# Chapter 15
# Uncertainty and Error

**Andrew Evans**

**Abstract** Errors in input data, parameterisation, and model form cause errors and uncertainty in model outputs. This is particularly problematic in non-linear systems where small changes propagate through models to create large output differences. This chapter reviews the issues involved in understanding error, covering a broad range of methodologies and viewpoints from across the spatial modelling sciences.

## 15.1 Introduction to Error and Its Terminology

> *There are known knowns. These are things we know that we know. There are known unknowns. That is to say, there are things that we know we don't know. But there are also unknown unknowns. There are things we don't know we don't know.*
>
> Donald Rumsfeld: February 12, 2002

The quote above outlines, as best as it can, an important truth in modelling the real world: that the ramifications of ignorance can be tempered by meta-information on the level of that ignorance. Whatever the appropriateness of Donald's statement at the time (on which, see Žižek 2004), the Rumsfeld 'Ladder of Ignorance' nevertheless summarises nicely that it is one thing not to know something, and it is quite another to be able to quantify that ignorance and to summarise it.[1] While there are things we know with perfect accuracy in modelling the real world, in general these are few and far between. It is much more the case that we know that there is some *error* in our understanding, and this leads to *assumptions* in our models and *uncertainty* about our model results that need to be communicated to users of the results. If we are

[1] Rumsfeld largely repeated the terminology of risk assessment in engineering, see, for example, Suter et al. (1987).

A. Evans (✉)
Centre for Applied Spatial Analysis and Policy (ASAP), University of Leeds, Leeds, UK
e-mail: a.j.evans@leeds.ac.uk

A.J. Heppenstall et al. (eds.), *Agent-Based Models of Geographical Systems*,
DOI 10.1007/978-90-481-8927-4_15, 

lucky, we can quantify this error and/or the resultant uncertainty. If we are very unlucky we either can't do this, or we don't know about the error in the first place: we have an unknown unknown, a situation to be avoided at all costs.[2]

Generally in agent-based modelling we have a difficult job, as we tend to deal with concealed and non-linear systems which may be influenced by multiple variables. Some of these variables we may not recognise as important (an error of understanding, that is, an *epistemic error*; see Faber et al. 1992). Indeed, we are often uncertain as to how closely the broader model form which these variables slot into replicates the system in the real world. It is often the case that we have data and need to infer at least part of the model from it. It may be that other models would do the same quality job against the data we have, and a better job against new data: the so-called model *equifinality* problem (Beven and Binley 1992). Even if we have the right form and variables, we may have multiple options for the weights of particular variables within the model (the *inverse problem*). These difficulties have led many to suggest that such models should be regarded as grand "thought experiments", not so much designed to predict accurately as to allow us to reflect on the systems we are studying and our understanding of them (Di Paolo, Bullock and Noble 2000).

Traditionally, however, modellers tend to feign confidence in their model forms and concentrate on error issues associated with another feature of dealing with real-world multivariate systems: that some of the variables we don't want to use will cause *noise* in the real data records of those we do. Noise is essentially variation in our variables of interest around the values we expect to represent their 'true' or 'important' value, but it is difficult to define objectively. At best 'true' in this context means, tautologically, uninfluenced by the variables causing the noise, while 'important' means, equally tautologically, the signal that we need to understand the system given the variables we've chosen to include. Noise produces a *sampling error*, as we hope sampling for our model inputs will generally give us the 'true' or 'important' behaviour of a sampled system, but what we get instead is varied by outside influences, including the mechanics of the measurement process. If we use the data as the foundation of a model prediction, such an error will plainly cause problems.

Noise is frequently treated as an *aleatory* [i.e. random] *error* (which may be regarded as, a type of *ontological* (Walker et al. 2003) or *phenomenological* (Faber et al. 1992) *uncertainty*), added to an underlying signal. The apparently random nature of noise is both problematic and of use. More often than not our definition of something as 'random' is an admission of our ignorance of the influences on a system, or our inability to model them directly and deterministically. However, even though the acknowledgement of noise represents something of an admission of failure, if we know something of the form of the errors involved we can build a description of them into our model. If our model is also a perfect representation of

[2]Here we will largely deal with ignorance from the viewpoint of uncertainty. For more detailed discussions of wider types of ignorance in modelling see: Faber et al. (1992), Walker et al. (2003), and Brown (2004).

the bit of the system we are interested in, this gives us the so-called *Perfect Model Scenario*. As noise-based errors can usually be treated as random, one simple way we can include such errors is by developing *stochastic models*, which include some randomisation of the key variables within strictly controlled ranges or distributions. This is usually achieved through *Monte Carlo testing*: the distribution of each input variable and/or parameter in the model is sampled randomly, but with an emphasis on more probable values appearing proportionally more often (so-called *Monte Carlo sampling*); the model is then run and the process repeated multiple times. Such stochastic models will give a distribution of results if run enough times, and this is often treated probabilistically (for a clear agent-based example centred on generating uncertainty statistics, see Bobashev and Morris 2010). However, for social modellers at least, the top-down analysis of final aggregate results isn't facilitated by the fact that, by-and-large, we lack the very large samples over time other modelling disciplines have and therefore struggle to understand whether the final probabilistic results match the real world well or poorly. This data scarcity sometimes perversely encourages social modellers to abandon randomisation and make one-number 'dart-board' predictions that attempt to hit the few real-world points we have as closely as possible. The alternative to top-down probabilistic assessments of results are bottom-up attempts to delimit the effects of different sources of error as they *propagate* through the system. Unfortunately these are far from simple in non-linear systems. There is a large body of literature on understanding the propagation of error from model inputs, through the model, to outputs/predictions, and for linear/linearisable mathematical models there are well-trodden solutions. However, these solutions usually rely on us being able to characterise the distribution of the noise or other error involved. For social modellers the lack of data highlighted above often makes this problematic. Moreover, many of the techniques assume the distribution is Gaussian/normal. For non-linear systems like ours this need not be the case – indeed, noise may additionally have a changing character (*heteroscedasticity*), and the system may have inputs that vary in importance (i.e. be *non-stationary*) – all of which render many of the traditional methods for dealing with the propagation of errors problematic.

In some cases, then, we may be limited to following some traditional non-linear systems analysis (e.g. Smith et al. 2010b) in bounding worst-case scenarios. This is the position that non-linear uncertainty analysts have endeavoured to move away from for years, not least because identifying and quantifying a "worst" case is usually difficult (Suter et al. 1987). Moreover, it may be that the error propagating through our systems renders even that approach problematic. Non-linearities tend to accentuate small initial data fluctuations (Lorenz 1963) until the small differences between our noisy model input data and the 'true' signal we were hoping for at the start have exploded to cause wild behaviour in our final model results. In such situations, the resultant uncertainty range swamps the range of predictions we suspect the system might have given if presented with the 'true' data. Such errors bloom equally where the *digital precision* with which we can deal with numbers in computer systems fails us and we get initially small changes to our figures through, for example, truncation. If our model is predictive over time, such exploding differences will only

increase as we move further from the starting conditions. As such, a nuanced approach to error at different stages of the model process seems critical. However, the peculiarly untraditional architecture of agent-based systems, and their complicated interactions and iterations, do create difficulties in applying the techniques developed for segregating and quantifying errors in more traditional non-linear systems modelling. For example, one standard technique used to more easily understand how errors are propagated is to linearize non-linear models at particular points of equilibrium (for example, with a Taylor expansion), under the assumption that these equilibria are the key modes of operation of the system. The view is taken that the loss of accuracy at these points due to (often low-order) linearization is a worthwhile sacrifice to make to understand the propagation. In agent-based systems, however, the large number of interactions between elements with mixed-method rulesets make any such approach difficult, on top of which most agent-based modellers see non-linear dynamics and a lack of equilibrium in a far more positive light than those working in alternative modelling paradigms.

Along with errors of the types above, uncertainty is also produced by *biases* in our system or inputs: systematic shifts in our model or results away from the 'true' picture. Although traditional modellers make a clear distinction between bias and error, for most models of any complexity the distinction is not always so clear – a missing variable from a model, for example, may be an error of understanding and a systematic bias, but may display as a set of variations that appear to be noise; each problem is related but often handled separately before the overarching issue is appreciated.

This chapter will outline some of the errors and uncertainties associated with modelling the real world, and introduce some of the techniques to deal with such issues. It is worth noting that the chapter only really deals with error from the point of view of uncertainty (the assessment of error in the calibration, verification, sensitivity testing, and validation of models is dealt with more fully in Ngo and See 2012). This chapter is broadly divided into the sources of error and uncertainty, following through the modelling process from inputs to outputs, and ends on an optimistic note with a discussion of why we stand some chance of dealing with this difficulty. It is probably beholden of the author to note that the size and scope of the uncertainty literature is significant beyond the limitations of a single chapter in a book, so this review, by necessity, is more selective in some areas than others.

## 15.2 Uncertainty Associated with Input Data

Most agent models are based, in some manner, on the real world. Even the most abstract models contain rulesets built on qualitative or quantitative data collection. Real world data can be directly used as an input during formation of a model's structure, the calibration of parameters, or for driving the model. This section looks at the errors that result from the recording process for this data; having insufficient data; missing data within a dataset of otherwise continuous values; and errors that

result from the pre-processing of data, such as classification binning. Generally it has to be noted that the complexities of dealing with sources of data mean that we often assume little error in our input data, any prediction error being usually attributed to our parameterisation. This is far from ideal.

### *15.2.1 Data Measurement and Transcription Errors*

Plainly, however data is measured there will generally be errors associated with the process, including transcription errors.

Input error is most successfully quantified for instrumental noise, where the instrument can be checked against multiple readings of the same physical property. In this situation, errors are represented by metrics of *accuracy* (closeness of the sample or a derived statistic to the real value) and *precision* (the tightness or lack of variance of a sample repeated under the same set of conditions). Provided there is no consistent bias in the sample, levels of accuracy will largely be determined by the measurement precision. Standard measures of variance will provide a representation of the error associated with a lack of precision, and, as most instrument errors are Gaussian, the usual figures reported are the standard deviation of the sampling distribution (the *standard error*) and the sample mean, in the form mean ±SD (see Nagele 2001). Such reported figures may be useful in the remaining modelling process provided the error distribution is Gaussian or the figures adapted to a reported alternative distribution. The JCGM/ISO GUM methodology (http://www.bipm.org/en/publications/guides/gum.html) is the standard in this area, and utilises a probabilistic treatment of the belief one might hold in a measurement and standard propagation of error techniques (see Sect. 15.5.1, below).

For spatially located data, particularly data that arrives without a clear error distribution associated with it, more care has to be taken that the data error is not heteroscedastic. That said, information about the spatial field, for example that all points within an area should be the same, or there is positive spatial autocorrelation (i.e. points should be more similar to nearer neighbours), can allow estimates of the distribution of errors to be made. Heuvelink (1998) gives details (see also, for spatio-temporal autocorrelation, Powers 2005). Spatial/locational uncertainties in spatial data are covered by an extensive literature, but Zhang and Goodchild (2002) provide a very comprehensive overview of standard techniques in raster, vector, and spatial object uncertainty modelling. Research issues in semantic uncertainty associated with objective and subjective spatial data are reviewed in Fisher et al. (2004) and Evans and Waters (2007), respectively.

Measurement errors that are non-instrumental, for example errors encouraged by qualitative survey design, are complicated issues of psychology and semiotics. They are one of the most important areas of concern for agent-based modelers wishing to deal with qualitative rulesets and decision making. Good survey design can go a long way – a good starting point on minimizing errors in quantitative judgments is Poulton (1989), while Groves et al. (2009) concentrate on minimizing

errors in surveys more generally. In addition, the use of fuzzy sets defined from surveys as model inputs can at least acknowledge and embrace the problem (Zadeh 1975, 1976; Verkuilen 2005; see Evans and Waters 2007, for a spatial example).

If we are lucky, such issues are simple and systematic biases we can recognize and may, infact, be of interest: for example, a bias from mis-understanding the intended levels of a Likert scale survey, or a genuine attempt at fraud. Issues of genuine fraud during data collection might be revealed by comparison with normal (or other) distributions, or through comparison of chosen digits in the data with the Benford distribution (Kiesling undated; Cho and Gaines 2007; Mebane and Kalinin 2009), but more usually they require detailed stakeholder knowledge and trust to be developed during the modelling project to solve them.

Transcription errors should become increasingly rare as more data is collected electronically at source. Most will be treated as noise, unless we are lucky enough to have a consistent bias, though some will be recognisable as outliers. The standard source on recognising and dealing with outliers is Barnett and Lewis (1994). For an updated treatment in multivariant space, see Cerioli and Farcomeni (2011), while López (1997) and Rogers et al. (2009) give good starting points for recognising geographical/spatio-temporal outliers.

