Infrastructure Recovery Curve Estimation Using Gaussian Process Regression on Expert Elicited Data

Quoc Dung Cao[a], Scott B. Miles[b], Youngjun Choe[a,*]

[a] *Department of Industrial & Systems Engineering, University of Washington, Seattle*
[b] *Department of Human Centered Design & Engineering, University of Washington, Seattle*

Abstract

The U.S. National Institute of Standards and Technology (NIST)'s Community Resilience Planning Guide uses recovery times of infrastructure functions as key metrics for disaster resilience. Although estimating the recovery times is critical to measuring and improving disaster resilience, this process remains challenging in the pre-event planning due to lack of historical data. To address this challenge, we consider a situation where infrastructure experts are asked to estimate the time for different infrastructure systems to recover to certain functionality levels after a scenario hazard event. We propose a methodological framework to use expert-elicited data to estimate the expected recovery curve of an infrastructure system. This framework uses the Gaussian process regression (GPR) to capture the experts' estimation-uncertainty and satisfy known physical constraints of recovery processes. The framework is designed to balance between the data collection cost of expert elicitation and the prediction accuracy of GPR. We evaluate the framework on simulated expert-elicited data concerning two case study events, the 1995 Great Hanshin-Awaji Earthquake and the 2011 Great East Japan Earthquake. It is shown that the framework is robust against different configurations such as the number of experts, how the quantities of interest are elicited, and uncertainty in the experts' estimates.

Keywords: Gaussian process regression, expert elicitation, infrastructure

*Corresponding author
Email addresses: milessb@uw.edu (Scott B. Miles), ychoe@uw.edu (Youngjun Choe)

restoration, disaster recovery, resilience planning

1. Introduction

This work is motivated by a recent trend of resilience planning initiatives. Our current ability to estimate infrastructure recovery trajectories is limited, as revealed in the recent resilience planning efforts of U.S. communities, which started in San Francisco, CA [1] and became state-wide initiatives in Washington State [2] and Oregon [3]. These efforts inspired the U.S. National Institute of Standards and Technology (NIST)'s Community Resilience Planning Guide [4] as a model for other jurisdictions. The current estimation practice is largely ad hoc. Although there is a growing body of literature on computational modeling of recovery [5, 6, 7, 8, 9], most models are often viewed as resource-intensive black-box approaches and not utilized by communities on the ground.

The NIST Guide defines time to recovery of function as "a measure of how long it takes before a building or infrastructure system is functioning" and "uses time to recovery of function as the primary metric for community resilience." This echoes the widely-recognized importance of characterizing disaster recovery for assessing community resilience [10, 11, 12, 13]. As the quote by Lord Kelvin says "if you cannot measure it, you cannot improve it," the first step to resilience improvement should be the reliable method to estimate the current system resilience. However, there are many challenges in this estimation exercise. There is usually not sufficient historical data on extreme events, in both magnitude and variation, that put infrastructure systems under challenge [14]. In addition, the lack of rigorous and sound estimation methods for recovery time impedes the measurable progress of resilience improvement.

The above needs and constraints are the major motivation for this work. This paper proposes a statistical framework to estimate infrastructure recovery curves (e.g., see Figure 1) for a hazard scenario using a combination of expert elicitation and Gaussian process regression (GPR). The result will facilitate the resilience planning process by providing estimations that reflect the domain ex-

perts' opinions on the current resilience of infrastructure systems. GPR model and experts' estimate complement each other to provide satisfactory solutions to this problem. Estimates gathered from experts will provide initial guidelines on how long it will take for a particular infrastructure to recover to some intermediate functionality levels. GPR will then use these estimates to predict the full recovery curve while capturing potential uncertainty in its prediction, as well as the uncertainty in the experts' estimates. GPR is also flexible enough to enforce important constraints on its predictions to allow the predicted curve to follow the physical behaviour of the actual recovery curve (e.g., monotonically increasing and bounded between 0 and 100%). The framework aims to be extensible to various types of infrastructure, while being intuitive and easy to be interpreted by the stakeholders.

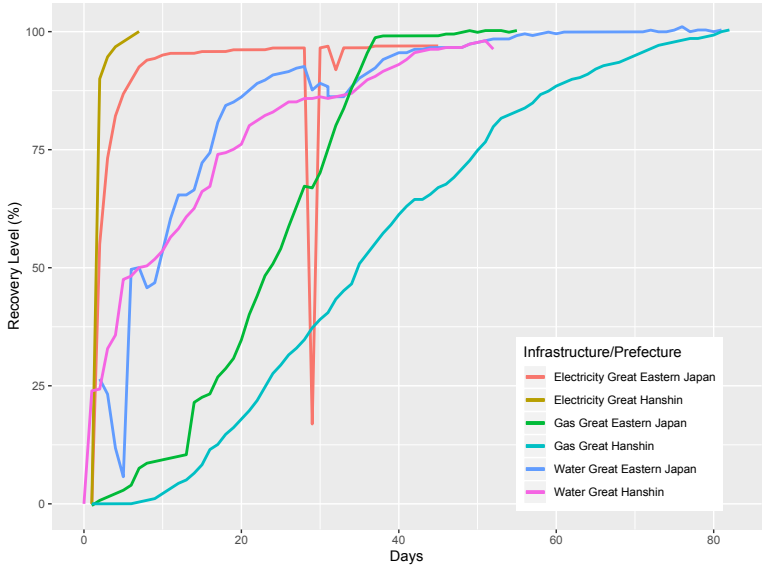


Figure 1: Empirical restoration curves of the 1995 Great Hanshin-Awaji Earthquake Disaster and the 2011 Great East Japan Earthquake Disaster.

While more data would generally yield a more accurate estimate, there is a practical limitation on collecting expert-elicited data. We study how to balance between the cost of collecting data from expert elicitation and the estimation

accuracy of GPR. We consider multiple expert elicitation schemes to identify the best way to estimate the recovery curve with a reasonable cognitive burden on experts while maintaining good estimation accuracy.

We simulate expert-elicited data by randomly generating expert estimates, which are assumed to be generally close to the empirical recovery curve observed in a case study event. We evaluate the proposed estimation method based on different empirical recovery curves from different prefectures and infrastructures after the 1995 Great Hanshin-Awaji Earthquake and the 2011 Great East Japan Earthquake [15]. We chose these events as they are the two major, extensively studied disasters that affected major infrastructure systems in Japan. Although they are different in quantities of damage, causes of damage, and spatial extents of damage, the recovery patterns are noteworthily similar. Their recovery data is also widely available, which facilitates our intensive modelling experiments and sensitivity analyses.

The rest of this paper is organized as follows. Section 2 briefly reviews relevant literature on expert elicitation and GPR. Section 3 presents the proposed estimation methodology. Section 4 shows the performance of this method through extensive numerical studies and sensitivity analyses. Section 5 draws insights for potential users of this method and concludes the paper.

