Received: 26 April 2022 | Revised: 20 May 2022 | Accepted: 31 May 2022

DOI: 10.1002/env.2745

OPINION

WILEY

Uncertainty: Nothing is more certain

Sally Cripps[1] | Hugh Durrant-Whyte[2]

[1]Data61, Commonwealth Scientific and Industrial Research Organisation (CSIRO), Sydney, New South Wales, Australia

[2]Office of the Chief Scientist and Engineer, Sydney, New South Wales, Australia

Correspondence
Sally Cripps, Data61, Commonwealth Scientific and Industrial Research Organisation (CSIRO), Canberra, ACT, Australia.
Email: sally.cripps@data61.csiro.au

Abstract

This articles discusses the importance and types of uncertainty in environmental science. Uncertainty is categorized into four types, three of which are familiar to statisticians and one of which is introduced by the authors as *knowledge uncertainty*. The article claims that the Bayesian paradigm is a logically consistent mechanism, and a useful framework for decision makers in environmental science, to manage and quantify the first three types of uncertainty. However *knowledge uncertainty*, common in environmental sciences, requires more detailed thought and more nuanced management strategies for robust decision making in the environment. The article concludes with the observation that only if we acknowledge the uncertainty inherent in inferring complex quantities from data can we make robust and explainable policy decisions that build trust with the public.

KEYWORDS

Bayesian methods, environmental science, uncertainty

1 | INTRODUCTION

Uncertainty is intrinsic to making decisions based on observations and models of the real world. Environmental sciences in particular employ complex models combined with sparse observations to both explain and predict the evolution of the world—so enabling decision making for management and stewardship of natural resources. Further, decision making in natural sciences is driven by model complexity and data diversity, rather than simply data volume—that is, the complexity of systems and models is much larger, potentially infinitely larger, than any amount of data that can be collected. Decision making in natural sciences is therefore not a "Big Data" problem, rather it is a small, sparse and diverse data problem—with uncertainty at its heart. Uncertainty exists both in measurements and models; uncertainty pervades explanations and predictions of complex natural worlds; and uncertainty lies at the heart of all decision making. How then should natural science describe uncertainty in measurements and models, build uncertainty quantified explanations and predictions, and make robust and honest decisions recognizing this uncertainty?

This article argues that there are four characterizations of uncertainty; inherent, parametric, model, and knowledge; and that each of these play an important role in the natural sciences. The first three of these—inherent, parametric, and model uncertainty—are familiar to the statistics community and are addressable through the machinery of Bayesian inference. However, in natural science, knowledge uncertainties usually dominate—large-model, sparse-data problems—and new approaches to Bayesian inference need to be developed to address these. Knowledge uncertainty covers those cases where there genuinely are no models and where there is often no data either—"unknown unknowns." Knowledge

Pliny the Elder 323 BC.

uncertainty is a real and practical issue in decision making in natural sciences and many other fields, yet there is little understanding as to how we should manage this uncertainty in an honest and robust manner. This article proposes that one way of practically managing knowledge uncertainty is through valuing information and using this to seek new observations, in a sequential manner, to actively resolve and bound knowledge uncertainties.

2 | UNCERTAINTY

Our categorization of the first three types of uncertainty is not new; inherent uncertainty is uncertainty that exists because outcomes are not perfectly predictable—even with perfect knowledge. Inherent uncertainty will not decrease with increasing data or information, and is often referred to as *aleatoric* uncertainty. In contrast parameter and model uncertainty is uncertainty that is attributable to a lack of information and is often referred to *epistemic uncertainty*, see Hacking (1975), Fox and Ülkümen (2011), Spiegelhalter (2017) for discussions on the topic. Knowledge uncertainty is also epistemic uncertainty, but an extreme version where models and data may be completely unknown. This article is concerned with epistemic uncertainty and its impact in decision making in the environmental context.