### *15.2.2 Appropriate Sample Size*

The inherent complexities of most of the systems agent-based modellers deal with mean than there is a complicated, multivariant, and non-linear relationship between variables of interest and system predictions. To capture the complete set of potential combinations of variables would involve considerable sampling efforts, in systems that are often hard, if not impossible, to sample well. In addition, we have the problem of noise distorting our samples. To understand noise we need repeated measurements of the same quantity/system state, with enough samples taken to define the distribution of the measurements under the influence of the noise sources. Once we have this distribution we may use it probabilistically (see above), or we may try to estimate what the noiseless data would look like. In general, to get as close to the noiseless value as possible, we have to pick a representative statistic to estimate that is as noiseless as possible: for example, if the noise is Gaussian, the arithmetic mean of the population. Where we want continuous data we may smooth out the noise. Keesman and van Straten (1989) summarise some of the opportunities for data smoothing, while Beck (1987) summaries some of the issues. However, it isn't especially clear whether treating data to remove noise is always appropriate. We generally try to minimise the effects of noise on inputs, especially with systems that explode errors non-linearly, as it is usually regarded as a detrimental influence from things we'd like to exclude from our models. However, this needs to be determined on a case-by-case basis; if the real systems suffer from such noise, are we correct to exclude it by, for example, statistical pre-processing? A better approach (Sect. 15.6) may be to build systems that show the same resilience to noise that we see in real world systems.

If we are going to try to remove noise, and we've identified our statistic of interest, we need to sample sufficiently highly that we can estimate that value in the population accurately from the samples. But how do we decide how large a sample is 'sufficient? Traditionally this has come down to trading off expensive increases in sample size against inherent risk that with small samples your value of interest may be unrepresentative. For situations where the value of interest has a well characterised, independent, and constant variation, we can directly calculate the size of sample needed for us to be able to make the estimate of the true value at some level of precision within some pre-defined levels of confidence. For example, with random independent samples, the standard error of the sample mean is the population standard deviation, divided by the square root of the sample size. By adjusting the sample size, we can reduce the error by a known degree. It is usual to trade off sample error and size for a specific confidence, such that if multiple samples were taken the number for which the true value of the statistic would fall within the range of the sample-based estimate ± the sample error would be, say, 95%.

However, this process is not so simple when the data/noise is non-normal and not independent, as it frequently is in non-instrumental noise cases. For basic non-normal distributions, appropriate sample sizes can be estimated for a given confidence using Kolmogorov-Rényi statistics (Spear 1970). However, for time series and spatial data, this process becomes more complicated. Spatial and temporal autocorrelation (where nearby points have related values) can have a significant effect on the apparent sample size of sampled datasets by introducing sample redundancy (Getis 2007). These issues become particularly important when datasets for validating models against are drawn from the same area by sample splitting (Araújo et al. 2005). Significance testing based on autocorrelated data should take autocorrelation into account (though rarely does). A summary of some of the general methods for dealing with spatial autocorrelation can be found in Legendre (1993) and Getis (2007). Kelejian and Prucha (2010) outline something of the size of the problem facing spatial modellers in their discussion of the relationships between sample size, spatial correlation, and missing data, in regression models with spatial lags.

Where non-spatial data suffers from heteroscedasticity or non-independence of noise, it can be treated (see Gallagher and Doherty 2007, for details) which helps with some issues. Spatial heteroscedasticity can further complicate the picture though (see Lauridsen and Kosfeld 2007); for example, positive autocorrelation in errors can falsely reduce error levels (Araújo et al. 2005; Getis 2007). With more complicated non-linear systems, we often have to take a slightly wider viewpoint and concentrate instead on how input variation affects the modelling we are trying to do. When we are trying to model a non-linear system, and the function that we are trying to estimate parameters for is known, it is possible to sample repeatedly to simultaneously build up a picture of the error and the resultant sample size needed. Methodologies can be found in the comparison provided by Malinarič and Ďuríšek (2004). If we know something of the variation in the model error at key points, we can sample these more frequently (O'Neill et al. 1980; Beck 1987). However, with complex systems and limited sampling budgets it is sometimes more practical to use more distribution-free methods, for example 'sampling to redundancy methods',

like species area curves, where sample novelty across multiple samples is plotted against sample numbers or size to determine when sampling is sufficient to capture all new elements in a population (Lyman and Ames 2007). While such methods give a poor statistical confidence, they do at least ensure a sample across the potential range of values has been taken. A final issue is that in many of the systems we study the relationships are non-Markovian, that is their future may be influenced by the specific pathway the system has taken historically, rather than just the current instantaneous, autocorrelated, state. This introduces considerable complications into the process of determining appropriate sample sizes.

Adequately sampling the set of potential combinations of variables and predictands is difficult in complex and non-linear systems with non-normal and interdependent variables, and it is doubly so if we also wish to understand the errors in the data. As such, we are generally thrown back on validating models at output, rather than trying to statistically validate the representativeness of the inputs.

### *15.2.3 Missing Data*

Missing data is usually a result of disrupted sampling, or the repurposing and combination of previously collected datasets. For traditional models missing data can often be problematic, especially where data is iteratively re-fed into the model. Artificial intelligence systems based around weight-adjusted learning (like artificial neural networks) and/or case-by-case decision making (like agent-based models) tend to respond better to missing data in both calibration and prediction than traditional models. Nevertheless, sometimes it is necessary to deal directly with missing data for computational or predictive purposes.

Where data is reasonably well distributed it is sometimes possible to interpolate new data into the gaps using a function based on the data we have. For simple datasets with well-known statistical properties, the techniques used for developing the functions (such as linear least-squares regression) have well-known error assessments that utilise all the data brought in to calculate the function. However, for complex non-linear datasets – especially spatial datasets – where the error and function are not easily characterised, it is more common to assess the error by rotational cross-validation (repeated removal of known data points, construction of the interpolation, and comparison of the interpolation against the removed points). This has the strengthening advantage of maintaining some distinction between the datasets used to construct and validate the function. A short but useful review of the relevant literature on missing spatial data can be found in Kelejian and Prucha (2010) and interpolation in Isaaks and Srivastava (1990). When a distribution of new data points is needed, rather than values on a continuous surface, popular techniques revolve around resampling what is already present to generate larger datasets. For example, in bootstrapping, systems are trained on data derived by sampling a distribution multiple times to generate a training set, unselected data giving a validation dataset. Such techniques are common when datasets are too

small to use as-is. In addition to generating new data with its own or inherited error, where data limitations are known resampling can be used to constrain errors, particularly where based on Bayesian or error-led assessments (Luoto et al. 2010).

Where data is poorly distributed, missing data can lead to biases. Such biases can be quite subtle, particularly when dealing with spatial autocorrelation. Where a surface is needed, it is sometimes possible to adjust the importance of samples to account for an inappropriate sample distribution. For example, spatially clustered data can be declustered to reduce the importance of over-sampled areas by weighting each value by a function of the distance to its neighbours (for techniques, see Dubois and Saisana 2002). When a distribution is required, resampling can remove some kinds of biases (for example, autocorrelation in sequential, or spatial, samples) if the sampling is carefully randomised (Luoto et al. 2010). However, ultimately biases caused by missing data usually necessitate additional data collection exercises to resolve the problem.

In the absence of good data, models often rely on strongly believed deterministic relationships or qualitative theory, where they might be better off including Bayesian entities in the relationships so that they can be updated as information comes in, and uncertainty can be properly quantified (Young et al. 1996). Bayesian approaches are, of course, only really worthwhile where we know more data may be forthcoming. This is not always the case in the kinds of systems agent-based modellers deal with, at least currently.

### *15.2.4 Classification Discretisation Error*

Almost all data is an aggregation or interpretation of facts about the world. Direct measurements of unique physical properties are very rare, even in such apparently concrete subjects as physics. There will, therefore, always be some loss of information in data recording and use. Even in the event that our instruments are recording at an accuracy/precision we are happy with, we generally add an additional uncertainty, or find one introduced in post-production, through data classification into bins.

Binning data into classifications can be problematic, especially where classification schemes are multivariate and prototypical (that is, very broadly, objects are classed by, and into, examples). Real-world membership of a set is usually fuzzy, and fuzzy sets are generally a more realistic way of dealing with the world. Where crisp sets are needed, entropy statistics can be used to represent multivariate classification uncertainty, and their relative simplicity provides a useful option for spatial mapping (van der Wel et al. 1996). The more common uncertainty, however, usually concerns the granularity of the bins and where the original data point fell within the bin. Data that has already been binned appropriately is not usually problematic where we have control over it (the uncertainty is easily quantified, and can be included in with other assessments of precision). It is only where we wish to use the data for other classification systems or as a continuous dataset, that binning causes issues. For numerical data, there is the possibility of shifting the data back into a continuous sequence by

stochastically distributing the data within each class to match an overall distribution fitted to the totality of the classified data. However, once such a distribution has been identified, sampling directly from the distribution becomes simpler.

Redistributing the original sample is only really worthwhile if the classified data is n-tuples, carrying ancillary data with the data that was originally binned. One common use of such a redistribution is within spatial microsimulation (Ballas et al. 2005), in which a population of individuals, which have been lumped together at some geographical scale (say a national bin) are redistributed to smaller areas (say electoral districts bins) such that their distribution matches some statistic (say, employment) in that area. If people can broadly be divided into socio-economic types, with correlated traits, we might expect ancillary traits (say, newspaper readership) to be recreated in the smaller areas (with an error associated with the strength of correlation between the two traits). Such models are increasingly used as the starting conditions for agent-based models where individual-level census data is unavailable, though assessing the accuracy of the recreation of ancillary variables is not easy without detailed new sampling, because we're usually trying to recreate distributions which are essentially unknown. Generally even where we are just trying to recreate the location of individuals with a set of traits which we have constraining distributions for, the geographical location is rarely accurate; commonly individuals are assigned to the smaller geographical boundary set itself or randomly allocated a home within the area. More sophisticated pycnophylactic (Tobler 1979), or other types of reallocation, are rarely completed, meaning there is also a considerable distribution error within each area.

A further major error during classification is caused by conversion between classification schemes, for example the placing of classified and geographically binned census data into new classes and geographical boundaries (Martin et al. 2002). Usually error can only be avoided by aggregating up bins or spatial boundaries to some common aggregate level (for example, Martin 2003).

## 15.3 Uncertainty Associated with Model Choice

As well as errors and uncertainty associated with data, we recognise that there are also epistemic uncertainties: those associated with our knowledge of the system. Essentially we may regard ourselves as being on a fruitless quest: languages (computer code included) are not the physical real world. Not only does this mean that we are unlikely to ever get a perfect computational representation of the real world (what Faber et al. 1992, call *hermeneutic ignorance*), but it also means we're unlikely to ever understand it properly, as we simply don't have the tools to do so (Faber et al.'s *axiomatic ignorance*), and those we do have may be fatally flawed (Faber et al.'s *logical ignorance*, following Gödel's incompleteness work). Not only this, but we essentially have to limit our modelling attempts in a way that the interconnected real systems are not limited (the *closure problem*: Lane 2001). Nevertheless, as languages, and mathematics in particular, have shown, we can get

a useful approximation that carries us forward. This is especially true for real systems that are mediated though language. In this section we examine some of the epistemic uncertainties we will have to deal with to do so. We shall assume a simple model where input variables are utilised via some kind of weighting against each other, or mathematical relationship, or ruleset, and the component *parameters* of these forms control the conversion of the variable inputs to one or more model outputs. The parameters are *calibrated*, that is fixed based on the real world to give as realistic output as possible. The parameterised section of the model may include actions by one or more agents.

### *15.3.1 Error in Choosing Variables*

Simultaneous with finding data for our models is the process of deciding which data we are going to use, and which we are going to exclude. The tendency to load a model with variables is a particular problem with those branches of agent-based modelling where the model is developed to accurately replicate reality. An increased number of variables may lead to a more realistic model, but it also leads to increased levels of error through the need to calibrate more parameters (the so-called *Information Paradox*, O'Neill 1973; Rowe 1977). Moreover, added detail often adds little to a model, and a shift from parsimony can obscure simpler models that perform just as adequately. Generally measures of model quality trade off accuracy of representation against model complexity (see Spiegelhalter et al. 2002, for a discussion of classical and Bayesian methods for achieving this tradeoff). This said, however, the option for adding additional variables is sometimes worth investing effort in early in the modelling process. Flexible code that allows for the addition and removal of variables through a well-structured object hierarchy and generic programming (parameterized types: Gamma et al. 1994) will pay considerable dividends on its investment in the longer term.

While we would hope that the choosing of variables was part of a linear progression, from thinking about the system we are interested in, to deciding how to model it, to picking data, it rarely works out so simply. Investigating our data often suggests we may have to settle for different, less than ideal, proxies for the data we would like to have, or, indeed, different data altogether. It may also be that we are using too small or too large a number of variables to represent the system (O'Neill 1973's *aggregation error*). To some extent having too many variables should reveal itself through covariance, but having too few variables, or the wrong type of variables, will result in errors or biases. In addition, there are often problems of scale: we may misunderstand the boundaries between objects in the real world (Suter et al. 1987), or, more simply, have the wrong time or spatial scale for the model.

There are broadly two sets of techniques for choosing/excluding variables. We can either examine the real system statistically, independent of the model, to see which variables might be appropriate, or we can run the model and use its ability to predict the real system to determine how well we've identified the variables needed.

The former methodologies have the advantage that we are dealing directly with the system. However, for agent-based models they have the issue that they tend to assume aggregate statistical tests on lumped data can identify variables acting at the individual level. With model-testing, we often assume our model form is appropriate, and any differences between the model outputs and the real world are due to poorly chosen variables/parameters, which is plainly untrue. However, the advantage with this approach is that testing is achieved at the same scale as the final model.

