https://doi.org/10.5194/nhess-2019-110
Preprint. Discussion started: 10 July 2019

Natural Hazards and Earth System Sciences
Open Access
Discussions

Spatial Seismic Hazard Variation and Adaptive Sampling of Portfolio Location Uncertainty in Probabilistic Seismic Risk Analysis

Christoph Scheingraber[1,2] and Martin Käser[2,1]

[1]Ludwig-Maximilians-Universität, Munich, Germany
[2]Munich Reinsurance, Munich, Germany

Correspondence: Christoph Scheingraber (scheingraber@geophysik.uni-muenchen.de)

Abstract. Probabilistic Seismic Risk Analysis is widely used in the insurance industry to model the likelihood and severity of losses to insured portfolios by earthquake events. Due to geocoding issues of address information, risk items are often only known to be located within an administrative geographical zone, but precise coordinates remain unknown to the modeler.

In the first part of this paper, we analyze spatial seismic hazard and loss rate variation inside administrative geographical zones in western Indonesia. We find that the variation of hazard can vary strongly not only between different zones, but also between different return periods for a fixed zone. However, the spatial variation of loss rate displays a similar pattern as the variation of hazard, without depending on the return period.

We build upon these results in the second part of this paper. In a recent work, we introduced a framework for stochastic treatment of portfolio location uncertainty. This results in the necessity to simulate ground motion on a high number of sampled geographical coordinates, which typically dominates the computational effort in Probabilistic Seismic Risk Analysis. We therefore propose a novel sampling scheme to improve the efficiency of stochastic portfolio location uncertainty treatment. Depending on risk item properties and measures of spatial loss rate variation, the scheme dynamically adapts the location sample size individually for insured risk items. We analyze the convergence and variance reduction of the scheme empirically. The results show that the scheme can improve the efficiency of the estimation of loss frequency curves.

1 Introduction

Seismic risk analysis is widely used in academia and industry to model the possible consequences of future earthquake events, but is associated with a wide range of deep uncertainties (Goda and Ren, 2010). The treatment and communication of uncertainties is highly important for informed decision making and a holistic view of risk (Tesfamariam et al., 2010; Cox, 2012; Bier and Lin, 2013). In the insurance industry, Probabilistic Seismic Risk Analysis (PSRA) is the means of choice to model the likelihood and severity of losses to insured portfolios due to earthquake events. In this context precise exposure locations are often unknown, which can have a significant impact on scenario loss as well as on loss frequency curves (Bal et al., 2010; Scheingraber and Käser, 2019).

In PSRA, uncertainty is usually taken into account by means of Monte Carlo (MC) simulation (e.g. Pagani et al. 2014; Tyagunov et al. 2014; Foulser-Piggott et al. 2017). This is a computationally intensive process, because the error convergence of MC is relatively slow and a high-dimensional loss integral needs to be evaluated with a sufficient sample size. In PSRA, the hazard component typically dominates the overall model runtime. As a result, stochastic treatment of portfolio location uncertainty can be particularly challenging - ground motion needs to be simulated on a large number of sampled risk locations. On the other hand, a fast model runtime is a key requirement for underwriting purposes in the insurance industry. Methods or sampling schemes to improve the error convergence of MC simulation are known as variance reduction techniques. MC simulation is ubiquitous in many areas of science and engineering and a wide variety of sampling schemes exists. Some well-known ideas are common random numbers and control variates (Yang and Nelson, 1991), importance-, stratified- and hypercube sampling, Quasi Monte Carlo Simulation (QMC) using low-discrepancy sequences, as well as adaptive sampling. The error convergence of different sampling schemes has been investigated for many different types of integrals and application areas (Hess et al., 2006; dos Santos and Beck, 2015). Some work has already been performed on variance reduction for PSHA and PSRA in the form of importance sampling, e.g. preferentially sampling the tails of the magnitude and site ground motion probability distributions (Jayaram and Baker, 2010; Eads et al., 2013). However, to our knowledge so far no study has specifically investigated variance reduction for location uncertainty in PSRA in a modern risk assessment framework. Building on a framework proposed in a recent study, in the present paper we describe a novel variance reduction scheme specifically designed to increase the computational efficiency of stochastic treatment of portfolio location uncertainty in PSRA.

The remainder of this paper is structured as follows. We outline the most important theoretical background in Section 2. Using a seismic risk model of western Indonesia, in Section 3 we explore spatial hazard and loss rate variation inside administrative zones. Based on this, in Section 4 we propose an adaptive location uncertainty sampling scheme and investigate its performance using several test cases in Section 5. In Section 6, we give some recommendations on how to apply the results in practice and conclude with possible future improvements.

2 Background

2.1 Probabilistic Seismic Hazard and Risk Analysis

PSRA is based on Probabilistic Seismic Hazard Analysis (PSHA; Cornell, 1968; Senior Seismic Hazard Committee, 1997; McGuire, 2004, where the exceedance rate λ of ground motion level y_0 at a site $\boldsymbol{r}_0$ is expressed by the hazard integral

$$\lambda(y_0,\boldsymbol{r}_0)[y \geq y_0] = \int\limits_V \int\limits_{m_{\min}}^{m_{\max}} P[y \geq y_0 | m,\boldsymbol{r},\boldsymbol{r}_0] \cdot \nu(m,\boldsymbol{r}) dm d\boldsymbol{r}, \tag{1}$$

with $\nu(m,\boldsymbol{r})dmd\boldsymbol{r}$ the seismic rate density which describes the spatio-temporal distribution of seismic activity, $P[y \geq y_0|m,\boldsymbol{r},\boldsymbol{r}_0]$ the conditional probability of exceeding ground motion y_0 at site $\boldsymbol{r}_0$ given a rupture of magnitude m at source location $\boldsymbol{r}$, and V the spatial integration volume containing all sources which can cause relevant ground motion at $\boldsymbol{r}_0$. Assuming that the occurrence of earthquake events is a temporal Poisson process, the probability of at least one exceedance of y_0 within time interval t_0 is given by

$$P(y_0,t_0,\bar{\lambda})[y \geq y_0] = 1 - e^{-\bar{\lambda}t_0}, \tag{2}$$

where $\bar{\lambda}$ is the mean annual recurrence rate.

For PSRA in the insurance industry, MC simulation is commonly used to obtain a set of stochastic ground motion fields $\hat{\mathbf{Y}}$ and to then compute the probability that a loss level ι_0 is exceeded as

$$P(\hat{\mathbf{Y}},\Theta)[\iota \geq \iota_0] = \sum_{i=1}^{n_\mathrm{e}} \int_{\iota_0}^{\infty} f_\iota(\iota|\hat{\mathbf{Y}}_i,\Theta)d\iota, \tag{3}$$

where $f_\iota(\iota|\hat{\mathbf{Y}}_i,\Theta)$ is the loss probability density function for a portfolio Θ given the *i*th ground motion field $\hat{\mathbf{Y}}_i$. Summing up the contribution of all n_e events yields the total loss exceedance probability. A Probable Maximum Loss (PML) curve, showing loss against mean return period T (with $T = 1/\bar{\lambda}$), can be obtained from the loss exceedance probability curve (Equation 3) using a first order Taylor approximation of Equation 2:

$$T = \frac{t_0}{P(y_0,t_0,\lambda)[y \geq y_0]}. \tag{4}$$

Here, t_0 is the period of interest (time interval), which is 1 year for most reinsurance contracts.

2.2 Portfolio Location Uncertainty

Perhaps surprisingly, in the insurance industry, portfolios frequently lack precise coordinate-based location information. Obtaining this information is often not possible, e.g. because geocoding engines are not used systematically or can not reliably obtain coordinates from the policy address of the insured risk. Especially for large treaty portfolios with thousands or millions of risks, it apparently is simply too much effort for the primary insurer or the insurance broker to obtain and provide this information. Unfortunately, this is also not uncommon for smaller portfolios consisting only of a few hundred high-value risks. However, administrative zones, such as postal codes, can easily be obtained from the insurance policy.

Exposure uncertainty has previously been identified as an important area of research (Crowley, 2014), and we already introduced a framework for stochastic treatment of location uncertainty in a recent paper (Scheingraber and Käser, 2019). In our framework, locations of risk items without precise coordinate location information are sampled with replacement from a weighted irregular grid inside their corresponding administrative zone. The grid weights are used to preferentially sample locations in areas of assumed high insurance density, e.g. based on population density or on commercial and industrial inventory data depending on the type of risk (Dobson et al., 2000). An example of such a weighted grid is shown in Figure 1.

In MC simulation, the choice of a pseudo-random number generator is of particular importance. In this study we use *MRG32K3a*, a combined multiple recursive generator which efficiently generates random number sequences with low memory requirements and excellent statistical properties (L'Ecuyer, 1999). MRG32K3a supports up to $1.8 \cdot 10^{19}$ statistically independent *substreams*. Each substream has a *period*[1] of $7.6 \cdot 10^{22}$. These properties make MRG32K3a well suited for a large scale parallel MC simulation of seismic risk.

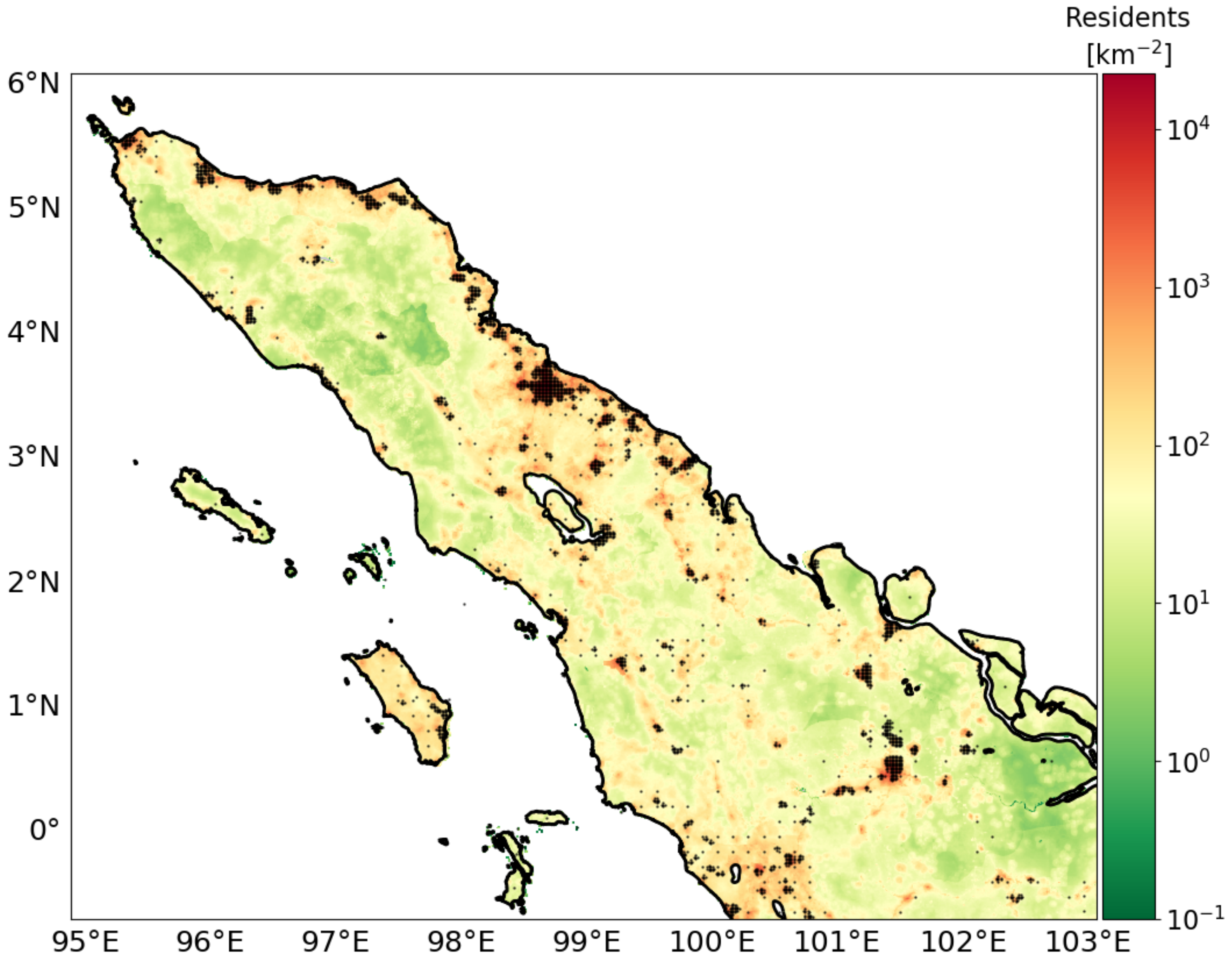

Figure 1. An example of a weighted grid used as an insurance density proxy for the location uncertainty framework. This shows northern Sumatra. Color indicates population density (residents per km^2) as a proxy for insured exposure density. Black markers depict grid points of the weighted grid. The population data in this plot is based on a free dataset (Gaughan et al., 2015).

