), 2007.

P293L1-5: It is not clear to me how the pairs are simulated. Do the Authors simulate 100 pairs from the p-level curve, i.e. from the Volpi-Fiori's conditional distribution? Please, add some technical detail.

REPLY: Yes, 100 pairs of peak and volume were simulated from the p-level curve (p=0.8,0.98 and 0.99 in this study) according to Volpi-Fiori's conditional distribution. We will illustrate this clearer in the revised manuscript.

Sections 6-7: The inundation maps are a bit difficult to read. The overall patterns seem to be coherent but does the pixel-wise calculations guarantee that the values in neighbor pixels are coherent even in light of the time of propagation along the drainage network?

REPLY: This part will be dropped.

P296L17-25: I do not agree with the Authors. The uncertainty of extrapolating beyond the observed frequencies is so high that every model can only provide guess estimates which are statistically indistinguishable. Assessing which extrapolation is more correct is rather difficult [Klemeˇs, 2000b] without further data or other sources of information. In my opinion the best strategy is to apply robust techniques complemented by a sound assessment of the sampling and (distribution) parameter uncertainty (the Bayesian procedures mentioned by Reviewer 1 are an option).

REPLY: As mentioned earlier, we will drop the extrapolation of the trends. We will also don't give any conclusion about higher quantiles/extreme events other than noting that due to the uncertainties involved no robust statement can be made.

P298L5-10: I do not agree with the Authors. I believe that there is not any theoretical justification for the use of nonstationary models. They can be used if we have some evidence that they perform better than the stationary distribution according to some design criterion. By complementing the estimates of the extreme quantiles with suitable confidence intervals, it will be evident that the only inference we can do (by 86 values) about e.g. the flood with 0.999 (univariate, conditional or joint) probability of exceedance is just that it will be surely larger than the largest observation. Of course this holds if no additional data (and information) is available.

REPLY: Following the answer given above, we will no longer advocate for the non-stationary model based on guesses related to higher quantiles.

P298L13-18: I believe that these statements are incorrect. As shown by the figures reported below, the combination of the sampling and parameter uncertainty is larger than the inherent variability described by the pairs that fall over a p-level curve, which on the contrary are just a subsample of the uncertain p-level scenarios. The main source of uncertainty is exactly epistemic. Resorting to univariate distributions as an example, a distribution

describes the inherent uncertainty while the confidence bands describe the epistemic uncertainty. For small sample, such as that in the present case, the uncertainty of the extrapolated quantiles is dominant. Data contain information, which is the basis of the specific knowledge (synthetic a posteriori prepositions, just to use classical Kant's terminology), which in turn provides a better understanding of the world when it is combined with the general knowledge (a priori synthetic prepositions).

REPLY: What we want to express with this statement is that the uncertainty in hydrological input we considered for the hazard maps is just the aleatory part. As we did not quantify the epistemic part, i.e. the sampling and parameter estimation uncertainty, the hazard maps show the natural variability only. However, as this part will be dropped, we will consider this comment in the foreseen following publication on the hazard maps.

References: please, check typos throughout the reference list.

REPLY: will be done.