2. Background

2.1. Expert elicitation for disaster recovery estimation

Participatory methods, especially expert elicitation, have been used extensively in disaster research especially in the areas where empirical data are scarce [16, 17]. The study in [18] elicits from experts infrastructure recovery estimates (at 0 hours, 72 hours, and 2 weeks from a hypothetical event) and qualitative inter-dependencies between those infrastructures. However, the study limits itself to short-term restoration and does not factor uncertainties into the recovery time estimation.

Expert elicitation itself is a well-established research domain [19, 20]. One of the most well-known elicitation approaches is the Delphi method [21, 22] characterized by its iterative, anonymous approach for developing consensus among experts. This method has been used widely in governments and industries [23, 24]. Another approach is the Cooke Classical Model [19, 25], also known as Cooke's method, which is one of the most established methods in expert elicitation literature. This method uses calibration questions, for which true values are known to the facilitator, to measure both accuracy and informativeness of an individual expert's judgement. These performance measurements, called calibration score and information score, respectively, are used as weights for aggregating multiple experts' judgements. Although developing calibration questions requires extra efforts, this performance-based weighting scheme has empirically proven effective [26] and represents the state-of-the-art among various weighting schemes [27, 28, 29]. In this paper, we propose to elicit data from the expert panel using both Delphi and Cooke's methods. The Delphi method is used to estimate a crucial quantity that needs a consensus across experts. The Cooke's method is used to aggregate recovery estimates across experts according to performance-based weights.

Although many studies elicit point estimates or probability distributions from experts, there are only a few studies on eliciting functions (e.g., recovery curve) from experts [30, 31, 32]. Arguably, the most systematic expert elicitation approach to functional estimation is developed in [33]. This study estimates seismic collapse fragility functions by eliciting quantiles of probability distributions, which encompass uncertainties of both seismic shaking intensity and resulting building collapse, from earthquake-engineering professionals. The reported estimates therein are created by first fitting lognormal distributions to the elicited probability estimates and then aggregating the distributions using Cooke's method. While this approach using the lognormal distribution (often used to model collapse fragilities) is defensible for this study, generalizing the approach to other functional estimation (especially recovery time estimation) has a major modelling drawback. Using a parametric distribution like lognormal is

too restrictive to reflect the uncertainties underlying the complex recovery processes being modeled, especially when we are using expert-elicited data, which will at least contain within-expert and across-expert uncertainties. Thus, this study uses GPR, which allows us to nonparametrically model recovery curves and the associated uncertainties.

Integration of expert judgements and empirical data is briefly mentioned in the NIST Guide [4], but no specific guideline is provided on the integration. The Oregon Resilience Plan [3] was the only resilience planning initiative that explicitly used both expert judgements and past event data, but the estimation process was still ad-hoc. Currently, to our best knowledge, there is no systematic statistical inference method being used for expert-based recovery time estimation in practice. This gap inspired us to develop the proposed method.

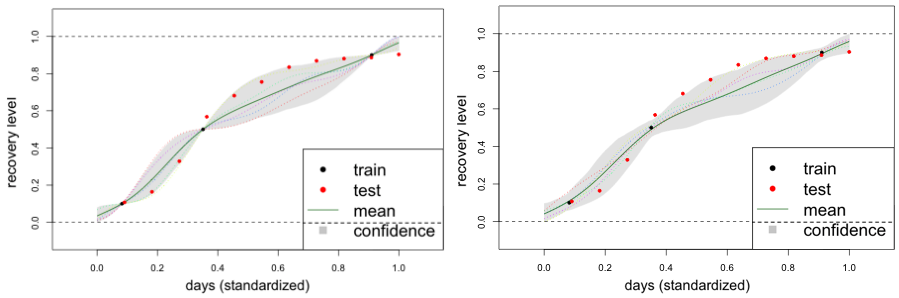
2.2. Gaussian process regression

Gaussian Process Regression (GPR) is a nonparametric model that offers the flexibility to model a stochastic process. It has been used successfully in many applications, such as engineering, physics, biology, economics, or other fields, in both regression and classification problems [34, 35, 36, 37]. In contrast to more parametric models where assumptions are more rigidly made about the data such as linear or polynomial regression, GPR specifies a prior distribution over function spaces, where the relationships over data are encoded in the covariance functions $k(x_1, x_2)$ of multivariate Gaussian distributions. The covariance between function values $f(x_1)$ and $f(x_2)$ only depends on the distance $\|x_1 - x_2\|$ between the data points instead of their vector coordinates. Once the input data is available, GPR can model the posterior over function spaces. The covariance will determine properties or constraints of the process, such as characteristic length scale, smoothness, or variance [38]. One commonly used covariance function is the squared exponential function,

$$k(x_1, x_2) = \sigma_f^2 \exp\left(\frac{-(x_1 - x_2)^2}{2l^2}\right),$$

where σ_f is the variance, which specifies how noise the data can be, and l is the characteristic length scale parameter, the higher of which will make the function smoother. In this study, we are estimating recovery curves, so we will focus on GPR where $x \in \mathbf{R}^1$ is the time for recovery to achieve a certain functionality level, and $f(x)$ is the functionality level.

Besides its low bias towards any functional form, GPR is also more suitable to our task than other parametric methods. It can capture both the uncertainty in the region where training data is not available and the variability in the training data itself. As a well-known issue in judgement-based forecasting, no matter how rigorous the elicitation process is, the results still depend on the experts' ability to estimate the quantity of interest. Because the expert estimates are noisy, GPR will capture the variability as an extra source of uncertainty during the inference step. Figure 2 shows two different ways to fit GPR to estimate a recovery curve, with or without noise in the training data. The grey bands show the 95% confidence interval to capture the uncertainty around the predicted curves.



(a) GPR with Noise-free observations (b) GPR with Noisy observations

Figure 2: GPR fitting with noise-free and noisy observation for Fukushima prefecture electricity recovery.