[1] The period of a pseudo-random number generator refers to the minimum length of a generated sequence before the same random numbers are repeated cyclically.

2.3 Evaluation of the Proposed Sampling Scheme

2.3.1 Standard Error

Because MC simulation is a stochastic method, there are no strict error bounds for statistics of interest obtained from a sample of finite size n. The error is therefore usually estimated using the standard deviation of the sampling distribution of the respective statistic, which is referred to as its standard error (E_{SE}). If the sampling distribution is known (e.g. normal), standard errors can often be obtained using a simple closed-form expression (Harding et al., 2014). For the statistics estimated in this study, e.g. PML at a specific return period, we can however not make a valid distribution assumption when taking location uncertainty into account. We therefore use repeated simulation to evaluate the performance of the proposed sampling scheme. The standard error can then be estimated as

$$\mathrm{E}_{\mathrm{SE}}(\hat{\Phi}_R) = \sqrt{\mathrm{Var}(\hat{\Phi}_R)}, \tag{5}$$

where $\hat{\Phi}_R$ denotes a set of estimations of a statistic obtained from R repeated simulations and $\mathrm{Var}(\cdot)$ the variance operator. The corresponding relative standard error E_{RSE} can be obtained by dividing E_{SE} by the estimated statistic. To estimate confidence intervals of standard errors, we use bootstrapping with the bias-corrected accelerated percentile method (Efron, 1979; Efron and Tibshirani, 1986).

2.3.2 Bias and Convergence Plots

The bias of an estimator $\hat{\theta}$ is defined as

$$\mathrm{Bias}(\hat{\theta}_n) = \mathbb{E}_\theta(\hat{\theta}_n) - \theta, \tag{6}$$

where $\hat{\theta}_n = f(x_1, x_2, \ldots, x_n)$ is the estimator depending on the n members of the sample and $\mathbb{E}_\theta$ its expected value. Because deriving the bias analytically is infeasible for a complex numerical simulation such as performed by our framework, we use simple MC[2] with a large sample size as empirical reference and approximation for θ. In addition we use convergence plots, which are a simple yet powerful method to monitor and verify the results (Robert and Casella, 2004). The values estimated using simple MC and the adaptive variance reduction scheme are plotted against increasing sample size n.

2.3.3 Variance Reduction, Convergence Order and Speedup

To quantify the performance of the proposed scheme at a particular sample size n, we use the following well-known definition of variance reduction VR:

$$\mathrm{VR} = \frac{\sigma^2_{\mathrm{MC}}}{\sigma^2_{\mathrm{LSS}}}, \tag{7}$$

where σ^2_{MC} is the variance using simple MC and σ^2_{LSS} the variance using the proposed location sampling scheme (MacKay, 2005; Juneja and Kalra, 2009)

[2] For simple MC, the strong law of large numbers guarantees an *almost certain* convergence for $n \to \infty$.

To describe asymptotic error behavior for growing n, we use the big O notation ($\mathcal{O}$; Landau, 1909; Knuth, 1976). For example, the error convergence order of simple MC is always $\mathcal{O}(n^{-0.5})$ independent of the dimensionality of the integrand (Papageorgiou, 2003).

To compare the real runtime required by simple MC and the proposed scheme to reach a specific relative standard error level ε_{RSE}, we use the speedup S defined as

$$\text{S} = \frac{t_{\text{MC}}}{t_{\text{LSS}}}, \tag{8}$$

where t_{MC} the runtime required by simple MC and t_{LSS} the runtime required by the proposed location sampling scheme.

2.4 Generation of Synthetic Portfolios

In this work, we use synthetic portfolios in western Indonesia modeled after real-world counterparts in terms of spatial distribution of risk items as well as value distribution among risk items.

2.4.1 Value Distribution

The total sum insured (TSI) is kept constant for all portfolios:

$$\text{TSI} = \text{const.} = 1 \cdot 10^6. \tag{9}$$

However, the TSI is distributed among a varying number of risk items (portfolio size). For this study, we use portfolio sizes n_r of 1, 10, 20, 50, 100, 1000 and 10000 risk items.

The value distribution observed in many real residential portfolios can be approximated well by a randomly perturbed flat value distribution:

$$\text{VI}^*_{\text{flat},i} = \frac{\text{TSI}}{n_\text{r}} \cdot X_i, \tag{10}$$

$$\text{VI}_{\text{flat},i} = \frac{\text{TSI}}{\sum_{i=1}^{n_\text{r}} \text{VI}^*_{\text{flat},i}} \cdot \text{VI}^*_{\text{flat},i}\,. \tag{11}$$

$\text{VI}_{\text{flat},i}$ (“value insured”) is the value assigned to the *i*th risk item and n_r denotes the number of risk items. X_i is a uniform random number in the interval $[1-p, 1+p]$, where p is a perturbation factor set to 0.2, which is consistent with the characteristics of many real portfolios. Equation 11 normalizes the n_r randomly perturbed insured values to ensure $\sum_{i=1}^{n_\text{r}} \text{VI}_{\text{flat},i} = \text{TSI}$.

2.4.2 Geographical Distribution

For each portfolio size, we created a set of 6 portfolios with an increasing fraction of unknown coordinates: 0%, 20%, 40%, 60%, 80%, and 100% of the risk items have unknown coordinates and are only known on the basis of their administrative zone (Indonesian provinces, or regencies and cities, see Section 3).

The geographical distribution of the exposure locations follows the weighted irregular grid described in Section 2.2. For each portfolio size, a portfolio with 0% unknown coordinates is initially created by choosing exposure locations from the irregular grid according to the grid point weights. For the other portfolios with the same number of risk items but a higher fraction of unknown coordinates, coordinate-based location information is then removed stepwise from the initial portfolio. In each step, 20% of the risk items are randomly selected for the removal of coordinates until all risk items have unknown coordinates.

3 Case Study: Spatial Seismic Hazard Variation in Western Indonesia

3.1 Hazard Model

We use a proprietary seismic risk model based on the South-East Asia hazard model of the United States Geological Service (USGS) by Mark Petersen et al. (2007). Site conditions are based on topographic slope (Wald and Allen, 2007). The geometry of the Sumatra subduction zone is a complex fault representation based on the three-dimensional *Slab 1.0* model (Hayes et al., 2012). For events on the complex fault, we use a rupture floating mechanism similar to the implementation of OpenQuake (Pagani et al., 2014), a free and open-source seismic hazard and risk software developed as part of the Global Earthquake Model initiative (Crowley et al., 2013). The model is described in greater detail in a recent paper (Scheingraber and Käser, 2019).

3.2 Spatial Seismic Hazard Variation

For this analysis, we compute seismic hazard on a regular grid using a resolution of 0.3°. We investigate the coefficient of variation (CV) of hazard inside administrative geographical zones for different levels of resolution, corresponding to provinces and regencies or cities in Indonesia. The CV is defined as

$$\mathrm{CV} = \frac{\sigma}{\mu}, \tag{12}$$

where σ is the standard deviation and μ the mean.

3.2.1 Dependence on Resolution Level of Geographical Zones

Figure 2 shows the CV of peak ground acceleration with an exceedance probability of 10% in 50 years per province in Indonesia. There is a noticeable decrease of the CV from west to east. The subduction modeled by the complex fault and the Sumatra Fault Zone (SFZ) result in the highest CV on Sumatra (most values 0.2 - 0.3). The CV is also relatively high on Java (around 0.15). The CV is the lowest in Kalimantan (< 0.1) due to the absence of any known or modeled crustal faults. As only gridded seismicity is used in this area, the hazard variation is very small. Furthermore, zones with a large extent perpendicular to the SFZ show a larger CV than zones with a smaller extent along the direction of the steepest hazard gradient. An example of this are the provinces of *Jambi* and *Bengkulu* in Figure 2. Arguably, location uncertainty is more important in *Jambi* than in *Bengkulu*.

Figure 3 shows the CV per regency or city for the same exceedance probability. Due to the smaller spatial extent of the administrative zones, the CV is in general lower at this more granular resolution of administrative geographical zones. Another

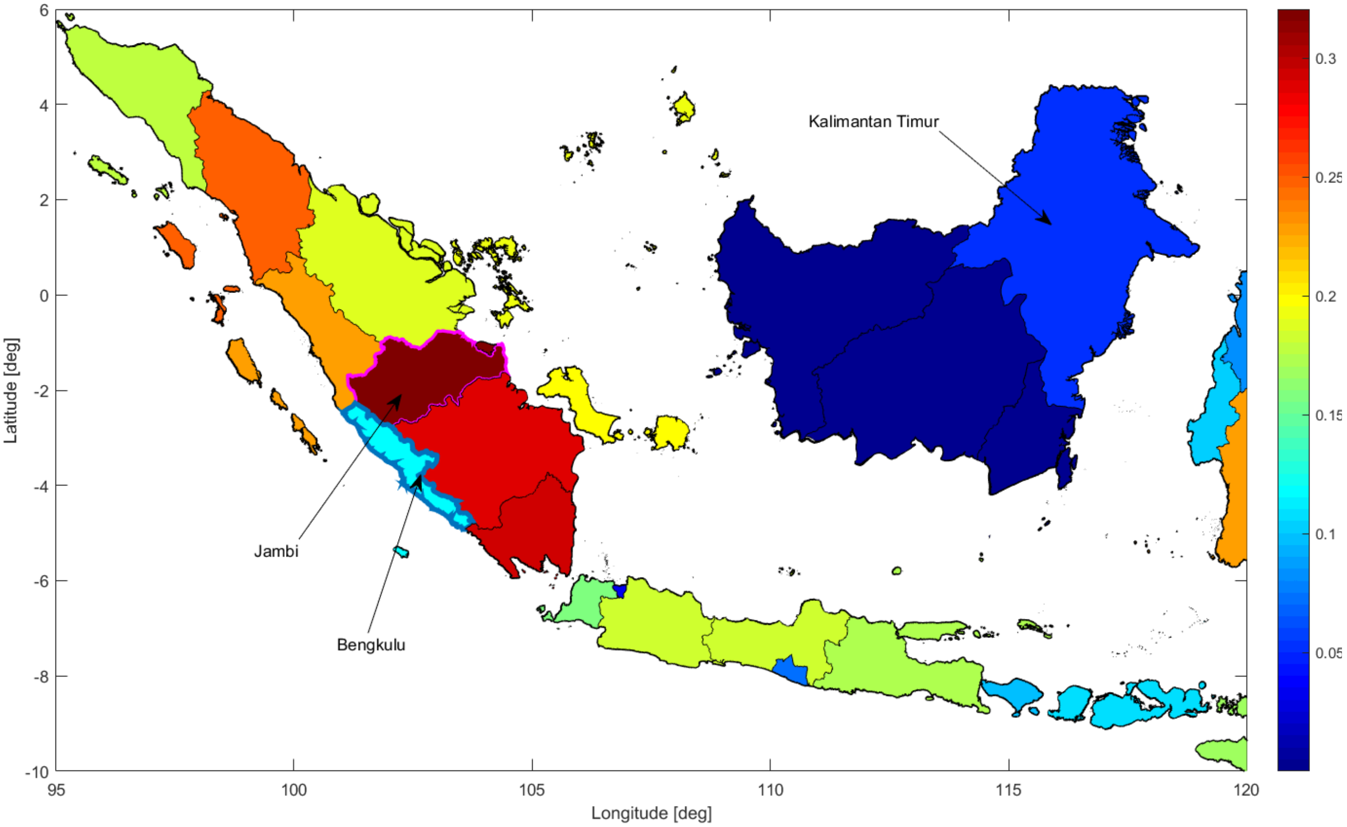

Figure 2. Coefficient of Variation (CV) of Peak Ground Acceleration (PGA) with an exceedance probability of 10% in 50 years inside provinces in western Indonesia. Color denotes the CV. Note how the CV is higher in provinces that have a large extent perpendicular to the Sumatra Fault Zone, such as *Jambi* (outlined in pink color), than in provinces with a small extent in that direction, such as *Bengkulu* (outlined in blue).

observation is that the influence of individual seismo-tectonic features emerges; the CV is higher in the vicinity of modeled faults. While the Sumatra subduction only has a weak influence, the SFZ has a pronounced effect. Near the SFZ, the CV has values of about 0.1 - 0.2. Perpendicular to the SFZ, the CV quickly drops below 0.1.

