Review of Uncertainty Evaluation Practices of Computer Simulation Models

1. Introduction

Deterministic agricultural and agro-ecosystem systems models have an important role worldwide. Their role centres around understanding how a set of real world processes behave and interact with one another. They are used in a range of settings to inform and support farm management practice, breeding strategies and government policy. These applications depend on research programs prioritizing food security and climate change adaptation (Boote, Jones et al. 1996, Sinclair and Seligman 2000, Jamieson, Brooking et al. 2007, Cooper, van Eeuwijk et al. 2009, Hochman, Van Rees et al. 2009, Bezlepkina, Adenäeur et al. 2010, Holzworth, Huth et al. 2014).

Although these models are mathematically deterministic, there are many possible sources of uncertainty that can propagate through the model. There is recognition in the agricultural (and wider) modelling community that the impact of uncertainty needs to be considered (Hammer, Kropff et al. 2002, O'Hagan 2008, Rotter, Carter et al. 2011, O'Hagan 2012, Holzkämper, Klein et al. 2015, Uusitalo, Lehikoinen et al. 2015).

The development of tools to evaluate uncertainty in deterministic models is an active area of research both within (Wallach 2011, Chichota, Snow et al. 2013, Clifford, Pagendam et al. 2013, Stanfill, Clifford et al. 2014, Wallach, Makowski et al. 2014) and outside of the agricultural sector. There are a number of options for model uncertainty evaluation that have been discussed in the literature. However, when carrying out an uncertainty evalution, there can to difficulties in both a) identifying the most appropriate techniques and b) in confirming that sufficient work has been done. The objectives of this review paper are threefold:

1. To describe a formalised state-space framework within which to describe the types and sources of uncertainty that arise in computer simulation models (Section 2).
2. To provide a framework to carry out a robust uncertainty evaluation of a computer simulation model (Section 3).
3. To summarise a selection of relevant (to the agricultural modelling community) sampling (Section 4) and analysis (Section 5) techniques for the uncertainty evaluation of computer simulation models.

2. Computer simulation models: a state-space framework for uncertainty allocation

2.1 Definition

A generic designation for a deterministic agricultural systems model is a 'computer simulation model', or 'simulator'. Bayarri and Berger et.al (2009) defined a simulator as 'a computational representation of a complex real-world process.' A simulator is usually developed to approximately describe and allow direct simulation of the real-world process.'

A simulator is defined by a series of equations, decisions and input information that aims to characterize a real world process (Saltelli, Chan et al. 2000, McFarland 2008). When run, the outputs of such a simulator are a simplified prediction of the real world phenomena. Often these models are dynamic. They update through time whilst responding to environmental information such as rainfall or nutrient management input. Many modules of APSIM, an agricultural simulator that has been described recently by Holzworth, Huth et al. (2014) are such models. An important characteristic is that although outputs are not really uncertain because they are a deterministic function of the inputs, in practice the simulator can be sufficiently complex that the outcome cannot be known prior to simulation (Kennedy and O'Hagan 2001).

2.2 Notation

Simulators are built for many different purposes and have many forms, but a single model can be formally represented as:

$$\boldsymbol{y = f(x)} \tag{1}$$

where $\boldsymbol{x}$ is a vector of inputs, and $\boldsymbol{y}$ a vector of outputs. The model structure, $\boldsymbol{f(\cdot)}$, defines (mathematically or computationally) how the characteristics of $\boldsymbol{y}$ are determined by those of $\boldsymbol{x}$. It can therefore be conceptualised as a formal statement of assumptions about the real world process (McKay and Morrison 1997). Strong (2012) additionally identified an extra discrepancy term $\boldsymbol{\delta}$ as a linear, additive term to quantify the effect of structural error[1] on the model's ability to predict the true, unknown target quantity $\boldsymbol{t: t = f(x)+ \delta}$**.** We extend this to include instead a more general term $\boldsymbol{\varepsilon}$ that does not assume either linearity or additivity, and is used to 'catch' any form of uncertainty that cannot be otherwise allocated (i.e as discussed in Section 3). Thus building on **(**1), we define:

$$\boldsymbol{t = f(x, \varepsilon)} \tag{2}$$

In the next section we describe sources of uncertainty in simulators, and allocate them to one of the three components on the right hand side of this simple representation.

2.3 Components of a model

A clear partitioning of the components of the model is an important step for any uncertainty evaluation of a simulator. Some of the partitions may seem somewhat artificial, but we believe they are necessary to enable the objectives of the uncertainty evaluation to be stated without confusion. Given that many agricultural and agro-ecosystem simulators are dynamic

[1] Structural and other components of uncertainty in computer simulation model outputs **y** are discussed in Section 3.

in nature, we adopt a state-space framework and follow the notation of authors such as (Gordon, Salmond et al. 1993, Cressie and Wikle 2011) to compartmentalise the model. Each component defined in this section can introduce uncertainty in the simulator, and this is discussed in Section 3.

2.3.1 Input parameters

We denote input parameters **θ.** **θ** represents independent input information that does not change during the sequential updating process of a dynamic simulator. Examples of input parameters in an agricultural setting could be soil type, cultivar or other 'scenario' indicators as discussed by (e.g.) (Holzkämper, Klein et al. 2015).

2.3.2 Observation data

Data which are observed are denoted $\boldsymbol{D_k} = \begin{pmatrix} \boldsymbol{Q_k} \\ \boldsymbol{E_k} \end{pmatrix}$**.** Here the subscript *k* identifies the time step the dynamic simulator operates on (e.g. week/day/hour). $\mathbf{Q_k}$ represents dynamic response or calibration data (possibly for multiple scenarios hence the matrix notation) this could also be available only as a single **Q** i.e yield at the end of the simulation process for a selection of scenarios or Q for a single scenario. $\mathbf{E_k}$ similarly represents updating environmental or managerial inputs such as rainfall/irrigation or temperature.

2.3.3 State equations

State equations $\mathbf{Z_k}$ jointly define the structure of the model. They represent either experimentally derived relationships or theoretical constructs. State equations are mathematical equations that describe the underlying scientific processes of the model. Although the coefficients of these equations may have been derived via a calibration process during the model building phase i.e. (O'Hagan 2006), these coefficients and the equations to which they relate are distinct from the input parameters as they are usually hard-coded into

the software of the simulator. Examples of state equations and their accompanying coefficients could be the vernalisation requirement for wheat.

2.3.4 State variables

State variables $\mathbf{Y_k}$ are now defined as the value of each state equation at each time-step *k* given the input parameters **θ,** the observation data $\boldsymbol{D_k} = \begin{pmatrix} \boldsymbol{Q_k} \\ \boldsymbol{E_k} \end{pmatrix}$. An important corollary of the structure of the model being allowed to depend upon theoretical constructs is that some components of $\mathbf{Y_k}$ may be latent, or unable to be observed in practice.

2.4 Types and sources of uncertainty and their allocation to model components

Simulators represent detailed scientific understanding of real-world systems. However, although models are a vital part of research and development, they are imperfect. Imperfections in the outputs of a deterministic model may be due to incorrect specification of state equations and input parameters, or to inherent stochasticity in observed data (Montanari, Shoemaker et al. 2009). Imperfections are therefore due to uncertainty – the lack of exact knowledge in at least some components of the model (Refsgaard, van der Sluijs et al. 2007). For example, in almost all simulators some components of the model are empirically determined (Sinclair and Muchow 2001, O'Hagan 2006). Many authors have discussed types and sources of uncertainty in simulators, including (O'Hagan, Kennedy et al. 1999, Kennedy and O'Hagan 2001, Katz 2002, Spiegelhalter and Best 2002, O'Hagan 2006, Cressie and Wikle 2011, Gupta, Clark et al. 2012). A classification system defining and allocating epistemic and aleatory sources of uncertainty might then be as is defined next. Here, elements of equation (2) are underscored to identify the source of each type of uncertainty is allocated.

2.4.1 Epistemic uncertainty

Epistemic uncertainty refers to uncertainty in events that is due simply to our lack of knowledge of them. Some sources of epistemic uncertainty in simulators are:

- Structural uncertainty: $\boldsymbol{t = \underline{f}(x, \varepsilon)}$; also known as model inadequacy (Gupta, Clark et al. 2012), state equation uncertainty, or 'ignorance', and refers to our basic lack of knowledge concerning the appropriate structure of the model.
    - The obvious symptom is the difference between the true mean value of the real world process, and the model output at the true values of the inputs.
    - This is either a component of methodological uncertainty or,
    - The real process may itself exhibit random variability, so model structure can itself be considered as an unknown state of the world and be subject of probabilistic sensitivity analysis (Spiegelhalter and Best 2002, Strong, Oakley et al. 2012, Strong and Oakley 2014) .
- Input parameter uncertainty: $\boldsymbol{t = f(\underline{x}, \varepsilon)}$; also known as 'state of the world' uncertainty and refers to uncertainty about the appropriate values input parameters describing the scenario to be modelled.
- Code Uncertainty: $\boldsymbol{t = f(x, \underline{\varepsilon})}$; uncertainty due to unknown possible response when it is not possible to completely sample the response surface of model outputs (Kennedy and O'Hagan 2001).
- Scaling/Aggregation: $\boldsymbol{t = \underline{f}(\underline{x}, \underline{\varepsilon})}$; spatial or temporal 'support' of model or data (Katz 2002, Cressie and Wikle 2011) .
    - Potentially one of the more complex types of uncertainty to evaluate.
    - Possible approaches to assess uncertainty are likely to be specific to the model under study.
    - Will not be discussed further in this paper.

2.4.2 Aleatory uncertainty

Aleatory uncertainty refers to uncertainty in repeatable events, which arises from their intrinsic randomness and unpredictability. This is usually thought of as residual, random stochastic uncertainty, describing the variation of a real world process even when the conditions are fully specified. The true process ***t*** is then defined as the mean value averaged over this intrinsic, random varaiton. However, we cannot always differentiate this true stochasticity from ignorance about some detail that would allow us to discriminate between conditions that actually lead to different process values as defined above.

- Observation uncertainty: There are two types of observation that can introduce uncertainty:
    - Response Data: $\boldsymbol{t = f(x, \underline{\varepsilon})}$; when calibrating or updating predictions with actual observations.
    - Environmental input data: $\boldsymbol{t = f(\underline{x}, \underline{\varepsilon})}$; can introduce uncertainty both as random data as above, but also it can be thought of in the same way as input parameter uncertainty. The uncertainty entering any model estimates due to environmental variables is a representation of measured weather conditions, and its variability and error in measurement effects in addition to in being incorrectly specified for entry into the model.