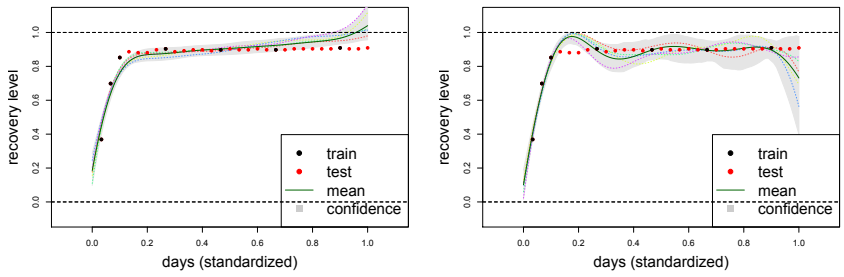
Due to the physical nature of the recovery curve, we also need to impose some constraints on the GPR model. First, the functionality level should be between 0% and 100%. Therefore, we will bound the prediction of the GPR model to be strictly between 0 and 1. Second, although it is possible that functionality

level may temporarily decrease in reality (e.g., due to an aftershock), it should generally increase over time. Hence, to capture this behaviour and reduce the prediction error, we also enforce that the curve is monotonically increasing with respect to time. Montonicity and boundedness are the linear inequality constraints actively researched in the GP framework [37, 39, 40, 41, 42]. In Figure 3, we show the effects of imposing only monotonicity, only [0,1] boundedness, and both constraints in the model for the Fukushima prefecture electricity recovery using the R package `lineqGPR` [41, 42]. It is helpful to have both constraints in the model. Otherwise, the model may behave in contrast to the expected physical behaviour of infrastructure recovery. In addition, the constraints will help to reduce the variance of the prediction. However, imposing these constraints may be potentially too rigid to capture the flat region near 0% and 90% of recovery. We can alleviate this issue by eliciting the boundary points so that the GPR is only interpolating between the elicited data points. Furthermore, the recovery curve can be constructed up to a functional level below 100% (e.g., 90%) as suggested by the NIST Guide [4]. We will apply these measures in Section 4 for the numerical studies.

3. Method

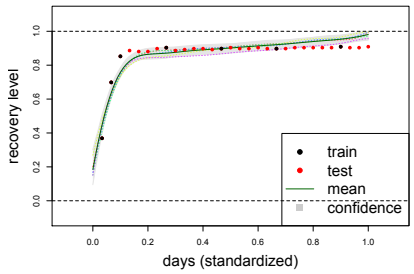
3.1. Consideration in recovery curve estimation

Our goal in this study is to estimate the infrastructure recovery curve from point estimates given by experts using GPR. The two steps (i.e., expert elicitation and GPR) are not designed independently. We carefully design the whole framework considering the logistical, computational, and theoretical constraints of both steps. The curve is characterized by two dimensions, namely, the recovery level measured in percentage (100% means the system is fully functional) and the recovery time measured in either days or hours from the disruption. GPR, similarly to other regression methods, is more suitable for interpolation between training data (as opposed to extrapolation). To achieve better performance, it is desired for the expert-elicited data to possess two properties. First, it should



(a) GPR with monotonicity constraint only.

(b) GPR with boundedness constraint only.



(c) GPR with both monotonicity and boundedness constraints.

Figure 3: Different GPR constraints for the Fukushima prefecture electricity event.

be as evenly distributed as possible so that the interpolated prediction does not exhibit too much uncertainty. Second, it should cover the boundary values to avoid predicting values beyond the range of the given data. It may seem best from a statistical perspective to elicit the data in both the range of recovery level (e.g 10%, 30%, 50%, 70%, 90%) and recovery time (e.g $2.5D/10$, $5D/10$, $7.5D/10$, $9D/10$), where D is the (estimated) earliest time for the infrastructure to recover to 100% or another high functional level (e.g. 90%) depending on what kind of recovery curve we want to construct, to follow the NIST Planning Guide [4]) given by the expert. However, it may not be very intuitive to elicit recovery levels at some certain time, e.g. "What is the estimated recovery level

at day 11 after the event?". Therefore, in our proposed method below and the numerical studies in Section 4, we only elicit the recovery time at some certain recovery levels. This choice is also consistent with the NIST Planning Guide [4].

In addition to eliciting recovery times at different recovery levels, we also want to elicit D, as introduced above, so that we can normalize the recovery time to be in the range of $[0, 1]$ for the following reasons:

1. The constructed curve could be more generalizable to future disaster events. If we face another similar events in the future, where similarity is defined by dominant characteristics of the events (e.g., Richter magnitude and earthquake resilience of the area, or Saffir–Simpson scale and hurricane resilience of the area), we can significantly reduce the elicitation effort by either using the existing recovery curves or simply eliciting the earliest full recovery time D and scaling the recovery time based on the particular D values of new events.

2. It is easier to compare different recovery curves of different natures on the same scale in the range of $[0, 1]$.

3. We can offer some insights from the shape or pattern (e.g., for hurricane category 1 vs. 5; magnitude 6 vs. 8; power vs. water; urban vs. rural) of recovery, which has formed consensus across many communities, so that other communities lacking the opportunity/resource to conduct extensive elicitation procedure can still use these curves as references of possible recovery trends.

4. In terms of GPR modelling, we want to have both axes in the process to be between $[0, 1]$ in the fitting and inference following the implementation in [41, 42]. The actual unit of recovery time can be easily scaled back to days or hours after the inference procedure.

3.2. Challenges in expert data elicitation and modelling

From the design considerations above, we anticipate some challenges in the elicitation process as follows:

1. Obtaining the earliest time to full recovery D: D should be universal across all experts. One way to obtain this is to have an open discussion among experts until they reach a general consensus on how long D should be. Another way is to employ point-based expert elicitation methods, for example in [43], to estimate the probability distribution of D. Another possibility is to use the individual expert's D value to normalize their own recovery time estimate.

2. Obtaining the input noise level σ: This is required for the statistical modelling process. This can be interpreted as how uncertain the experts' estimates are. We may gather the data and estimate the uncertainty based on their data after the elicitation process. This σ will account for both within-expert and across-expert uncertainty. The GPR framework assumes that one type of noise is present in the data, which accounts for all the uncertainty, and that the noise level is constant across all levels of input. In case we want to decompose the uncertainty further, it is more straightforward to estimate the across-expert uncertainty since we have different expert data at each recovery level. However, within-expert uncertainty estimation is tricky. One way to estimate it is through Cooke's method where calibration questions are used to measure the inherent estimation uncertainty. However, one can challenge the underlying assumption that the estimated uncertainty based on calibrating questions remains the same as the uncertainty for main questions. Regardless, it is reasonable to assume that within-expert variability is negligible compared with across-expert variability.

3. Obtaining more elicited data: While we may drive down the estimation uncertainty by collecting more data, this would impose more logistical

burden to the experts. In addition, the experts may have some cognitive difficulty to distinguish between smaller difference in recovery levels, e.g 10% and 20%.

On the other hand, there are also some challenges in the modelling and inference process:

1. If we ask each expert to give an estimate of D, it is challenging to determine which D to use and how to normalize the time.

2. If we have noise/uncertainty in both dimensions (input and output), it does not follow the conventional GPR framework, in which $y = f(x) + \epsilon$, where ϵ follows $N(0, \sigma^2)$. To further elaborate this point, in the GPR framework, we assume that the input is fixed, i.e. if we want to predict the recovery time at each functionality level, we fix the functionality level and the prediction of recovery time will exhibit some level of uncertainty. This is consistent with our experiment implementation in Section 4.