In both cases, the choice of variables is often (though not always) compared with a single dependent predictand, with the strength of the relationship being used to exclude variables. It should be noted that is not necessarily ideal. Utilisation of a single output statistic (or, indeed, multiple statistics) is always going to be problematic, as it will fail to calibrate the system to the nuances of the detailed individual characteristics of the system (Wagener et al. 2003) even if the model is at the individual level. Optimisation against a single output may only be sufficient to identify between three and five parameters with any accuracy. It may be necessary to consider multiple outputs to gain any further distinction (Wagener et al. 2003). Moreover, following Benjamini and Hochberg (1995), there is an argument that more attention should be given to the false-positive (Type I) errors when variables are kept, to ensure than random variation doesn't allow in variables that could be trimmed out (Green and Babyak 1997). The probability of Type I and II errors in multi-model assessment can usefully be balanced with reference to the costs to policy makers that result from the different errors (Hartley et al. 2006).

In the first category of techniques, examining the real system, the simplest method is just to examine the size of the variables. For linear models, variables can be removed on the basis that smaller variables are less likely to have an effect than larger ones, and small co-varying variables, particularly those on the same time-cycles, can be removed or aggregated (O'Neill and Rust 1979). However, this is less possible for non-linear models, where small variations in variables can have large effects. Looking in more detail at the relationships rather than the size, Stepwise Linear Regression has been used since the 1940s to exclude insignificant variables (Glahn and Lowry 1972). Although the core technique is broadly distribution-insensitive, it does assume variables are uncorrelated and related to the final dependent variable linearly. Stepwise variables really need to be on a common range to avoid size effects. While there are issues with this (see King 1986), a range transformation can aid when working with some non-linearities.

Where there is co-linearity between variables it may be that an underlying variable or process may be responsible. While we may be able to tease apart the relationships with an instrumental variable approach, the usual method for proceeding in such cases is to use Principle Components Analysis (PCA) to combine variables into independent components representing the latent variables. This can both indicate variables that are essential/non-essential and provide combined-variable components that represent the missing 'true' variable influencing the system. PCA analysis of model parameter sets following calibration runs can additionally reveal potential points of investigation for new processes not directly captured by the model (Keesman and van Straten 1990). Plainly, we may also find ourselves in the situation of having 'known

unknowns' – knowing a variable is missing, but being unable to discover what it is. Provided we know something of the part played by the variable we may still be able to represent such unknowns as latent variables within an agent based model, as they are within Bayesian (Kavetski et al. 2006), Hierarchical Bayesian (Clark 2005), or Structural Equation Modelling. In these techniques the explicit representation of uncertainty usefully shifts the models away from only assessing uncertainty at the level of inputs (through Monte Carlo sampling) and outputs (Clark and Gelfand 2006). However, embedding Bayesian techniques themselves, for example, is not always simple in agent-based models, not least because Bayesian assessments of any detail often rely on an assumption of independent Gaussian output noise (see, for example, Kavetski et al. 2006).

In general, for non-linear systems that are sensitive to small variable changes, it is usually the case that attempts to identify variables statistically from the original data are of limited success. For such systems we really need to consider all possible variable combinations and their effects on model runs, though generally a subset of the combinotronic space is used. There is a large literature on variable selection that utilises models. George (2000) provides an overview of the key issues. Statistical representations of the model may suggest the number of parameters that can reasonably be extracted from the data (e.g. Young et al. 1996), but more usually selection proceeds by running the model with a set of variables and assessing how well it runs, either through significance testing (for example, in Structural Equation Modelling: Green and Babyak 1997) or, more commonly, by ranking the errors associated with different selections.

The spread of values of parameters that match model inputs to model results can tell us if the associated variables are important to the sensitivity of the model. If we are confident in our model structure, parameters which vary a great deal between calibrations while still producing viable results may not be especially *important* to the detailed behaviour of a system (Spear 1970; though see below) and might be discarded. The Generalised/Regional Sensitivity Analysis (GSA/RSA) Hornberger-Spear-Young Algorithm utilises this *rejection sampling* and Monte Carlo testing of inputs and parameters to determine which variables a model should contain (Hornberger and Spear 1981; see Beck 1987; Young et al. 1996 for summaries of developments). Although such techniques tend to be tied to statistical models, the general principles are applicable in agent-based systems. A popular alternative to GSA, sometimes merged with it in hybrid methodologies, is to allow the weighting of variables to be dynamically set during single model runs, and to prune weightings associated with the variables dynamically as model calibration moves towards highlighting some variables over others (essentially the non-linear equivalent of the above pruning of small-sized variables). This pruning can, for example, be done with a Bayesian approach (George 2000). Of course, the danger with this is that parameters extracted from the real-world system may not be stable, and the relationships as represented may vary (Matott et al. 2009). In one hybrid example, Wagener et al. (2003) suggest that by splitting up the parameters' range and different modelling time-windows it should be possible to identify which parameters are important at specific model periods. This also allows an assessment of the sensitivity of specific model components formed by combining parameters.

Looking at parameter variation has another useful by-product: variation over time may also tell us whether variables are *missing*. When we think that variables should be related in a stable fashion, variation may result from the current parameters adapting to make up for missing parameters (Beck 1987). Moreover, Beck (1987) and (for an agent-based system) Heppenstall et al. (2007) have suggested that for recursive estimation and Genetic Algorithm based parameter calibration respectively, trajectories through parameter space may reveal underlying processes in the real data. Beck notes that calibration can often clash with model forms, suggesting adjustment is necessary.

### *15.3.2 Model Representation – Is This the Right Functional Form?*

Even if we can correctly identify the variables involved in our model, we still have the potential for *model error*, that is, error in our final outputs resulting from a structural problem with our model. We need to tackle the *identifiability problem*, for both variables and the relationships between them captured in the model. In general this is not an area of error much investigated by agent-based modellers. This perhaps reflects our general feeling that we are better placed than most modellers to claim our models match reality and are directly representative of true objects and relationships in the world. Even if we believe our agents represent active agencies in the real world, we can be much less certain that we have no *functional error* (van der Sluijs et al. 2003), that is, that we are using the correct relationships between them.

In general agent rulesets will be built up from other studies that generate/test hypotheses about relationships in the real world, and give them a significance value that, broadly, represents the likelihood that the relationships are not falsely identified as real. Plainly there are potential errors here associated with identifying the incorrect hypothesized relationships, and most statistical tests include terms to allow for sample size and degrees of freedom, and will have a particular power representing the likelihood of false positives and false negatives. The question then, really, is how these sub-models/rulesets are combined when no, or relatively little, information on the combination process exists. Frequently this combination in agent systems is achieved through choosing weighted elements based on a ranking process, or combining them arithmetically, but there are many alternatives (see, for examples, Wooldridge 2009). This problem arises beyond areas of, for example, decision making – we may lack a coherent understanding of even relatively deterministic elements of the model.

On the simplest level, we can examine the performance of a single model run under different starting conditions and parameterisations to gain an idea of the range of probabilistic outcomes. Differences between the space of model responses and the real data may allow us to explore model deficiencies and even go some way to separating out model structural error from input uncertainties (Keesman and van

Straten 1989). Alternatively, we can build our models by evolving them to have the right components, through Genetic Programming, with sub-models as genes (see Poli et al. 2008 as a starting point).

However, multiple model testing is now becoming the preferred option in many modelling fields. Indeed, if one looks at subject areas where models are entrenched in the testing of hypotheses, multiple model testing is replacing single model vs null hypothesis testing as the standard methodology (Johnson and Omland 2004), with the likelihood of gaining a correct hypothesis considerably enhanced by multiple hypothesis testing as significances can be ranked and filtered, and likelihoods enhanced through Bayesian techniques (see Farcomeni 2008).

A general methodology for multiple model testing of parameters was developed by Hornburger, Spear, and Young (see Sect. 15.3.1, above). We shall come back to examine this in detail when we look at calibration, however, the basic idea is that multiple models with different parameters are run and only those models that can hit a given set of targets are kept (so called *rejection sampling*). This algorithm was developed into the GLUE (Generalised Likelihood Uncertainty Estimation) procedure by Bevan and Binley (1992). This utilises multiple model runs which may vary in form or parameterisation, and assigns a likelihood to each. Results can then be ranked by likelihood and/or summary statistics generated by weighted combinations of the predictions. Poor models sets can be removed when new data from the real world is available to validate against. O'Neill et al. (1980) have suggested that by filtering out model runs by validation criteria at different stages of the model evolution (e.g. days 10, 20, and 30 of a model run) it is possible to constrain the error of the final models that survive. The days for this filtering are best taken when the inter-model variation is high (O'Neill et al. 1980). By adjusting the parameters on the basis of their co-variance it is possible to reduce their error further (O'Neill et al. 1980). Gupta et al. (1998) extend these broad techniques to multi-objective (~output) models and review alternative developments.

The potential for combining model results under GLUE marks it out as an early basic example of a broader set of methodologies for *Multimodel Ensemble Forecasting*. With ensemble modelling the issue of which model to run is avoided, to an extent, by running multiple models and then selecting the best or combining their results. In ensembles, one can either run very different models, or the same model can be run multiple times with a variety of initial states drawn from the potential distribution of real conditions and their potential errors. Once ensemble models have run, they can be combined to give an overall prediction including an uncertainty measure dictated by not only within-forecast variation, but between forecasts as well, for example, using Bayesian Model Averaging (Raftery et al. 2005). In general, combining multiple predictions will improve forecast reliability in the same way that generating the mean of noisy data is usually a better estimate of the true value than picking a single sample (Leith 1974). The combination of predictions means that forecast *sharpness* (closeness of forecasts) can be assessed as an additional uncertainty measure (Gneiting and Raftery 2005). A good review of multi-model selection criteria and combination techniques can be found in Johnson and Omland (2004). Generally multiple-model ensemble methods are most frequently used in climate/weather studies and hydrology.

They are rarer elsewhere, where single models with randomisation of key components and a probabilistic assessment are more likely (Brown 2010; for a review in meteorology, see Gneiting and Raftery 2005). This reflects the considerable costs involved in multiple model development and the limited number of researchers working in very specific fields, particularly in the social sciences.

### *15.3.3 Picking Scale*

One of the problems with data-driven identification of models/variables is that the system explicitly represents the spatial and temporal scales at which the data is sampled, rather than that most appropriate for the system (Young et al. 1996). To an extent this is mitigated in agent-based systems which have the potential for modelling different components at the appropriate spatio-temporal scales with less of the cross-scale errors that creep into other kinds of models. Multi-scale modelling and validation where there was any doubt would be an ideal solution, but data and computational effort are strongly limiting factors in this. To an extent the issue can be investigated by using cross-scale validation techniques (Costanza 1989) both during calibration and to examine key scales at which the model best represents the system (Malleson 2010).

### *15.3.4 Model Fitting – Picking Parameters*

For any given sub-component of a model there is usually a need to estimate parameters from the real world as represented in training datasets. Such parameters are almost certainly going to be a 'fudge' on real-world processes, and therefore be associated with errors of verisimilitude, and there will be additional errors associated with accurately estimating them: inversion errors of picking the correct weights from the vast number that may model training datasets, and accuracy errors associated with picking them well. A large number of parameters can potentially lead to cryptic equifinality, with erroneous models matching training data simply by providing so many tunable parameters that they can match any function. Having a wide variety of parameters may also enhance overfitting unless care is taken to prevent it, that is, model weights may be adjusted to the point that they very accurately model training datasets, but don't have the flexibility to capture the alternative behaviours of the real system. There is a perverse relationship between the number of parameters and overfitting because poor models with lower levels of parameters won't overfit, whereas those with the right number are more likely to do so. The usual solution to overfitting is to reserve some data as a test set not involved in training, but this is often difficult to justify where data is thin on the ground or critical and unique, as it tends to be in social sciences, and, as mentioned above, spatial and temporal autocorrelation can cause problems in determining the appropriate size of dataset necessary to do a good calibration job.

More generally the errors associated with parameters are adjusted largely by minimising the error of the output of the model, either assuming that the error is entirely due to the parameters being mis-calibrated, or trying to segregate the errors from different sources, i.e. inputs, parameter calibration, and the model form. A key element of this may be *sensitivity testing*: perturbing inputs and/or parameters to see what the result is on the final model output. This allows an assessment of the importance of the input/parameter on the model behaviour, and, if the perturbation is drawn from an input error distribution using Monte Carlo sampling, an idea of how much those errors change the range of results (*uncertainty testing*). The standard text on sensitivity testing is Saltelli et al. (2000), but a good introductory review is Hamby (1994).

Once errors are assessed they can be used to adjust the parameters to improve the match, either statically, at the end of the model run, or dynamically as the model is running. The estimate/adjustment of unknown weights associated with variables can be achieved in a variety of ways:

1. through expert/stakeholder advice,
2. real-world experimentation in aggregate,
3. or automatic fitting to known input–output data.