2.5 Terminology

Aleatory sources of uncertainty are usually seen as irreducible, whereas epistemic sources of uncertainty can often be quantified and sometimes reduced (Refsgaard, van der Sluijs et al. 2007, Uusitalo, Lehikoinen et al. 2015). It is therefore fortunate that almost all the uncertainties in the analysis of process model outputs are epistemic O'Hagan (2006). In this review we will use the term ‘model uncertainty evaluation’ or ‘uncertainty evaluation’ (UE) to refer to exploration of any of the sources of uncertainty described in this section. Note that

specific subsets of these have their own name that is ubiquitous across the literature; for example *sensitivity analysis* refers to exploration of the effect of changes in input parameters ***x*** on the outcomes ***y***.

3. Robust computer simulation model uncertainty evaluation

3.1 The life of a model

One thing to consider when approaching an uncertainty evaluation of a model is the phases of model development and use. The model's life is usually a continuum with movement in both directions between conceptualisation and implementation in code (Model Building), testing (Model Assessment) and use (Model Application). It may be useful to conceptualise a simplified schematic of the life of a simulator as shown in Figure 1 below. Awareness of which phase the model is in during the uncertainty evaluation process is an important piece of information when defining the objective of the UE. This will then help identify the most appropriate model UE techniques since some may be more suited to some phases than to others.

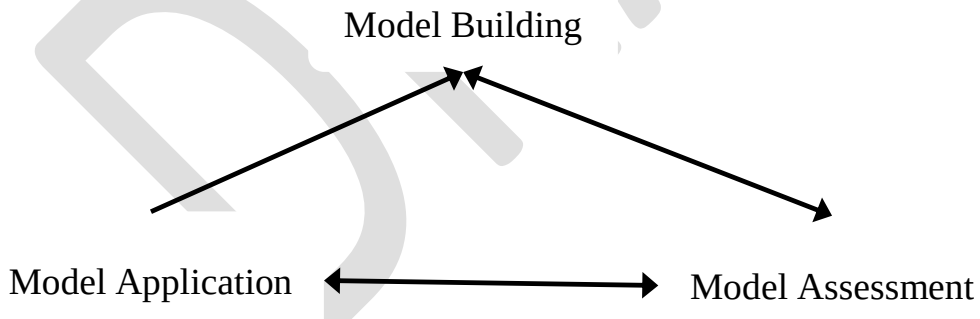

Figure 1: Simplified schematic of phase of a simulators life

3.2 Outline of a robust uncertainty evaluation

Several sampling (Section 4) and analysis (Section 5) techniques are reviewed in this paper. Whilst many are complementary, not all will be suitable for all applications. The choice of techniques to utilise is dependent upon what resources are available, and the objectives (see Section 3.2.5) of the UE. Figure 2 outlines seven steps that, if followed, can help ensure a robust UE of a model. Figure 2 is not dissimilar to the framework put forward by (Refsgaard, van der Sluijs et al. 2006). Each step in this outline is expanded upon in Sections 3.2.1 – 3.2.7.

Figure 6 in Section 3.2.8 then demonstrates an exemplar of a possible process by which to proceed. It steps the reader through an elicitation of the objectives of the evaluation and the available information. It then provides suggestions for appropriate sampling and analysis techniques.

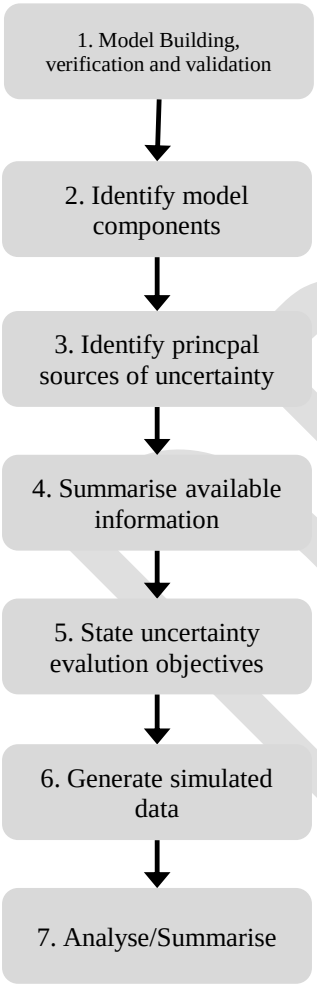

Figure 2: Seven steps for simulator uncertainty evaluation

This paper does not explore methods for collecting observational data $\mathbf{D_k}$ if this is desired and/or possible. Interested readers may find many introductory statistical texts for optimal

data collection in many fields of research are available. An important feature of such real-world calibration data is that it is often difficult or impossible to obtain for many possible computer simulation scenarios. For obvious reasons in these situations only simulation data can be obtained, and are analysed Sensitivity Analysis. The elicitation of expert opinion has been well covered by e.g. (Refsgaard, van der Sluijs et al. 2007, O'Hagan 2012). Specific application of such expert information is given in the relevant sections below.

3.2.1 Verify and validate the model

Figure 3 below displays the iterative process that is integral to model building. As with any modelling exercise, it begins with an observation, from which hypotheses are derived, and then implemented in code. The joint processes of verification and validation, which are integral components of model building, are defined below.

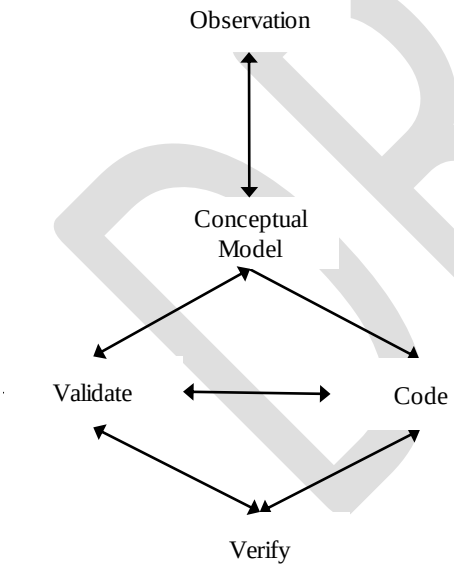

Figure 3: The model building process

Verification

Verification is defined as the process of determining whether the model implemented in computer code accurately represents the algorithms that were intended (Carson 2002, Trucano, Swiler et al. 2006) Since verification is usually an integral part of the coding process it will not be discussed further in this paper.

Validation

Validation is more complex, with authors such as Oberkampf and Roy (2010) pointing out that different communities view validation from different perspectives:

- Quantification of the accuracy of the model results by comparing model outputs with experimental data.
- Use of the model to make predictions corresponding to the model's domain of intended use.
- Determination of whether the estimated accuracy of the model results satisfies the some specified accuracy requirements.

That is, in the first case some experimental data is required, whereas in the second two it is not necessarily expected. However, as long as the process of validation is kept conceptually as 'confirmation of fit for purpose' then we enjoy the definition of Sargent (2005) 'Validation is the process of determining the degree to which a simulation model and its associated data are an accurate representation of the real world from the perspective of the intended uses of the model'. Techniques for model UE during the validation stage are discussed in greater detail in Section 5.1.

3.2.2 Identify components of the model

Each component of the model defined in Section 2 should be able to be assigned to one of the columns in Table 1. If not, another column should be added to allow for other types of components.

Table 1: Model component matrix

Title	Input Parameter	Observation Variable		State Equation	State Variable
		Calibration Variable	Environmental Variable		
Symbol	θ	Q_k	E_k	Z_k	Y_k
Model component 1					
Model component 2					
…					

3.2.3 Identify principal sources of uncertainty in the model

Given the components identified in Step 2, assess which of the sources of information in Table 2 are likely to be of primary importance.

Table 2: Sources of uncertainty

Source of uncertainty	Notation
Structural uncertainty	$t = f(x, \varepsilon)$
Input parameter uncertainty	$t = f(\underline{x}, \varepsilon)$
Calibration data uncertainty	$t = f(x, \underline{\varepsilon})$
Environmental data uncertainty	$t = f(\underline{x}, \underline{\varepsilon})$
Code uncertainty	$t = f(x, \underline{\varepsilon})$
Scaling/Aggregation	$t = f(x, \varepsilon)$
Stochastic uncertainty	$t = f(x, \underline{\varepsilon})$

3.2.4 Summarise available information

Data, expert opinion, phase of the models life, model components, sources of uncertainty and other relevant information can be identified in Table 3. Extra rows should be added as required for each individual component; i.e. if there are 10 state equations then there should be 10 rows in the first segment. If the model is too large to easily enable this process a first step should be to identify which sub-component of the model structure is to be evaluated.

Table 3: Information matrix adjusted from (Refsgaard, van der Sluijs et al. 2007)

Context	Notation	Model components likely to be involved		Data	Expert opinion	other information
Structural uncertain due ignorance, scaling/aggregation etc.	$y = f(x) + \varepsilon$	Z_k				
Model Inputs or code uncertainty	$y = f(x) + \varepsilon$	E_k	θ			
Stochastic uncertainty or incorrect calibration data	$y = f(x) + \varepsilon$	Q_k				

3.2.5 State objectives of evaluation

Any attempt to parameterise uncertainty cannot be general or universal; rather, it is an exploration that will be specific not only to the model but to the environment/scenario for which it is to be used (McFarland 2008). In most situations a selection of techniques will be likely to be helpful and should be combined to provide a heuristic view of the model; hence our use of the term 'uncertainty evaluation'. Which techniques will prove most insightful will vary depending on the phase of the model's life (Figure 1) that is under study. Further, during model building, assumptions and simplifications are made. The techniques chosen to describe the uncertainty in a given model depend on specific properties of that model (Wallach, Makowski et al. 2014) and these include:

- Principal sources of uncertainty for that particular model,
- Assumptions made during the model building process,
- What information (data, expert opinion) is available.

We believe the key to any simulator uncertainty evaluation is to clearly state the objective of the analysis, allowing for the resources available as summarised in tables 1-3. Once the objectives have been written down, sampling and analysis can begin.

3.2.6 Generate simulation data

Acquiring an appropriate sample of simulated data from the model is an important aspect of model evaluation. This simulated data provides information about how the model responds

to complex combinations of inputs, and is not synonymous with real-world, observed data. Simulation data generated from a simulator using an technique such as those described in Section 4 below can be used either to explore $\boldsymbol{y = f(x)}$ (sensitivity analysis) or $\boldsymbol{t = f(x, \varepsilon)}$ (calibration) (see Figure 5). The difference lies in whether there is real-world data available to provide information about $\boldsymbol{t}$, and hence $\boldsymbol{\varepsilon}$.

Given a large amount of time and computer resources for a particular problem, the ideal approach to data generation would be to sample evenly over all possible combinations of parameter values. However, this is usually impractical or even impossible, and the objective is therefore to reduce computational load whilst ensuring an appropriate representation of the response surface is obtained. The objectives of the evaluation defined in the previous step should help guide the sampling technique taken.