In discussing uncertainty the Bayesian paradigm is extremely useful; it has the advantage of quantifying inherent, model and parameter uncertainty in a logically consistent fashion. Parameter and model uncertainty is introduced by placing a joint distribution over these quantities, $p(\theta, \mathcal{M}) = p(\theta|\mathcal{M})p(\mathcal{M})$, where θ denotes parameters, and $\mathcal{M}$ denotes a model. These quantities represent the prior belief about θ given a model $\mathcal{M}$, $p(\theta|\mathcal{M})$, and the prior belief of a model $p(\mathcal{M})$. Upon observing some data, $\mathcal{D}$, these beliefs are updated by a factor known as a likelihood function, $p(\mathcal{D}|\theta, \mathcal{M})$ to give a joint posterior given by

$$p(\theta, \mathcal{M}|\mathcal{D}) = \frac{p(\mathcal{D}|\theta, \mathcal{M})p(\theta|\mathcal{M})p(\mathcal{M})}{p(\mathcal{D})}. \tag{1}$$

The quantity $p(\mathcal{D})$, is the marginal likelihood of the data after integrating over all possible parameters and models, where the integration is w.r.t the prior distribution of these quantities.

The uncertainty surrounding a possible event, E, given some data, $\mathcal{D}$, is also quantified by its posterior distribution and given by[1]

$$P(E|\mathcal{D}) = \sum_{\mathcal{M}\in M} \int P(E|\mathcal{D}, \theta, \mathcal{M})p(\theta|\mathcal{M}, \mathcal{D})d\theta p(\mathcal{M}|\mathcal{D}). \tag{2}$$

In Equation (2), the inherent uncertainty, that is, the uncertainty which exists even if we knew θ and $\mathcal{M}$, is given by $P(E|\mathcal{D}, \theta, \mathcal{M})$. To account for the uncertainty in the parameter, θ, and the model, $\mathcal{M}$, in the prediction of the event E, we need to integrate over these quantities where the integration is w.r.t the posterior distribution, $p(\theta, \mathcal{M}|\mathcal{D}) = p(\theta|\mathcal{M}, \mathcal{D})p(\mathcal{M}|\mathcal{D})$, which is given by Equation (1).

We will discuss parameter, model, and knowledge uncertainty with reference to two examples. The first example concerns the estimation of the three dimensional subsurface geology, at a depths of up to 5 km below the Earth's surface, specifically for geothermal exploration at the Moomba gas field in the Cooper Basin, South Australia, see Beardsmore et al. (2016) for a full description of the problem. The second example is concerned with understanding the impact of non-native species on complex ecosystems, and in particular the impact of wild horses on the ecology in the Kosciuszko national park, Australian Broadcasting Commission (2022).

Both these problems have common attributes; sparseness of data, complexity of natural processes and decisions which, once made, are not easy to reverse and have substantial impact on the environment. However, despite these similarities, our knowledge of the processes which govern these systems is very different for these two examples and highlights the point we wish to make about knowledge uncertainty.

In the geothermal example, the decision at hand is where to drill to maximize the probability of finding granite intrusions above 270°C for generation of geothermal energy. Despite the spareness of observational data on the geology at depth, this is a relatively well defined problem; a lot is known *a priori* about the system. There has been previous research about the subsurface geology in Moomba, and previous research about the properties of various rock types, such as

[1] Assuming the set of all possible models, given by M, is discrete and countable. The integral sign in Equation (2) is replaced by a summation sign for discrete parameter spaces.

density and magnetic susceptibility. Furthermore much is known about the natural processes of the systems, often referred to as forward models; in this case the physical equations which map rock properties at depth, not directly observed, for example, density and magnetic susceptibility, to the surface sensor measurements, for example, gravity and magnetotellurics. There is still parameter and model uncertainty, all of which must be quantified and taken into consideration in any decision, but decision makers are operating with a reasonable amount of knowledge.