#### 15.3.4.1 Expert Advice

There are a wide range of methodologies for involving experts and stakeholders in model design and assessment. At the simplest, this involves expert *face validation* of parameters determined automatically, that is getting experts to agree the model looks ok. Seminal work on the process and problems of eliciting uncertainty assessments from experts was presented by Spetzler and von Holstein (1975). More recently a sophisticated expert-analysis process, which includes quantitative sensitivity testing, was developed by Funtowicz and Ravetz (1990). Their NUSAP (Numeral Unit Spread Assessment Pedigree) methodology builds up a 'pedigree' for a model based on evidence including expert opinion on proxy use, the empirical basis of parameters, theoretical understanding, methodological rigour, and model validation (Van der Sluij et al. 2002; http://www.nusap.net). Alternatively expert advice can be incorporated at one remove from the assessment process, by getting experts to design the metrics for uncertainty assessment, rather than completing the assessment themselves (Bevan and Binley 1992).

At the other end of the scale, experts can directly choose parameters. Because of the complication of most models and the lack of absolute verisimilitude, it is rare for experts to choose the values that parameters are fixed at. It is more usual for expert advice to be used in initialising weights that are then adjusted through calibration against the real world. For example, expert advice can be: incorporated into the development of priors in Bayesian treatments of parameters/parameter uncertainty (for a summary, see Clark 2005; for a clear discussion on options for very non-informed priors, see Kavetski et al. 2006); used to constrain the ranges parameters

are sampled from (Lutz et al. 1996); or alternatively incorporated though the development of inputs or parameters as fuzzy sets (Janssen et al. 2010). The balance between automatic calibration and expert input can vary considerably, with attempts made to integrate expert calibration into an otherwise automatic procedure (Gupta 1999) and to replicate the actions of experts automatically (Boyle et al. 2000).

#### 15.3.4.2 Real-World Experimentation in Aggregate

Sadly governments seem strangely unwilling to give agent-based modellers the complete control over national policies they need and deserve. Real-world experimentation in aggregate is more common in the physical sciences, where ethical issues play out less. For social science modellers, such parameters are usually taken from the quantitative literature outlining statistical treatments of society, but these more rarely generate laws and sets of parameters that can be built directly into larger-scale models in the same way. Large scale experiments to derive rulesets are rare, even in these days of internet data collection and citizen scientists.

#### 15.3.4.3 Fitting to Known Input–Output Data

Most commonly models follow a process of *data assimilation*, in which forecasts (or, more rarely, backcasts) are generated and compared with real-world data, with the model being adjusted automatically on the basis of the difference. With agent-based models this adjustment is commonly a static process – the model runs to some completion and then the adjustment takes place. This is because agent-based systems are generally initiated and allowed to run on their internal dynamics without the injection of external driving data as the model progresses. However, sequential/dynamic data assimilation (that is, adjustment as the model runs) is common in other fields and likely to become an increasingly important element of agent-based modelling as it attempts to take on predicting large scale and dynamic socio-economic systems (as we shall see, machine learning does represent a middle-way taken by many agent-based systems).

The calibration process has to find optimal parameter weights in a variable space of potential solutions. For simple mathematical functions with a limited number of variables, the technique used is usually to assume the function includes one or more error terms, and then to fit the function to the data by minimising the error term. The classic example of this is linear least-squares fitting, which seeks to place a line representing data through scattered data points by minimising the residual error between the line and the points along its length. Such techniques make a number of assumptions, not least that the errors are random and limited to specific variables. For example, the standard least-squares method assumes there is only an error on the independent variable, not the dependent variable that is being predicted. This is rarely the case where two datasets are being used to derive model rules.

Unfortunately for most agent-based models the non-linearities and considerable interactions involved render mathematical treatments impossible for almost all components. The solution spaces involved are complicated and too extensive to try all parameter combinations. In the absence of expert advice and experimental results, we are usually left with imputing the parameters from data. The worse-case scenario is where we have clear input data, but only a very qualitative understanding of what potential outputs might look like (for example, in predicting urban form). Choosing parameters by manually manipulating their values to see what gives appropriate-looking results is generally to be avoided. The inversion problem plays out particularly badly against researchers with limited time on their hands and it is likely that local or sub-optima will be chosen. Nevertheless, this technique is frequent in agent-based modelling, as the computational resources needed for model runs are high (removing automated checking as an option), and the variables are often interdependent in non-linear manners (rendering mathematical optimisations inappropriate/impossible). Experts should always be involved in the face validation process where it can't be avoided to limit the potential errors.

Where the computational demands are less restrictive, but still prevent a full characterisation of the solution space, we have the option of adjusting the parameters through either a greedy algorithm (adjusting the weights by some rule and keeping those changes that improve the final fit) or some mathematical equivalent (distributing the error to individual components and adjusting them to reduce the local error). As part of this data-led process we usually have to identify some optimisation function to minimise (usually the error between reality and the model output, but not always), and heuristics to control the selection of adjustments.

Standard treatments in non-agent-based models are, at their simplest, recursive greedy treatments with parameters updated on the basis of new data (Gupta et al. 1998, review standard methods for multi-input/multi-output calibration). Many modelling techniques rely on transfer functions to convert between input sources and output objectives (one can visualise a matrix that stores the functions that convert between the two). Given output errors, it is possible, if we know the form of the relationship between input parameters and outputs, to estimate the error in the functions' parameters (Beck 1987), in a manner similar to back-propagation in neural-networks (though with a more flexible set of functional relationships). A great many techniques rely on linearising these functions through Taylor expansions for key conditions or dynamically, as a precursor to allocating error to the parameters. As one can imagine, the mathematics of updating the associated parameters becomes quite complicated. Many models rely on Bayesian methodologies to cope with the updating process, though this is still far from simple. Furthermore as many inputs and parameters are non-normal and cross-correlated, inputs and parameters are often sampled using Monte Carlo techniques when looking at the error in the model due to noise and calibration issues. Generally the sensitivity of non-linear models to small changes in parameters means that multiple parameter sets need to be tested uniquely (Smith et al. 2010a). Multiple runs of the same model utilising different starting conditions and parameter sets allows for the quantification of the error and its effects and the use of this information in the updating process. The full adjustment process is therefore

of considerable complexity. While most such techniques are heavily embedded in statistical modelling, it is nevertheless worth considering the application of their core ideas to agent-based modelling.

The standard technique used is the *Extended Kalman Filter*. The idea behind an Extended Kalman Filter is essentially that we know the output (real predicted values plus model-caused error) is a function of the model components, i.e. the model inputs and parameters, along with errors associated with both. Knowing the output, real values, input values, parameters, and the error associated with the input measurements, we can estimate the remaining missing element, the parameter errors, and adjust them on this basis. The parameter error is only an estimate, and (with an adaptive Kalman filter: see Evensen 1992 or Young 2002 for an introduction) will change with each new input/output pair, but if we know the parameter error, even roughly, we can adjust the parameters to remove that error. When, as usually is the case, there are multiple outputs from the model and multiple parameters, the adjustment is in the form of a Kalman gains matrix, which is used to adjust the parameters' actions in the next iteration along with the error value. The process generally moves recursively. The uncertainty is usually represented through Bayesian-like probabilities (as the error cannot be assumed Gaussian, these are usually dealt with through Monte Carlo methods: see Young 2002 for an introduction), and the adjustment takes place preferentially when we know more about the real world than we do about the model (i.e., there's no adjustment if we're more sure about the model than the current real-world values). Beck (1987) gives a summary of both this technique, and recursive estimation techniques in general, along with a summary of the issues with Extended Kalman Filters, chief of which, from our point of view, is the usual assumption of Gaussian input noise throughout. To gain a best estimate of parameters where there is error, assumptions must be made about the variables the error relates to directly and the error distribution (Smith et al. 2010a), but this is frequently not well characterised for social-science models.

Generally when multiple model runs are used with an algorithm from the Hornburger/Spear/Young family the spread of results gives a minimal estimate of the parameter uncertainty (Gupta et al. 1998). However, with multiple models there is the potential for intelligently utilising cross-model comparison to further limit the parameter uncertainty. The Ensemble Kalman filter, after Evensen (1994), can be utilised on single-model multiple-run ensembles to reduce the combinatorial load needed to characterise the parameter change. It uses Monte Carlo sampling (commonly Markov Chain Monte Carlo) to take an initially naive distribution for each parameter and update it using a Bayesian treatment of new data to gain a better parameter distribution. An alternative methodology by Toth and Kalnay (1993) utilises the differences between perturbed and unperturbed ensemble models to adjust the unperturbed models, removing potential errors caused by specific system instabilities. Of promise is also the SIMEX methodology (Cook and Stefanski 1994) in which a system that has well-understood input errors has increments of those errors added to the inputs across multiple model runs, and the output error assessed. As the output error increases tell us about the relationship between the stepped input and output errors, the remaining output error due to poor parameterisation can be identified as the equivalent of the

intercept on a graph of input vs. output errors (in a perfectly modelled system). By building a relationship between the final and input errors, it is therefore possible to estimate the parameter error, and, thereafter, to correct the parameters. Chowdhurya and Sharma (2007) review the literature on this technique, the adjustments necessary under a variety of conditions, and compare it with methodologies like GLUE. More generally, ensemble re/starting conditions can be subjected to a variety of algorithms, including evolutionary algorithms, to constraint the errors and lower computational effort (see NRC 2006, for a review).

In general, however, agent-based social and ecological modellers don't tend to follow the techniques generated in other fields of similar complexity. Instead, they are turning to artificial intelligence (AI) mechanisms to calibrate their models. In part this is because most agent models are extremely computationally expensive to run, but the subject area doesn't have the computational, personnel, or data resourcing seen in, for example, climate modelling. AI represents a sound and relatively fast method of calibration. It is usual for most models to be a mix of parameters fixed on the basis of the literature, parameters fixed by an AI method like a Genetic Algorithm, and parameters that vary stochastically across a distribution, picked with Monte Carlo sampling. Such multi-method models are difficult to assess for parameter quality except by validation of their outputs, though there is no reason some of the algorithms above could not be applied to elements of the models.

In addition to fixed parameters, most agent-based techniques include some form of machine learning, essentially doing the same job as dynamic data assimilation for a limited sub-set of parameters; parameters are derived by experiencing the system and optimising behaviour based on one or more objective functions. These objective functions are generally more internalised than simply the error between model outputs and the real world. In many senses agent-based modellers would rather see a model that learns well, than one that minimises an output error, but which has unrealistic internal dynamics. The problem is, of course, that such learning is hard to assess for reliability, except to the degree to which the overall model works.

Where information comes from experimentation or the literature, rather than model testing, confidence intervals are usually used to represent input and parameter uncertainty because inter-relationships are rarely known (Young et al. 1996). Confidence intervals for model parameters are more difficult to calculate properly when there is covariance between inputs/parameters, when the solution surface is complicated, and where input errors are poorly understood (Gallagher and Doherty 2007; who give some indications of ways forward). Under these conditions, and for relatively simple parameterisations, uncertainty associated with inputs can be represented through a sensitivity coefficient matrix – more detail on these will be given below, but essentially they are the covariance matrix showing how the output/s change as each input varies. In terms of parameter uncertainty, for large numbers of more complicated tests statistical significances can be generated to reduce parameter errors, with significances adjusted to pare down the potential for Type I and II errors, which would be high using traditional one-test *p*-values (Farcomeni 2008). The variation during parameterisation can also be used to give uncertainty statistics (see Matott et al. 2009, for a review). Equally, some calibration tests, notably those

based on Fuzzy or Baysian and/or Monte Carlo techniques can give uncertainty estimates (Keesman and van Straten 1990; Kennedy and O'Hagan 2001; Clancy et al. 2010) and assign likelihoods to parameter sets (Mitchell et al 2009). However, while such tools exist, it is nevertheless more common in model calibration to simply take the best result without considering the potential identification error.

So far we've dealt with uncertainty as something that is outside of the model and to be assessed for reduction. A more realistic way of dealing with it may be to build it into the model, so that the model reacts as more information comes online (through Baysian probabilities or more general Dempster–Shafer methodologies, or by including a more explicit error distribution in the model) or, furthermore, to assume that such an uncertainty is inherent in reality through the use of Fuzzy Sets and Logic (see Hassan et al. 2010, on agent-based systems; also Zadeh 2005 which goes further in handling modelling uncertainty explicitly using methods including Fuzzy Logic).

More generally, however, there is a fundamental question to be asked about many of these calibration techniques, including those used currently by many agent-based modellers. Many traditional model calibration/inversion techniques fail to cope with agent-based systems simply on the basis that they adjust parameter weightings to an average across a system, which isn't what an individual agent would respond to. Ideally each agent needs calibrating separately, rather than picking up average behaviours. However, if traditional calibration is to be utilised, the space to explore for individual calibration is considerable and the number of parameters fitting the system very large. In this sense, giving each agent some degree of machine learning may be the closest we can get to appropriate parameterisation in agent-based systems.

## 15.4 Model Mechanics – Errors Generated by Running the Model

In general, agent-based modellers assume models run well, not least because processor time renders multiple-platform runs difficult. Our confidence in this matter may be misplaced. *Model-fix Errors* come in when elements are added to the model that are not in the real system, either for simplification or because an element of the real system is not understood (van der Sluijs et al. 2003). These errors can be distinguished from *Process Error*, in which a complex element of the real system is simplified to make calculation tractable (van der Sluijs et al. 2003). On top of such accepted errors, it may be that our software is not well formed, either because of software bugs, or because the digital precision needed is not sufficient.