The technique used to generate data from the simulator will strongly influence the direction of the analysis toward evaluating uncertainty either due to input parameters or to structural uncertainty. Some techniques will allow more options than others, however. A selection of sampling techniques to explore structural or input parameter uncertainty are shown in Figure 4. These and others will be discussed in Section 4. In general, simulated data arising from any of the sampling techniques could be used for most analysis techniques described in Section 5. An exception is the analysis technique for code uncertainty that arises from a Gaussian Process (GP) emulator sample.

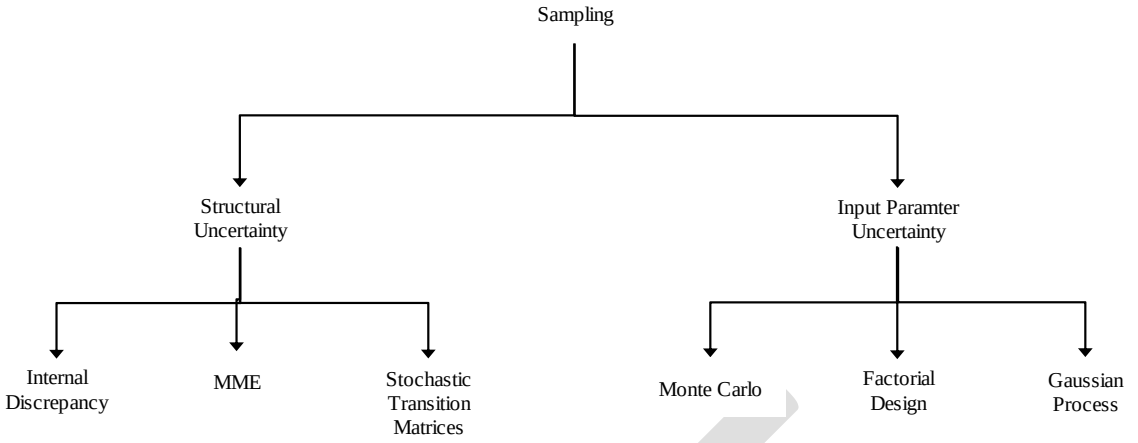

Figure 4: Some sampling techniques

3.2.7 Analyse/Summarise data

Once the simulation data has been generated, analysis and summary of the information can begin. Depending on the objectives defined in step 5 (Section 3.2.5), the data will be analysed either to identify areas in need of further research (Model Assessment), or to predict or smooth with confidence ranges representing the desired sources of information (Model Application).

Figure 5 shows a possible classification of uncertainty evaluation techniques in this phase. Some techniques based on real-world observation data (calibration) or not (sensitivity analysis) are shown. These and others are described in Section 5. Graphical methods are placed apart from either because they should be a part of any uncertainty evaluation, regardless of the presence or absence of observational data. The calibration techniques are further split into static and dynamic techniques in the discussion in Section 5.

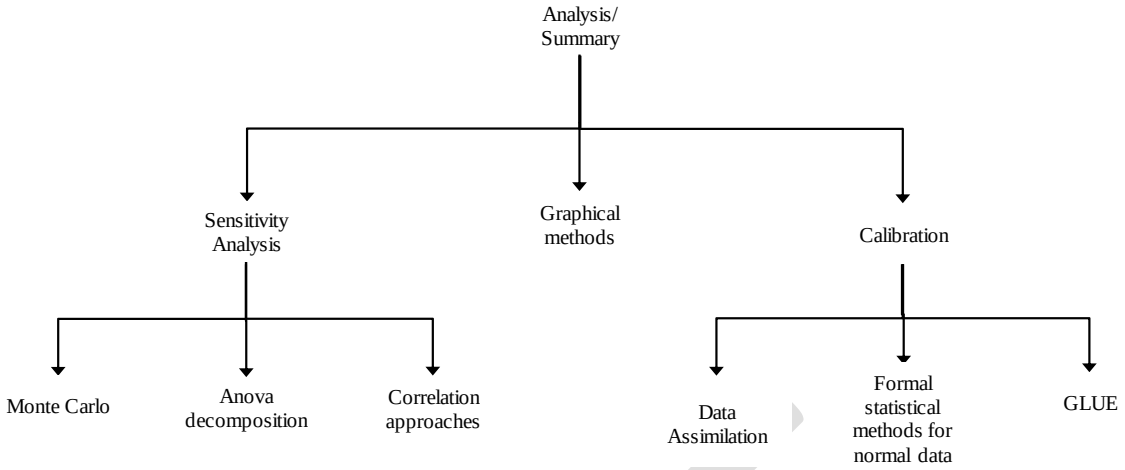

Figure 5: Analysis techniques with and without observed data

3.2.8 An uncertainty evaluation examplar

An example of the process that could be followed during a model evaluation is given in

Figure 6. This is just one of many possible techniques, and some of the techniques might be equally applicable at other termination points; for example multi-model ensembles could be used with observations if desired. Neither is this figure complete, there are many techniques that have not been discussed in this paper. This figure is merely an indication of the types of questions that should be asked during the evaluation, and the class of techniques that could provide insight as a result.

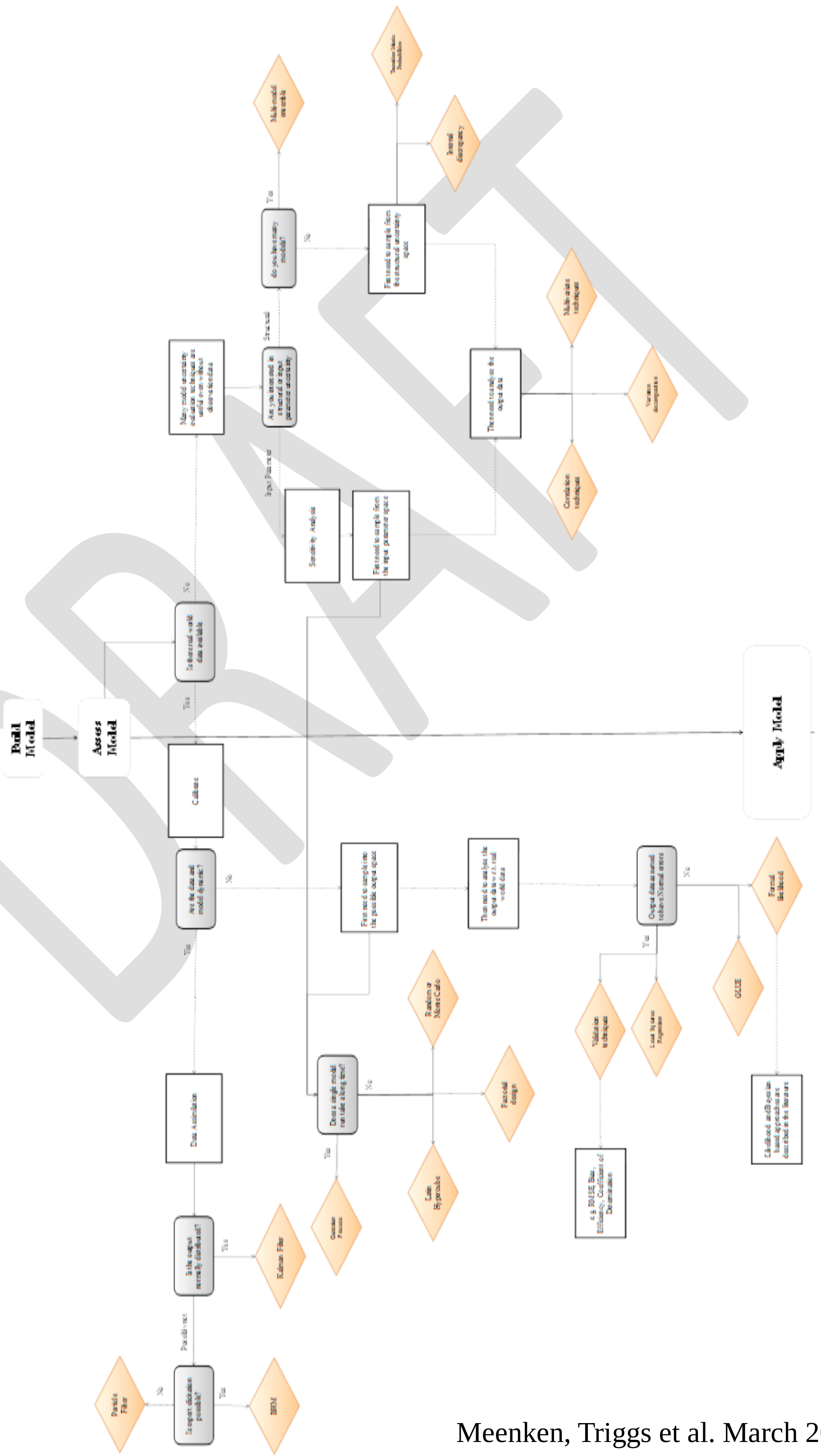

Figure 6: An uncertainty evaluation exemplar

4. Sampling

4.1 Generation of data to represent input parameter uncertainty

4.1.1 Simple sampling

Since there are usually many possible values for input parameters $\boldsymbol{x}$ of varying levels of plausibility (Helton 1997), its uncertainty can be characterized by assigning a distribution probability distribution function (pdf), and thus defining a sampling space that is practically reasonable. These pdfs characterise a degree of belief with respect to where the appropriate input to use in the analysis is located, and this translates to a belief about the appropriate values of the distribution of outcomes $\boldsymbol{y}$. According to (Helton and Davis 2000), it is usually most helpful to elicit expert opinion to characterise the distributions of $\boldsymbol{x}$. The process of expert elicitation is discussed by (O'Hagan and Oakley 2004, Strong 2012) amongst others. One straightforward approach is then to use simple random sampling from the appropriate range, or Monte Carlo sampling to simulate the pdf of the input parameters; however this may not allow thorough investigation of interactions and correlated input parameters.

4.1.2 Factorial experiments

The well-known designed experiment (e.g. (Fisher 1926, Cochran and Cox 1957, Mead 1988)) can be readily extended to computer simulation experiments by selecting the combinations of factor values that will be actually simulated (Sacks, Welch et al. 1989) based on the pdfs described above. Factorial and fractional factorial experiments can be used when there are relatively few factors or variates that can be summarised by a manageable number of sensible factor levels. Analysis of the data is straightforward using ANOVA like decompositions, and can help

- Identify less important terms, so that they can be set to their nominal values and other terms more fully explored.
- Identify interactions between variables (Santner, Williams et al. 2003).

Note that in deterministic computer experiments the lack of random error leads to a number of differences from traditional design of experiments:

- The absence of random error ensures the complexity of the computer code is not disguised.
- The adequacy of a response-surface model fitted to the observed computer data is determined solely by systematic bias.
- No need for blocking.
- Concepts of experimental unit, replicate and randomization are irrelevant. (Sacks, Welch et al. 1989).