The wild horses of the Kosciusko National Park example is not a well defined problem. The decision at hand in this example is how many wild horses should be allowed to exist in the national park. Again data is sparse, to our knowledge no experiments have been run to establish *a priori* beliefs on the extent to which wild horses damage national parks. However, in contrast to the geothermal example, there are few, if any, forward models which map the number of horses, to tree damage. This is an example where all levels of uncertainty, parameter, model and knowledge uncertainty, play an active and important role; how many horses are there, how much damage is caused by horses (the environment is impacted from many sources), what would be the impact of reduced numbers of horses?

One of the authors was significantly involved in the development of the recent plan to manage wild horses in the Kosciuszko national park. The issue of wild horses in the Kosciuszko—the ecological damage they cause, the heritage values they represent to parts of the community, the emotions generated by potential culling of these animals—has been hugely divisive across the community and politics for over half a century and more, Australian Broadcasting Commission (2022).

2.1 | Parameter uncertainty

The quantification of parameter uncertainty in environmental models, whether they are physical or statistical models, or combinations of both, is a well researched area, but still challenging because the data are often equally well explained by different possible combinations of parameter values, resulting in likelihood surfaces, and therefore often posterior surfaces, which are multimodal. In such instances measures of uncertainty in any local mode do not adequately reflect the uncertainty of the parameters, so care needs to be taken to ensure the entire posterior surface is explored. This is not as straightforward as one might first assume.

2.1.1 | Parameter uncertainty; Geothermal exploration

An illustration of the difficulty in exploring a posterior distribution is given by Scalzo et al. (2019), which is an extension of Beardsmore et al. (2016). In this example the parameter of interest, θ, is the subsurface geology, which includes the location of the geological layers, and the type of rock and its properties in any given layer, up to a depth of 5 km. The implicit assumed model, $\mathcal{M}$, is that the geology is made of layers and that the sensor readings have a Gaussian distribution. The information available to the researchers is in three forms. The first is sensor measurements such as gravity and magnetotellurics at the surface. The second type of information is the forward models. These two types of information were incorporated into the likelihood function, where it was assumed that, given a subsurface geology, the sensor readings had a Gaussian distribution, with a mean given by the forward models of the assumed geology and an unknown variance. The third type of information is expert prior knowledge of the location of the layers, and this was incorporated through the prior distributions.

There is no closed form expression for the posterior distributions of subsurface geology and Beardsmore et al. (2016) used Parallel Tempering Markov chain Monte Carlo (PTMCMC) to obtain samples from the posteriors. Later work by Scalzo et al. (2019) examined the sensitivity of the estimated posterior distributions to different transition kernels in the Markov chain, used to explore the posterior parameter space.

One component of θ is the rock density in the third layer and Figure 1, modified from Scalzo et al. (2019), displays the resulting marginal posterior distributions of this quantity for three different transition kernels used for the within chain moves of the PTMCMC scheme: an isotropic Gaussian random walk Metropolis Hastings, (iGRWMH), a preconditioned Crank-Nicholson Metropolis Hastings (pCNMH) and an adaptive Gaussian random walk Metropolis Hasting (aGRWMH). Panels (a)–(c) correspond to the transition kernels, iGRWMH, pCNMH, and aGRWMH respectively. Theoretically these distributions should be identical and yet Figure 1 shows very different estimates for rock density in the third layer. Clearly some, if not all, of the chains have not converged. This example shows that even given a well defined problem, quantifying parameter uncertainty is not as easy as it is often assumed to be.