### *15.4.1 Model Bugs*

It is an unpleasant truth that many of our models probably contain coding errors. Les Hatton produced a devastating report on the quality of coding in industrial programs influenced by academia (Hatton 1997). He noted that on average the C

programs he looked at contained 8 lines hiding serious faults per 1,000 lines of code. Programs written in the academic favourite Fortran were generally so over-parameterised and poorly written that the average rose to 12 lines in 1,000. Moreover, the situation with Fortran wasn't helped by the fact that software written in Fortran contained 2.5 times as many lines as the equivalent C software. When a single algorithm for seismic processing was tested with the same data across multiple programs and platforms, Hatton found that the results were only comparable to within one significant figure. Before anyone gets too smug about this never happening in their code, let us not forget that these programs, albeit starting out as academic software, were finalised by software houses who work to specific quality standards and testing regimes. Some recent changes in programming will have reduced these issues: the removal of some error prone areas of code, such as pointers, from languages like Java will have helped considerably, as will the rise of Programming by Contract, Unit Testing, and the inclusion of Assertions. However, it remains true that most academic code, particularly that written in older versions of languages like Fortran, is likely to be replete with issues. Galán et al. (2009) offer practical advice, for agent-based modellers specifically, on model verification and code-checking.

### *15.4.2 Uncertainty Due to Representation*

Computers can only usually hold memory-limited binary representations of numbers. As such, some numbers are, by necessity, stored as approximations. Such digital imprecision can, if unchecked and/or propagated, result in catastrophic macro-scale errors (see, for an example, Hayes 2003). Good, programmer-centred, discussions on mitigating this issue can be found in Warren (2002) or Hyde (2004), while Izquierdo and Polhill (2006) and Polhill et al. (2006) provide sound practical advice and concentrate specifically on the propagation of these errors in agent-based modelling. Ultimately, however, the issue is constraining. The unification of most platforms around IEEE 754 as the standard for floating-point arithmetic has helped coders at least tackle the issue consistently (though utilising IEEE 754 standard routines in some languages is still far from direct – yes, Java, I'm talking about you). Nevertheless, one still has to take care with the transfer of code involving other data types between platforms (for example, the maximum integer size can change considerably). In general it is good practice to assess model error due to differences in processor, complier, and memory architecture, by transferring models to different platforms. However, the implementation of such transfers is limited by the lack of common code representation schemes of sufficient detail and the coding time needed. Common runtime environments such as the Java and .Net virtual machines mitigate the effort required to some extent, but don't stress-test code to a great enough degree as some issues that usually play out more apparently on different platforms are ameliorated at the virtual machine level. For problems of representation specifically, efforts to work using *Interval Computation* (essentially arithmetically treating the potential upper and

lower bounds of representations as they interact) seem promising, if at an early stage for complex models. A good introduction can be found in Hayes (2003), while further material can be found at http://www.cs.utep.edu/interval-comp/.

Related issues for spatio-temporal modellers include the granularity within which space and time are represented, which controls the *resolution* of the data – the size of the smallest useful object. These issues stretch from the appropriateness of different styles and sizes of neighbourhood in Cellular Automata (see Moreno et al. 2008), through to the synchronous or asynchronous updating of agent states (see Schönfisch and Roos 1999). This is a vast area of potential error; however, in general, the most recognised response is to build up models across a variety of tested landscapes, starting with abstract plains, and to test models on multiple systems as above.

## 15.5 Output Uncertainties

More often than not, the problems involved in quantifying input and parameter uncertainties mean that agent-based modellers deal with uncertainty at the point of model output. While outputs can be assessed for overall uncertainty, it is also at output that we most often consider the representation of uncertainty to stakeholders, and the recording of uncertainty in metadata.

### *15.5.1 Assessing Overall Uncertainty*

In general a large number of agent-based studies either make no direct comparison with the real world (in the sense that they are abstract behavioural models), or treat the error between predictions and reality as the single expression of model uncertainty. If this error is low, the assumption is that inputs are realistic and parameters well estimated. While there is some truth to this, such characterisations give us little idea of how a model will respond to change, or where the model or data needs investment. If, instead, we can examine the contribution of specific input, parameter, and model-form errors to the final prediction we stand a better chance of commenting on, and tackling, these issues. Of course, if a model isn't sensitive to errors, it matters less if they are present; but if a model changes in a strongly non-linear fashion under error, then that has important ramifications for its predictive power.

Traditionally the contribution of errors in mathematical models is examined by tracking the noise from the inputs and using the difference between model outputs (including the noise) and the real world (the, so-called, *prediction error*) to estimate the errors due to parameters. Generally a traditional error propagation/sensitivity analysis utilises the following formula (commonly after Ku 1966), which gives the standard deviation of the results $Y$ of a function, based on the standard deviations

($s$)/ variances ($s^2$) of the input variables ($X$, $Z$…), and the relationship between each variable and Y:

$$s_y = \sqrt{\left(\frac{\partial Y}{\partial X}\right)^2 s_x^2 + \left(\frac{\partial Y}{\partial Z}\right)^2 s_z^2 + \cdots + \left(\frac{\partial Y}{\partial X}\right)\left(\frac{\partial Y}{\partial Z}\right) s_{xz}^2 + \cdots}$$

where $s_{xz}{}^2$ is the estimated covariance between the variables $X$ and $Z$, and $\frac{\partial Y}{\partial X}$ is the partial derivative of the function $Y$ with respect to $X$, known in this context as the *sensitivity coefficient* (see NIST/SEMATECH 2010 for a summary). Where more than one output value is predicted, the equation needs expanding to Jacobean matrices (that relate each variable to each output via a partial derivative). Even when the relationship between Y and each variable is poorly characterised, as long as the variables can be shown to be independent (i.e. with no covariance), input variation can be empirically correlated with outputs individually to give the sensitivity coefficients (O'Neill et al. 1980). For independent variables and relatively simple relationships this leads to reasonably simple predictors for error which can be used within models and which can give rankable information on the importance of variables to the model sensitivity and confidence intervals (Walker 1982).

However, there are considerable issues in applying this methodology in the kinds of systems agent-based modellers deal with, and the kinds of models they generate. Variables are rarely completely independent in non-linear systems, and in such cases a more sophisticated development based around variance-covariance matrices is necessary (O'Neill et al. 1980; van Straten 1985; for developed details, see Beck 1987). In combination the error terms can gain strange distributions if the same variables link together multiple mathematical representations within a model (Tang and Wang 2001 – see references therein). If these relationships vary with time, the matrices may need updating with new input data iteratively (see parameter estimates, Sect. 15.3.4). In addition, spatial systems have their own problems, both with spatial autocorrelation of errors, and with large combinotronic spaces when multiple spatial locations contribute to a final prediction at one or more points. Heuvelink (1998) details the use of this technique when mathematically modelling simple spatial systems with well-known input errors, however, there are considerable issues with more complex non-linear spatial models.

In general, non-linear relationships are usually linearized in such treatments through Taylor expansions. This may be limited to points of assumed equilibrium, that is, where it is assumed that if there is no change in inputs there is no change in outputs (Young et al. 1996); but, as mentioned above, this is not always appropriate in the kinds of systems agent-based modellers tackle. Alternatively the linearization may be around dynamic model points, but such schemes do not cope well with the kinds of relationships modelled by agent-based systems, which tend not to be continuously differentiable, if, indeed, they can be represented mathematically at all, and where function-changing relationships between two variables and a third can make partial differentials difficult to work with consistently. Either way, for more complicated functions under large variances the first-order linear approximations generally

used introduce their own errors. Replacements for this technique, which take more account of the non-linear nature of most systems, still tend to rely on an overly mathematical treatment in which noise is regarded as an additional component to a signal (see, for example, Smith et al. 2010a, and for a review Matott et al. 2009).

Given these problems, for models of any level of complication it is usual to resort to Monte Carlo Sensitivity (MCS) testing, in which the model is run multiple times with input data perturbed in some fashion. In uncertainty testing the perturbations are usually drawn from the error distribution of the appropriate inputs, and the parameter distributions are usually also sampled to provide the parameters for each run. Although sensitivity testing can proceed by targeted 'manual' manipulation of the inputs, automated Monte Carlo sample selection based on input/parameter distributions is needed for full output distribution uncertainty testing. There are some broad variations on the scheme: Bayesian models, for example, generally explore uncertainty by sampling their parameter distributions, and then adding white noise to the inputs, while GLUE simply varies the parameter values (Kavetski et al 2006). Either way, multiple model runs using such carefully selected inputs and/or parameters allow for an assessment of the variation in the model outputs on the basis of their errors, and statistical summaries can be generated, along with confidence statistics. In this way, the technique avoids the middle stage of traditional error assessments: the stage of directly calculating the error propagation. A good introduction to Monte Carlo techniques in a spatial context can be found in Walker et al. (2003), along with references to work on sensitivity testing, while a clear detailing of the technique from the point of view of tracking input errors can be found in JCGM (2008a). A more generic study of uncertainty testing, concentrating on statistical summaries, can be found in Bobashev and Morris (2010).

Because the run-time of models can be long, Monte Carlo simulation of thousands of runs length may be inappropriate, even with parallel processing. Some spatial analysts have claimed that much of a distribution can be determined with a small number of runs (up to a hundred: Openshaw 1989; Bobashev and Morris 2010), but this is of considerable contention (Heuvelink 1998). Given this, more restricted tests have been devised which control the sampling of inputs to ensure a small but representative sample of their distributions is taken into account. One could, for example, sample the parameter space regularly: a so-called *Grid Sample*. While this has the advantage that it is easy to see the sensitivity of one parameter against others (Urban and Fricker 2010), this still generates large numbers of runs. More notable is the Latin Hypercube sampling technique (see McKay et al. 1979), in which each input is divided up into $n$ number of sections, each with an equal probability of occurring (i.e. for a normal distribution the sections are larger at the distribution limbs, where probabilities are generally lower). Each section is then Monte Carlo sampled once and only once. Each value from the series of sections for one variable is then combined with an equivalent value from each of the other variables, generating $n$ sets of input values. Essentially this ensures the full range of sample distributions will be sampled, but only generates $n$ tests. The combination of samples from each distribution can be random, or the combination can be chosen to enhance or dampen correlations between the values (for a discussion, see Wyss and

Jorgensen 1998; Urban and Fricker 2010). An excellent summary and pointers to the literature on sensitivity testing can be found in Wyss and Jorgensen (1998) which is associated with software for generating both Monte Carlo and Latin Hypercube input sample datasets. Of course, the use of this technique in testing output variation based on input error assumes the modeller has some idea of the distribution of errors in their inputs, which may not always be true.

Increasingly, in climate modelling, researchers are avoiding the use of full models in parameter sweeps. Instead, they train an emulator (for example, an Artificial Neural Network) on the outputs of a sub-set of parameter sweeps, and then use the emulator to predict the results of a more comprehensive sweep through parameter space with appropriate significance values for the predictions. Of particular interest is Gaussian Process emulators, which use the equivalent of Bayesian kriging to estimate the form of a solution space, in the same way that kriging can be used to estimate missing data in geographical space (Urban and Fricker 2010; see also Young et al. 1996, for a statistical approach that provides a statistical linearization of complex deterministic models). As with kriging, it may be appropriate to feed in training samples in areas of particular variation, worrying less about other areas (Urban and Fricker 2010).

In addition to quantifying uncertainty and error using the outputs, it is also possible to process outputs to reduce the uncertainty by redefining the objective function we are aiming at. For example, where thresholds are involved, uncertainty in models can be reduced by predicting event occurrences rather than continuous probabilities; indeed, generally the prediction of statistical aggregations of outputs, or aggregations related to model outputs can reduce the uncertainty if the relationships are more robust to variance (Glahn and Lowry 1972). More generically, *Forecast Post-Processing* can include interpolation and adjustment for biases and local conditions (NRC 2006). If the outputs are to be used in sequential/dynamic data assimilation (i.e. as the model runs and real data comes in) they will plainly have an effect on the non-linear behavior of the model, and filtering results to remove small-scale instabilities can stop non-linearities getting out of hand (Evensen 1992).

Finally, it is worth noting that in agent-based systems there is interesting work to be done at the meta-assessment level. One direction here is the push towards more objective and automatic hands-off model assessment by allowing meta-agents to assess the models (Li et al. 2007). A second area of interesting potential is the broadening of our criteria of assessment. It is worth noting, with Mearns (2010), that even with the best models, metrics of uncertainty may well increase in some modelling efforts before they decrease. The improvement of models is not always about improving very specific error metrics; structural change may bring greater verisimilitude and future error constraints, without these resulting immediately. We have to be wary of measuring success on the basis of error metrics. Indeed, it may be that with our software, like any engineered solution, we might actually wish to trade error off against alternative values, such as model versatility, adaptability, evolvability or interoperability, and there are a number of techniques from engineering that allow us to manually examine these trade-offs on a cost basis (see Hasings and McManus 2004, for an introduction). This may be an interesting area for meta-agents to additionally explore.