4.1.3 Gaussian Process Emulators

The sampling techniques described above usually demand a very large number of model runs, and when a single model run takes several minutes or more, these methods quickly become impractical. One way to reduce the CPU load in the optimization is to use response surfaces as proxies to the true model response. A statistical representation of the simulator, known as a meta-model or *emulator*s. It is analogous to regression modelling or multivariate neural networks, but more flexible, accurate and efficient than these methods in challenging problems where there is limited information about the simulator. A full mathematical treatment of this technique is given by (Kennedy and O'Hagan 2001, Oakley and O'Hagan 2002, Oakley and O'Hagan 2004). A tutorial for non-statisticians is given in (O'Hagan 2006).

The use of Gaussian process emulators has been a very large area of research across many disciplines, both extending the methodologies via research in different areas of mathematical

complexity e.g. (Oakley and O'Hagan 2004, Bhattacharya 2007, Rougier 2008, Conti and O'Hagan 2010, Johnson, Gosling et al. 2011, Wilkinson, Vrettas et al. 2011) and across a wide range of biological, medical, oceanographic, climate, economic and engineering applications e.g. (O'Hagan, Stevens et al. 2001, Stevens, O'Hagan et al. 2003, O'Hagan 2005, Kennedy, Anderson et al. 2006, Rougier, Guillas et al. 2009, Becker, Rowson et al. 2011, Fricker, Oakley et al. 2011, Hall, Manning et al. 2011, O'Hagan 2012, Strong, Oakley et al. 2012, Cripps, O'Hagan et al. 2013).

4.1.4 Other input parameter sampling plans

When there are many nonlinear input variables that may interact to form complex response surfaces, the choices of inputs that will adequately describe the simulation space is less straightforward. There had been a large amount of research in this area, and texts by (Santner, Williams et al. 2003) and (Saltelli, Chan et al. 2000) provide summaries of these. If two or more input variables are correlated then it is necessary that the appropriate correlation structure be incorporated into the sample if meaningful results are to be obtained in subsequent analyses (Iman and Davenport 1982, Jacques, Lavergne et al. 2006, Da Veiga, Wahl et al. 2009).

4.2 Generation of data to represent structural uncertainty

4.2.1 Internal Discrepancy Approach

This technique, put forward by Strong, Oakley et al. (2012), is based on specifying a distribution for the model structural error. There is no attempt to make assessments about the adequacy of the model structure in relation to alternative structures as in ensemble methodologies (Section 4.2.2); instead we assess how large an error might be due to the structure of the model. The method requires the ability to decompose the model into 'subfunctions'. Where there is thought to be potential structural error, a discrepancy term is

introduced. The key idea is that in some applications it is easier to make judgements about internal discrepancies than about the external discrepancy which results from inadequacies throughout the whole model (Strong 2012, Strong, Oakley et al. 2012, Strong and Oakley 2014). (Strong, Oakley et al. 2012) introduced this technique for simple discrete subfunctions, however in a more complex (e.g. dynamic) case, then we may want to introduce discrepancies at each time step, most likely leading to a correlated structure. This idea was explored in a state-space context by (Strong and Oakley 2014). A similar idea has been explored in the context of multiple models by (Goldstein and Rougier 2009).

4.2.2 Ensemble Methods

Also known as model averaging, multi-model ensembles (MME's) is usually defined as a technique that incorporate outputs across more than one model. In this technique the predictions or probability statements of a number of plausible models are averaged, with weights based either on some measure of model adequacy or some measure of the probability that the model is true (Draper 1995, Kadane and Lazar 2004, Strong, Oakley et al. 2012, Strong and Oakley 2014). Because by incorporating the views of many research groups/scientists and experts, the structural effects are said to be better described than if only one model is used (Gal, Makler-Pick et al. 2014)). Model averaging is an technique implemented for simulators e.g. (Rougier 1996, Kalnay 2003, Raftery, Gneiting et al. 2005, Rotter, Carter et al. 2011) and both frequentist and Bayesian approaches to model averaging can be used to allocate weights to the outputs from different models based on data e.g. (Bernardo and Smith 1994, Claeskens and Hjort 2008, Montgomery, Hollenbach et al. 2012).

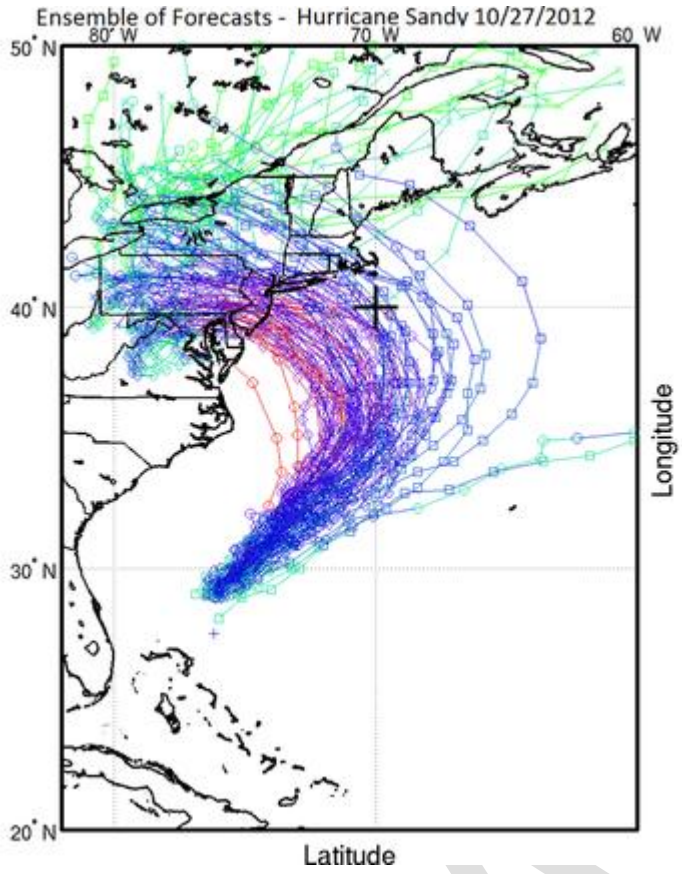

Figure 7: Aggregate of many ensemble model forecasts for Hurricane Sandy 60 hours before landfall

An example can be seen in Figure 7 which was taken from (http://comap.weebly.com/ensemble-forecasting-and-post-processing.html). Each line denotes a model forecast track for the centre of Hurricane Sandy.

(Van Ittersum, Ewert et al. 2008, Ewert, van Ittersum et al. 2009)

This technique has also applied in the hydrological modelling area (Hsu, Moradkhani et al. 2009, Rings, Vrugt et al. 2012). Here Clark, Slater et al. (2008) attempted to quantify the uncertainty in model structure by using a method they called 'Framework for Understanding Structural Errors ' (FUSE). This technique constructs many unique model structures by combining components from a smaller number of parent models.

4.2.3 Transition Matrix probabilities

A very common implementation is to incorporate stochasticity within the transition matrices of a state space model e.g. as described by (Spiegelhalter and Best 2002) for a discrete state-space medical cost-effectiveness model.

5. Analysis/Summary at each phase of the model life

This section discusses techniques for analysis and summary of simulation data at each of the three phases of the model life (Figure 1) consecutively: Techniques for Model Building are discussed in Section 5.1; for Model Assessment in Section 5.2 and for Model Application in Section 5.3. However, often techniques may be useful at other phases than the one to which it has been allocated here.

5.1 Uncertainty evaluation during the model building phase

Some commonly used (smith and smith, Edith;s paper?) are described next, and together help understand how the modelled data and the observed data are related.

The total difference between the simulated and measured values as calculated by the root mean square error: RMSE:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(O_i - P_i)^2}{n}}$$

Where $i = 1,2\ldots,n$ observed values. P represents a simulated value and O an observed value with mean $\bar{O}$.

The mean bias in the total difference between simulations and measurements is determined by:

$$Bias = \frac{\sum_{i=1}^{n} O_i - P_i}{n}$$

The coefficient of determination (CD), is in some ways equivalent to the well-known r^2 used to assess the fit of a simple ordinary least square regression model. The key difference is that the total variation in the predicted values from the mean of the measurements may be greater than that of the observed data. This could occur for example if the model were very biased. CD is then defined:

$$CD = \frac{\sum_{i=1}^{n}(P_i - \overline{O})^2}{\sum_{i=1}^{n}(O_i - \bar{O})^2}$$

CD describes the proportion of the total variance in the observed data that is explained by the predicted data. The lowest value of CD is 0. A value below 1 indicates that the model describes the measured data better than the mean of the measurements, and a value of 1 or above indicates it is worse.

If the model describes the measured data better than the mean of the measurements, then an indication of the efficiency, EF, of the model can then be found by calculating the 1 minus the CD. This is because then (and only then) the following equality holds:

$$\sum_{i=1}^{n}(O_i - \bar{O})^2 = \sum_{i=1}^{n}(O_i - P_i)^2 + \sum_{i=1}^{n}(P_i - \overline{O})^2$$

And then:

$$EF = \frac{\sum_{i=1}^{n}(O_i - P_i)^2}{\sum_{i=1}^{n}(O_i - \bar{O})^2} = 1 - CD$$

If the model is not better than the mean of the measurements, then values for EF can still be calculated, but the result may be positive or negative with a maximum value of 1. A negative value indicates the simulated values describe the data less well than a mean of the observations as for the CD.

Pearson's correlation coefficient, denoted *r*, is useful to assess how well the shape of the simulation data matches the shape of the measured data (i.e. is the relationship monotonic). This value will be between 0 and 1, with one indicating perfect correlation.

$$r = \frac{cov(O,P)}{var(O),var(P)} = \frac{\sum_{i=1}^{n}(O_i - \bar{O}) \quad \sum_{i=1}^{n}(P_i - \bar{P})}{\sqrt{\sum_{i=1}^{n}(O_i - \bar{O})^2}\sqrt{\sum_{i=1}^{n}(P_i - \bar{P})^2}}$$

Finally, simple least squares regression modelling testing the following hypotheses will help describe straight-line departures from the 1:1 line using F-statistics to test the following hypotheses:

Taken together, RMSE, Bias, CD, EF, *r* and linear regression can help explore the model more thoroughly than one measure alone.

5.2 Uncertainty evaluation during the model assessment phase

5.2.1 Not dependent on Data - Sensitivity Analysis

Background

Sensitivity Analysis (SA) assumes the form of the model as defined by the state equations is adequate (Saltelli, Chan et al. 2000). SA, like calibration techniques, is dependent on the generation of a reliable simulated data set. SA is discussed in detail in in many places; see (Chatfield and Collins 1980, Sacks, Welch et al. 1989, Koehler and Owen 1996, Krzanowski 2000, Saltelli, Chan et al. 2000, Santner, Williams et al. 2003, Cacuci, Ionescu-Bujer et al. 2005, Kurowicka and Cooke 2006, Saltelli, Ratto et al. 2006) for details on the design and

analysis of computer experiments, multivariate analysis techniques, partitioning of variance, and other useful tools for sensitivity analysis.