In general, the CV is highest in zones close to modeled faults of shallow depths, as they result in a higher spatial hazard gradient than compared to areas where hazard is dominated by rather regularly distributed gridded seismicity. A reasonable assumption is that location uncertainty can be particularly high in such zones.

3.2.2 Dependence on Return Period

Analysis of the CV across different return periods for individual zones revealed a similar pattern for most administrative zones. The CV is small for short return periods, and reaches a relatively stable level above a certain return period. An example of

https://doi.org/10.5194/nhess-2019-110
Preprint. Discussion started: 10 July 2019

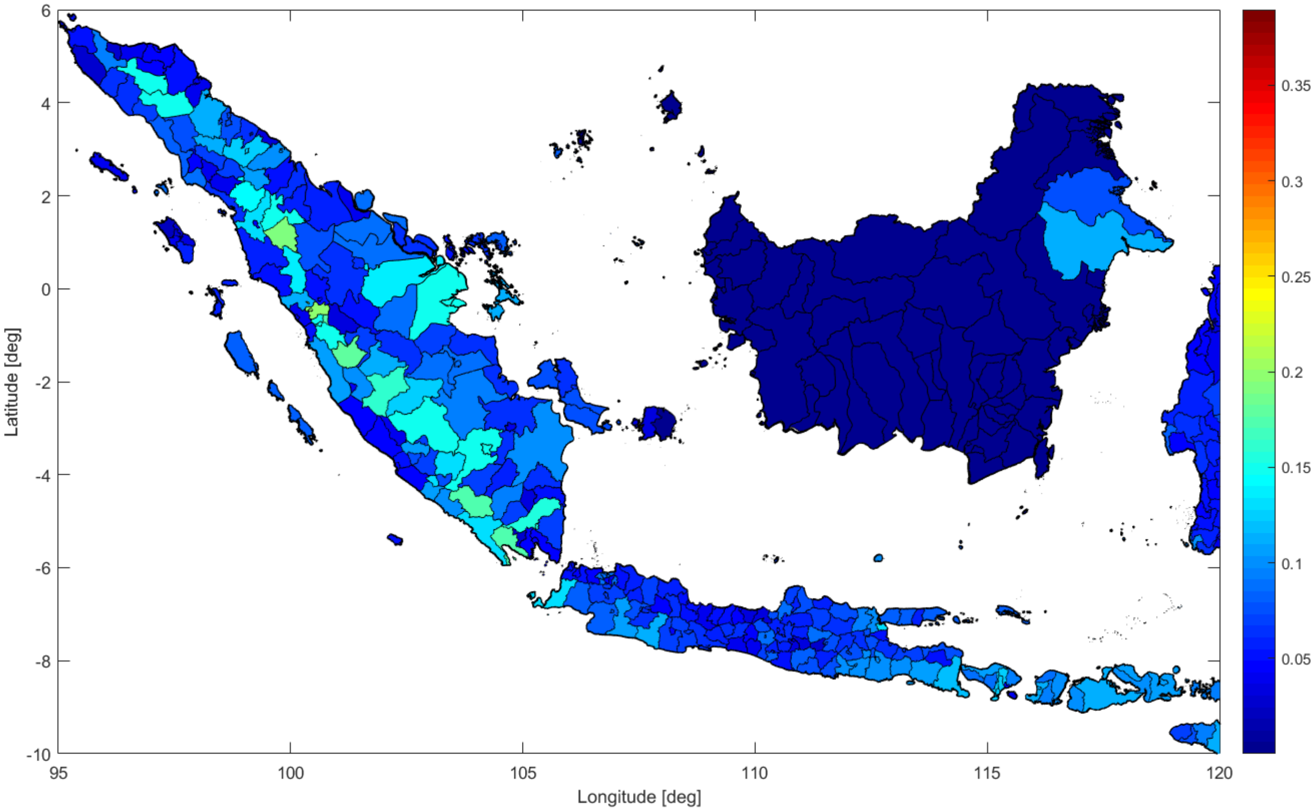

Figure 3. Coefficient of Variation (CV) of Peak Ground Acceleration (PGA) with an exceedance probability of 10% in 50 years inside regencies and cities in western Indonesia. Color denotes the CV. At this geographical resolution the CV is lower than for provinces (see Figure 2), and the influence of individual seismo-tectonic features, such as the Sumatra fault zone, becomes apparent.

this behavior is shown in Figure 4 for the province of *Jambi*. However, the CV does not show this pattern in all administrative zones. For some zones, especially at the level of regencies and cities, we could not determine a range of return periods for which the CV is roughly constant, as for example in the province of *Kalimantan Timur* shown in Figure 5.

3.3 Loss Rate Variation

The variability of the CV over return periods for certain zones makes it difficult to choose a general return period suitable for assessing the spatial variation of hazard inside a zone. To avoid the subjectivity introduced by a manual decision process for a suitable return period, we use the CV of the loss rate per zone, as it considers all return periods. Figure 6 shows the CV of the loss rate for Indonesian provinces. The overall pattern agrees with the pattern of the spatial hazard variation in Figure 2, but the range of values is much higher, from about 0.1 to 0.9.

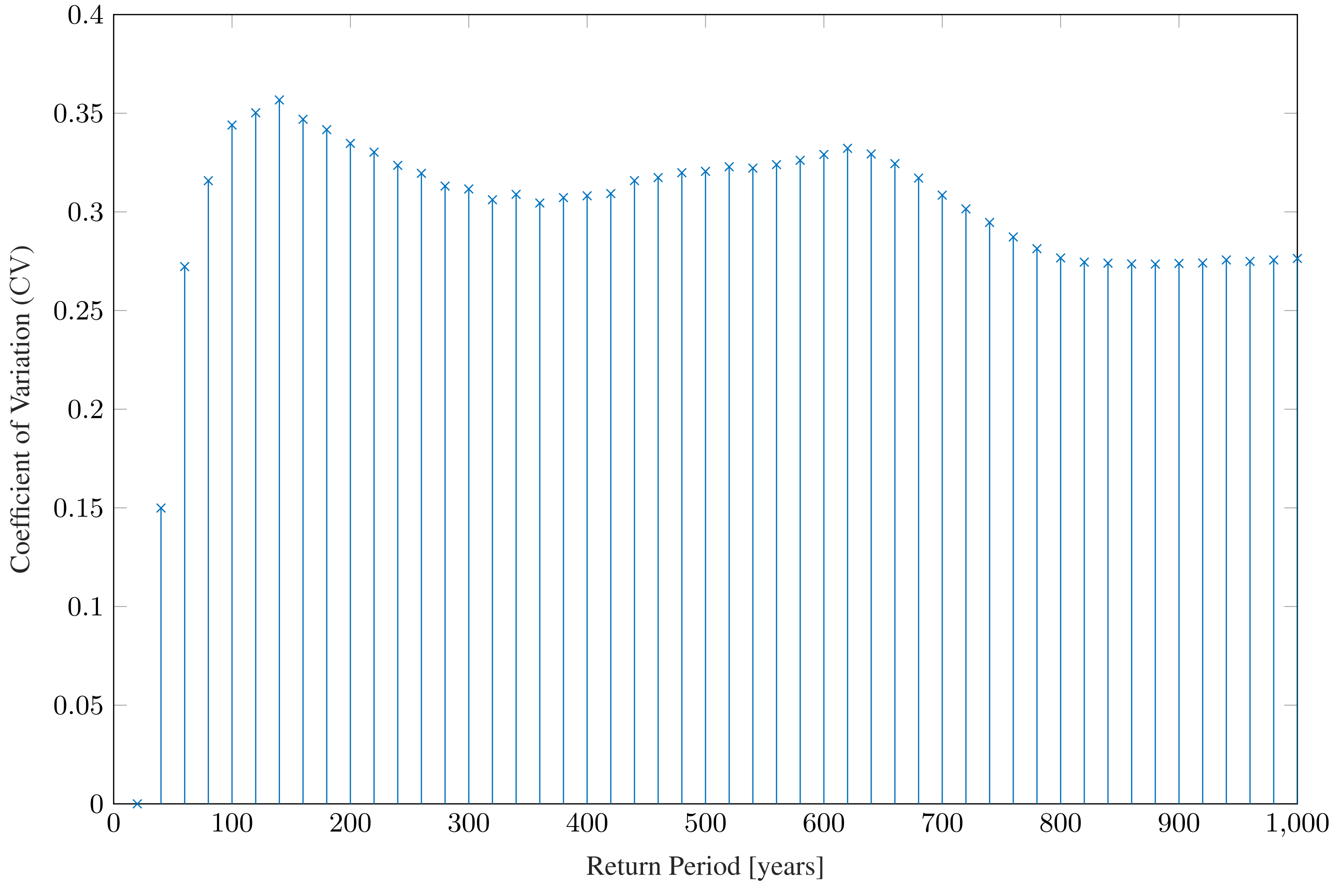

Figure 4. Coefficient of variation (CV) of ground motion predicted to be exceeded at various return periods for the *Jambi* province (see Figure 2). The CV remains quite stable over a large range of return periods.

4 A Framework for Adaptive Sampling of Portfolio Location Uncertainty

To increase efficiency, in our framework ground motion is jointly simulated on all unique locations of all sampled location sets. Since the computation of hazard dominates the overall runtime of PSRA, it is worthwhile to explore possibilities to distribute the number of locations on which hazard is computed in a smart way among risk items. To this end, we introduce three sampling criteria to determine the location sample size individually per risk item. A large location sample size is used for risk items for which at all three criterions indicate that location uncertainty has a strong influence. If any of the three criteria predicts that location uncertainty has a lesser effect, a smaller sample size is used. In this way, more computational effort is invested where it is important and a better estimation of the PML curve associated with a lower variance is obtained for a given number of used hazard locations. To not add noticeable overhead to the calculation, a key requirement is that all criteria can be evaluated very efficiently. To keep the computational overhead small, another design goal is that the framework is adaptive in a sense

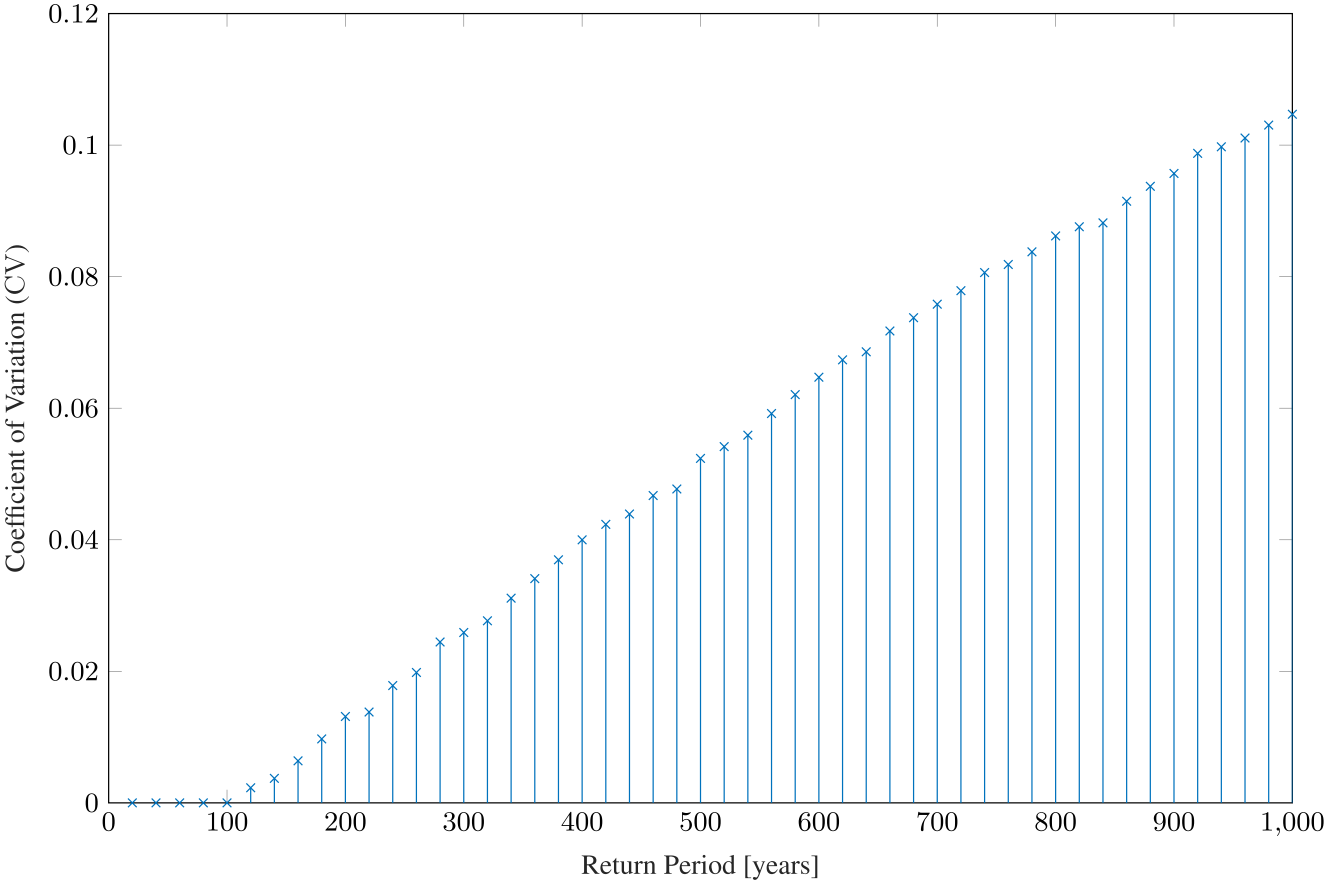

Figure 5. Coefficient of variation (CV) of ground motion predicted to be exceeded at various return periods for the *Kalimantan Timur* province (see Figure 2). In this case, it is not possible to determine a range of return periods for which the CV remains in a stable range.

that it depends directly on properties of the portfolio and a precalculated hazard variability (see Section 3), but does not require on-the-fly integral presampling such as used by some general purpose adaptive variance reduction schemes (Press and Farrar, 1990; Jadach, 2003).