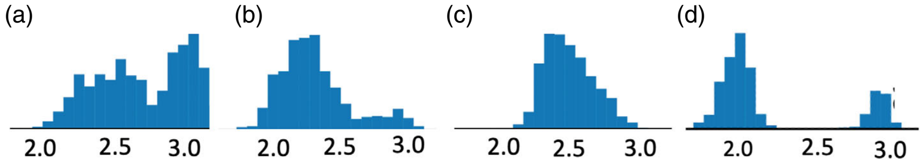

FIGURE 1 Panels (a)–(c) are sampling estimates of the marginal posterior distributions of rock density in the third layer using within transition kernels iGRWMH, pCNMH, aGRWMH respectively, for a Gaussian likelihood. Panel (d) shows the same marginal posterior distribution for a Cauchy likelihood using pCNMH for the with-in chain transition kernel, Scalzo et al. (2019)

2.1.2 | Parameter uncertainty; Wild horses in Kosciusko

The issue of parameter uncertainty in the wild horses example is even more difficult. There are many parameters which contribute to damage in national parks, the number of wild horses being only one, and it is surprisingly difficult to even get a prior distribution for this quantity; there is significant uncertainty even in simply counting the number of horses. Typically surveys are run by flying a helicopter over “representative” transects, counting sightings of horses and then using established statistical methods to estimate the total number of horses in an area, NSW Department of Planning and Environment (2020a). Uncertainties in this method can be significant—the most recent (2020) survey estimated there were between 9000 and 22,000 horses (with 14,000 being the quoted number)—a significant error bar.

2.2 | Model uncertainty

Model uncertainty refers to the uncertainty about the validity or correctness of a mathematical or statistical model from which data are assumed to be generated. All models are just approximations and the validity of any model is usually judged by whether the assumptions underpinning the model are reasonable and also by how well the model fits observed data and predicts new outcomes.

Quantifying the uncertainty over models can also be done via Bayes theorem, again via the posterior distribution of the model,

$$p(\mathcal{M}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathcal{M})P(\mathcal{M})}{P(\mathcal{D})}.$$

However, the computation of $P(\mathcal{D})$ requires integration over all possible models. For this reason model uncertainty is often approached by comparing one model, $\mathcal{M}_1$, say, with another, $\mathcal{M}_2$, say, via the ratio of the respective posterior distributions,

$$\frac{p(\mathcal{M}_1|\mathcal{D})}{p(\mathcal{M}_2|\mathcal{D})} = \frac{p(\mathcal{D}|\mathcal{M}_1)P(\mathcal{M}_1)}{p(\mathcal{D}|\mathcal{M}_2)P(\mathcal{M}_2)}, \tag{3}$$

so that the computation of the marginal distribution $P(\mathcal{D})$ is not required.

2.2.1 | Model uncertainty; Geothermal exploration

In Bayesian modeling, assumptions are made both about the likelihood function and the prior. An example of a prior assumption used in Beardsmore et al. (2016) is that the subsurface geology had a layered structure, intrusions were allowed but folds were not.

An example of a likelihood model is one based on the assumption that the sensor observations have a particular type of distribution. Figure 1 panels (b) and (d), from Scalzo et al. (2019) show the posterior distribution of rock density in the third layer corresponding to Gaussian and Cauchy likelihoods, respectively, and show how inference is affected by these assumptions. We can then use Equation (3) to compute which likelihood function is a better approximation, and hence which estimate of the posterior best summarizes the information about the density of rock in the third layer.

Similarly we could define models for prior geologies; for example a layer structure versus a structure which allows for folds; computations could be performed to determine which model was more likely given the data. Several other choices for priors, likelihoods and forward models could have been made, and the beauty of the Bayesian framework is that prior assumptions are explicitly encoded, allowing the user to explore the impact of assumptions on inference.

However for these quantities to be meaningful we must have some knowledge that the set of priors, likelihoods and forward models under consideration contain models which are reasonable approximations to the "truth," and that we are not omitting large classes of more likely models—it is of little use to know which of two models is more likely, if both are wildly inaccurate.