3.3. Recovery curve estimation framework

In this section, we present a few potential elicitation schemes for consideration. Each scheme has its own advantage and disadvantage. In terms of workshop design, we can adopt the Cooke Classical Model [19] to perform elicitation of expert judgments. In this work, experts are mathematically defined as those who can provide informed estimates of infrastructure recovery times such that the across-expert mean estimates are arbitrarily close to the true recovery curve when a sufficiently large number of experts are elicited. In practice, experts may include but not limited to utility operators, emergency managers, and infrastructure researchers. There are several ways to aggregate experts' estimates, such as linear pooling or performance based weighting. Linear pooling, although with its least logistical cost of designing calibration questionnaire, is shown to under-perform other performance based methods [26]. Although questionnaire design is beyond the scope of this work, we outline one way to perform calibration on the expert judgments following the performance based

weighting methods. There should be a set of calibration questions, which is closely related to the quantity of interest we are trying to estimate. An example question could be, given a functionality level, what is the estimated time that the expert thinks an infrastructure can take to achieve. An expert will be asked to give different quantile estimates on the quantity, and they form their subjective probability mass about such quantity. Under the Cooke Classical Model, there will be two types of scores being generated from this calibration exercise. The first is an information score, or how confident an expert is about her estimates. The second score is a calibration score, which is the likelihood that her judgement corresponds to the actual results. A product of the two quantities can be used as a general score to determine the performance weight, which is then used to take the weighted average of experts' estimates. To further optimize for performance, we can vary the selection threshold, below which will render an expert's weight to 0, to get the best performance metric on the calibration questions. Then, that set of optimized weights can be used to elicit the quantity of interest.

Scheme 1: Maximum elicitation on two dimensions.

1. Ask each expert for the earliest time to full recovery D, recovery times at fixed functionality levels (10%, 30%, 50%, 70%, 90%), and functional levels at fixed recovery times (e.g $2.5D/10$, $5D/10$, $7.5D/10$, $9D/10$).

2. Use the sample mean/median (across experts) of all elicited data as the training data, with Cooke's method weighting if necessary.

3. Use all estimates (across experts) at each level to estimate the noise level.

4. Fit GPR and construct a recovery curve with its estimation uncertainty.

Advantage: Full range of data over both dimensions. Impose less burden in elicitation logistics than scheme 2. Elicitation can finish in one stage.

Disadvantage: Uncertainty in D estimation can lead to erroneous and high uncertainty in prediction. Furthermore, as mentioned above, the GPR frame-

work assumes one dimension as fixed input. Eliciting in both dimensions violates this assumption.

Scheme 2: Two-stage elicitation. The earliest full recovery time D will be iteratively discussed among the experts until reaching consensus.

Stage 1:

1. Ask each expert for the earliest time to full recovery D.
2. Show all the experts the (range of) elicited D values.
3. Ask experts to revise their D estimate until reaching agreement.

Stage 2:

4. Ask each expert for a full range of fixed recovery level (10%, 30%, 50%, 70%, 90%) and recovery time (e.g $2.5D/10$, $5D/10$, $7.5D/10$, $9D/10$), with Cooke's method weighting if necessary.
5. Use the mean/median of across-expert estimates as the training data.
6. Use all elicited data to estimate the noise level.
7. Fit GPR and construct a recovery curve with its estimation uncertainty.

Advantage: Full range of data over both dimensions. Reduce uncertainty in the earliest full recovery time D.

Disadvantage: Two-stage elicitation will require more effort from the expert.

Scheme 3: Using either scheme 1 or scheme 2 but with smaller elicited data (e.g., 3 points for each dimension)

Advantage: Less burden on the expert.

Disadvantage: May result in a sub-optimal fit and prediction.

Scheme 4: Elicitation on only one dimension, fixing recovery levels and ask for recovery times.

1. Obtain estimate of D, following either scheme 1 or scheme 2. (The numerical studies in Section 4 will use scheme 4 with each expert's estimate

of recovery time being normalized by her own estimate of D.) In case it is hard to reach a consensus D among the experts, the recovery time estimates at each functionality level can be aggregated (e.g., via equal-weighting) across the experts first, and normalized by the maximum value of D.

2. (OPTIONAL) Repeat the process for the 3 scenarios (worst, best, most likely)

Advantage: Straightforward in modelling. Simple to interpret and implement.

Disadvantage: Data may be sparse. In some events (e.g the Fukushima electricity recovery in Section 4), the recovery is expected to be very fast in the first few hours. The expert may say the recovery is up to 70% in the first day and 90% the next day. In this case, the GPR model may not provide much additional values to stakeholders in recovery planning.

Scheme 1 will speed up the elicitation process since we can elicit on both dimensions. However, the question is whether we need to elicit in both ways (fix the level then elicit the time, and fix the time then elicit the level). Scheme 2 is almost identical to scheme 1, except with the elicitation of the earliest full recovery time D to reach either 100% or 90% to normalize the time axis. Scheme 3 is a less resource-demanding version of Scheme 1 and 2. In Section 4, we will study the optimal number of elicitation levels through sensitivity analysis. Although we can try to elicit in both ways, to be consistent with the GPR framework, we can only use one dimension (either recovery time or recovery level) as input and predict the remaining dimension. Instead of spending experts' resources on eliciting in both ways, we can use their effort to elicit more recovery time at higher granularity of functionality level or elicit more scenarios (best, worst, most likely). In view of the above considerations, we will demonstrate the framework of Scheme 4 in the numerical studies in Section 4.

4. Numerical Studies

To demonstrate the performance of the framework, we evaluate it on different empirical recovery curves from different prefectures and infrastructures after the 2011 Great East Japan Earthquake and the 1995 Great Hanshin-Awaji Earthquake. The framework is designed to be applied where an expert elicitation workshop is run in conjunction with statistical modelling. For demonstration purpose in this paper, we will simulate the expert opinion. Assuming that the experts are capable of estimating the true recovery curve with a reasonable accuracy, we use the entire available empirical data (such as those in Figure 1) to fit the polynomial regression model as a surrogate to the expert opinion. The simulated expert can be queried for recovery time given a functionality level and vice versa. In Scheme 4, we provide a functionality level as an input to the simulated expert and obtain the recovery time estimate as the output. Each expert can be modelled using Eq. (1):

$$days = f_{poly}(recovery) + \epsilon_{days}, \tag{1}$$

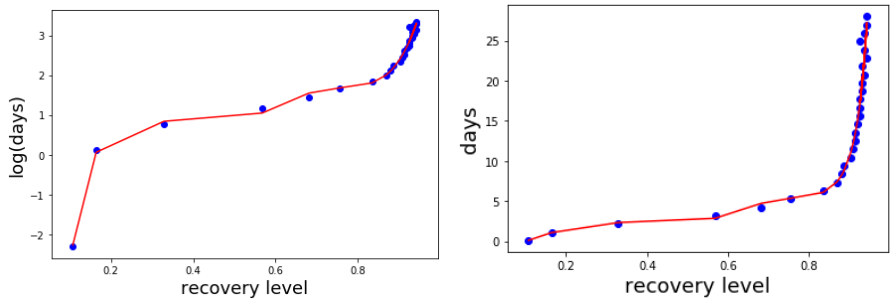
where $days$ is the number of days from the beginning of disruption (e.g., earthquake, hurricane landfall), $recovery$ is the functionality level of the system that is recovering, f_{poly} is the polynomial regression function, and ϵ_{days} is assumed to follow a normal distribution with mean 0 and variance σ^2 that captures the estimation variability.