4.1 Risk Location Index Mapping Table

We store an array containing all unique geographical locations on which ground motion is simulated, and another array storing the sampled location indices per risk item. Table 1 illustrates the concept. Each column of the table corresponds to a location set representing a valid realization of location uncertainty for the entire portfolio. To combine unequal sample sizes for risk items without introducing bias due to overemphasis of a subset of a sample, we restrict the sample size to powers of two. The full sample can then be repeated in the mapping table.

https://doi.org/10.5194/nhess-2019-110
Preprint. Discussion started: 10 July 2019

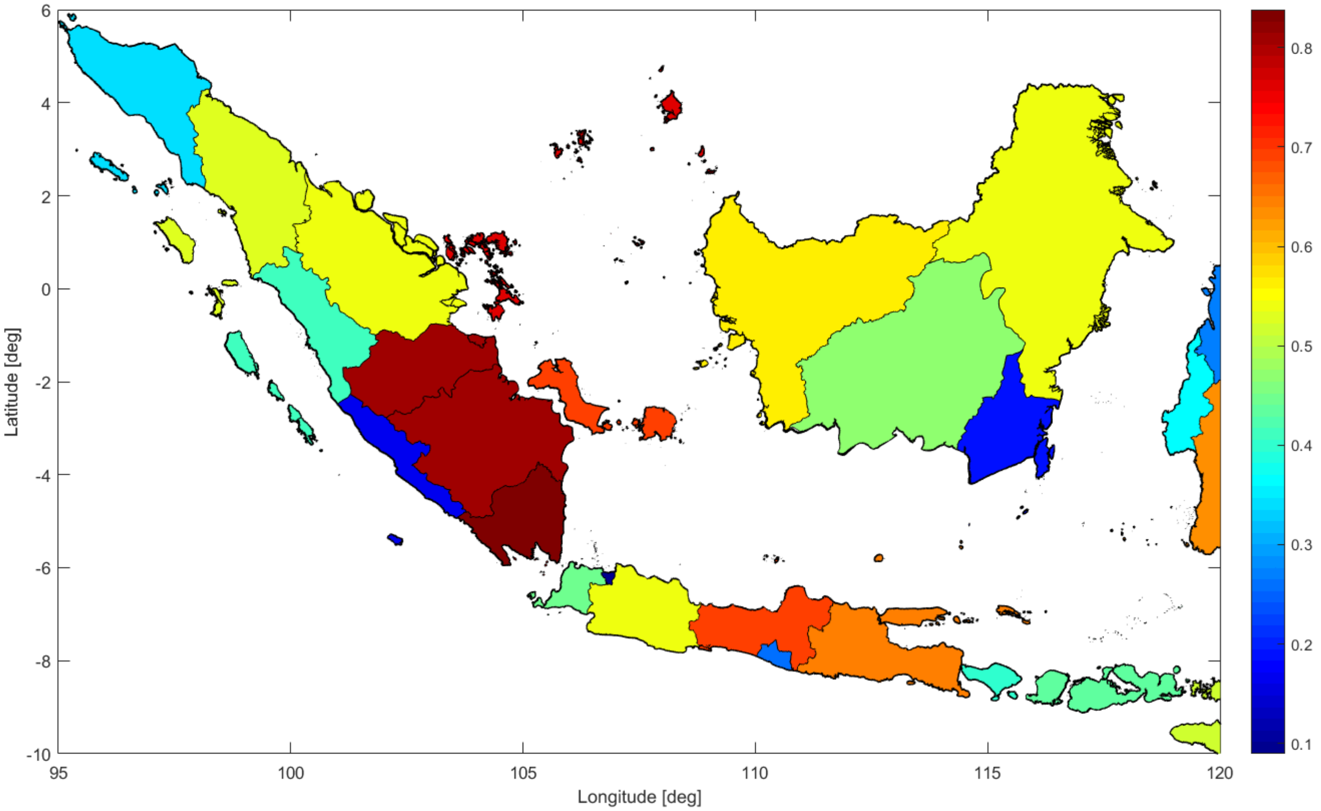

Figure 6. Coefficient of variation (CV) of the loss rate inside provinces in western Indonesia. Color denotes the CV.

Table 1. Risk Location Index Mapping Table. Rows correspond to individual risk items, showing sampled grid point indices. Each column represents a possible spatial distribution of the portfolio. Risk item 1 has the maximum location sample size of $n_{\max} = 4$, but risk items 2 and 3 only have a sample size of 2 and 1, respectively.

Risk Item Index	**Sample Size**	**Sample 1**	**Sample 2**	**Sample 3**	**Sample 4**
1	*4*	43	13	31	51
2	*2*	23	18	23	18
3	*1*	51	51	51	51

4.2 Criterion I: Coefficient of Variation of Loss Rate

The first criterion is based on the CV of loss rate within a zone (see Section 3), hereafter denoted by $\mathrm{CV_z}$. The values of $\mathrm{CV_z}$ can be precomputed for all administrative geographical zones, and therefore the evaluation of this criterion can be implemented in a very efficient manner.

The number of samples n_{L} due to criterion I is defined piecewise:

$$n_{\mathrm{L}}^* = \begin{cases} 1, & \text{if } \mathrm{CV_z} \leq t_{\mathrm{l}}, \\ \frac{n_{\max}-1}{t_{\mathrm{u}}-t_{\mathrm{l}}} \cdot \mathrm{CV_z}+1, & \text{if } \mathrm{CV_z} \in (t_{\mathrm{l}}, t_{\mathrm{u}}), \\ n_{\max}, & \text{if } \mathrm{CV_z} \geq t_{\mathrm{u}}. \end{cases} \tag{13}$$

Here, t_{l} and t_{u} are lower and upper threshold values. $n_{\max}$ represents the maximum used sample size. We round n_{L}^* up to the next higher power of two to obtain the final n_{L}. The criterion is shown in Figure 7 for the example $t_{\mathrm{l}} = 0.1$, $t_{\mathrm{u}} = 0.4$ and $n_{\max} = 16$. In our final implementation, t_{l} and t_{u} are chosen adaptively as empirical quantiles of the CV distribution ($\mathrm{CV}_{0.4}$ for t_{l} and $\mathrm{CV}_{0.6}$ for t_{u}, i.e. the 40% and 60% percentiles) of the loss rate of all administrative zones of a model (see Section 3), which was found to be a reasonable choice for our test cases with the aid of an extensive parameter study (see Section 5.1).

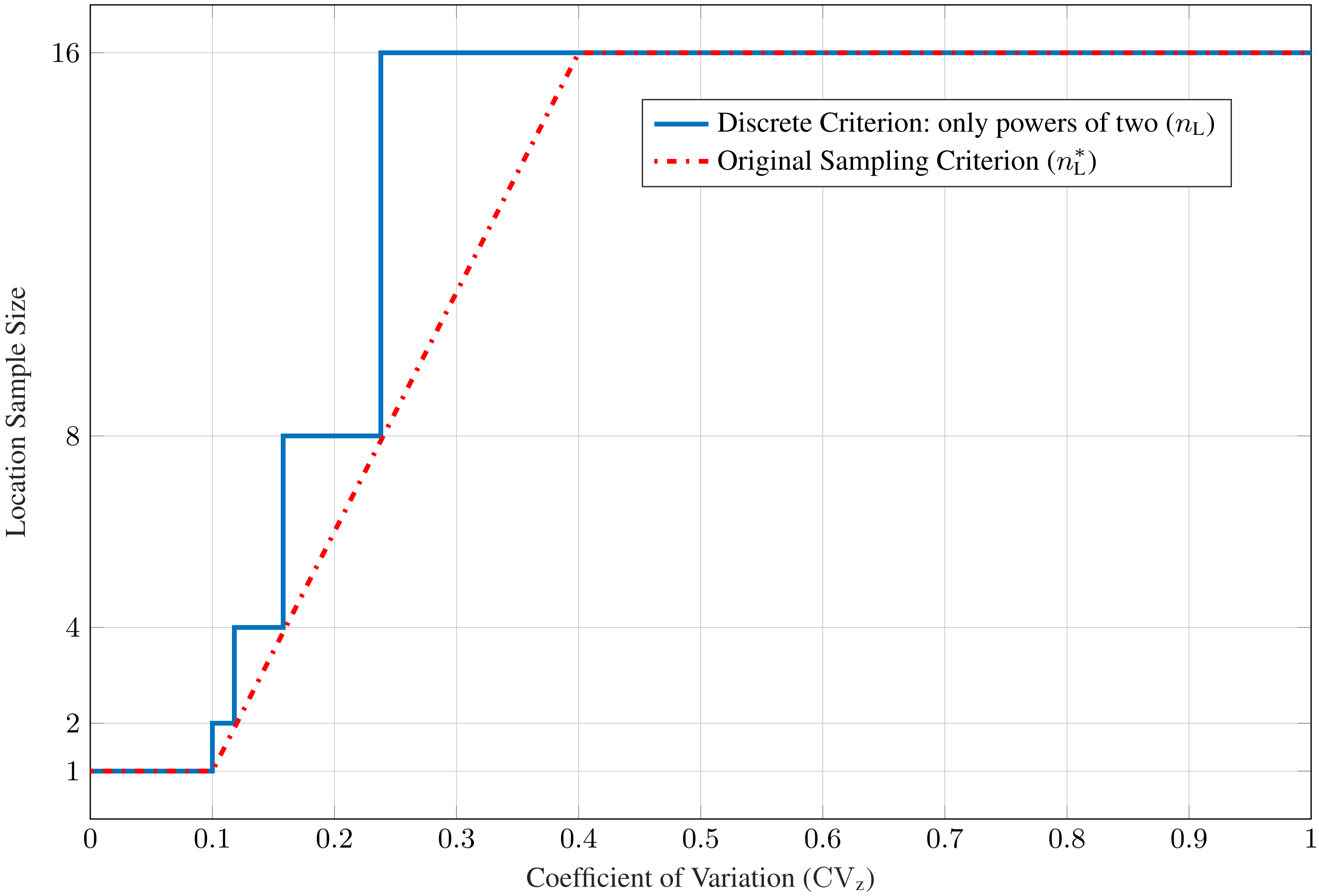

Figure 7. Criterion I: Number of samples per zone depending on coefficient of variation. The discrete realization of the criterion (limited to powers of two) is shown in blue, while the red line represents the theoretical linear behavior.