Variance Decomposition

Variance based methods are a particularly useful class of global sensitivity analysis techniques. Sensitivity indices (and importance ratios arising from them) are based on an ANOVA type decomposition of the function **f(x)** and can be used to assess the sensitivity of the output to individual variables or combinations of variables, even when the effects are not linear (Santner, Williams et al. 2003, Wallach, Makowski et al. 2014). This is a large area of research that is still very active i.e. (Prieur 2014), so a very brief overview follows (Saltelli, Chan et al. 2000):

- The use of variance as an indicator of importance for input factors also underlie regression based methods,
- These techniques are useful for situations with non-linearity and or non-monotonicity in **y(x)**,
- Variance decomposition same as that for standard Design of Experiments if orthogonal inputs (Archer, Saltelli et al. 1997),
- Correlation ratio, (McKay 1995) and importance measures, (Hora and Iman 1986), are equivalent and based on conditional variance of model outputs using on a simple description of uncertainty using probability distributions. Regression-based methods are special cases of this type of method (i.e. can use sums of squares derived from analysis of variance for estimation when inputs are independent and orthogonal),
- However, these analyses are not appropriate when inputs are not independent and orthogonal e.g. for Monte Carlo type designs.

Correlation based methods

There are many standard statistical texts that provide detailed descriptions of this methods (e.g. (Draper and Smith 1981)), (Saltelli, Chan et al. 2000) is an excellent resource. Nonlinear regression extends linear least squares for use with a larger and more general class of functions, i.e. (Trucano, Swiler et al. 2006) where linearity is not a reasonable assumption.

Multivariate statistics

(Chatfield and Collins 1980, Krzanowski 2000, Harding and Payne 2011) are three of many resources on multivariate data analysis.

5.2.2 Dependent on data: Dynamic data calibration

Data assimilation is the process by which observations are incorporated into the estimates from a simulator, using all available information for optimal prediction. It is distinct from calibration techniques discussed in section 5.2.3 next in that it makes use of observed data as it becomes available through time. Data assimilation is based on a two-step process and depends on a state-space model formulation, the forecast step, and the update (filtering) step. Figure 8 provides a slightly altered representation of a view of data assimilation provided by (Lewis, Lakshmivarahan et al. 2006) that shows how the state evolution model, observations and input parameters are unified to provide filtered estimates that can be used to explore model sensitivity, uncertainty and predictability.

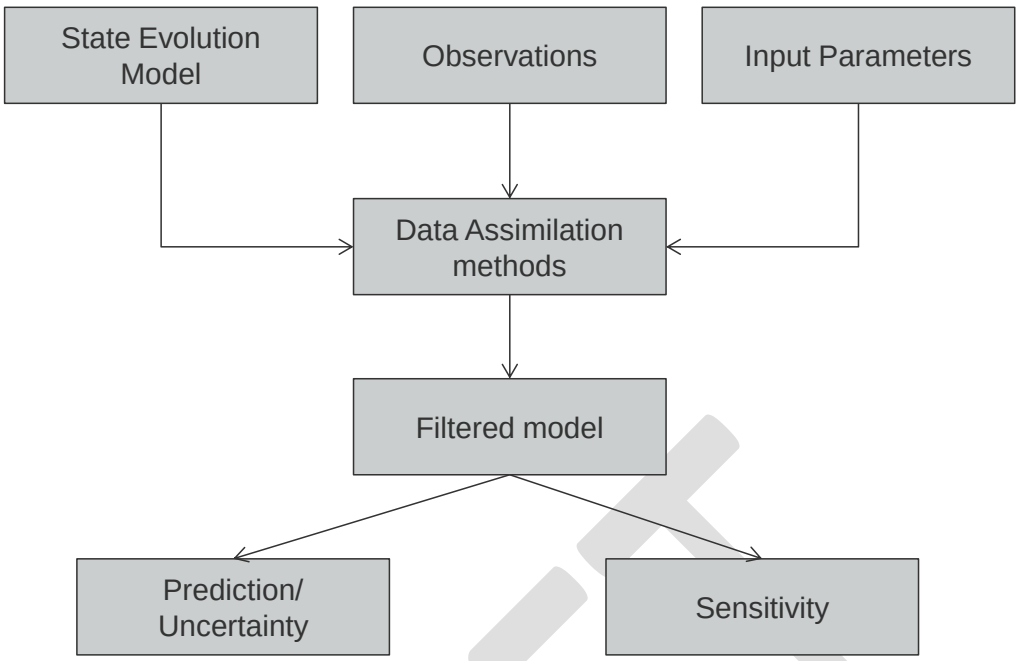

Figure 8: A view of data assimilation slightly altered from (Lewis, Lakshmivarahan et al. 2006)

Bayesian data assimilation

The Bayesian data assimilation approach takes the conditional probabilities inherent in Bayesian hierarchical models i.e. (Carlin and Louis 2000, Gelman, Carlin et al. 2006) and incorporates the ability to update (filter) state predictions using recursive Bayesian model properties. It independently allows the incorporation of a probability distribution function describing input parameter uncertainty when appropriate, perturbation of structural state equations if desired and up-to-date information from data when available. It is best paradigm to date in which to partition variability and quantify input parameter, data and structural uncertainty, in addition to addressing any problems with initial conditions (Cressie and Wikle 2011). The joint pdf of all the quantities in the model (i.e. the state model, and expert prior knowledge of input parameters) results from multiplying together the conditional pdfs to provide an estimate of the process under study at time *k* based on all of the data available at that time (Gordon, Salmond et al. 1993, Higdon 2007, Candy 2009, Vrugt, Braak et al. 2009, Vrugt, ter Braak et al. 2009, Cressie and Wikle 2011, Murray 2013). This

technique allows an enhanced, dynamic signal with associated performance statistics to be estimated. An example is given in accompanying paper.

Kalman Filter

The Kalman Filter is the most well-known of all the Bayesian data assimilation techniques. It is optimal in the very specific case that assumes normality in the noise component of the model outputs and the observation data. There are many books and tutorials describing the Kalman Filter e.g. (Thacker and Lacey 1996, Lewis, Lakshmivarahan et al. 2006).

The Kalman filter uses a series of (noisy) measurements observed over time to produce updated estimates of state variables. Use of the Kalman filter requires the same matrices need to be specified as discussed in Section 2 with regards to the state-space model framework: $\mathbf{Z}_k$, the state-transition model; $\mathbf{H}_k$, the observation model. We further specifically define $\mathbf{CovP}_k$, the covariance of the process noise and $\mathbf{CovObs}_k$, the covariance of the observation noise. Sometimes $\mathbf{B}_k$, the control-input model for each time-step, k, is also included, but we have not done so. Therefore, in the linear form of the Kalman filter model assumes the true state **t** at time k is evolved from the state at (k−1) according to the **state equation:**

$$\boldsymbol{t}_k = \boldsymbol{Z}_k \boldsymbol{t}_{k-1} + \boldsymbol{w}_k$$

Where $\mathbf{Z}_k$ is the state transition model which is applied to the previous state $\mathbf{x}_{k-1}$ with $\mathbf{w}_k$ being the additive, linear process noise which is assumed to be drawn from a zero mean multivariate normal distribution with covariance $\mathbf{CovP}_k$.

$$\boldsymbol{w}_k \sim \boldsymbol{N}(\mathbf{0}, \boldsymbol{CovP}_k)$$

We cannot directly observe the true state vector $\mathbf{t}_k$. At time k an observation (or measurement) $\mathbf{Q}_k$ of the true state $\mathbf{t}_k$ is made according to the **observation equation:**

$$\boldsymbol{Q_k = H_k t_k + v_k}$$

Where $\mathbf{H_k}$ is the observation model which maps the true state space into the observed space and $\mathbf{v_k}$ is the additive, linear observation noise which is assumed to be zero mean Gaussian noise with covariance **CovObs**$_k$**.**

$$\boldsymbol{v_k \sim N(0, CovObs_k)}$$

Extended and ensemble Kalman Filter

These techniques are extensions to when the linearity requirement of the Kalman Filter cannot be met (Evensen 2007, Fowler 2012).

Particle Filter

Particle filters (also known as Sequential Monte Carlo (SMC)). This is not a Bayesian approach. Particle filters use a grid-based approach, and use a set of particles to represent the posterior density. The particle filter works by propagating and then updating a set of random samples (particles) to approximate the required continuous probabilistic distribution. Figure 9 below shows a visual representation of the process and is taken from (Bando, Shibata et al. 2007). The state-space model can be non-linear and the initial state and noise distributions can take any form required i.e. (Arulampalam, Maskell et al. 2002, Künsch 2013).

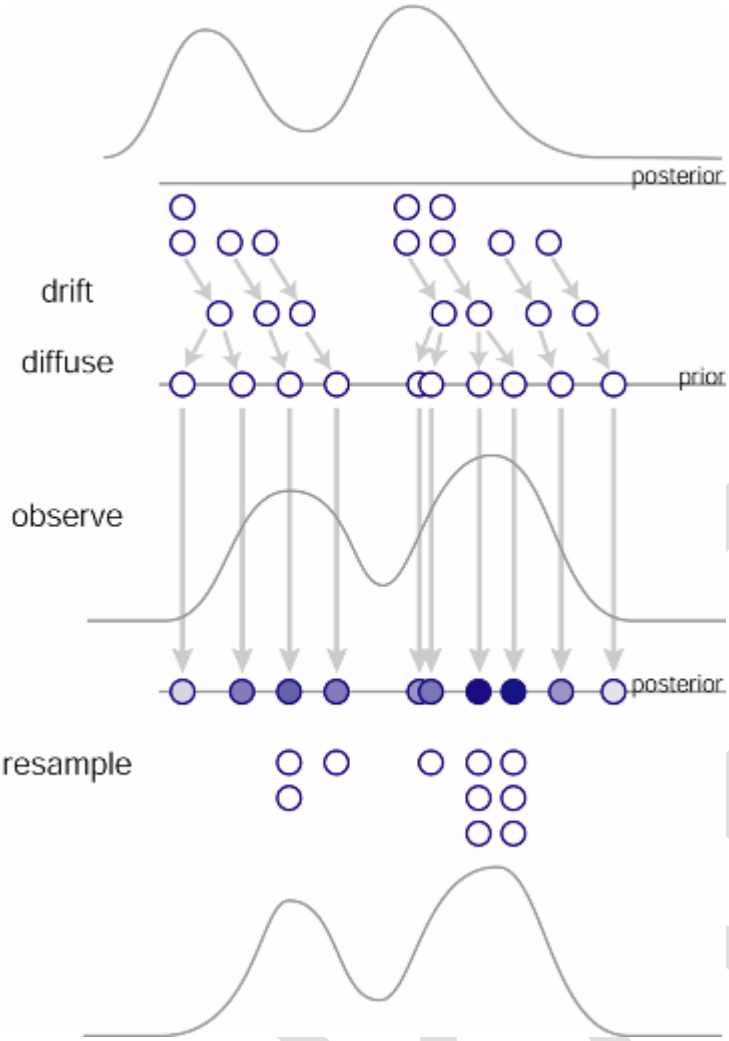

Figure 9: Visual representation of a particle filtering sample-resample cycle.