However, it is better for the model to utilize the fact that the output (i.e., recovery time) is always positive by taking log transformation on the output variable. Thus, the expert model in Eq. (1) becomes

$$\log(days) = g_{poly}(recovery) + \epsilon_{log(days)} \tag{2}$$

The fitted polynomial regression function g_{poly} is demonstrated in Figure 4.

Furthermore, it is desirable to model two distinct sources of the estimation variability $\epsilon_{log(days)}$. Thus, a layer of Gaussian noise ϵ_1 is added to the output to model within-expert variability. A second layer of Gaussian noise ϵ_2 is then



(a) Polynomial fit in the log scale of the days/recover time

(b) Predicting recovery time in the original scale

Figure 4: Expert simulated model is built based on all the available data of a past event using polynomial regression, which will allow the sampling from the curve will be very close to the actual values. The model represents an average prediction across multiple replications and multiple expert. In other words, if we have infinite amount of experts, we assume that their average prediction will converge to the fitting curve or the actual data.

added to model across-expert variability. The two noise terms are additive in the log-transformed model because we model the errors to be multiplicative in the original scale. The log transformation will then make the multiplicative errors become additive, to be consistent with the polynomial regression framework. The multiplicative errors are intuitive. For example, consider a scenario event that makes the recovery estimation challenging for all experts (i.e., high across-expert variability, $Var(\epsilon_2)$). Then, the individual expert's large uncertainty perhaps due to lack of experience (i.e., high within-expert variability, $Var(\epsilon_1)$) will *amplify* the effect of the challenging estimation problem, thus resulting in highly variable recovery time estimates. In summary, the elicited recovery time estimates are simulated using

$$days_{simulated} = \exp(g_{poly}(recovery) + \epsilon_1 + \epsilon_2). \tag{3}$$

As an implementation note, due to the randomness from ϵ_1 and ϵ_2, sometimes the sequence of simulated expert's estimates could be non-monotonic. How

likely it happens depends on the variance of the errors. Since we assume the experts are only providing estimates for a monotonic recovery curve (i.e., no deterioration of infrastructure functionality in the midst of recovery, for example, due to aftershocks), they will only provide monotonically increasing recovery time estimates with respect to the functionality level. In our simulation, to ensure that the simulated recovery estimates satisfy this assumption, we reject the non-monotonic estimate paths until a monotonic sequence is generated.

Using the simulated data, the GPR model with monotonicity and boundedness is fit as follows:

$$recovery = GPR(days_{simulated}). \tag{4}$$

The GPR models in the numerical studies use a squared exponential kernel, with variance parameter $\sigma_f^2 = 0.1$ and characteristic length scale $l = 1$. These choices of hyperparameters are based on the physical nature of the problem and the workshop. The variance is picked to model reasonable noise in the experts' estimate, which could be estimated empirically during the elicitation workshop. The length scale is chosen for a smooth recovery process, which is generally the case for infrastructure recovery. Since the purpose of the model is to provide resilience estimate when historical data is not available, we did not perform cross validation to optimize for the hyperparameters or kernel functions as our modelling framework is less intended for a forecasting or predictive exercise. Figure 5 shows the performance of this modeling framework on the same infrastructure sector in different prefectures (electricity recovery of Miyagi, Fukushima, and Iwate) and Figure 6 shows different infrastructures within the same prefecture (water and gas recovery in Great Hanshin). We simulate the process of eliciting from 5 experts, asking for recovery time at 10%, 30%, 50%, 70%, 90% functionality levels, with within-expert and across-expert noise variance to be $Var(\epsilon_1) = Var(\epsilon_2) = 0.1$. We then take the average of their estimates with equal weighting, and construct the GPR curve.

It is observed that the method is very flexible. In Fukushima electricity recovery, although the actual recovery started at about 40% in day 1, we can

still capture the rest of the recovery curve simply by eliciting from 30% onward. This translates to some freedom to the experts in actual workshops. They can skip some levels if they think it does not make sense to estimate when they think the recovery actually will happen quickly initially.

In Figure 6, it may seem that the model does not capture the initial recovery stage (e.g., below 10% functionality) very well. This often happens with infrastructures whose recovery tends to follow others, such as gas, which is usually recovered after electricity and water. The model still captures the majority of the recovery curve (between 10% and 90%) quite well with high confidence.

We also investigate how sensitive the estimation framework is to the number of experts by monitoring the root mean square error (RMSE) of prediction on the available test data (different from the recovery levels elicited from the experts). We first perform simulation to measure the performance in terms of RMSE of the framework with 1, 3, 5, 7, 9, 11 experts based on Miyagi electricity recovery to see if there is an "elbow" of performance change point to balance the logistics of elicitation and accuracy, as shown in Figure 7. In Table 1, we vary the number of simulated experts to be 3, 5, or 10. Given a fixed noise level within and across experts, it seems that the result is quite stable with 5 experts. We acknowledge that in this simulation, all the experts are modelled to exhibit the same level of uncertainty, which is not realistic in practice. In fact, in usability testing experiments in [44], where the participatory performance, involving both expert and novice users, is measured in a group of 5 and beyond, the study shows that some randomly selected group of 5 participants can perform relatively well although the risk is that the performance variance is high. However, in actual workshops, there could be more than 5 experts (among whom, the expertise level is theoretically more consistent than the study in [44]), and as long as their opinions converge to some underlying quantity, the estimation still can provide a reasonable recovery curve.

We conduct another analysis to measure how the framework performs with different levels of elicitation. Our initial hypothesis is that performance will improve as we elicit more data, which may increase more logistical burden to

the expert. The hypothesis is generally confirmed from Figure 8. It also does not penalize performance very much to have custom spacing of levels, so we can focus more on asking the experts at more intuitive recovery levels.

Table 1: Sensitivity analysis on the framework performance to the number of experts. In this table, the experts are simulated to have equal weights to their estimate, and the simulated noise variance in Eq. (3) is $Var(\epsilon_1) = Var(\epsilon_2) = 0.1$. Note that the unit for RMSE is the fraction of recovery level. The RMSE presented is the average across 100 simulation replications.