4.3 Criterion II: Number of Risk Items

The second criterion involves two steps. The first step defines a maximum sample size for the entire portfolio depending on the total number of risk items n_r in the portfolio and a threshold t_p as

$$n_\mathrm{R}^\dagger = \begin{cases} -\frac{n_\mathrm{max}-1}{\log(t_\mathrm{p}-1)} \cdot \log(n_\mathrm{r}-1) + n_\mathrm{max}, & \text{if } n_\mathrm{r} < t_\mathrm{p}, \\ 1, & \text{if } n_\mathrm{r} \geq t_\mathrm{p}, \end{cases} \tag{14}$$

which is then used to obtain a maximum sample size per zone, depending on the number of risk items in a zone n_z and a threshold t_z:

$$n_\mathrm{R}^* = \begin{cases} -\frac{n_\mathrm{R}^\dagger-1}{t_\mathrm{z}-1} \cdot (n_\mathrm{z}-1) + n_\mathrm{R}^\dagger, & \text{if } n_\mathrm{z} < t_\mathrm{z}, \\ 1, & \text{if } n_\mathrm{z} \geq t_\mathrm{z}. \end{cases} \tag{15}$$

We round n_R^* up to the next higher power of two to obtain the final n_R. Figures 8 and 9 illustrate this criterion for $t_\mathrm{p} = 10000$, $t_\mathrm{z} = 100$ and $n_\mathrm{max} = 16$. In this study, t_p is chosen to be 10000 and t_z is set adaptively to equal the number of grid points of the weighted location uncertainty sampling grid (see Section 2.2) inside each administrative zone. The design of this criterion is based on the results of a previous study, in which we systematically investigated the effect of location uncertainty and loss aggregation due to spatial clustering of risk items for a large range of different portfolios. It was found that location uncertainty typically has a neglectable effect for very large portfolios and a roughly flat value distribution (Scheingraber and Käser, 2019).

4.4 Criterion III: Value Distribution

The third criterion depends on the relative insured values of risk items (“sum insured”, SI). Risk items are sorted with respect to their SI, and the index of their sorted order I_r is used along with a threshold index t_i to determine the maximum sample size per risk item:

$$n_\mathrm{V}^* = \begin{cases} -\frac{n_\mathrm{max}-1}{t_\mathrm{i}-1} \cdot (\mathrm{I}_\mathrm{r}-1) + n_\mathrm{max}, & \text{if } \mathrm{I}_\mathrm{r} < t_\mathrm{i}, \\ 1, & \text{if } \mathrm{I}_\mathrm{r} \geq t_\mathrm{i}. \end{cases} \tag{16}$$

We round n_V^* up to the next higher power of two to obtain the final n_V. Figure 10 illustrates this criterion for $t_\mathrm{i} = 6$ and $n_\mathrm{max} = 16$. In this study, for t_i we adaptively set the index of the first risk item which has a SI higher than the mean of all risk items.

4.5 Combination of Criteria

The final sample size for a specific risk item is then given by the minimum of the three criteria:

$$n = \min\{n_\mathrm{L}, n_\mathrm{R}, n_\mathrm{V}\}. \tag{17}$$

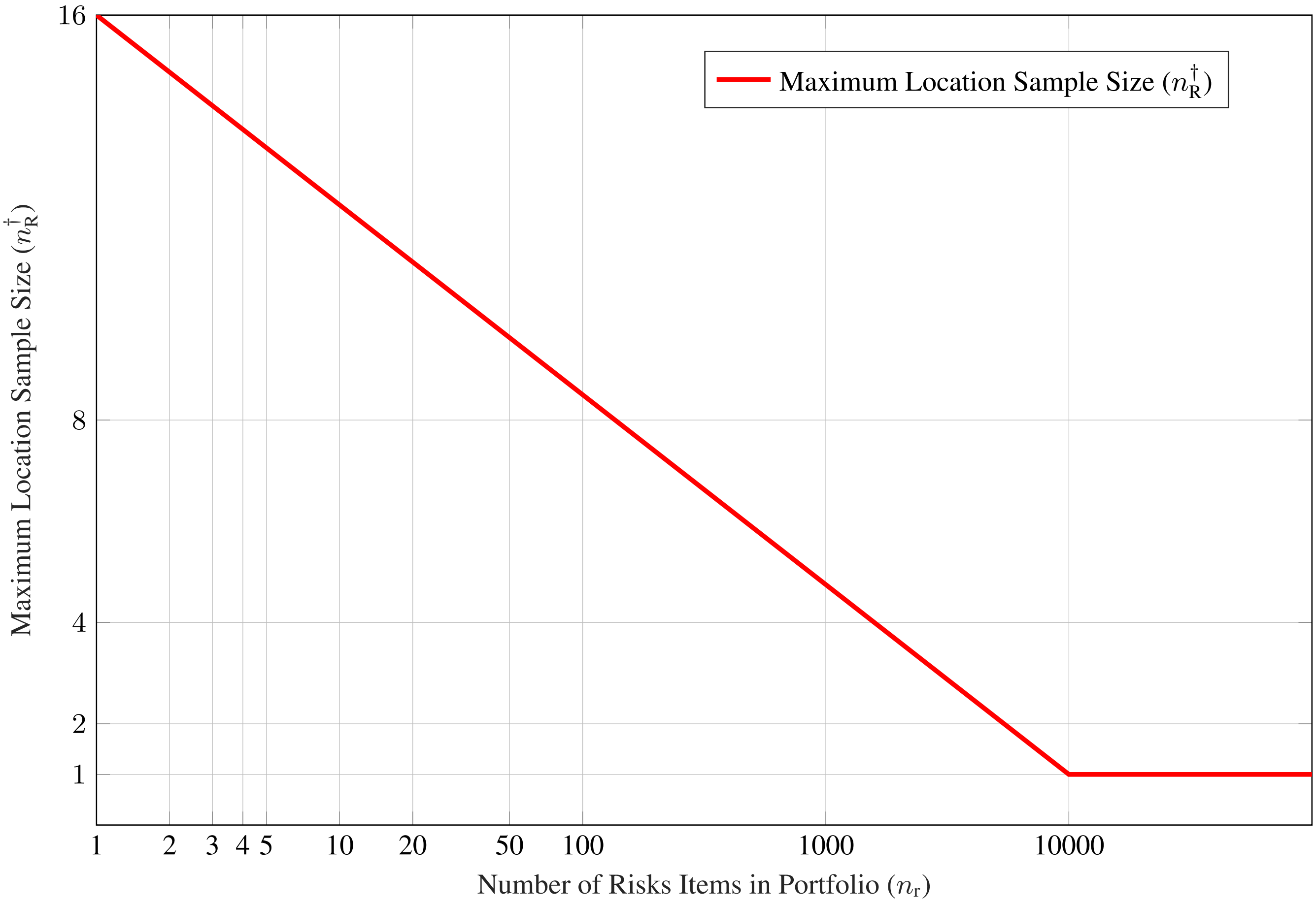

Figure 8. Criterion II.a: Number of samples per zone depending on the number of risk items in the portfolio. Note that we do not round up to the next higher power of two, since this plot illustrates Equation 14, which is an intermediate step.

The rationale behind this decision is that any of the criteria can separately predict that a particular risk item has a low impact on loss uncertainty. For example, if a risk item with an unknown coordinate has a low insured value, it has a relatively low impact on loss uncertainty even if the variation of hazard or loss rate within the corresponding administrative zone is high, and thus a small location uncertainty sample size can be used. Vice versa, the impact of location uncertainty is limited if a risk item with an unknown coordinate has a high insured value but the hazard within the corresponding administrative zone is relatively flat. Furthermore, loss uncertainty is also limited if a portfolio contains a very high number of total risk items or the number of risk items belonging to an administrative zone is high compared to the number of grid points within this zone.

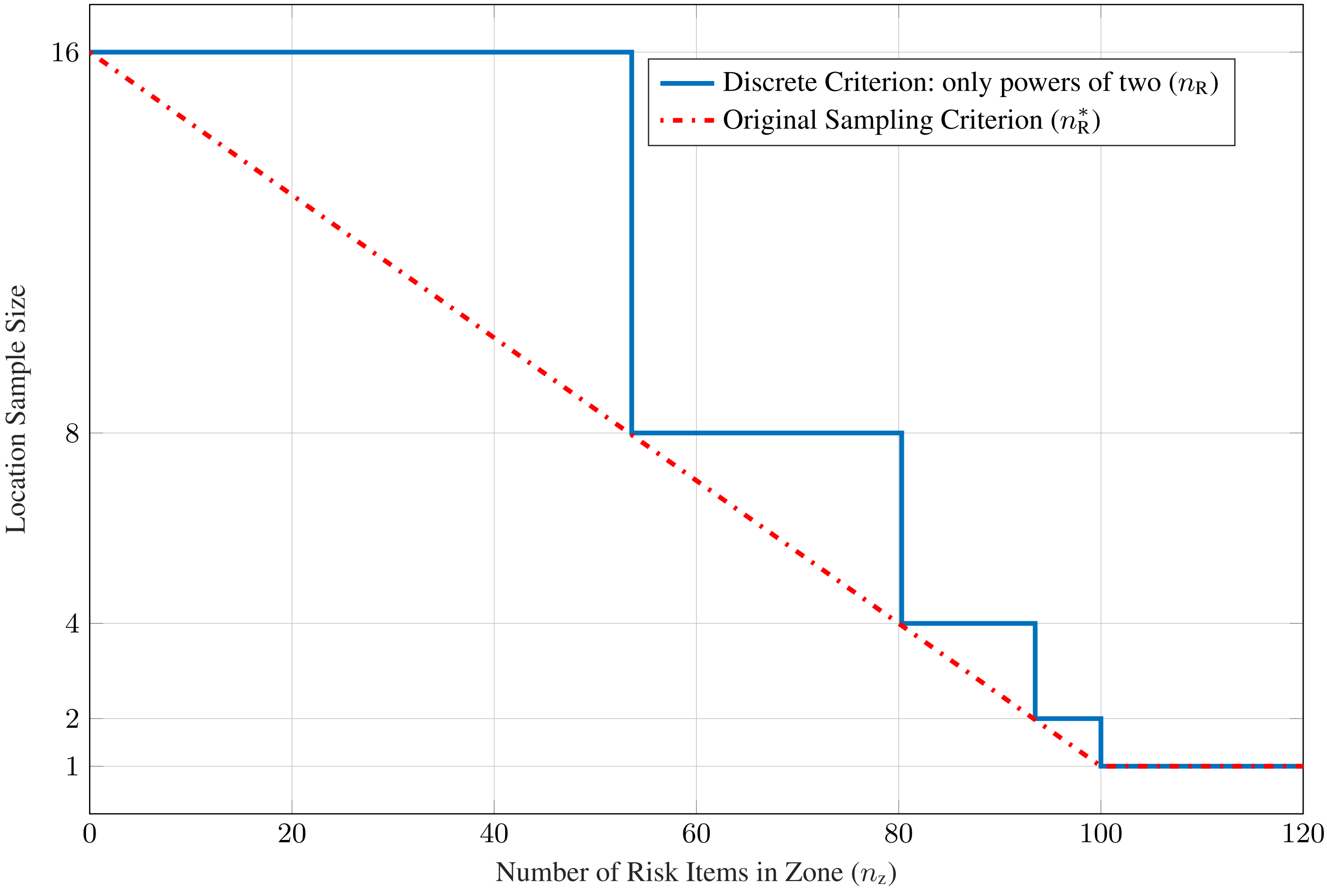

Figure 9. Criterion II.b: Maximum number of samples per zone depending on the number of risk items in an administrative zone. The discrete realization of the criterion (limited to powers of two) is shown in blue, while the red line represents the theoretical linear behavior.