2.2.2 | Model uncertainty; Wild horses in Kosciusko

In the wild horses of the Kosciusko, there is substantially more model uncertainty than in the geothermal example. First, there are no well-known forward models to map the number of horses to the damage to the environment, NSW Department of Planning and Environment (2020b). To address this, test cases are set-up to build models—areas are fenced off, river banks and other special areas are monitored—but all of these cases yield small samples in an immensely complex, dynamic and fragile environment.

Even if one were to build a model which mapped the number of horses to environmental damage, there is also the major issue of understanding the damage caused by wild horses versus damage caused by deer or other invasive species. In short, it is highly unlikely that we have a set of models which, collectively, contain models which are reasonable approximations to the truth. In such instances, we can turn the handle of the Bayesian machinery and compute the relative likelihoods of various models, but it is unlikely that these models will be useful in predicting the amount of future damage which horses will cause.

2.3 | Knowledge uncertainty

Knowledge uncertainty refers to that uncertainty when there are no useful existing models, either because the phenomena has never been, or is rarely, observed; or because the system under study is very complex and dynamic. This is case for the wild horses of the Kosciusko, and is the situation in which many environmental decisions must be made. How to make decisions in these situations is of the utmost importance, because, handled poorly, these situations give all members of the community—including scientists—an opportunity to push opinions rather than evidence, as answers to these complex problems.

So how does one proceed in these problems? First, start with what is known, however trivial it may be, and then sequentially acquire data which provides the most useful information. To illustrate this idea it is instructive to consider the geothermal exploration problem. Suppose we are interested in the location (latitude and longitude) of the smallest depth at which hot granite is found in the Moomba region, call this quantity θ. Given current information we approximate the distribution $p(\theta|\mathcal{D})$ and find there is still considerable uncertainty surrounding θ. Suppose further that we have a finite set of K possible surface locations, denoted by l_k, for $k = 1, \ldots, K$, at which we are able to take another sample of the subsurface geology at depth of z, say. These samples are costly, both financially and environmentally, and we wish to minimize these sampling costs while maximizing the expected gain in information which will result from the additional sample. Denote by x_{l_k} the realization of a sample taken at location l_k (ignoring the dependence on z for ease of exposition) and consider the case where the subsurface geology is either sandstone or granite, with $x_{l_k} = 1$ if the geology is sandstone and $x_{l_k} = 2$ if it is granite. One metric is the expected gain in information in regard to θ (where the expectation is w.r.t the joint distribution $p(\theta, x|\mathcal{D})$) for a sample is taken at location, l_k, $E(I_{l_k}(\theta))$. We choose the next location to sample from which maximizes this quantity, call it $l_k^{\max}$. That is

$$l_k^{\max} = \arg\max_{l_k} E(I_{l_k}(\theta)) = \arg\max_{l_k} \int \sum_{j=1}^{2} \left\{ \log p(\theta|\mathcal{D}, x_{l_k} = j) - \log p(\theta|\mathcal{D}) \right\} \Pr(x_{l_k} = j|\theta) p(\theta|\mathcal{D}) d\theta.$$

This procedure is an example of a technique known as Bayesian optimization, Brochu et al. (2010) where the acquisition function is based on information theory, see, Shahriari et al. (2016).

The wild horses example is more complex but the principles guiding decision makers in both examples are the same. The new management plan for the wild horses, NSW Department of Planning and Environment (2020b) tries to address many of the areas of uncertainty in arriving at a decision around horse numbers—and draws on the key ideas in this article.

First, establish what is known. It is fairly well agreed that about 350 horses can be re-homed each year, NSW Department of Planning and Environment (2020b). Based on horse reproduction, 350 new births each year equates to approximately 3000 wild horses. This is the maximum number of horses whose offspring can be rehomed off park—that is, the maximum number of horses in the national park that will not grow over time.

Second the horses will be counted now on a continuous basis using drone technology—rather than using periodic counts from a manned helicopter. The horse count will be available in real-time to the community in an effort both to quantify and reduce (parametric) uncertainty and to build trust in the count.