Prefacture/ Infrastructure	Number of Experts	RMSE
Fukushima electricity	3	0.0637
	5	0.0567
	10	0.0524
Miyagi electricity	3	0.0405
	5	0.0340
	10	0.0297
Iwate electricity	3	0.0457
	5	0.0398
	10	0.0373
Great Hanshin water	3	0.0447
	5	0.0352
	10	0.0334
Great Hanshin gas	3	0.0542
	5	0.0516
	10	0.0497

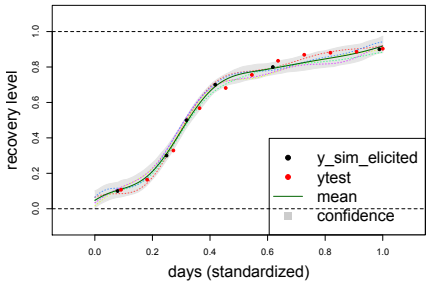
5. Conclusion

We demonstrated in this research a framework to assist the community resilience planning through estimating potential infrastructure recovery curves. The framework combines experts' opinions and Gaussian process regression to unify domain knowledge and uncertainty quantification in the estimated curves.

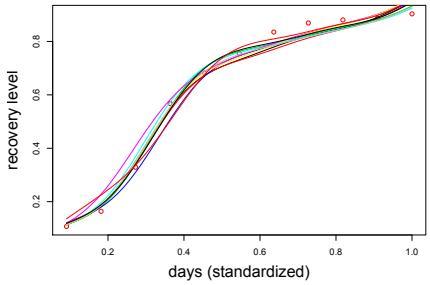
Table 2: Sensitivity analysis on the framework performance to the uncertainty in expert estimation ($Var(\epsilon_1), Var(\epsilon_2)$ in Eq. (3). In this table, data is simulated from 5 experts for 100 simulation replications.

Prefacture/ Infrastructure	$Var(\epsilon_1), Var(\epsilon_2)$	RMSE
Fukushima electricity	0.1	0.0567
	0.3	0.0600
	0.5	0.0985
Miyagi electricity	0.1	0.0340
	0.3	0.0583
	0.5	0.0811
Iwate electricity	0.1	0.0398
	0.3	0.0549
	0.5	0.0686
Great Hanshin water	0.1	0.0352
	0.3	0.0427
	0.5	0.0546
Great Hanshin gas	0.1	0.0484
	0.3	0.0636
	0.5	0.0916

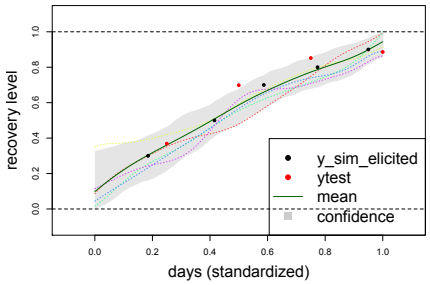
We performed extensive sensitivity analyses to draw insights into various elicitation schemes and the effects of number of experts, number of elicited points, and elicitation levels on the predictive performance. Although the framework was developed for modeling post-event infrastructure recovery, it can be generalized to other recovery modeling, such as for different capitals and services that are important for community resilience [45]. We do not explicitly consider dependencies between infrastructures in this study. Future work may directly model their dependencies to improve the predictive performance and/or reduce the reliance on expert estimates.



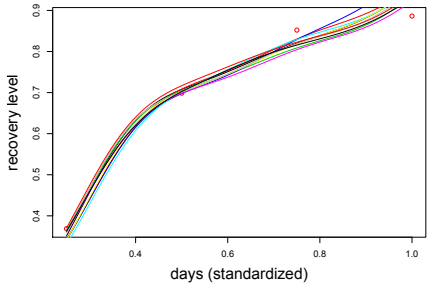
(a) Miyagi electricity recovery curve with 95% confidence interval



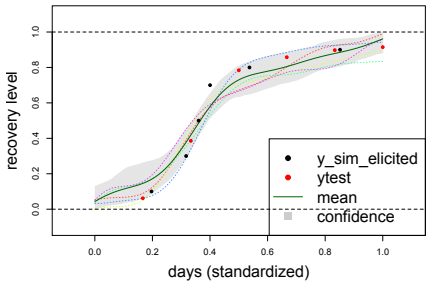
(b) Miyagi electricity recovery curve with 10 mean predictions



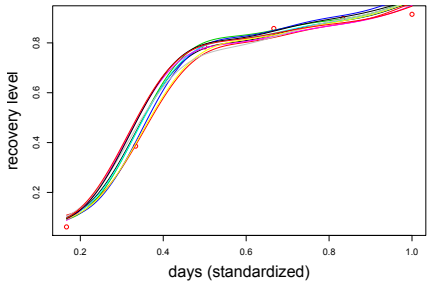
(c) Fukushima electricity recovery curve with 95% confidence interval



(d) Fukushima electricity recovery curve with 10 mean predictions

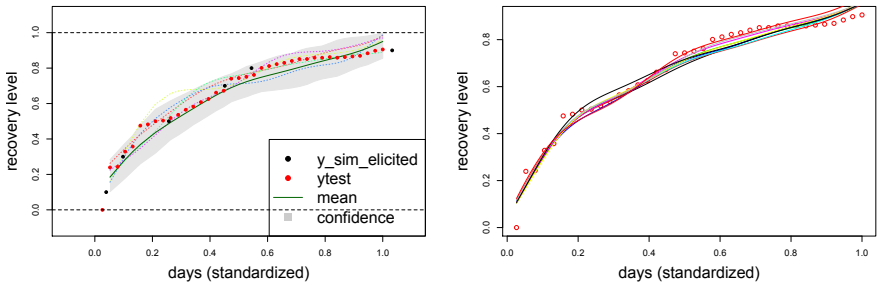


(e) Iwate electricity recovery curve with 95% confidence interval



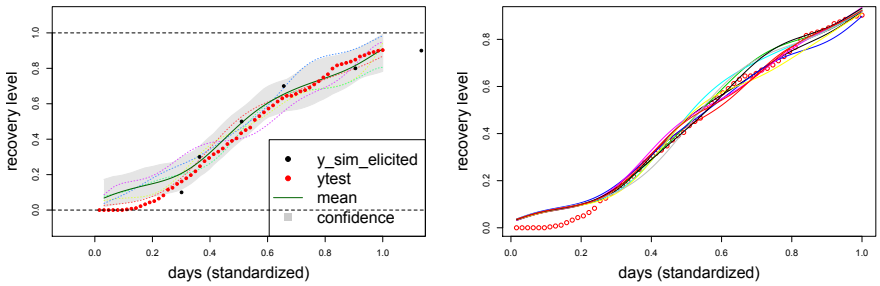
(f) Iwate electricity recovery curve with 10 mean predictions

Figure 5: Numerical results on different prefectures (Miyagi, Fukushima, and Iwate). The figures on the left column show the result of GPR model built on one simulated draw of expert opinion. The grey bands show the 95% confidence interval to capture the uncertainty around the predicted curves. The figures on the right column show different mean predictions based on different simulated draws of expert opinion. In all cases, we simulate the process of elicitation from 5 experts.