### *15.5.2 Representing Uncertainty*

In general model error is calculated as the total absolute difference between the real and predicted values, normalised in a variety of ways. The bias of a model can be examined by using the total non-absolute difference, as this will highlight consistently great or lesser predictions (if our inputs are reliable, a consistent bias is usually indicative of an issue with the model form). Despite these relatively simple outlines, where the error between a model and reality is given, the statistic used in detail needs considerable and careful thought, especially where the error is heteroscedastic. Gupta et al. (1998) summarise some of the measures commonly used for aspatial value predictions, especially multi-objective predictions that need combining, while Knudsen and Fotheringham (1986) discuss errors in the context of spatial predictions.

Uncertainty itself is usually reported as an estimated statistic (like the mean of model runs) and an uncertainty or set of confidence intervals. For Gaussian sample data, for example, this is usually the sample mean ± standard error. As Smith et al. (2010a) point out, this type of representation is appropriate for linear systems where behaviour varies predictably and slowly with a shift from the mean, but means considerably less in sensitive non-linear systems. In addition error measures like the standard deviation of a sampling distribution drawn from a Gaussian population are well understood for standard statistical estimators like sample means, and the biases between them and population figures are well characterised. The biases in the statistics can therefore be taken into account by readers or augmented when reporting results. For complex and novel model errors, however, this is less easy, and generally it is simpler to quote the distribution-free summaries of model runs. For example, model 95% output ranges are quoted more often than formal 95% confidence intervals (for reasonably clear details of generating confidence intervals from Monte Carlo runs, see Lodwick 1989; Heuvelink 1998; or Bobashev and Morris 2010). However, almost all simple metrics can hide considerable useful information; for example, with Bayesian predictions summary statistics usually hide the fact that forecasts are influenced by the prior belief used to initialise the system.

The relationships between model inputs and outputs can be represented, as discussed above, by sensitivity coefficients. Where the relationship is linear, standard regression between the inputs and outputs, along with a correlation coefficient, is useful, but this becomes more complicated with non-linear non-normal data. For non-linear but independent variables there are less powerful representations of the relationships between inputs and outputs that allow the contribution of the inputs to be quantified, such as the Importance Index, and Relative Deviation Ratio. For more co-linear variables, there is the partial correlation coefficient. A wide range of such basic sensitivity statistics are reviewed in Hamby (1994).

For spatial modellers, it is key to understand the distribution of uncertainty in space and time. Uncertainty can, therefore, usefully be displayed on maps. For example, based on output confidence limits, maps displaying all possible results within 95% confidence limits can be displayed (e.g. Hartley et al. 2006).

For ensemble predictions, Bayesian Model Averaging (Raftery et al. 2005) will produce uncertainty maps that take into account both intra and inter-model variation. Laffan (1999), Reinke and Hunter (2002), Drecki (2002) and Kardos et al. (2003) explore some of the theoretical issues and solutions associated with communicating uncertainty using 2D maps. Uncertainty representation in 3D spatial datasets is explored by Viard et al. (2011).

However, it is plainly important that we consider not only the display of uncertainty to other scientists, but also to policy makers and the public at large. This is equally plainly problematic, and an area in which contentions about the relationship between science, decision-making, the public, and trust are extremely likely to arise (see Brown 2010 for a review). Scientific uncertainty can be converted into policy reticence, even when the science points strongly to action. Equally, however, the exposition of uncertainty can lead to increasingly targeted investment in areas with high uncertainty (Suter et al. 1987). Agent-based systems, with their individual-level processes, may be well placed to bring policy-centred discussions of uncertainty back to a more detailed level of treatment (Zellner 2008), arguably lost for non-linear systems since the move to Monte Carlo assessments.

Shackley and Wynne (1996) discuss some of the mechanisms by which scientists mitigate the effects of uncertainty, while Walker et al. (2003) subdivide some of the uncertainties in ways more pertinent to the interaction of modellers and policy-makers (for example, they identify *scenario uncertainty*, in which it is not clear what scenario is going to occur). Specifically spatial uncertainties and decision making are examined from a policy-makers' viewpoint by Cornélis and Bruet (2002). Morss et al. (2008) give a useful template study for determining how the public understand uncertainty and want it displayed, while a detailed discussion of stakeholder engagement (a very large area) is provided by Dewulf et al. (2005). A formal approach to uncertainty in decision making may be formulated by embedding uncertainty representation within the demands of Quality Assurance (QA) guidelines (see, for examples, Refsgaard et al. 2005; van der Sluijs et al. 2003; JCGM 2008b), potentially including schemes designed under the ISO 9000 standards family. Such guidelines can also include detailed frameworks for decision-making under uncertainty (for an example of a formal quantitative decision-making framework centred on risk and uncertainty see Marin et al. 2003). On the flip-side, Brown (2004, 2010) and Couclelis (2003) provide usefully discussions on the place of uncertainty in science as a social process, and uncertainty's place in scientific self-reflection and knowledge production, while some of the more cognitive uncertainties associated with science-led decision making are described in van der Sluijs et al. (2003).

While we have dealt here with uncertainty associated with the advancement of knowledge, there is one further uncertainty or error that doesn't impact the quality of knowledge advancement, but is nevertheless important for scientists and society because it reduces the speed of progress: the uncertainty that scientific work is novel. Smithson (1989, 3) identified the difficulty of constructing models in a world in which scientific pursuits are becoming increasingly swamped by knowledge and separated into different areas. Recent developments have suggested that scientists

are already starting to "re-invent the wheel" (Monaco and Anderson 1994). With over 1,350,000 scientific papers a year published (Björk et al. 2009), this novelty error or (at best) uncertainty can only increase, and represents a very real threat to modelling, if not a significant barrier science needs to avoid as it becomes a mature human endeavour.

### *15.5.3 Metadata Systems*

A structured meta-framework for uncertainty may be built into the model itself, as it is in Bayesian treatments (see Clark 2005) or, for example, through Zadeh's (2005) GTU (Generalized Theory of Uncertainty). However, there are increasing efforts to develop separate metadata systems that focus on uncertainty assessments (Dassonville et al. 2002; Gan and Shi 2002). Such efforts are key to the chained interoperability of models, and the transmission of uncertainty with results. In particular, eXtensible Markup Language schemata that encapsulate uncertainty promise to take uncertainty recording and manipulation from the current level of the dataset down to the specific datum, storing detailed uncertainty information with each data point. A notable example for spatial modellers is UncertML (Williams et al. 2008; http://www.uncertml.org/) which has the potential to be used with the Geographical Markup Language (Cornford 2011), along with the web-based framework supplied to aid in its more general use, UncertWeb (http://www.uncertweb.org/).

## 15.6 Conclusions

> *To be able to predict only that all things are more or less equally probable is not a useful basis for decision making*
>
> M.B.Beck (1987)

All the above may seem terribly depressing. We work with non-linear, non-normal, high-combinatronic-space, models, highly demanding of computing power and memory storage. Why, then, do we believe we can do any better at modelling the world than astrology or the I-Ching? Are we not generating just the same kinds of largely random outputs and imputing meanings to them beyond rational boundaries? I don't think so, and part of the reason for this confidence comes down to the way the world works. By and large, at the scale at which we deal with it, the world is not completely random; rivers do not leap 50 m into the air and turn into a shower of goldfish; economic systems do not contrive to feed everyone bullion from ATMs. Systems are, generally, very stable compared with the wide range of potential states they could be in. Self-regulatory elements in the systems act to dampen the effect of noise and constrain the propagation of errors. However, my confidence rests in mod-

elling that concentrates to a far greater degree than we currently do in identifying elements that act to dampen systems and drive them towards attractors; elements like social negotiation, averaging, buffering, and thresholding. If we can centre our investigations of the real world on these, and then represent them in our models, we stand a much greater chance of building reasonable models of our apparently highly unpredictable systems.

Finally, it is also worth highlighting an alternative viewpoint put forward by Beck (1987), who suggests that rather than asking what the future will be, given parameters now, we instead ask what parameters would be necessary now, to create a reasonable future. It is generally true that as agent-based modellers we usually model current systems to predict what they might be like in the future, with very little reflection on our duty as academics to imagine a better world, and critique the fundamental components of the systems we are modelling. It is easy to point out when current policies will be disastrous, and even to see how small tweaks may mitigate this, but it is much harder for us to consider wholesale changes that might make the world a notably better, if stranger, place.

## 15.7 Further Reading

For a good overview of the subject area, which weights scientific methodology and stakeholder engagement, see Refsgaard et al. (2007). NRC (2006: Climate models) reviews uncertainty assessment and control methods, with good sections on uncertainty and decision making, while a formal strategy for conveying uncertainty to policymakers can be found in van der Sluijs et al. (2003). Funtowicz and Ravetz (1993) provide a solid attempt to place uncertainty in the context of both critical theory and the democratization of science.

The Royal Society special issue "Ensembles and probabilities: a new era in the prediction of climate change" (Collins and Knight 2007) gives an insight into the state of the art in much of the field of complex systems modelling outside of agent-based modelling, including the use of emulators, while Brown (2010: Physical models) and Matott et al. (2009: Environmental models) provide good technical overviews of these areas.

Matott et al. (2009) additionally give a very complete review of uncertainty software, broken down into data analysis, identifiability analysis, parameter estimation, uncertainty analysis, sensitivity analysis, multimodel analysis, and Bayesian networks. This is supplemented by an ongoing website at: http://www.epa.gov/athens/research/modeling/modelevaluation/index.html

A good review of Monte Carlo techniques, with a meta-review of other sensitivity and uncertainty testing techniques, is Helton et al. (2006), and Hamby (1994) gives a good review of sensitivity statistics. Finally, Bobashev and Morris (2010) provide a very clear walkthrough of one such sensitivity/uncertainty analysis for an agent-based system.

## References


Araújo, M., Pearson, R. G., Thuiller, W., & Erhard, M. (2005). Validation of species-climate impact models under climate change. *Global Change Biology, 11*, 1504–1513.

Ballas, D., Rossiter, D., Thomas, B., Clarke, G. P., & Dorling, D. (2005). *Geography matters: Simulating the local impacts of national social policies*. York: Joseph Rowntree Foundation.

Barnett, V., & Lewis, T. (1994). *Outliers in statistical data* (3rd ed.). Chichester: Wiley. pp.604.

Beck, M. B. (1987). Water quality modeling: A review of the analysis of uncertainty. *Water Resource Research, 23*(8), 1393–1442.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B, 57*, 289–2300.

Beven, K. J., & Binley, A. M. (1992). The future of distributed models: Model calibration and uncertainty prediction. *Hydrological Processes, 6*, 279–298.

Björk, B-C., Roos, A., & Lauri, M. (2009). Scientific journal publishing: Yearly volume and open access availability. *Information Research 14*(1): Paper 391. http://InformationR.net/ir/14-1/paper391.html. Accessed 7 Feb 2011

Bobashev, G. V., & Morris, R. J. (2010). Uncertainty and inference in agent-based models. In *Proceedings of the 2010 Second International Conference on Advances in System Simulation*. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5601895. Accessed 31 Mar 2011.

Boyle, D. P., Gupta, H. V., & Sorooshian, S. (2000). Towards improved calibration of hydrologic models: Combining the strengths of manual and automatic methods. *Water Resources Research, 36*, 3663–3674.

Brown, J. D. (2004). Knowledge, uncertainty and physical geography: Towards the development of methodologies for questioning belief. *Transactions of the Institute of British Geographers NS, 29*, 367–381.

Brown, J. D. (2010). Prospects for the open treatment of uncertainty in environmental research. *Progress in Physical Geography, 34*(1), 75–100.

Cerioli, A., & Farcomeni, A. (2011). Error rates for multivariate outlier detection. *Computational Statistics & Data Analysis, 55*(1), 544–553.

Cho, W. K. T., & Gaines, B. J. (2007). Breaking the (Benford) Law: Statistical fraud detection in Campaign Finance. *The American Statistician, 61*(3), 218–223.

Chowdhury, S., & Sharma, A. (2007). Mitigating parameter bias in hydrological modelling due to uncertainty in covariates. *Journal of Hydrology, 340*, 197–204.

Clancy, D., Tanner, J. E., McWilliam, S., & Spencer, M. (2010). Quantifying parameter uncertainty in a coral reef model using Metropolis-Coupled Markov Chain Monte Carlo. *Ecological Modelling, 221,* 1337–1347.

Clark, J. S. (2005). Why environmental scientists are becoming Bayesians. *Ecological Letters, 8*, 2–14.

Clark, J. S., & Gelfand, A. E. (2006). A future for models and data in environmental science. *Trends in Ecology and Evolution, 21*(7), 375–380.

Collins, M., & Knight, S. (2007). Theme issue 'Ensembles and probabilities: A new era in the prediction of climate change'. *Philosophical Transactions of the Royal Society A, 365*, 1957–2191.

Cook, J. R., & Stefanski, L. A. (1994). Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association, 89*(428), 1314–1328.

Cornélis, B., & Brunet, S. (2002). A policy-maker point of view on uncertainties in spatial decisions. In W. Shi, P. F. Fisher, & M. F. Goodchild (Eds.), *Spatial data quality* (pp. 168–185). London: Taylor and Frances.