5 Results

In this section, the variance reduction and speedup obtained with the proposed adaptive location uncertainty sampling scheme is analyzed using the western Indonesia hazard model described in Section 3.1 in conjunction with a vulnerability model for regional building stock composition. To this end, loss frequency curves are computed for the synthetic portfolios described in Section 2.4 with simple MC as well as the adaptive scheme. The convergence and relative standard errors are evaluated against the number of unique hazard locations used for the loss calculation by either approach and the associated required runtime is compared.

https://doi.org/10.5194/nhess-2019-110
Preprint. Discussion started: 10 July 2019

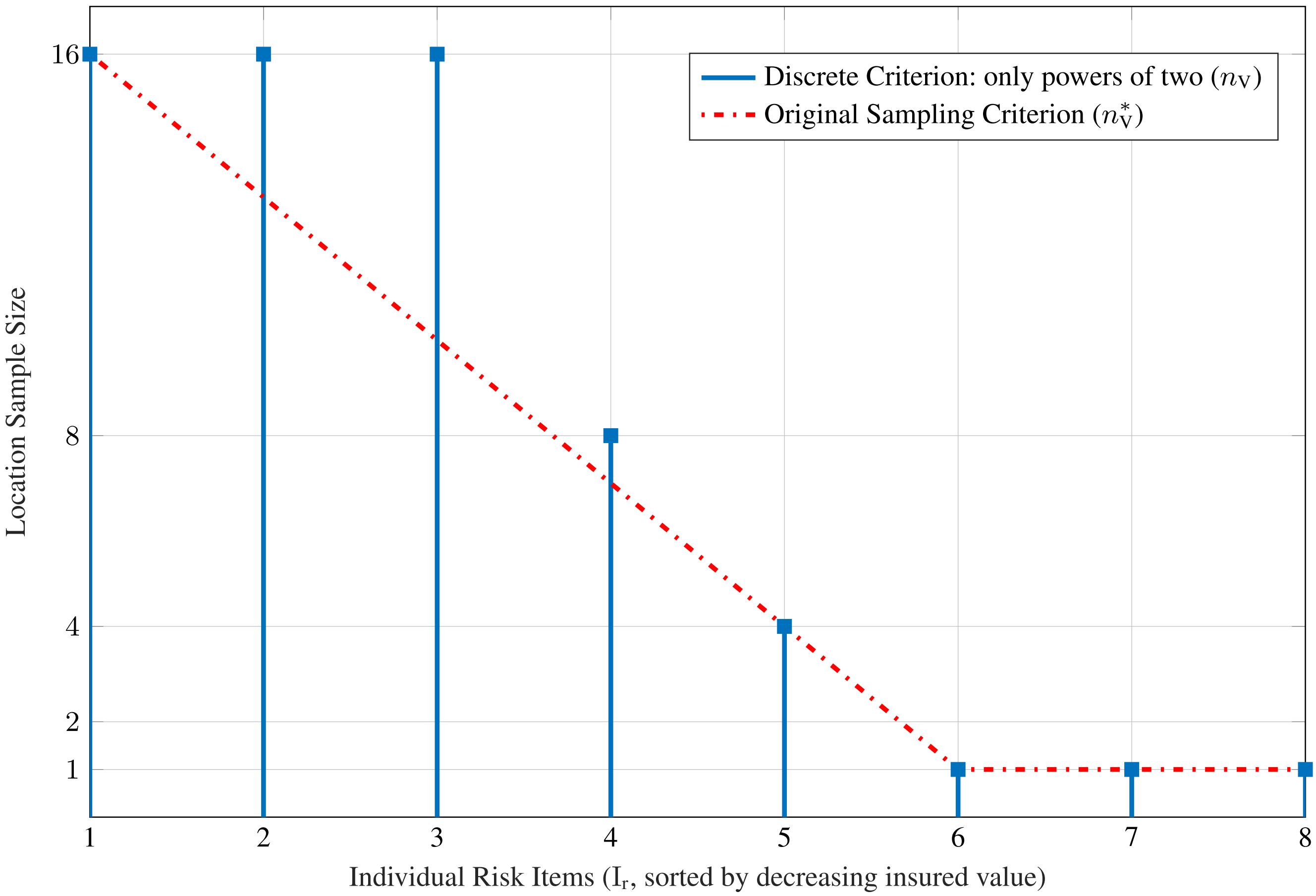

Figure 10. Criterion III: Number of samples per zone depending on the insured values of the risk items. The discrete realization of the criterion (limited to powers of two) is shown in blue, while the red line represents the theoretical linear behavior.

5.1 Spatial Variation Parameter Study

We first analyze the performance of the adaptive sampling scheme for different values of the lower (t_{l}) and upper (t_{u}) threshold parameters for the spatial variation of loss rate in an administrative zone in comparison to simple sampling. In simple MC, all risk items get the same location uncertainty sample size $n_{\max}$ and there is not restriction to powers of two. For this parameter study, we use values of $n_{\max} = 32, 64, 96, 128, 160, 192, 224, 256$ in order to obtain a smooth curve with a high number of support points.

For the adaptive variance reduction scheme, the sample size is restricted to powers of two and is determined for each risk item individually - potentially smaller than the maximum allowed location uncertainty sample size $n_{\max}$ (see Section 4 and Table 1). Since the sample size varies between risk items, for a meaningful comparison with simple MC it is necessary to use a measure of the total effort spend for the treatment of location uncertainty of all risk items. We use the total number of unique hazard locations (n_{hazard}) and the runtime spent for the computation of hazard (t_{hazard}). While for simple MC all risk items get the maximum sample size $n_{\max}$, the adaptive location sampling scheme reduces the sample size for risk items for which location uncertainty likely has a smaller influence. This means that the adaptive location sampling scheme results in a smaller n_{hazard} than simple MC for the same portfolio and $n_{\max}$. Therefore, in order to obtain a comparable values for n_{hazard}, a larger maximum sample size $n_{\max}$ has to be employed for the adaptive scheme than for simple MC. Here, we use $n_{\max} = 2^i$ with $i = 5, 6, \ldots, 8$.

For each sample size, the spatial variation threshold parameters are varied over the distribution of CV values, picking quantiles in constant steps of 0.2. The lower threshold t_{l} is varied from $\text{CV}_{0.0}$ to $\text{CV}_{0.8}$, and the upper threshold t_{u} from $\text{CV}_{0.2}$ to $\text{CV}_{1.0}$. For each combination of t_{l} and t_{u}, $R = 20$ repeated simulations were performed for each sample size to estimate the respective relative standard error E_{RSE}.

In general, for our test cases the scheme works well around $t_{\text{l}} \in [\text{CV}_{0.2}; \text{CV}_{0.4}]$ in combination with $t_{\text{u}} \in [\text{CV}_{0.6}; \text{CV}_{0.8}]$. For example, for a portfolio of 20 risk items and 100% unknown coordinates, Figure 11 shows a logarithmic plot of the relative standard error E_{RSE} of PML at a return period of 100 years against the number of used hazard locations n_{hazard} for some combinations of t_{l} and t_{u}. The error curves for all combinations of t_{l} and t_{u} have the same slope as the curve for simple MC and thus the same convergence order of $\mathcal{O}(n^{-0.5})$. For certain combinations, the error curve is below the curve for simple MC, meaning that in these cases the scheme successfully reduces the variance of the estimation and therefore the associated standard error.

For the final implementation, we used $t_{\text{l}} = \text{CV}_{0.4}$ and $t_{\text{u}} = \text{CV}_{0.6}$, which performed best in this parameter study.

5.2 Performance of the Final Implementation

We now evaluate the performance of the final implementation of the adaptive scheme, checking if it results in any unwanted systematic bias and investigating variance reduction and speedup for the calculation of PML for different portfolios.

5.2.1 Convergence and Bias

Figures 12 and 13 show convergence plots of PML at 100 years return period against the number of used hazard locations n_{hazard} for portfolios with $n_{\text{r}} = 10$ and $n_{\text{r}} = 100$ risk items, respectively. The left plots depict the results for portfolios with 60% unknown coordinates, the right plots the results for portfolios with 100% unknown coordinates. Simple sampling is shown in blue, the adaptive scheme in red. For all portfolios, the sample size n was varied as $n = 2^i$ with $i = 3, 4, \ldots, 9$. For each sample size and both sampling schemes $R = 20$ repeated simulations are shown as semi-transparent circles, with solid lines highlighting one individual repetition.

The results show that empirically the adaptive scheme converges to the same result as simple MC for our test cases, meaning that the scheme does not result in any systematic bias. It is also apparent that for a given number of used hazard locations n_{hazard}, the relative PML values obtained with the adaptive scheme scatter less than those estimated with simple MC.

https://doi.org/10.5194/nhess-2019-110
Preprint. Discussion started: 10 July 2019

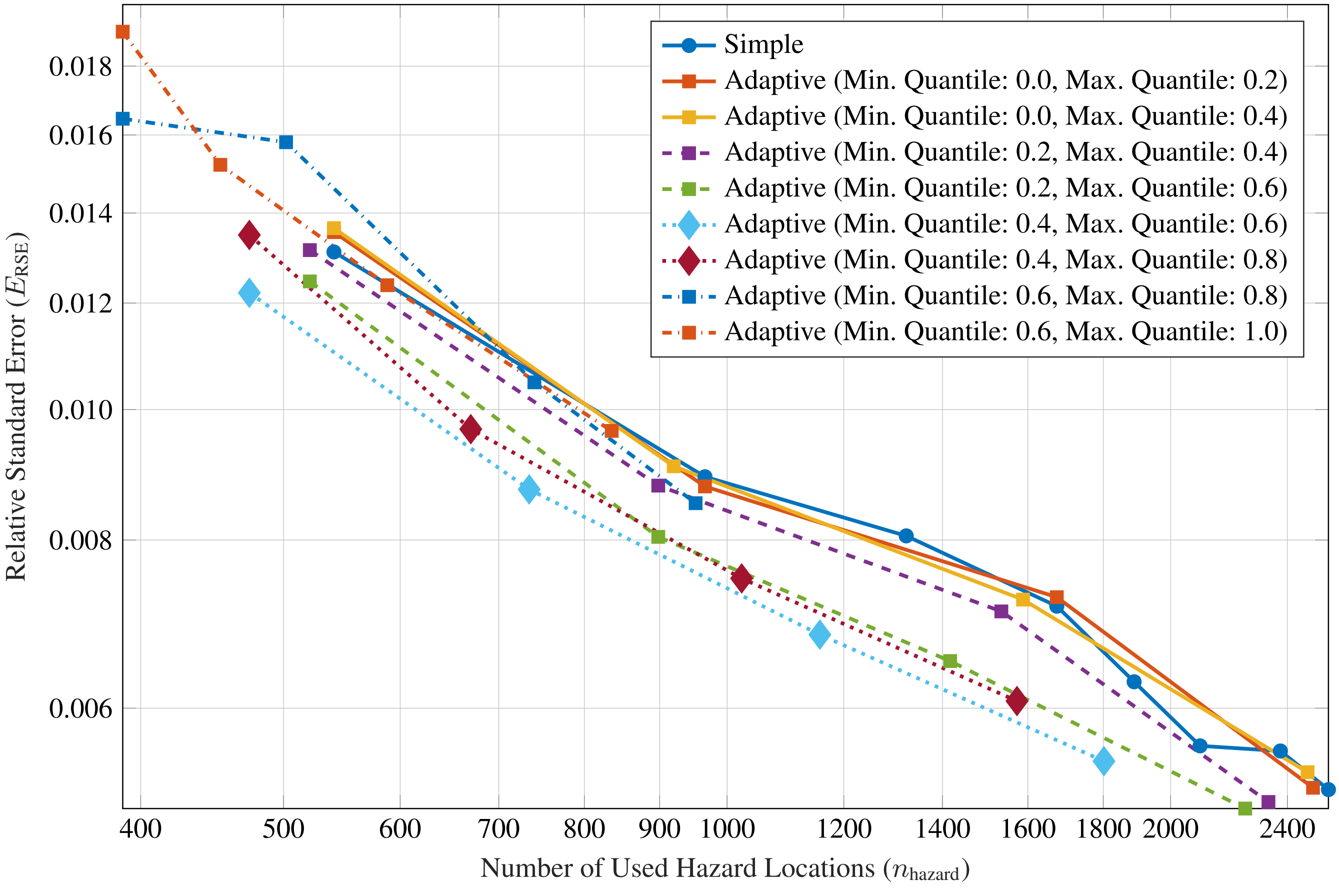

Figure 11. Results of a systematic parameter study with the goal of finding good values for the lower threshold (t_{l}) and upper threshold (t_{u}) parameters of Criterion I of the adaptive location uncertainty sampling scheme (see Section 4), based on the distribution of the Coefficient of Variation (CV) of loss rate in administrative zones. This shows a logarithmic plot of relative standard error (E_{RSE}) of Probable Maximum Loss (PML) at a return period of 100 years against the number of used hazard locations (n_{hazard}) for a portfolio of 20 risk items with 100% unknown coordinates. Color indicates different combinations for the threshold parameters t_{l} and t_{u}. Quantiles of the CV distribution around $t_{\mathrm{l}} \in [\mathrm{CV}_{0.2}; \mathrm{CV}_{0.4}]$ in combination with $t_{\mathrm{u}} \in [\mathrm{CV}_{0.6}; \mathrm{CV}_{0.8}]$ work best.