Bayesian model with Particle Filter

Bulygina and Gupta (2009) used a Bayesian data assimilation approach to directly construct the form of the input parameters, outputs and state variables such that they are statistically consistent with data measurements of the system, and then incorporated the method of particle filtering to construct efficient estimates of the pdfs of the internal model structure.

5.2.3 Dependent on Data: Static Data Calibration

Calibration for output data assumed to be Normally distributed

Depending on whether optimisation or uncertainty assessment is required, there are many options at this stage for model evaluation. These may include the Efficiency, Bias, RMSE

and CD estimates described above in the validation section. Linear and nonlinear least-squares regression are the most common example of calibration methods in practice. However, there are some fairly stringent assumptions (Trucano, Swiler et al. 2006) usual to analyses assuming the normal distribution.

There are situations in which the data are correlated, nonstationary or non-Gaussian, and researchers publishing in the area of hydrological models in particular have explored formal and informal statistical techniques when normality assumptions about residuals may be inappropriate. One informal technique is discussed in this paper, however the following authors have discussed other, formal techniques (Kavetski, Franks et al. 2002, Kavetski, Kuczera et al. 2006, Kuczera, Kavetski et al. 2006, Montanari and Grossi 2008, Thyer, Renard et al. 2009, Renard, Kavetski et al. 2010, Schoups and Vrugt 2010)

Calibration for output data not assumed to be Normally distributed

One well-known informal method was put forward by (Beven and Binley 1992), with further work done by (Beven and Freer 2001, Beven 2006). The methodology, called Generalised Likelihood Uncertainty Estimation (GLUE) has been extremely popular (the 1992 paper has received greater than 1,000 citations since its publication e.g. (Blazkova and Beven 2009, Juston, Andrén et al. 2010)). The methodology works by as follows:

1. Sample from the space (as discussed in Section 4.1.1) of each input parameter to generate a model scenario.
2. Fitting the model using the scenario.
3. Use the resulting simulated data to assess how well the model fits against some observed data using a pre-selected rule.
4. Accept or reject that scenario.
5. Uncertainty bounds are set at the desired percentage of the accepted models.

This technique and others have been appraised by authors such as (Vrugt, Gupta et al. 2003, Pappenberger 2006, Blasone, Vrugt et al. 2008, Stedinger, Vogel et al. 2008, Vrugt, Braak et al. 2009).

Calibration allowing for code uncertainty

The techniques outlined above are used in calibrating parameters of models, taking into account uncertainty in the observation data but assuming no uncertainty in the model itself (structural uncertainty). The Bayesian statistics community has developed formal statistical methods that address code uncertainty, one important technique is that of (Kennedy and O'Hagan 2001) which is an analysis method that builds upon the emulator described in section 3.2.3 The representation given by (Kennedy and O'Hagan 2001) equation (5) describes the relationship between the observations, the true process and the computer model output:

$$z_i = \varsigma(\boldsymbol{x}_i) + e_i = \rho\eta\left(\boldsymbol{x}_{i,}\theta\right) + \delta(\boldsymbol{x}_i) + e_i$$

Where e_i is the observation error for the *i*th observation, ρ is an unknown regression parameter and $\delta(\cdot)$ is a model inadequacy function that is *independent* of the code output $\eta(\cdot,\cdot)$.

There are as yet some limitations to this approach; the error term e_i is assumed to be normally distributed without systematic error as N(0,λ), and the constant regression parameter ρ implies that the underlying observation process $\delta(\boldsymbol{x}_i) + e_i$ is stationary, which may be unrealistic in some applications (Trucano, Swiler et al. 2006, McFarland 2008).

5.2.3 Graphical tools

As with any statistical analysis, exploratory data analysis (EDA) is an important step to help understand patterns in the data. Problems with linearity and monotonicity can be identified, and help guide selection of appropriate analysis techniques (Kurowicka and Cooke 2006).

5.3 Uncertainty evaluation during the model application phase

Figure 10 summarises the options for describing uncertainty in the model as part of the forecasting/smoothing phase of the models life. This phase can utilise the techniques discussed in Sections 4 and 5 to provide a range for predictions rather than point estimates only. As in previous sections, this figure underscores the type of uncertainty that can be evaluated with that technique.

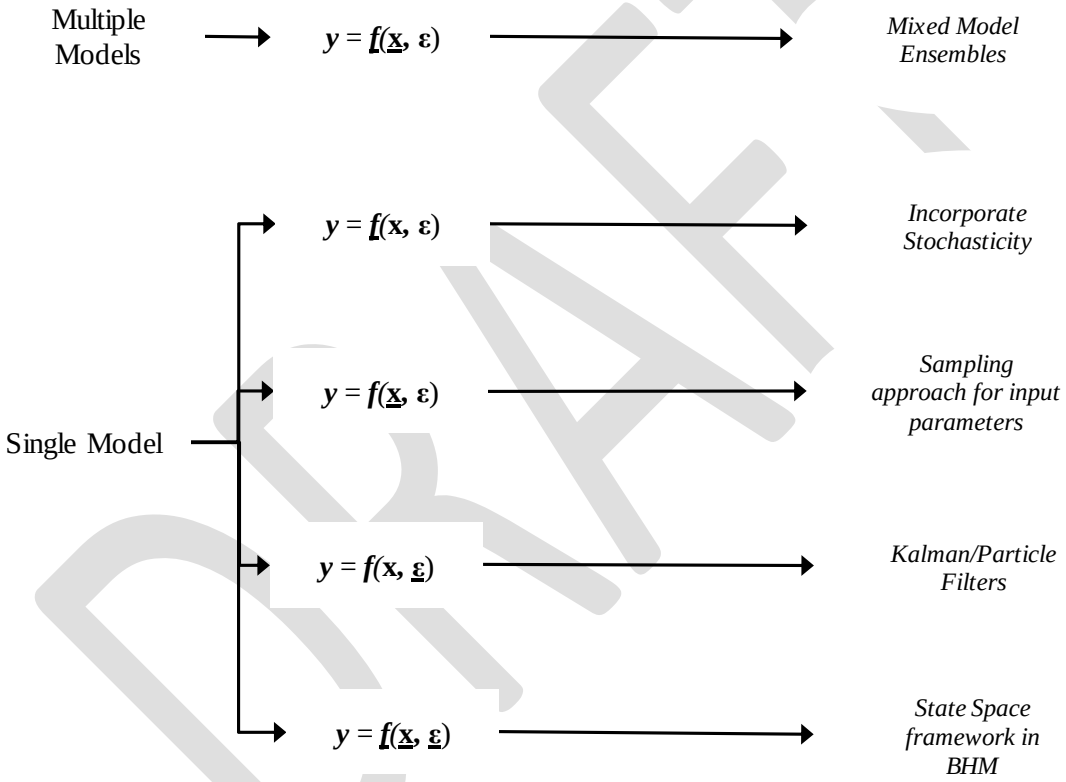

Figure 10: Tools to evaluate simulator uncertainty during the forecasting/smoothing phase. Ovals represent evaluation techniques, and rectangles define types of uncertainty explored with the technique in question.

6. Summary

References

Archer, G., A. Saltelli and I. Sobol (1997). "Sensitivity measures, ANOVA-like techniques and the use of bootstrap." Journal of Statistical Computation and Simulation **58**(2): 99-120.

Arulampalam, M. S., S. Maskell, N. Gordon and T. Clapp (2002). "A Tutorial on Particle Filters for Online Nonlinear/Non-Gaussian Bayesian Tracking." IEE Transactions on Signal Processing, **50**(2).

Bando, T., T. Shibata and S. Ishii. (2007). "On-line Variational PCA fro Adaptive Visual Traching." 2014.

Bayarri, M., J. Berger and D. M. Steinberg (2009). "Special issue on computer modeling." Technometrics **51**(4): 353-353.

Becker, W., J. Rowson, J. E. Oakley, A. Yoxall, G. Manson and K. Worden (2011). "Bayesian sensitivity analysis of a model of the aortic valve." Journal of Biomechanics **44**(8): 1499-1506.

Bernardo, J. and A. Smith (1994). Bayesian Theory.

Beven, K. (2006). "On undermining the science?" Hydrological Processes **20**(14): 3141-3146.

Beven, K. and A. Binley (1992). "The future of distributed models: Model calibration and uncertainty prediction." Hydrological Processes **6**(3): 279-298.

Beven, K. and J. Freer (2001). "Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology." Journal of Hydrology **249**(1–4): 11-29.

Bezlepkina, I., M. Adenäeur, M. Kuiper, S. Janssen, R. Knapen, A. Kanellopoulos, F. Brouwer, J. Wien, J. Wolf and M. van Ittersum (2010). "Using the SEAMLESS integrated framework for ex-ante assessment of trade policies." Towards effective food chains: models and applications. Wageningen Academic Publishers, Wageningen, the Netherlands: 251-271.

Bhattacharya, S. (2007). "A simulation approach to Bayesian emulation of complex dynamic computer models." Bayesian Analysis **2**(4): 783-815.

Blasone, R.-S., J. A. Vrugt, H. Madsen, D. Rosbjerg, B. A. Robinson and G. A. Zyvoloski (2008). "Generalized likelihood uncertainty estimation (GLUE) using adaptive Markov Chain Monte Carlo sampling." Advances in Water Resources **31**(4): 630-648.

Blazkova, S. and K. Beven (2009). "A limits of acceptability approach to model evaluation and uncertainty estimation in flood frequency estimation by continuous simulation: Skalka catchment, Czech Republic." Water Resour. Res. **45**(12): W00B16.

Boote, K. J., J. W. Jones and N. B. Pickering (1996). "Potential Uses and LImitation of Crop Models." Agronomy Journal **88**: 704-716.

Bulygina, N. and H. Gupta (2009). "Estimating the uncertain mathematical structure of a water balance model via Bayesian data assimilation." Water Resour. Res. **45**(12): W00B13.

Cacuci, D. G., M. Ionescu-Bujer and I. M. Navon (2005). Sensitivity and Uncertainty Analysis. Boca Raton, Chapman & Hall/CRC.