Cornford, D. (2011). Uncertainty and the OGC: An Aston University take. Presented 03/03/2011 at OGC TC Meeting, Universities DWG Session, Bonn. http://www.uncertweb.org/documents. Accessed 1 Mar 2011.

Costanza, R. (1989). Model goodness of fit: A multiple resolution procedure. *Ecological Modelling, 47*, 199–215.

Couclelis, H. (2003). The certainty of uncertainty: GIS and the limits of geographic knowledge. *Transactions in GIS, 7*(2), 165–175.

Dassonville, L., Vauglin, F., Jakobsson, A., & Luzet, C. (2002). Quality management, data quality and users, metadata for geographical information. In W. Shi, P. F. Fisher, & M. F. Goodchild (Eds.), *Spatial data quality* (pp. 202–215). London: Taylor and Frances.

Dewulf, A., Craps, M., Bouwen, R., Taillieu, T., & Pahl-Wostl, C. (2005). Integrated management of natural resources: Dealing with ambiguous issues, multiple actors and diverging frames. *Water Science & Technology, 52*, 115–124.

Di Paolo, E. A, Noble, J, & Bullock, S (2000) Simulation models as opaque thought experiments. In *Seventh International Conference on Artificial Life* (pp. 497–506) Cambridge: MIT Press,. http://eprints.ecs.soton.ac.uk/11455/

Drecki, I. (2002). Visualisation of uncertainty in geographical data. In W. Shi, P. F. Fisher, & M. F. Goodchild (Eds.), *Spatial data quality* (pp. 140–159). London: Taylor and Frances.

Dubois, G., & Saisana, M. (2002). Optimal spatial declustering weights — Comparison of methods. In *Proceedings of the Annual Conference of the International Association for Mathematical Geology*, 15–20 Sept Berlin, 479–484. http://composite-indicators.jrc.ec.europa.eu/Document/Optimizing Spatial Declustering Weights – Comparison of Methods.pdf. Accessed 31 Mar 2011.

Evans, A. J., & Waters, T. (2007). Mapping vernacular geography: Web-based GIS tools for capturing "fuzzy" or "vague" entities. *International Journal of Technology, Policy and Management, 7*(2), 134–150.

Evensen, G. (1992). Using the extended Kalman filter with a multilayer quasi-geostrophic ocean model. *Journal of Geophysical Research, 97*(C11), 17905–17924.

Evensen, G. (1994). Sequential data assimilation with nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *Journal of Geophysical Research, 99*(C5), 143–162.

Faber, M., Manstetten, R., & Proops, J. (1992). Humankind and the environment: An anatomy of surprise and ignorance. *Environmental Values, 1*, 217–241.

Farcomeni, A. (2008). A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion. *Statistical Methods in Medical Research, 17*, 347–388.

Fisher, P., Wood, J., & Cheng, T. (2004). Where is Helvellyn? Fuzziness of multi-scale landscape morphometry. *Transactions of the Institute of British Geographers, 29*(1), 106–128.

Funtowicz, S. O., & Ravetz, J. R. (1990). *Uncertainty and quality in science for policy*. Dordrecht: Kluwer Academic Publishers.

Funtowicz, S. O., & Ravetz, J. R. (1993). Science for the post-normal age. *Futures, 25*, 739–755.

Galán, J. M., Izquierdo, L. R., Izquierdo, S. S., Santos, J. I., del Olmo, R., López-Paredes, A., & Edmonds, B. (2009). Errors and artefacts in agent-based modelling. *Journal of Artificial Societies and Social Simulation 12*(1): 1. http://jasss.soc.surrey.ac.uk/12/1/1.html. Accessed 25 Mar 2011.

Gallagher, M., & Doherty, J. (2007). Parameter estimation and uncertainty analysis for a watershed model. *Environmental Modelling and Software, 22*, 1000–1020.

Gamma, E., Helm, R., Johnson, R., & Vlissides, J. (1994). *Design patterns: Elements of reusable object-oriented software*. Reading, Mass: Addison Wesley. pp.416.

Gan, E., & Shi, W. (2002). Error metadata management systems. In W. Shi, P. F. Fisher, & M. F. Goodchild (Eds.), *Spatial data quality* (pp. 251–266). London: Taylor and Frances.

George, E. I. (2000). The variable selection problem. *Journal of the American Statistical Association, 95*(452), 1304–1308.

Getis, A. (2007). Reflections on spatial autocorrelation. *Reg Sci Urban Econ, 37*, 491–496.

Glahn, H., & Lowry, D. (1972). The use of Model Output Statistics (MOS) in objective weather forecasting. *Journal of Applied Meteorology, 11*, 1203–1211.

Gneiting, T., & Raftery, A. E. (2005). Weather forecasting with ensemble methods. *Science, 310*(5746), 248–249.

Green, S. B., & Babyak, M. A. (1997). Control of type I errors with multiple tests constraints in structural equation modeling. *Multivariable Behavioural Research, 32*, 39–51.

Groves, R. M., Fowler, F. J., Jr., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey methodology*. Hoboken: Wiley-Blackwell. pp.488.

Gupta, H. V. (1999). Status of automatic calibration for hydrologic models: Comparison with multilevel expert calibration. *Journal of Hydrologic Engineering, 4*(2), 135–143.

Gupta, H. V., Sorooshian, S., & Yapo, P. O. (1998). Toward improved calibration of hydrologic models: Multiple and noncommensurable measures of information. *Water Resources Research, 34*(4), 751–763.

Hamby, D. M. (1994). A review of techniques for parameter sensitivity analysis of environmental models. *Environmental Monitoring and Assessment, 32*, 135–154.

Hartley, S., Harris, R., & Lester, P. J. (2006). Quantifying uncertainty in the potential distribution of an invasive species: Climate and the Argentine ant. *Ecology Letters, 9*, 1068–1079.

Hasings, D., & McManus, H. (2004). A framework for understanding uncertainty and its mitigation and exploitation in complex systems. In *2004 Engineering Systems Symposium*, 29–31 Mar 2004, MIT. http://esd.mit.edu/symposium/pdfs/papers/hastings.pdf. Accessed 28 February 2011.

Hassan, S., Garmendia, L., & Pavon, J. (2010). Introducing uncertainty into social simulation: Using fuzzy logic for agent-based modelling. *International Journal of Reasoning-based Intelligent Systems, 2*(2), 118–124.

Hatton, L. (1997). The T experiments: Errors in scientific software. *Computational Science & Engineering, IEEE, 4*(2), 27–38.

Hayes, B. (2003). A Lucid interval. *American Scientist, 91*(6), 484–488.

Helton, J. C., Johnson, J. D., Salaberry, C. J., & Storlie, C. B. (2006). Survey of sampling based methods for uncertainty and sensitivity analysis. *Reliability Engineering and System Safety, 91*, 1175–1209.

Heppenstall, A. J., Evans, A. J., & Birkin, M. H. (2007). Genetic algorithm optimisation of a multi-agent system for simulating a retail market. *Environment and Planning B, 34*(6), 1051–1070.

Heuvelink, G. B. M. (1998). *Error propagation in environmental modelling with GIS*. London: Taylor and Francis. pp.127.

Hornberger, G., & Spear, R. (1981). An approach to the preliminary analysis of environmental systems. *Journal of Environmental Management, 7*, 7–18.

Hyde, R. (2004). *Write great code 1: Understanding the machine*. San Francisco: No Starch Press. pp. 440.

Isaaks, E. H., & Srivastava, R. M. (1990). *Applied geostatistics*. Oxford: Oxford University Press. pp.592.

Izquierdo, L. R., & Polhill, J. G. (2006). Is your model susceptible to floating-point errors? *Journal of Artificial Societies and Social Simulation 9*(4): 4. http://jasss.soc.surrey.ac.uk/9/4/4.html. Accessed 25 Mar 2011.

Janssen, J. A. E. B., Krol, M. S., Schielen, R. M. J., Hoekstra, A. Y., & de Kok, J.-L. (2010). Assessment of uncertainties in expert knowledge, illustrated in fuzzy rule-based models. *Ecological Modelling, 221*, 1245–1251.

JCGM (2008a) Evaluation of measurement data – Supplement 1 to the "Guide to the expression of uncertainty in measurement" – Propagation of distributions using a Monte Carlo method, Joint Committee for Guides in Metrology 100. http://www.bipm.org/utils/common/documents/jcgm/JCGM_101_2008_E.pdf. Accessed 15 Feb 2011.

JCGM (2008b) Evaluation of measurement data – Guide to the expression of uncertainty in measurement, Joint Committee for Guides in Metrology 100. http://www.bipm.org/utils/common/documents/jcgm/JCGM_100_2008_E.pdf. Accessed 15 Feb 2011.

Johnson, J. B., & Omland, K. S. (2004). Model selection in ecology and evolution. *Trends in Ecology & Evolution, 19*, 101–108.

Kardos, J. D., Moore, A., & Benwell, G. L. (2003). Visualising uncertainty in spatially-referenced attribute data using hierarchical spatial data structures. In *Proceedings of the 7th International Conference on GeoComputation*. University of Southampton, 8–10 Sept 2003. http://www.geocomputation.org/2003/. Accessed 25 Mar 2011

Kavetski, D., Kuczera, G., & Franks, S. W. (2006). Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory. *Water Resources Research, 42*, W03407.

Keesman, K., & van Straten, G. (1989). Identification and prediction propagation of uncertainty in models with bounded noise. *International Journal of Control, 49*, 2259–2269.
Keesman, K., & van Straten, G. (1990). Set-membership approach to identification and prediction of lake eutrophication. *Water Resources Research, 26*, 2643–2652.
Kelejian, H., & Prucha, I. (2010). Spatial models with spatially lagged dependent variables and incomplete data. *Journal of Geographical Systems, 12*(3), 241–257.
Kennedy, M. C., & O'Hagan, A. (2001). Bayesian calibration of computer models. *Journal of the Royal Statistical Society B, 63*(3), 425–464.
Kiesling, J. B. (undated) Charting Electoral Fraud: Turnout Distribution Analysis as a Tool for Election Assessment. *Diplomacy Lessons*. http://www.bradykiesling.com/election_fraud_analysis.htm. Accessed 10 Jan 2010.
King, G. (1986). How not to lie with statistics: Avoiding common mistakes in quantitative political science. *American Journal of Political Science, 30*, 666–687.
Knudsen, C. D., & Fotheringham, A. S. (1986). Matrix comparison, goodness-of-fit and spatial interaction modeling. *International Regional Science Review, 10*(2), 127–147.
Ku, H. (1966). Notes on the use of propagation of error formulas. *Journal of Research of National Bureau of Standards C. Engineering and Instrumentation, 70*(4), 263–273.
Laffan, S. W. (1999). Spatially assessing model error using geographically weighted regression. In *Proceedings of the 4th International Conference on GeoComputation.* Fredericksburg: Mary Washington College, 5–28 July 1999. http://www.geocomputation.org/1999/. Accessed 25 Mar 2011.
Lane, S. N. (2001). Constructive comments on D Massey 'Space-time, "science" and the relationship between physical geography and human geography'. *Transactions of the Institute of British Geographers, 6*, 243–256.
Lauridsen, J., & Kosfeld, R. (2007). Spatial cointegration and heteroscedasticity. *Journal of Geographical Systems, 9*(3), 253–265.
Legendre, P. (1993). Spatial Autocorrelation: Trouble or New Paradigm? *Ecology, 74*, 1659–1673.
Leith, C. E. (1974). Theoretical skill of Monte Carlo forecasts. *Monthly Weather Review, 102*(6), 409–418.
Li, Y., Brimicombe, A. J., & Li, C. (2007). Agent-based services for validating multi-agent models. In *Proceedings of the 9th International Conference on GeoComputation*, Maynooth: Eire.
Lodwick, W. A. (1989). Developing confidence limits on errors of suitability analyses in geographical information systems. In M. Goodchild & S. Gopal (Eds.), *Accuracy of spatial databases* (pp. 69–80). London: Taylor and Francis.
López, C. (1997). Quality of geographic data – Detection of outliers and imputation of missing values (Unpublished PhD thesis, Universidad ORT, Uruguay). http://www.thedigitalmap.com/~carlos/papers/PhDthesis/phdthesis.pdf. Accessed 1 Mar 2011.
Lorenz, E. N. (1963). Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences, 20*, 130–141.
Luoto, M., Marmion, M., & Hjort, J. (2010). Assessing spatial uncertainty in predictive geomorphological mapping: A multi-modelling approach. *Computers & Geosciences, 36*, 355–361.
Lutz, W., Sanderson, W., & Scherbov, S. (1996). Probabilistic population projections based on expert opinion. In W. Lutz (Ed.), *The future population of the world. What can we assume today?* (pp. 397–428). London: Earthscan Publications. Chapter 16.
Lyman, R. L., & Ames, K. M. (2007). On the use of species-area curves to detect the effects of sample size. *Journal of Archaeological Science, 34*, 1985–1990.
Malinarič, S., & Ďuríšek, P. (2004). Sensitivity coefficients analysis. In *Proceedings of Thermophysics 2004*. http://www.tpl.fpv.ukf.sk/engl_vers/thermophys/2004/Mal-Dur.pdf. Accessed 14 Feb 2011.
Malleson, N. (2010). Agent-based modelling of Burglary (Unpublished PhD thesis, School of Geography, University of Leeds, Leeds). http://www.geog.leeds.ac.uk/fileadmin/downloads/school/people/postgrads/n.malleson/thesis-final.pdf. Accessed 1 Mar 2011.
Marin, C. M., Guvanasen, V., & Saleem, Z. A. (2003). The 3MRA risk assessment framework – A flexible approach for performing multimedia, multipathway, and multireceptor risk assessments

under uncertainty. *Human and Ecological Risk Assessment: An International Journal, 9*(7), 1655–1677.