(a) Great Hanshin water recovery curve with 95% confidence interval

(b) Great Hanshin water recovery curve with 10 mean predictions



(c) Great Hanshin gas recovery curve with 95% confidence interval

(d) Great Hanshin gas recovery curve with 10 mean predictions

Figure 6: Numerical results on water supply and natural gas infrastructures in Great Hanshin. The data simulates the process of elicitation from 5 experts.

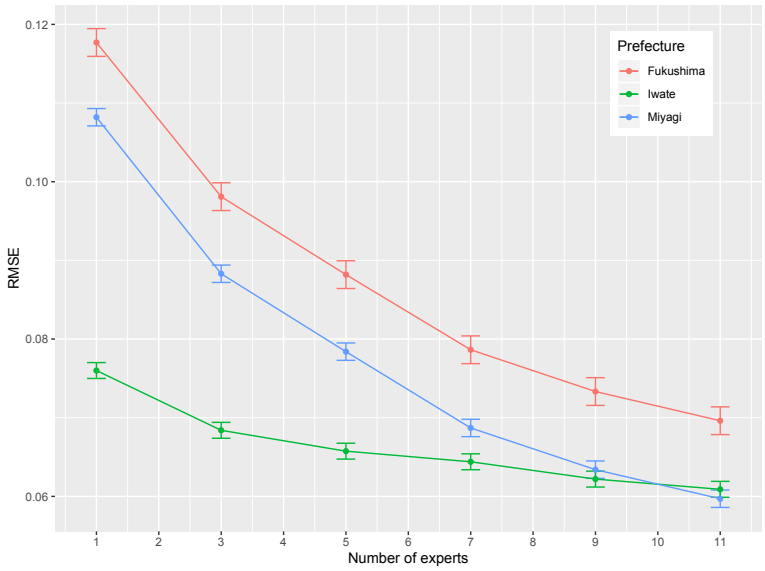


Figure 7: The plot shows the performance of the framework for electricity recovery at Fukushima, Miyagi, and Iwate prefactures as a function of the number of experts. The error bar at each number of experts shows the 95% confidence interval of test RMSE in 100 simulation replications. Although the more number of experts involved in the elicitation process results in better performance, it is observed that there is a diminishing marginal return as the number of experts increases in 2 out of 3 prefactures. The rate of performance gain is fastest when engage from 1 to 3 experts. The rate is slower from 3 to 7 experts. It drops to the slowest rate if we increase from 7 to 11 experts.

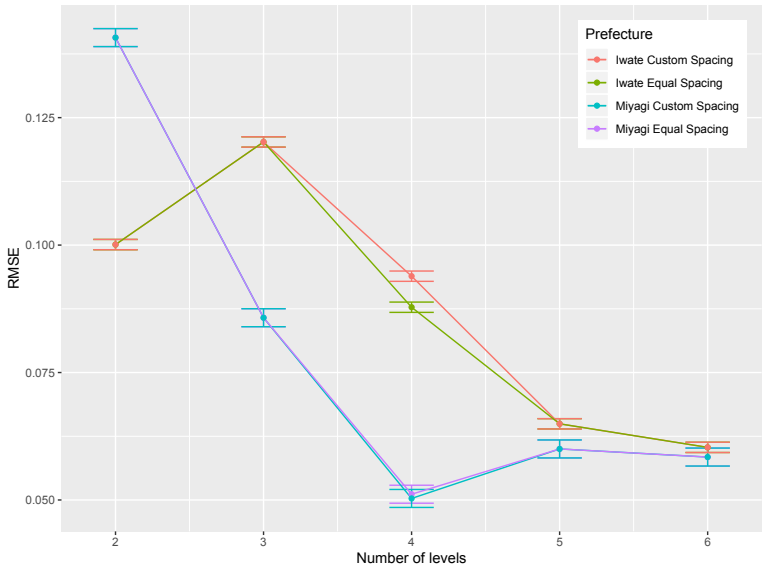


Figure 8: The plot shows the performance of the framework in terms of test RMSE in 100 simulation replications for electricity recovery at Miyagi and Iwate prefectures as a function of the number of elicitation levels. We evaluate the performance when eliciting 2, 3, 4, 5, 6 levels from the experts. Custom spacing means we fix the elicited recovery levels at intuitive levels to the experts such as 10%, 30%, 50%, etc, regardless of the number of levels. Equal spacing means we get the levels by equally dividing the range from 10% to 90% by the number of levels, which results in some odd levels, such as 10%, 36.67%, 63.33%, 90% at 4 elicitation levels. The plot shows some general trends that at 4 to 5 elicitation levels, the performance can be satisfactory.

Acknowledgements

This work was supported by the National Science Foundation (NSF grant CMMI-1824681).

References

[1] SPUR, The resilient city: Defining what San Francisco needs from its seismic mitigation policies, The San Francisco Planning and Urban Research Association (2009).

[2] WSSSC, Resilient Washington State: A Framework for Minimizing Loss and Improving Statewide Recovery After an Earthquake: Final Report and Recommendations, November 2012, Washington State Seismic Safety Committee, 2012.

[3] OSSPAC, The oregon resilience plan: Reducing risk and improving recovery for the next cascadia earthquake and tsunami (2013).

[4] Community resilience planning guide for buildings and infrastructure systems I (2016).
URL https://dx.doi.org/10.6028/NIST.SP.1190v1

[5] W. Liu, Z. Song, M. Ouyang, J. Li, Recovery-based seismic resilience enhancement strategies of water distribution networks, Reliability Engineering & System Safety (2020) 107088.

[6] E. M. Hassan, H. Mahmoud, An integrated socio-technical approach for post-earthquake recovery of interdependent healthcare system, Reliability Engineering & System Safety (2020) 106953.

[7] M. Monsalve, J. C. de la Llera, Data-driven estimation of interdependencies and restoration of infrastructure systems, Reliability Engineering & System Safety 181 (2019) 167–180.

[8] B. Cassottana, L. Shen, L. C. Tang, Modeling the recovery process: A key dimension of resilience, Reliability Engineering & System Safety 190 (2019) 106528.

[9] R. Guidotti, P. Gardoni, N. Rosenheim, Integration of physical infrastructure and social systems in communities' reliability and resilience analysis, Reliability Engineering & System Safety 185 (2019) 476–492.

[10] M. Bruneau, S. E. Chang, R. T. Eguchi, G. C. Lee, T. D. O'Rourke, A. M. Reinhorn, M. Shinozuka, K. Tierney, W. A. Wallace, D. Von Winterfeldt, A Framework to Quantitatively Assess and Enhance the Seismic Resilience of Communities, Earthquake Spectra 19 (4) (2003) 733–752. doi:10.1193/1.1623497.

[11] S. E. Chang, Urban disaster recovery: A measurement framework and its application to the 1995 Kobe earthquake, Disasters 34 (2) (2010) 303–327. doi:10.1111/j.1467-7717.2009.01130.x.