Candy, J. V. (2009). Bayesian Signal Processing: Classical, Modern and Particle Filtering Methods. New Jersey, John Wiley and Sons.

Carlin, B. P. and T. A. Louis (2000). Bayes and Empirical Bayes Methods for Data Analysis. New York, Chapman & Hall.

Carson, J. S. (2002). Model verification and validation. Simulation Conference, 2002. Proceedings of the Winter, IEEE.

Chatfield, C. and A. J. Collins (1980). Introduction to Multivariate Analysis. New York, Chapman and Hall.

Chichota, R., V. O. Snow and F. M. Kelliher (2013). Sensitivity analysis to investigate the factors controlling the effectivesness of a nitrificaiton inhibitor in the soil. 20th International Congress on Modelling and Simulation. Adelaide, Australia.

Claeskens, G. and N. L. Hjort (2008). Model Selection and Model Averaging. Cambridge, Cambridge University Press.

Clark, M. P., A. G. Slater, D. E. Rupp, R. A. Woods, J. A. Vrugt, H. V. Gupta, T. Wagener and L. E. Hay (2008). "Framework for Understanding Structural Errors (FUSE): A modular framework to diagnose differences between hydrological models." Water Resour. Res. **44**: W00B02.

Clifford, D., D. Pagendam, D. Baldock, N. Cressie, R. Farquaharson, M. Farrell, L. Macdonald and L. Murray (2013). Bayesian Hierarchical modeling of soil carbon dynamics. 20th International Congress on Modelling and Simulation. Adelaide, Australia.

Cochran, W. G. and G. M. Cox (1957). Experimental Designs, John Wiley & Sons.

Conti, S. and A. O'Hagan (2010). "Bayesian emulation of complex multi-output and dynamic computer models." Journal of Statistical Planning and Inference **140**(3): 640-651.

Cooper, M., F. A. van Eeuwijk, G. L. Hammer, D. W. Podlich and C. Messina (2009). "Modeling QTL for complex traits: detection and context for plant breeding." Current Opinion in Plant Biology **12**(2): 231-240.

Cressie, N. and C. K. Wikle (2011). Statistics for Spatio-Temporal Data. Hoboken, New Jersey, John Wiley & Sons.

Cripps, E., A. O'Hagan and T. Quaife (2013). "Quantifying uncertainty in remotely sensed land cover maps." Stochastic Environmental Research and Risk Assessment **27**(5): 1239-1251.

Da Veiga, S., F. Wahl and F. Gamboa (2009). "Local polynomial estimation for sensitivity analysis on models with correlated inputs." Technometrics **51**(4): 452-463.

Draper, D. (1995). "Assessment and propagation of model uncertainty." Journal of the Royal Statistical Society. Series B (Methodological): 45-97.

Draper, N. R. and H. Smith (1981). "Applied regression analysis 2nd ed."

Evensen, G. (2007). Data Assimilation: The Ensemble Kalman Filter. Berlin, Springer.

Ewert, F., M. K. van Ittersum, I. Bezlepkina, O. Therond, E. Andersen, H. Belhouchette, C. Bockstaller, F. Brouwer, T. Heckelei and S. Janssen (2009). "A methodology for enhanced flexibility of integrated assessment in agriculture." Environmental Science & Policy **12**(5): 546-561.

Fisher, R. A. (1926). "The arrangement of field experiments." Journal Minist. Agric **33**: 503-513.

Fowler, A. (2012) "Data Assimilation tutorial on the Kalman filter."

Fricker, T. E., J. E. Oakley, N. D. Sims and K. Worden (2011). "Probabilistic uncertainty analysis of an FRF of a structure using a Gaussian process emulator." Mechanical Systems and Signal Processing **25**(8): 2962-2975.

Gal, G., V. Makler-Pick and N. Shachar (2014). "Dealing with uncertainty in ecosystem model scenarios: Application of the single-model ensemble approach." Environmental Modelling & Software **61**(0): 360-370.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari and D. B. Rubin (2006). Bayesian Data Analysis. Boca Raton, Florida, Chapman & Hall/CRC.

Goldstein, M. and J. C. Rougier (2009). "Reified Bayesian modelling and inference for the physical systems." Journal of Statistical Planning and Inference.

Gordon, N. J., D. J. Salmond and A. F. M. Smith (1993). "Novel approach to nonlinear/non-Gaussian Bayesian state estimation." Radar and Signal Processing, IEE Proceedings F **140**(2): 107-113.

Gupta, H. V., M. P. Clark, J. A. Vrugt, G. Abramowitz and M. Ye (2012). "Towards a comprehensive assessment of model structural adequacy." Water Resources Research **48**(8): W08301.

Hall, J. W., L. J. Manning and R. K. S. Hankin (2011). "Bayesian calibration of a flood inundation model using spatial data." Water Resources Research **47**.

Hammer, G. L., M. J. Kropff, T. R. Sinclair and J. R. Porter (2002). "Future contributions of crop modelling - from heuristics and supporting decision making to understanding genetic regulation and aiding crop improvement." European Journal of Agronomy **18**: 15-31.

Harding, S. and R. W. Payne (2011). A Guide to Multivariate Analysis in GenStat. Rothamsted, VSN International Ltd.

Helton, J. C. (1997). "Uncertainty and sensitivity analysis in the presence of stochastic and subjective uncertainty." Journal of Statistical Computation and Simulation **57**(1-4): 3-76.

Helton, J. C. and F. J. Davis (2000). Sampling-Based Methods. Sensitivity Analysis. A. Saltelli, K. Chan and E. M. Scott. Chichester, John Wiley & Sons.

Higdon, D. (2007). A Primer on Space-Time Modeling from a Bayesian Perspective. Statistical Methods for Spatio Temporal Systems. Boca Raton, FL, Chapman & Hall/CRC.

Hochman, Z., H. Van Rees, P. Carberry, J. Hunt, R. McCown, A. Gartmann, D. Holzworth, S. Van Rees, N. Dalgliesh and W. Long (2009). "Re-inventing model-based decision support with Australian dryland farmers. 4. Yield Prophet® helps farmers monitor and manage crops in a variable climate." Crop and Pasture Science **60**(11): 1057-1070.

Holzkämper, A., T. Klein, R. Seppelt and J. Fuhrer (2015). "Assessing the propagation of uncertainties in multi-objective optimization for agro-ecosystem adaptation to climate change." Environmental Modelling & Software **66**(0): 27-35.

Holzworth, D. P., N. I. Huth, P. G. deVoil, E. J. Zurcher, N. I. Herrmann, G. McLean, K. Chenu, E. J. van Oosterom, V. Snow, C. Murphy, A. D. Moore, H. Brown, J. P. M. Whish, S. Verrall, J. Fainges, L. W. Bell, A. S. Peake, P. L. Poulton, Z. Hochman, P. J. Thorburn, D. S. Gaydon, N. P. Dalgliesh, D. Rodriguez, H. Cox, S. Chapman, A. Doherty, E. Teixeira, J. Sharp, R. Cichota, I. Vogeler, F. Y. Li, E. Wang, G. L. Hammer, M. J. Robertson, J. P. Dimes, A. M. Whitbread, J. Hunt, H. van Rees, T. McClelland, P. S. Carberry, J. N. G. Hargreaves, N. MacLeod, C. McDonald, J. Harsdorf, S. Wedgwood and B. A. Keating (2014). "APSIM – Evolution towards a new generation of agricultural systems simulation." Environmental Modelling & Software **62**(0): 327-350.

Hora, S. C. and R. L. Iman (1986). A comparison of maximus/bounding and the Bayes/Monte Carlo for fault tree uncertainty analysis, Sandia Nat. Lab.

Hsu, K.-l., H. Moradkhani and S. Sorooshian (2009). "A sequential Bayesian approach for hydrologic model selection and prediction." Water Resour. Res. **45**(12): W00B12.

Iman, R. L. and J. M. Davenport (1982). "Rank correlation plots for use with correlated input variables." Communications in Statistics - Simulation and Computation **11**(3): 335-360.

Jacques, J., C. Lavergne and N. Devictor (2006). "Sensitivity analysis in presence of model uncertainty and correlated inputs." Reliability Engineering & System Safety **91**(10-11): 1126-1134.

Jamieson, P. D., I. R. Brooking, M. A. Semenov, G. S. MeMaster, J. W. White and J. R. Porter (2007). "Reconciling alternative models of phenological development in winter wheat." Field Crops Research **103**(1): 36-41.

Johnson, J. S., J. P. Gosling and M. C. Kennedy (2011). "Gaussian process emulation for second-order Monte Carlo simulations." Journal of Statistical Planning and Inference **141**(5): 1838-1848.

Juston, J., O. Andrén, T. Kätterer and P.-E. Jansson (2010). "Uncertainty analyses for calibrating a soil carbon balance model to agricultural field trial data in Sweden and Kenya." Ecological Modelling **221**(16): 1880-1888.

Kadane, J. B. and N. A. Lazar (2004). "Methods and criteria for model selection." Journal of the American Statistical Association **99**(465): 279-290.

Kalnay, E. (2003). Atmospheric modeling, data assimilation, and predictability, Cambridge university press.

Katz, R. W. (2002). "Techniques for estimating uncertainty in climate change scenarios and impact studies." Climate Research **20**: 167-185.

Kavetski, D., S. W. Franks and G. Kuczera (2002). "Confronting input uncertainty in environmental modelling." Calibration of watershed models: 49-68.

Kavetski, D., G. Kuczera and S. W. Franks (2006). "Bayesian analysis of input uncertainty in hydrological modeling: 2. Application." Water Resources Research **42**(3).

Kennedy, M. C., C. W. Anderson, S. Conti and A. O'Hagan (2006). "Case studies in Gaussian process modelling of computer codes." Reliability Engineering & System Safety **91**(10-11): 1301-1309.

Kennedy, M. C. and A. O'Hagan (2001). "Bayesian calibration of computer models." Journal of the Royal Statistical Society: Series B (Statistical Methodology) **63**(3): 425-464.

Koehler, J. R. and A. B. Owen, Eds. (1996). Computer Experiments. Handbook of Statistics, 13: Design and Analysis of Experiments. Amsterdam.

Krzanowski (2000). Principles of Multivariate Analysis: A User's Perspective. Osford, Oxford University Press.

Kuczera, G., D. Kavetski, S. Franks and M. Thyer (2006). "Towards a Bayesian total error analysis of conceptual rainfall-runoff models: Characterising model error using storm-dependent parameters." Journal of Hydrology **331**(1): 161-177.

Künsch, H. R. (2013). "Particle filters." Bernoulli **19**(4): 1391-1403.

Kurowicka, D. and R. Cooke (2006). Uncertainty Analysis. Chichester, John Wiley & Sons.