Martin, D. (2003). Extending the automated zoning procedure to reconcile incompatible zoning systems. *International Journal of Geographical Information Science, 17*, 181–196.

Martin, D., Dorling, D., & Mitchell, R. (2002). *Linking censuses through time: problems and solutions Area, 34*, 82–91.

Matott, L. S., Babendreier, J. E., & Parucker, S. T. (2009). Evaluating uncertainty in integrated environmental models: A review of concepts and tools. *Water Resources Research 45*: WO6421. http://www.agu.org/journals/wr/wr0906/2008WR007301/2008WR007301.pdf. Accessed 2 Feb 2011.

McKay, M. D., Beckman, R. J., & Conover, W. J. (1979). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics, 21*(2), 239–245.

Mearns, L. O. (2010). The drama of uncertainty. *Climatic Change, 100*, 77–85.

Mebane, W. R Jr., & Kalinin, K. (2009). Comparative election fraud detection. In *Annual Meeting of the American Political Science Association*, Toronto 3–6 Sept 2009. http://www-personal.umich.edu/~wmebane/apsa09.pdf. Accessed 10 Jan 2010.

Mitchell, S., Beven, K., & Freer, J. (2009). Multiple sources of predictive uncertainty in modeled estimates of net ecosystem $CO_2$ exchange. *Ecological Modelling, 220*(23), 3259–3270.

Monaco, J. H., & Anderson, R. L. (1994). Tai's formula is the trapezoidal rule. *Diabetes Care, 17*(10), 1224–1225.

Moreno, N., Wang, F., & Marceau, D. J. (2008). An object-based land-use cellular automata model to overcome cell size and neighborhood sensitivity. In *Proceedings OF GEOBIA 2008 – Pixels, Objects, Intelligence GEOgraphic Object Based Image Analysis for the 21st Century*. Calgary, 5–8 Aug. http://www.isprs.org/proceedings/XXXVIII/4-C1/Sessions/Session6/6753_Marceau_Proc_pap.pdf. Accessed 25 Mar 2011.

Morss, R. E., Demuth, J. L., & Lazo, J. K. (2008). Communicating uncertainty in weather forecasts: a survey of the US public. *Weather and Forecasting, 23,* 974–91.

Nagele, P. (2001). Misuse of standard error of the mean (SEM) when reporting variability of a sample. *A critical evaluation of four anaesthesia journals. British Journal of Anaesthesia, 90*(4), 514–516.

National Research Council of the National Academies (NRC). (2006). *Completing the forecast: Characterizing and communicating uncertainty for better decisions using weather and climate forecasts*. Washington, DC: National Academies Press.

Ngo, T. A., & See, L. M. (2012). Calibration and validation of agent-based models of land cover change. In A. J. Heppenstall, A. T. Crooks, L. M. See, & M. Batty (Eds.), *Agent-based models of geographical systems* (pp. 181–196). Dordrecht: Springer.

NIST/SEMATECH (2010). *e-Handbook of Statistical Methods*, http://www.itl.nist.gov/div898/handbook/. Accessed 10 Jan 2011.

O'Neill, R. V. (1973). Error analysis of ecological models. In D. J. Nelson (Ed.), *Radionuclides in ecosystems. CONF-710501* (pp. 898–908). Springfield: National Technical Information Service.

O'Neill, R. V., & Rust, B. (1979). Aggregation error in ecological models. *Ecological Modelling, 7*(2), 91–105.

O'Neill, R. V., Gardner, R. H., & Mankin, J. B. (1980). Analysis of parameter error in a nonlinear model. *Ecological Modelling, 8*, 297–311.

Openshaw, S. (1989). Learning to live with errors in spatial databases. In M. Goodchild & S. Gopal (Eds.), *Accuracy of spatial databases* (pp. 263–276). London: Taylor and Francis.

Polhill, J. G., Izquierdo, L. R., & Gotts, N. M. (2006). What every agent-based modeller should know about floating point arithmetic. *Environmental Modelling & Software, 21*(3), 283–309.

Poli, R, Langdon, W. B., and McPhee, N. F., with Koza, J. R. (2008). *A Field Guide to Genetic Programming*. http://www.gp-field-guide.org.uk/. Accessed 28 February 2011.

Poulton, E. C. (1989). *Bias in quantifying judgments*. Hove: Erlbaum. pp.328.

Powers, A. C. (2005). Simulating patterns of uncertainty in postclassification change detection. In *Proceedings of the 8th International Conference on GeoComputation*. University of Michigan, 31 July–3 Aug 2005. http://www.geocomputation.org/2005/. Accessed 25 Mar 2011.

Raftery, A. E., Gneiting, T., Balabdaoui, F., & Polakowski, M. (2005). Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review, 133*, 1155–1174.

Refsgaard, J. C., Henriksen, H. J., Harrar, W. G., Scholten, H., & Kassahun, A. (2005). Quality assurance in model based water management – Review of existing practice and outline of new approaches. *Environmental Modelling & Software, 20*(10), 1201–1215.

Refsgaard, J. C., van der Sluijs, J. P., Etejberg, A. L., & Vanrolleghem, P. A. (2007). Uncertainty in the environmental modeling process – A framework and guidance. *Environmental Modelling and Software, 22*, 1543–1556.

Reinke, K., & Hunter, G. J. (2002). A theory for communicating uncertainty in spatial databases. In W. Shi, P. F. Fisher, & M. F. Goodchild (Eds.), *Spatial data quality* (pp. 76–101). London: Taylor and Frances.

Rogers, J. P., Barbara, D., & Domeniconi, C. (2009). Detecting spatio-temporal outliers with kernels and statistical testing. In *17th International Conference on Geoinformatics*. http://dx.doi.org/10.1109/GEOINFORMATICS.2009.5293481. Accessed 1 Mar 2011.

Rowe, W. D. (1977). *An anatomy of risk*. New York: John Wiley and Sons.

Saltelli, A., Chan, K., & Scott, E. M. (2000). *Sensitivity analysis*. Chichester: Wiley. pp.475.

Schönfisch, B., & de Roos, A. (1999). *Synchronous and asynchronous updating in cellular automata Biosystems, 51*(3), 123–143.

Shackley, S., & Wynne, B. (1996). Representing uncertainty in global climate change science and policy: Boundary ordering devices and authority. *Science Technology Human Values, 21*, 275–302.

Smith, L. A., Cuéllar, M. C., Du, H., & Judd, K. (2010a). Exploiting dynamical coherence: A geometric approach to parameter estimation in nonlinear models. *Physics Letters A, 374*, 2618–2623.

Smith, A. H. C., Ponci, F., & Monti, A. (2010b). Bounding the dynamic behavior of an uncertain system via polynomial chaos-based simulation. *Simulation, 86*, 31–40.

Smithson, M. (1989). *Ignorance and uncertainty: Emerging paradigms*. New York: Springer.

Spear, R. C. (1970). The application of Kolmogorov-Renyi statistics to problems of parameter uncertainty in systems design. *International Journal of Control, 11*, 771–778.

Spetzler, C. S., & von Holstein, S. (1975). Probability encoding in decision analysis. *Management Science, 22*(3), 340–358.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit (with Discussion). *Journal of the Royal Statistical Society B, 64*, 583–639.

Suter, G. W., Barnthouse, L. W., & O'Neill, R. V. (1987). Treatment of risk in environmental impact assessment. *Environmental Management, 11*, 295–303.

Tang, S., & Wang, Y. (2001). A parameter estimation program for error-in-variable model. *Ecological Modelling, 156,* 225–236.

Tobler, W. R. (1979). Smooth pycnophylactic interpolation for geographical regions. *Journal of the American Statistical Association, 74*, 519–530.

Toth, Z., & Kalnay, E. (1993). Ensemble forecasting at NMC: The generation of perturbations. *Bulletin of American Meteorological Society, 74*, 2317–2330.

Urban, N. M., & Fricker, T. E. (2010). A comparison of Latin hypercube and grid ensemble designs for the multivariate emulation of an Earth system model. *Computers & Geosciences, 36*, 746–755.

Van der Sluijs, J. P., Potting, J., Risbey, J., van Vuuren, D., de Vries, B., Beusen, A., Heuberger, P., Quintana, S. C., Funtowicz, S., Kloprogge, P., Nuijten, D., Petersen, A. C., & Ravetz, J. (2002). *Uncertainty assessment of the IMAGE = TIMER B1 CO2 emissions sScenario, using the NUSAP method.* (Report No. 410 200 104). Bilthoven: Dutch National Research Program on Climate Change.

Van der Sluijs, J. P., Risbey, J., Kloprogge,P., Ravetz, J. R., Funtowicz, S. O., Quintana, S. C., Pereira, A. G., De Marchi, B., Petersen, A. C., Janssen, P. H. M., Hoppe, R., & Huijs, S. W. F. (2003). RIVM/MNP guidance for uncertainty assessment and communication. http://www.nusap.net/downloads/detailedguidance.pdf. Accessed 14 Feb 2011.

van der Wel, F. J. M., van der Gaag, L. C., & Gorte, B. G. H. (1996). Visual exploration of uncertainty in remote sensing classifications. In *Proceedings of the 1st International Conference on GeoComputation*. Leeds: University of Leeds, 17–19 Sept 1996. http://www.geocomputation.org/1996/. Accessed 25 Mar 2011.

van Straten, G. (1985). Analytical methods for parameter-space delimitation and application to shallow-lake phytoplankton-dynamics modeling. *Applied Mathematics and Computation, 17*, 459–482.

Verkuilen, J. (2005). Assigning membership in a fuzzy set analysis. *Sociological Methods & Research, 33*(4), 462–496.

Viard, T., Caumon, G., & Lévy, B. (2011). Adjacent versus coincident representations of geospatial uncertainty: Which promote better decisions? *Computers & Geosciences, 37*(4), 511–520.

Wagener, T., McIntyre, N., Lees, M. J., Wheater, H. S., & Gupta, H. V. (2003). Towards reduced uncertainty in conceptual rainfall-runoff modelling: Dynamic identifiability analysis. *Hydrological Processes, 17*(2), 455–476.

Walker, W. W. (1982). A sensitivity and error analysis framework for lake eutrophication modeling. *Water Resources Bulletin, 18*(1), 53–60.

Walker, W. E., Harremoes, P., Rotmans, J., van der Sluijs, J. P., van Asselt, M. B. A., Janssen, P., et al. (2003). Defining uncertainty: A conceptual basis for uncertainty management in model-based decision support. *Integrated Assessment, 4*(1), 5–18.

Warren, H. S., Jr. (2002). *Hacker's delight*. Boston: Addison Wesley. pp. 306.

Williams, M., Cornford, D., Bastin, L., & Ingram, B. (2008). UncertML: an XML schema for exchanging uncertainty. In *Proceedings of GISRUK 2008*. http://www.uncertml.org/publications.php. Accessed 1 Mar 2011.

Wooldridge, M. (2009). *An introduction to multiagent systems* (2nd ed.). Chichester: Wiley. pp. 484.

Wyss, G. D., & Jorgensen, K. H. (1998). *A user's guide to LHS: Sandia's Latin Hypercube Sampling Software*. SAND98-0210 Albuquerque: Sandia National Laboratories.

Young, P. C. (2002). Advances in real-time flood forecasting. *Philosophical Transactions of the Royal Society of London A, 360*, 1433–1450.

Young, P. C., Parkinson, S., & Lees, M. J. (1996). Simplicity out of complexity in environmental modelling: Occam's razor revisited. *Journal of Applied Statistics, 23*, 165–210.

Zadeh, L. A. (1975). The concept of a linguistic variable and its application to approximate reasoning–1. *Information Sciences, 8*, 199–249.

Zadeh, L. A. (1976). A fuzzy-algorithmic approach to the definition of complex or imprecise concepts. *International Journal of Man-Machine Studies, 8*(3), 249–291.

Zadeh, L. A. (2005). Toward a generalized theory of uncertainty (GTU) – An outline. *Information Sciences, 172*, 1–40.

Zellner, M. L. (2008). Embracing complexity and uncertainty: The potential of agent-based modeling for environmental planning and policy. *Planning Theory & Practice, 9*(4), 437–457.

Zhang, J., & Goodchild, M. (2002). *Uncertainty in geographical information* (p. 266). London: Taylor and Francis.

Žižek, S. (2004). What Rumsfeld doesn't know that he knows about Abu Ghraib. *In These Times,* 21 May 2004. http://www.lacan.com/zizekrumsfeld.htm. Accessed 10 Jan 2010.