[12] G. P. Cimellaro, A. M. Reinhorn, M. Bruneau, Framework for analytical quantification of disaster resilience, Engineering Structures 32 (11) (2010) 3639–3649. doi:10.1016/j.engstruct.2010.08.008.
URL http://dx.doi.org/10.1016/j.engstruct.2010.08.008

[13] A. Barabadi, Y. Z. Ayele, Post-disaster infrastructure recovery: Prediction of recovery rate using historical data, Reliability Engineering & System Safety 169 (2018) 209–223.

[14] A. Mottahedi, F. Sereshki, M. Ataei, A. N. Qarahasanlou, A. Barabadi, Resilience estimation of critical infrastructure systems: Application of expert judgment, Reliability Engineering & System Safety (2021) 107849.

[15] N. Nojima, Restoration processes of utility lifelines in the great east japan earthquake disaster, 2011, in: 15th World Conference on Earthquake Engineering (15WCEE), 2012, pp. 24–28.

[16] S. B. Miles, S. E. Chang, Modeling community recovery from earthquakes, Earthquake Spectra 22 (2) (2006) 439–458. doi:10.1193/1.2192847.

[17] S. B. Miles, Participatory model assessment of earthquake-induced landslide hazard models, Natural Hazards 56 (3) (2011) 749–766. doi:10.1007/s11069-010-9587-5.

[18] S. E. Chang, T. Mcdaniels, J. Fox, R. Dhariwal, H. Longstaff, Toward disaster-resilient cities: Characterizing resilience of infrastructure systems with expert judgments, Risk Analysis 34 (3) (2014) 416–434. doi:10.1111/risa.12133.

[19] R. Cooke, Others, Experts in uncertainty: opinion and subjective probability in science, Oxford University Press on Demand, 1991.

[20] W. Gordon, Calculating catastrophe, World Scientific, 2011.

[21] N. C. Dalkey, The Delphi method: An experimental study of group opinion, Tech. rep., RAND CORP SANTA MONICA CALIF (1969).

[22] B. B. Brown, Delphi process: a methodology used for the elicitation of opinions of experts, Tech. rep., Rand Corp Santa Monica CA (1968).

[23] S. Dalal, D. Khodyakov, R. Srinivasan, S. Straus, J. Adams, ExpertLens: A system for eliciting opinions from a large pool of non-collocated experts with diverse knowledge, Technological Forecasting and Social Change 78 (8) (2011) 1426–1444.

[24] N. Dalkey, B. Brown, Comparison of group judgment techniques with short-range predictions and almanac questions, Tech. rep., RAND CORP SANTA MONICA CA (1971).

[25] R. T. Clemen, Comment on Cooke's classical method, Reliability Engineering & System Safety 93 (5) (2008) 760 – 765. doi:https://doi.org/10.1016/j.ress.2008.02.003.

[26] R. M. Cooke, L. L. Goossens, TU Delft expert judgment data base, Reliability Engineering & System Safety 93 (5) (2008) 657–674.

[27] W. Aspinall, R. Cooke, Quantifying scientific uncertainty from expert judgement elicitation, in, Risk and uncertainty assessment for natural hazards (2013) 64.

[28] R. T. Clemen, R. L. Winkler, Calibrating and combining precipitation probability forecasts, in: Probability and Bayesian statistics, Springer, 1987, pp. 97–110.

[29] R. M. Cooke, S. ElSaadany, X. Huang, On the performance of social network and likelihood-based expert weighting schemes, Reliability Engineering & System Safety 93 (5) (2008) 745–756.

[30] K. Zickfeld, A. Levermann, M. G. Morgan, T. Kuhlbrodt, S. Rahmstorf, D. W. Keith, Expert judgements on the response of the Atlantic meridional overturning circulation to climate change, Climatic Change 82 (3-4) (2007) 235–265.

[31] F. Beccacece, E. Borgonovo, G. Buzzard, A. Cillo, S. Zionts, Elicitation of multiattribute value functions through high dimensional model representations: Monotonicity and interactions, European Journal of Operational Research 246 (2) (2015) 517–527.

[32] I. Durbach, B. Merven, B. McCall, Expert elicitation of autocorrelated time series with application to e3 (energy-environment-economic) forecasting models, Environmental Modelling & Software 88 (2017) 93–105.

[33] K. Jaiswal, D. J. Wald, D. M. Perkins, W. P. Aspinall, A. S. Kiremidjian, Estimating structural collapse fragility of generic building typologies using expert judgment (2014).

[34] H. Nickisch, C. E. Rasmussen, Approximations for binary Gaussian process classification, Journal of Machine Learning Research 9 (2008) 2035–2078.

[35] K. P. Murphy, Machine learning: a probabilistic perspective, MIT press, 2012.

[36] S. Golchi, D. R. Bingham, H. Chipman, D. A. Campbell, Monotone emulation of computer experiments, SIAM-ASA Journal on Uncertainty Quantification 3 (1) (2015) 370–392. doi:10.1137/140976741.

[37] J. Riihimäki, A. Vehtari, Gaussian processes with monotonicity information, Journal of Machine Learning Research 9 (2010) 645–652.

[38] C. E. Rasmussen, C. K. I. Williams, Gaussian processes for machine learning, Vol. 2, MIT press Cambridge, MA, 2006.

[39] H. Maatouk, Finite-dimensional approximation of Gaussian processes To cite this version : HAL Id : hal-01533356 Finite-dimensional approximation of Gaussian processes with inequality constraints (2017). arXiv:arXiv:1706.02178v2.

[40] H. Maatouk, X. Bay, Gaussian Process Emulators for Computer Experiments with Inequality Constraints, Mathematical Geosciences 49 (5) (2017) 557–582. arXiv:1606.01265, doi:10.1007/s11004-017-9673-2.

[41] A. F. Lopez-Lopera, F. Bachoc, N. Durrande, O. Roustant, Finite-dimensional Gaussian approximation with linear inequality constraints, SIAM-ASA Journal on Uncertainty Quantification 6 (3) (2018) 1224–1255. arXiv:1710.07453, doi:10.1137/17M1153157.

[42] A. F. López-Lopera, F. Bachoc, N. Durrande, J. Rohmer, D. Idier, O. Roustant, Approximating Gaussian Process Emulators with Linear Inequality Constraints and Noisy Observations via MC and MCMC (Mc) (2019). arXiv:1901.04827.
URL http://arxiv.org/abs/1901.04827

[43] J. Oakley, A. O'Hagan, Shelf: the sheffield elicitation framework (version 2.0), Sheffield, UK: School of Mathematics and Statistics, University of Sheffield (2010).

[44] L. Faulkner, Beyond the five-user assumption: Benefits of increased sample sizes in usability testing, Behavior Research Methods, Instruments, & Computers 35 (3) (2003) 379–383.

[45] S. B. Miles, Foundations of community disaster resilience: Well-being, identity, services, and capitals, Environmental Hazards 14 (2) (2015) 103–121.