Lewis, J. M., S. Lakshmivarahan and S. K. Dhall (2006). Dynamic Data Assimilation: A Least Squares Approach. Cambridge, Cambridge University Press.

McFarland, J. M. (2008). Uncertainty Analysis for Computer Simulations Through Validation and Calibration. Doctor of Philosophy, Vanderbilt University.

McKay, M. D. (1995). Evaluating prediction uncertainty, Nuclear Regulatory Commission, Washington, DC (United States). Div. of Systems Technology.

McKay, M. D. and J. D. Morrison (1997). Structural model uncertainty in stochastic simulation, Los Alamos National Lab., NM (United States).

Mead, R. (1988). The Design of Experiments, Cambridge University Press.

Montanari, A. and G. Grossi (2008). "Estimating the uncertainty of hydrological forecasts: A statistical approach." Water Resour. Res. **44**: W00B08.

Montanari, A., C. A. Shoemaker and N. van de Giesen (2009). "Introduction to special section on Uncertainty Assessment in Surface and Subsurface Hydrology: An overview of issues and challenges." Water Resour. Res. **45**(12): W00B00.

Montgomery, J. M., F. M. Hollenbach and M. D. Ward (2012). "Improving Predictions Using Ensemble Bayesian Model Averaging." Political Analysis **20**(3): 271-291.

Murray, L. M. (2013). "Bayesian State-Space Modelling on High-Performance Hardware Using LibBi." arXiv preprint arXiv:1306.3277.

O'Hagan, A. (2006). "Bayesian analysis of computer code outputs: A tutorial." Reliability Engineering & System Safety **91**(10-11): 1290-1300.

O'Hagan, A. (2008). "Managing Uncertainty in Complex Models." 2010, from http://mucm.group.shef.ac.uk/index.html.

O'Hagan, A. (2012). "Probabilistic uncertainty specification: Overview, elaboration techniques and their application to a mechanistic model of carbon flux." Environmental Modelling & Software **36**: 35-48.

O'Hagan, A., M. C. Kennedy and J. E. Oakley (1999). Uncertainty analysis and other inference tools for complex computer codes.

O'Hagan, A. and J. E. Oakley (2004). "Probability is perfect, but we can't elicit it perfectly." Reliability Engineering & System Safety **85**(1-3): 239-248.

O'Hagan, A., J. W. Stevens and J. Montmartin (2001). "Bayesian cost-effectiveness analysis from clinical trial data." Statistics in Medicine **20**(5): 733-753.

O'Hagan, A. C. R. A. A. K. E. D. M. D. A. (2005). "Incorporation of Uncertainty in Health Economic Modelling Studies." PharmacoEconomics **23**(6): 529-536.

Oakley, J. and A. O'Hagan (2002). "Bayesian inference for the uncertainty distribution of computer model outputs." Biometrika **89**(4): 769-784.

Oakley, J. E. and A. O'Hagan (2004). "Probabilistic sensitivity analysis of complex models: a Bayesian approach." Journal of the Royal Statistical Society: Series B (Statistical Methodology) **66**(3): 751-769.

Oberkampf, W. L. and C. J. Roy (2010). Verification and validation in scientific computing, Cambridge University Press.

Pappenberger, F. (2006). "Ignorance is bliss: Or seven reasons not to use uncertainty analysis." Water Resources Research **42**(5).

Prieur, C. (2014). Recent inference approaches for Sobol' sensitivty indices. Uncertainty in Computer Models, Sheffield, UK.

Raftery, A. E., T. Gneiting, F. Balabdaoui and M. Polakowski (2005). "Using Bayesian model averaging to calibrate forecast ensembles." Monthly Weather Review **133**(5): 1155-1174.

Refsgaard, J. C., J. P. van der Sluijs, J. Brown and P. van der Keur (2006). "A framework for dealing with uncertainty due to model structure error." Advances in Water Resources **29**(11): 1586-1597.

Refsgaard, J. C., J. P. van der Sluijs, A. L. Højberg and P. A. Vanrolleghem (2007). "Uncertainty in the environmental modelling process – A framework and guidance." Environmental Modelling & Software **22**(11): 1543-1556.

Renard, B., D. Kavetski, G. Kuczera, M. Thyer and S. W. Franks (2010). "Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors." Water Resources Research **46**(5).

Rings, J., J. A. Vrugt, G. Schoups, J. A. Huisman and H. Vereecken (2012). "Bayesian model averaging using particle filtering and Gaussian mixture modeling: theory, concepts, and simulation experiments." Water Resources Research **48**(5): W05520.

Rotter, R. P., T. R. Carter and J. E. Olesen (2011). "Crop–climate models need an overhaul." Nature Climate Change **1**: 175-177.

Rougier, J. (2008). "Efficient emulators for multivariate deterministic functions." Journal of Computational and Graphical Statistics **17**(4): 827-843.

Rougier, J., S. Guillas, A. Maute and A. D. Richmond (2009). "Expert Knowledge and Multivariate Emulation: The Thermosphere-Ionosphere Electrodynamics General Circulation Model (TIE-GCM)." Technometrics **51**(4): 414-424.

Rougier, J. C. (1996). "Probabilistic inference for future climate using an ensemble of climate model evaluations." Climatic Change.

Sacks, J., W. J. Welch, J. M. Toby and H. P. Wynn (1989). "Design and Analysis of Computer Experiments." Statistical Science **4**(4): 409-423.

Saltelli, A., K. Chan and E. M. Scott, Eds. (2000). Sensitivity Analysis. New York, Wiley.

Saltelli, A., M. Ratto, S. Tarantola and F. Campolongo (2006). "Sensitivity analysis practices: Strategies for model-based inference." Reliability Engineering & System Safety **91**(10-11): 1109-1125.

Santner, T., B. Williams and W. Notz, Eds. (2003). The Design and Analysis of Computer Experiments. New York, Springer Verlag.

Sargent, R. G. (2005). Verification and validation of simulation models. Proceedings of the 37th conference on Winter simulation, Winter Simulation Conference.

Schoups, G. and J. A. Vrugt (2010). "A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic, and non-Gaussian errors." Water Resour. Res. **46**(10): W10531.

Sinclair, T. R. and R. C. Muchow (2001). "System Analysis of Plant Traits to Increase Grain Yield on Limited Water Supplies." Agronomy Journal **93**(2): 263-270.

Sinclair, T. R. and N. a. Seligman (2000). "Criteria for publishing papers on crop modeling." Field Crops Research **68**(3): 165-172.

Spiegelhalter, D. J. and N. G. Best (2002). "Bayesian approaches to multiple sources of evidence and uncertainty in complex cost-effectiveness modelling." Statistics in Medicine **22**(23): 3687-3709.

Stanfill, B., D. Clifford and P. Thorburn (2014). An efficient sensitivity analysis method for spatio-temporal data from an agricultural systems simulator. Australasian Applied Statistic Conference, Kangaroo Island, South Australia, Australia.

Stedinger, J. R., R. M. Vogel, S. U. Lee and R. Batchelder (2008). "Appraisal of the generalized likelihood uncertainty estimation (GLUE) method." Water Resour. Res. **44**: W00B06.

Stevens, J. W., A. O'Hagan and P. Miller (2003). "Case study in the Bayesian analysis of a cost-effectiveness trial in the evaluation of health care technologies: Depression." Pharmaceutical Statistics **2**(1): 51-68.

Strong, M. (2012). Managing Structural Uncertainty in Health Economic Decision Models. Doctor of Philosophy, University of Sheffield.

Strong, M. and J. E. Oakley (2014). "When Is a Model Good Enough? Deriving the Expected Value of Model Improvement via Specifying Internal Model Discrepancies." SIAM/ASA Journal on Uncertainty Quantification **2**(1): 106-125.

Strong, M., J. E. Oakley and J. Chilcott (2012). "Managing structural uncertainty in health economic decision models: a discrepancy approach." Journal of the Royal Statistical Society: Series C (Applied Statistics) **61**(1): 25-45.

Thacker, N. A. and A. J. Lacey (1996) "Tutorial: The Kalman Filter."

Thyer, M., B. Renard, D. Kavetski, G. Kuczera, S. W. Franks and S. Srikanthan (2009). "Critical evaluation of parameter consistency and predictive uncertainty in hydrological modeling: A case study using Bayesian total error analysis." Water Resources Research **45**(12).

Trucano, T. G., L. P. Swiler, T. Igusa, W. L. Oberkampf and M. Pilch (2006). "Calibration, validation, and sensitivity analysis: What's what." Reliability Engineering & System Safety **91**(10-11): 1331-1357.

Trucano, T. G., L. P. Swiler, T. Igusa, W. L. Oberkampf and M. Pilch (2006). "Calibration, validation, and sensitivity analysis: What's what." Reliability Engineering & System Safety **91**(10): 1331-1357.

Uusitalo, L., A. Lehikoinen, I. Helle and K. Myrberg (2015). "An overview of methods to evaluate uncertainty of deterministic models in decision support." Environmental Modelling & Software **63**(0): 24-31.

Van Ittersum, M. K., F. Ewert, T. Heckelei, J. Wery, J. A. Olsson, E. Andersen, I. Bezlepkina, F. Brouwer, M. Donatelli and G. Flichman (2008). "Integrated assessment of agricultural systems–A component-based framework for the European Union (SEAMLESS)." Agricultural Systems **96**(1): 150-165.

Vrugt, J., C. F. Braak, H. Gupta and B. Robinson (2009). "Equifinality of formal (DREAM) and informal (GLUE) Bayesian approaches in hydrologic modeling?" Stochastic Environmental Research and Risk Assessment **23**(7): 1011-1026.

Vrugt, J. A., H. V. Gupta, W. Bouten and S. Sorooshian (2003). "A Shuffled Complex Evolution Metropolis algorithm for optimization and uncertainty assessment of hydrologic model parameters." Water Resources Research **39**(8): 1201.

Vrugt, J. A., C. J. F. ter Braak, C. G. H. Diks, B. A. Robinson, J. M. Hyman and D. Higdon (2009). Accelerating Markov Chain Monte Carlo Simulation by Differential Evolution with Self-Adaptive Randomized Subspace Sampling. International Journal of Nonlinear Sciences and Numerical Simulation. **10:** 273.

Wallach, D. (2011). "Crop Model Calibration: A Statistical Perspective." Agronomy Journal **103**(4): 1144-1151.

Wallach, D., D. Makowski, J. W. Jones and B. Francois, Eds. (2014). Working with Dynamic Crop Models, Elsevier.

Wilkinson, R. D., M. Vrettas, D. Cornford and J. E. Oakley (2011). "Quantifying Simulator Discrepancy in Discrete-Time Dynamical Simulators." Journal of Agricultural Biological and Environmental Statistics **16**(4): 554-570.