Third, the damage caused by horses will be monitored over time using images taken by drones, analyzed with deep learning algorithms, to provide a better understanding of uncertainty in models of environmental damage.

Finally, with horse (parametric) and damage (model) uncertainty measured, there will be a continuing process of valuing and acquiring additional information to improve management decisions, using techniques such Bayesian optimization described above—and using this to better select sustainable numbers of horses allowed in the park.

3 | CONCLUSION

A major challenge in applying data science methodologies to natural sciences is the amount of uncertainty facing the decision maker at several levels. This article shows that the Bayesian paradigm is a coherent set of principles by which to categories and measure this uncertainty, as well as a methodology for resolving this uncertainty by the sequential acquisition of high value information. In situations where there are no reliable models and data is sparse, the authors recommend an adaptive decision making strategy, which allows for decisions to be adjusted as new information comes to light to ensure robust management.

For these adaptive strategies to be accepted by communities it is essential to be honest about the often high levels of uncertainties; failure to do so will undermine public trust in data driven decision making. However this is not an easy task when there are many stakeholders, all with different objective functions, who demand certainty, and often see measurements of uncertainty as failures of models, but it is up to us, the data science community, to remedy this. In attempting to do this it is uplifting to recall the words of Richard Feyman

> *We will not become enthusiastic for the fact, the knowledge, the absolute truth of the day, but remain always uncertain … In order to make progress, one must leave the door to the unknown ajar.*

ORCID

Sally Cripps https://orcid.org/0000-0003-3207-172X

REFERENCES

Australian Broadcasting Commission. (2022). The battle over Australia's brumbies intensifies in a clash of culture, colonialism and conservation. https://www.abc.net.au/news/2022-02-22/kosciuszko-brumby-battle-turns-feral-mountain-culture-war/100830536

Beardsmore, G., Durrant-Whyte, H., McCalman, L., O'Callaghan, S., & Reid, A. (2016). A Bayesian inference tool for geophysical joint inversions. *ASEG Extended Abstracts*, *2016*(1), 1–10.

Brochu, E., Cora, V. M., & de Freitas, N.. A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. *ArXiv, abs/1012.2599, 2010*.

Fox, C., & Ülkümen, G. (2011). Distinguishing two dimensions of uncertainty. *SSRN Electronic Journal*.

Hacking, I. (1975). *The emergence of probability: A philosophical study of early ideas about probability, induction and statistical inference*. Cambridge University Press.

NSW Department of Planning and Environment. (2020a). Kosciuszko national park; tracking the wild horses. https://www.environment.nsw.gov.au/topics/animals-and-plants/pest-animals-and-weeds/pest-animals/wild-horses/kosciuszko-national-park-wild-horse-management/tracking-the-wild-horse-population

NSW Department of Planning and Environment. (2020b). Kosciuszko national park wild horse management. https://www.environment.nsw.gov.au/topics/animals-and-plants/pest-animals-and-weeds/pest-animals/wild-horses/kosciuszko-national-park-wild-horse-management

Scalzo, R., Kohn, D., Olierook, H., Houseman, G., Chandra, R., Girolami, M., & Cripps, S. (2019). Efficiency and robustness in Monte Carlo sampling for 3-d geophysical inversions with obsidian v0.1.2: setting up for success. *Geoscientific Model Development*, *12*(7), 2941–2960.

Shahriari, B., Swersky, K., Wang, Z., Adams, R. P., & de Freitas, N. (2016). Taking the human out of the loop: A review of Bayesian optimization. *Proceedings of the IEEE*, *104*(1), 148–175.

Spiegelhalter, D. (2017). Risk and uncertainty communication. *Annual Review of Statistics and Its Application*, *4*(1), 31–60.

How to cite this article: Cripps, S., & Durrant-Whyte, H. (2023). Uncertainty: Nothing is more certain. *Environmetrics*, *34*(1), e2745. https://doi.org/10.1002/env.2745