5.2.2 Variance Reduction and Speedup

For the same portfolios as analyzed in the previous section, Figure 14 shows logarithmic plots of the relative standard error E_{RSE} obtained from $R = 20$ repeated simulations against the number of used hazard locations n_{hazard}. Vertical bars depict upper 95% confidence intervals estimated using bootstrapping with 1000 resamples. Simple MC is again shown in blue, the variance reduction sampling scheme in red. While the observed error convergence order of the adaptive scheme remains the same as for simple MC (i.e. $\mathcal{O}(n^{-0.5})$, compare Section 5.1), the error curves are below those for simple MC for all portfolios.

https://doi.org/10.5194/nhess-2019-110
Preprint. Discussion started: 10 July 2019

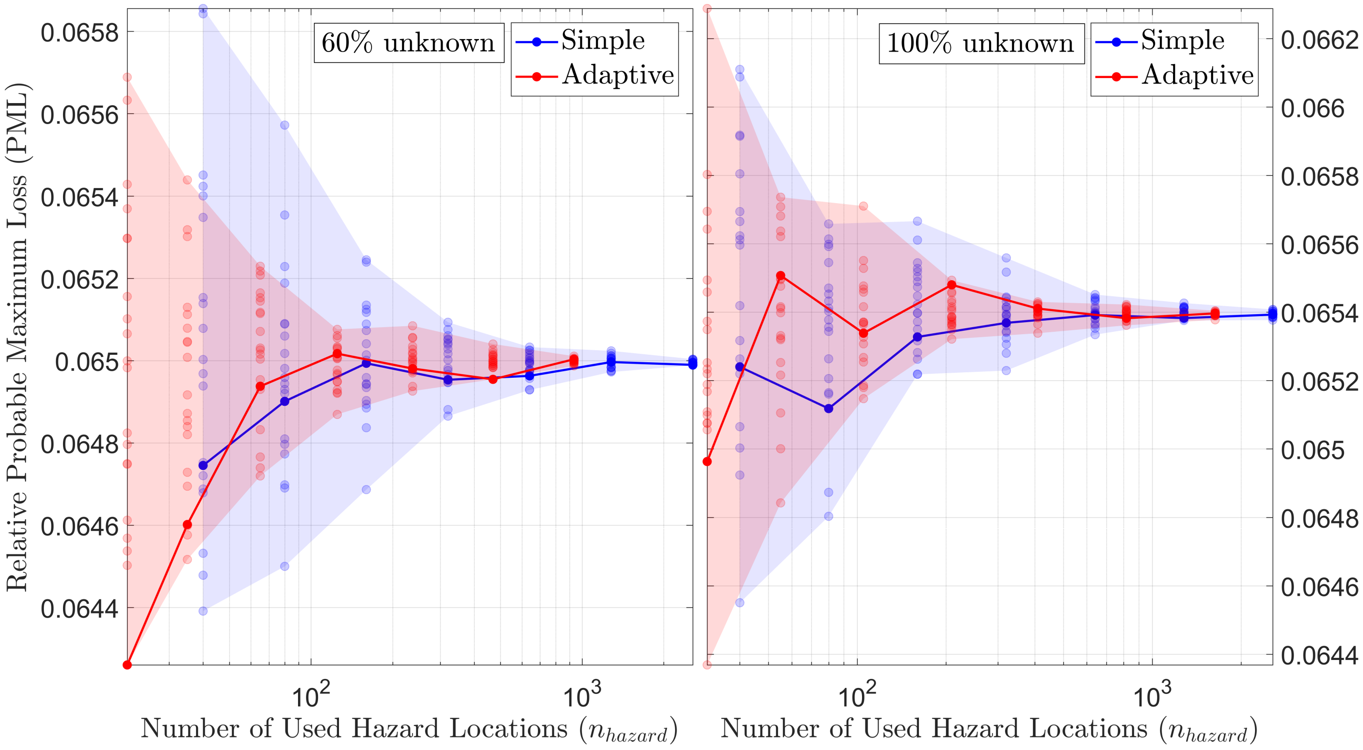

Figure 12. Convergence plots showing relative Probable Maximum Loss (PML) at a return period of 100 years against the number of used hazard locations n_{hazard} for portfolios of $n_{\text{r}} = 10$ risk items with 60% (left plot) and 100% (right plot) unknown coordinates using simple MC (shown in blue) as well as the adaptive scheme (shown in red). Semi-transparent circles depict $R = 20$ repeated simulations for each sample size, solid lines highlight one repetition. The transparently shaded background shows the entire range for each sampling scheme. The plots show that the adaptive scheme scatters less and converges faster to the same result as simple sampling.

The variance reduction quotient (VR, the ratio of the variances of the estimations obtained using simple MC and the adaptive scheme, see Equation 7) varies between portfolios with different number of risk items and fractions of unknown coordinates, but generally increases with growing n_{hazard}. For example, for the portfolio with 10 risk items and 60% unknown coordinates, VR is about 6.2 at $n_{\text{hazard}} = 10^2$ and increases to 13.2 at $n_{\text{hazard}} = 10^3$. For the portfolio with 10 risk items and 100% unknown coordinates, VR $\approx$ 1.8 at $n_{\text{hazard}} = 10^2$ and 2.2 at $n_{\text{hazard}} = 10^3$. For the portfolios with 100 risk items, the situation is similar. For 60% unknown coordinates, VR $\approx$ 2.4 at $n_{\text{hazard}} = 10^3$ and 3.7 at $n_{\text{hazard}} = 10^4$. For 100% unknown coordinates, VR $\approx$ 1.7 at $n_{\text{hazard}} = 10^3$ and 3.0 at $n_{\text{hazard}} = 10^4$.

The obtained variance reduction partially leads to a speedup of the computational runtime to reach a specific relative standard error level ε_{RSE}. Table 2 shows the speedup S of the scheme to reach relative standard error levels of $\varepsilon_{\text{RSE}} = 10^{-4}$ and $\varepsilon_{\text{RSE}} =$

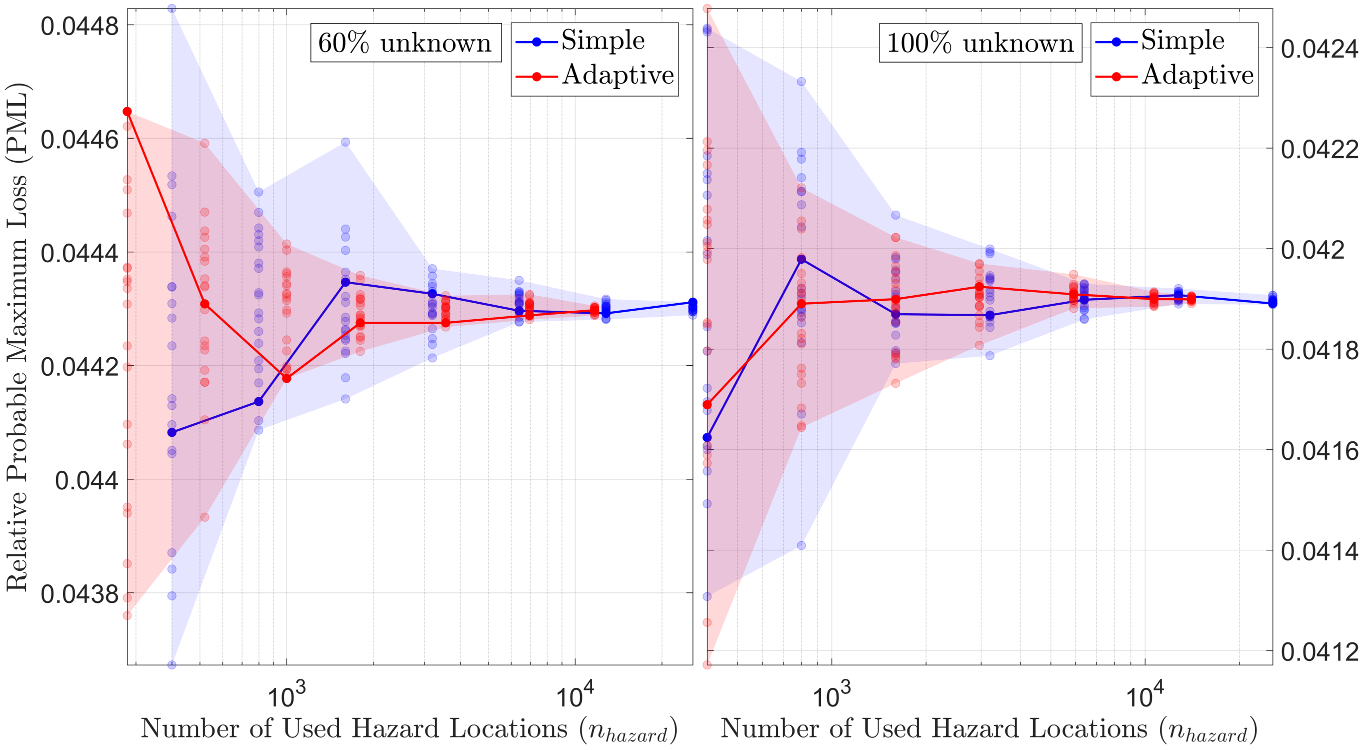

Figure 13. Convergence plots showing relative Probable Maximum Loss (PML) at a return period of 100 years against the number of used hazard locations n_{hazard} for portfolios of $n_{\text{r}} = 100$ risk items with 60% (left plot) and 100% (right plot) unknown coordinates using simple MC (shown in blue) as well as the adaptive scheme (shown in red). Semi-transparent circles depict $R = 20$ repeated simulations for each sample size, solid lines highlight one repetition. The transparently shaded background shows the entire range for each sampling scheme. The plots show that the adaptive scheme scatters less and converges faster to the same result as simple sampling.

10^{-5} for the same portfolios. Depending on the portfolio, the scheme achieves a speedup between 8% and 35% to reach $\varepsilon_{\text{RSE}} = 10^{-4}$, and between between 6% and 37% to reach $\varepsilon_{\text{RSE}} = 10^{-5}$. Note that we obtained these speedup values using a highly optimized seismic hazard and risk analysis framework. We suspect that the scheme can result in a significantly higher speedup for less optimized code, especially if the hazard simulation is not vectorized but contains a loop over locations.

5 6 Conclusions

In seismic risk assessment the exact location of risks is often unknown due to geocoding issues of address information. Therefore, in this paper we propose a novel adaptive sampling strategy to efficiently treat this location uncertainty using a seismic hazard and risk model for western Indonesia. The adaptive scheme considers three criteria to decide how often an unknown risk

Table 2. Mean runtime speedup and standard errors ($\mathrm{S} \pm E_{\mathrm{SE}}$) of the hazard computation achieved by the adaptive location uncertainty sampling scheme in comparison to simple sampling to obtain relative standard error levels of $\varepsilon_{\mathrm{RSE}} = 10^{-4}$ and $\varepsilon_{\mathrm{RSE}} = 10^{-5}$, estimated from $R = 20$ repeated simulations. Depending on the portfolio and $\varepsilon_{\mathrm{RSE}}$, the mean speedup ranges from 6% to 37%.

	Speedup (S)	
Portfolio	$\varepsilon_{\mathrm{RSE}} = 10^{-4}$	$\varepsilon_{\mathrm{RSE}} = 10^{-5}$
10 risk items, 60% unknown coordinates	1.24 ± 0.09	1.14 ± 0.04
10 risk items, 100% unknown coordinates	1.35 ± 0.06	1.37 ± 0.09
100 risk items, 60% unknown coordinates	1.08 ± 0.04	1.06 ± 0.03
100 risk items, 100% unknown coordinates	1.09 ± 0.03	1.08 ± 0.02

coordinate has to be sampled within a known administrative zone: (1) the loss rate variation within the zone, (2) the number of risks within the zone, and (3) the individual value of the risk. As the variation of hazard can vary quite strong not only between different administrative geographical zones, but also between different return periods, we use the spatial variation of loss rate which displays a similar pattern as the variation of hazard, but is independent of the return period. Furthermore, the total number of risks in the corresponding administrative zone, as well as the value (importance) of the risk with respect to the entire portfolio are considered by the adaptive scheme.

We investigated the performance of the scheme for a large range of sample sizes using different synthetic portfolios of different levels of unknown risk locations. We have found that the scheme successfully reduces the expected error, i.e. it reaches the same error levels as simple Monte Carlo with less samples of potential risk locations. This results in lower memory requirements and a moderate but appreciable runtime speedup to reach a desired level of reliability when computing loss frequency curves - a critical measure of risk in the insurance industry. The scheme could also be applied to other natural perils, such as probabilistic wind and flood models.

While the proposed scheme already successfully reduces the variance of loss frequency curve estimations, future improvements in the treatment of uncertainty in PSRA are conceivable. The computation might become yet more efficient by the application of variance reduction techniques to other uncertainties, for example in the ground motion and vulnerability models. Moreover, it would be essential to investigate the relative importance of location uncertainty in comparison to these other uncertainty types.

Competing interests. No competing interests are present.

Acknowledgements. This work was supported by a scholarship provided to the first author by Munich Re.

https://doi.org/10.5194/nhess-2019-110
Preprint. Discussion started: 10 July 2019

References

Bal, I. E., Bommer, J. J., Stafford, P. J., Crowley, H., and Pinho, R.: The influence of geographical resolution of urban exposure data in an earthquake loss model for Istanbul, Earthquake Spectra, 26, 619–634, https://doi.org/10.1193/1.3459127, 2010.

Bier, V. M. and Lin, S. W.: On the treatment of uncertainty and variability in making decisions about risk, Risk Analysis, 33, 1899–1907, https://doi.org/10.1111/risa.12071, 2013.

Cornell, C. A.: Engineering seismic risk analysis, Bulletin of the Seismological Society of America, 58, 1583–1606, 1968.

Cox, L. A.: Confronting deep uncertainties in risk analysis, Risk Analysis, 32, 1607–1629, https://doi.org/10.1111/j.1539-6924.2012.01792.x, 2012.

Crowley, H.: Earthquake Risk Assessment: Present Shortcomings and future directions, in: Geotechnical, Geological and Earthquake Engineering, vol. 34, pp. 515–532, https://doi.org/10.1007/978-3-319-07118-3_16, 2014.

Crowley, H., Pinho, R., Pagani, M., and Keller, N.: Assessing global earthquake risks: the Global Earthquake Model (GEM) initiative, in: Handbook of Seismic Risk Analysis and Management of Civil Infrastructure Systems, pp. 815–838, Woodhead Publishing Limited, https://doi.org/10.1533/9780857098986.5.815, 2013.

Dobson, J. E., Bright, E. A., Coleman, P. R., Durfee, R. C., and Worley, B. A.: LandScan: a global population database for estimating populations at risk, Photogrammetric engineering and remote sensing, 66, 849–857, 2000.

dos Santos, K. and Beck, A.: A benchmark study on intelligent sampling techniques in Monte Carlo simulation, Latin American Journal of Solids and Structures, 12, 624–648, https://doi.org/10.1590/1679-78251245, 2015.

Eads, L., Miranda, E., Krawinkler, H., and Lignos, D. G.: An efficient method for estimating the collapse risk of structures in seismic regions, Earthquake Engineering & Structural Dynamics, 42, 25–41, https://doi.org/10.1002/eqe.2191, 2013.

Efron, B.: Bootstrap methods: another look at the Jackknife, The Annals of Statistics, 7, 1–26, https://doi.org/10.1214/aos/1176344552, 1979.

Efron, B. and Tibshirani, R.: Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy., Statistical Science, 1, 54–75, 1986.

Foulser-Piggott, R., Bowman, G., and Hughes, M.: A framework for understanding uncertainty in seismic risk assessment, Risk Analysis, Advance Online Publication, https://doi.org/10.1111/risa.12919, 2017.

Gaughan, A. E., Stevens, F. R., Linard, C., Jia, P., and Tatem, A. J.: High resolution population distribution maps for Southeast Asia in 2010 and 2015, PLoS ONE, 8, e55 882, https://doi.org/10.1371/journal.pone.0055882, 2015.

Goda, K. and Ren, J.: Assessment of seismic loss dependence using copula, Risk Analysis, 30, 1076–1091, https://doi.org/10.1111/j.1539-6924.2010.01408.x, 2010.

Harding, B., Tremblay, C., and Cousineau, D.: Standard errors: A review and evaluation of standard error estimators using Monte Carlo simulations, The Quantitative Methods for Psychology, 10, 107–123, https://doi.org/10.20982/tqmp.10.2.p107, 2014.

Hayes, G. P., Wald, D. J., and Johnson, R. L.: Slab1.0: A three-dimensional model of global subduction zone geometries, Journal of Geophysical Research, 117, B01 302, https://doi.org/10.1029/2011JB008524, 2012.

Hess, S., Train, K. E., and Polak, J. W.: On the use of a Modified Latin Hypercube Sampling (MLHS) method in the estimation of a mixed logit model for vehicle choice, Transportation Research Part B: Methodological, 40, 147–163, https://doi.org/10.1016/j.trb.2004.10.005, 2006.

Jadach, S.: Foam: A general-purpose cellular Monte Carlo event generator, Computer Physics Communications, 152, 55–100, https://doi.org/10.1016/S0010-4655(02)00755-5, 2003.

Jayaram, N. and Baker, J. W.: Efficient sampling and data reduction techniques for probabilistic seismic lifeline risk assessment, Earthquake Engineering and Structural Dynamics, 39, 1109–1131, https://doi.org/10.1002/eqe.988, 2010.

Juneja, S. and Kalra, H.: Variance reduction techniques for pricing American options using function approximations, Journal of Computational Finance, 12, 79–101, https://doi.org/10.1.1.509.8171, 2009.

Knuth, D. E.: Big Omicron and big Omega and big Theta, ACM SIGACT News, 8, 18–24, https://doi.org/10.1145/1008328.1008329, 1976.

Landau, E.: Handbuch der Lehre von der Verteilung der Primzahlen, vol. 1, B. G. Teubner, Leipzig, 1909.

L'Ecuyer, P.: Good parameters and implementations for combined multiple recursive random number generators, Operations Research, 47, 159–164, https://doi.org/10.1287/opre.47.1.159, 1999.

MacKay, D. J. C.: Information theory, inference, and learning algorithms, vol. 100, Cambridge university press, https://doi.org/10.1198/jasa.2005.s54, 2005.

Mark Petersen, Stephen Harmsen, Charles Mueller, Kathleen Haller, James Dewey, Nicolas Luco, Anthony Crone, David Lidke, and Kenneth Rukstales: Documentation for the Southeast Asia seismic hazard maps, USGS Administrative Report, 2007.

McGuire, R. K.: Seismic hazard and risk analysis, Earthquake Engineering Research Institute, 1st edn., 2004.

Pagani, M., Monelli, D., Weatherill, G., Danciu, L., Crowley, H., Silva, V., Henshaw, P., Butler, L., Nastasi, M., Panzeri, L., Simionato, M., and Vigano, D.: OpenQuake engine: An open hazard (and risk) software for the Global Earthquake Model, Seismological Research Letters, 85, 692–702, https://doi.org/10.1785/0220130087, 2014.

Papageorgiou, A.: Sufficient conditions for fast quasi-Monte Carlo convergence, Journal of Complexity, 19, 332–351, https://doi.org/10.1016/S0885-064X(02)00004-3, 2003.

Press, W. and Farrar, G.: Recursive stratified sampling for multidimensional Monte Carlo integration, Computers in Physics, 4, 190–195, 1990.

Robert, C. P. and Casella, G.: Monte Carlo statistical methods, Springer Texts in Statistics, Springer New York, https://doi.org/10.1007/978-1-4757-4145-2, 2004.

Scheingraber, C. and Käser, M.: The Impact of Portfolio Location Uncertainty on Probabilistic Seismic Risk Analysis, Risk Analysis, 39, 695–712, https://doi.org/10.1111/risa.13176, 2019.

Senior Seismic Hazard Committee: Recommendations for probabilistic seismic hazard analysis: Guidance on uncertainty and use of experts, NUREG/CR-6372, 1997.

Tesfamariam, S., Sadiq, R., and Najjaran, H.: Decision making under uncertainty - an example for seismic risk management, Risk Analysis, 30, 78–94, https://doi.org/10.1111/j.1539-6924.2009.01331.x, 2010.

Tyagunov, S., Pittore, M., Wieland, M., Parolai, S., Bindi, D., Fleming, K., and Zschau, J.: Uncertainty and sensitivity analyses in seismic risk assessments on the example of Cologne, Germany, Natural Hazards and Earth System Sciences, 14, 1625–1640, https://doi.org/10.5194/nhess-14-1625-2014, 2014.

Wald, D. J. and Allen, T. I.: Topographic slope as a proxy for seismic site conditions and amplification, Bulletin of the Seismological Society of America, 97, 1379–1395, https://doi.org/10.1785/0120060267, 2007.

Yang, W.-N. and Nelson, B. L.: Using common random numbers and control variates in multiple-comparison procedures, Operations Research, 39, 583–591, 1991.

https://doi.org/10.5194/nhess-2019-110
Preprint. Discussion started: 10 July 2019

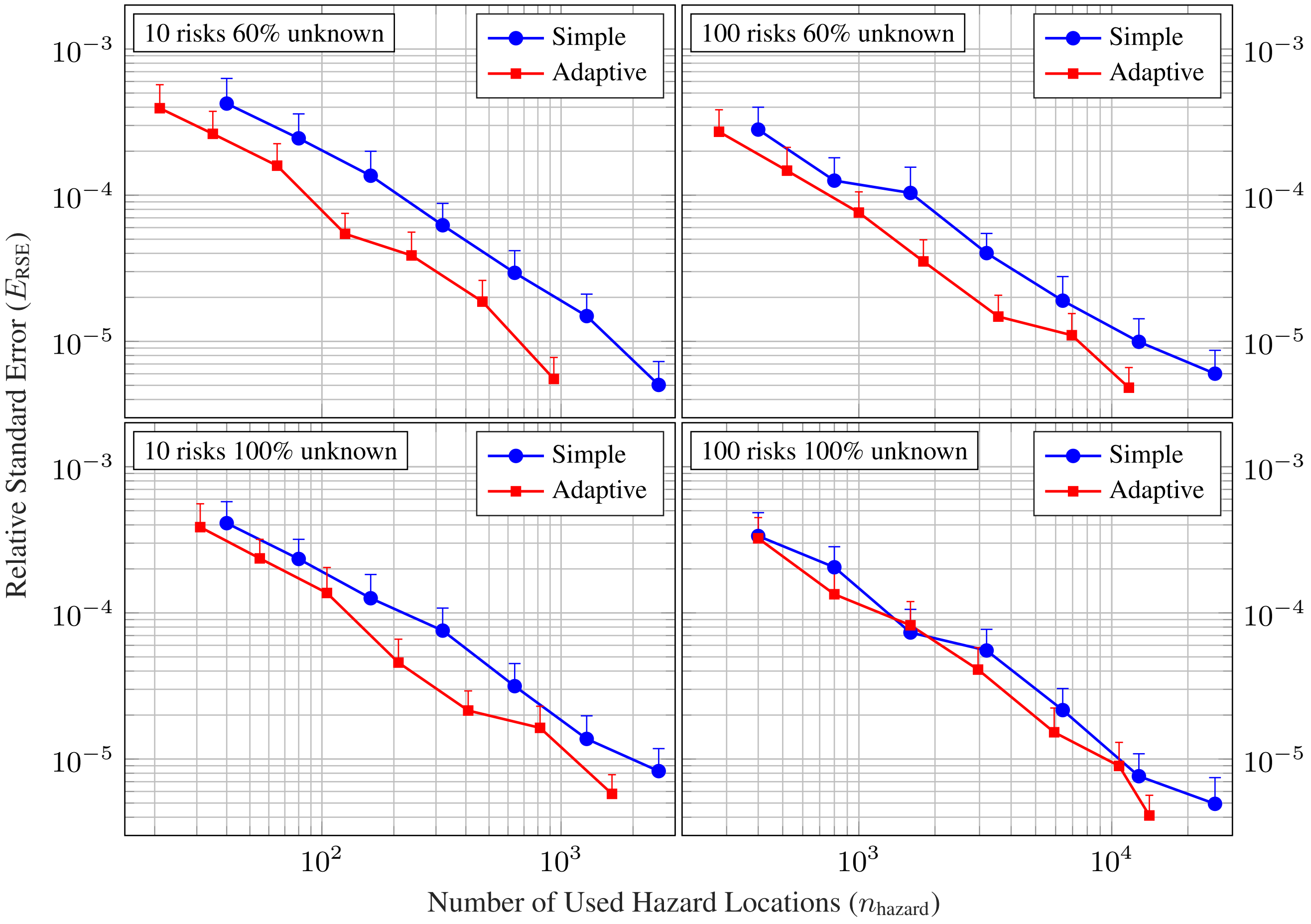

Figure 14. Logarithmic plot of relative standard errors E_{RSE} of Probable Maximum Loss (PML) at a return period of 100 years against the total number of used hazard locations n_{hazard} for different portfolios with $n_{\mathrm{r}} = 10$ (left plots) and $n_{\mathrm{r}} = 100$ (right plots) risk items and 60% (upper plots) and 100% (lower plots) unknown coordinates. Simple MC is shown in blue, the adaptive variance reduction scheme in red. All E_{RSE} have been obtained from $R = 20$ repeated simulations, vertical error bars depict upper 95% confidence intervals estimated using bootstrapping with 1000 resamples.