Chapter 6
Invalidation of Models and Fitness-for-Purpose: A Rejectionist Approach

Keith Beven and Stuart Lane

> I am….an almost orthodox adherent of unorthodoxy: *I hold that orthodoxy is the death of knowledge since the growth of knowledge depends entirely on the existence of disagreement.*
> (Karl Popper 1994, p. 34)

Abstract This chapter discusses the issues associated with the invalidation of computer simulation models, taking environmental science as an example. We argue that invalidation is concerned with labelling a model as not fit-for-purpose for a particular application, drawing an analogy with the Popperian idea of falsification of hypotheses and theories. Model invalidation is a good thing in that it implies that some improvements are required, either to the data, to the auxiliary relations or to the model structures being used. It is argued that as soon as epistemic uncertainties in observational data and boundary conditions are acknowledged, invalidation loses some objectivity. Some principles for model evaluation are suggested, and a number of potential techniques for model comparison and rejection are considered, including Bayesian likelihoods, implausibility and the GLUE limits of acceptability approaches. Some problems remain in applying these techniques, particularly in assessing the role of input uncertainties on fitness-for-purpose, but the approach allows for a more thoughtful and reflective consideration of model invalidation as a positive way of making progress in science.

Keywords Epistemic uncertainty · Model equifinality · Bayes · GLUE · Limits of acceptability · Behavioural models

K. Beven (✉)
Lancaster Environment Centre, Lancaster University, Lancaster, UK
e-mail: k.beven@lancaster.ac.uk

S. Lane
Institute of Earth Surface Dynamics, Université de Lausanne, Lausanne, Switzerland

C. Beisbart and N. J. Saam (eds.), *Computer Simulation Validation*,
Simulation Foundations, Methods and Applications,
https://doi.org/10.1007/978-3-319-70766-2_6

6.1 Setting the Scene for Model Evaluation

In this chapter, we discuss the problems in applying a scientific methodology to computer models and, in particular, to the issue of rejection or invalidation of computer simulation models. We use invalidation of a simulation model structure and falsification of any of its component hypotheses here equivalently, to indicate that a simulation model has been shown to fail in some important respect and should consequently not be considered fit-for-purpose in making predictions. We do so from the point of view of practical environmental modellers in the domains of hydrology and hydraulics, who have an interest in the philosophical underpinnings of the modelling process and its role in the development of the associated science. Computer simulation is widely used in this domain, with (often complex) models being constructed to represent environmental systems with elements that sometimes have a good theoretical basis (e.g. mass and energy balance principles); that sometimes are derived from empirical studies (e.g. roughness relationships to represent bulk energy losses); and that sometimes have a purely conceptual basis (e.g. canopy resistance for transpiration from a vegetated surface). In such models, many of the functional relationships involve parameters that need to be identified for particular applications. These are often considered to be constant at a particular location and through time (especially when calibrating parameter values to past data) but have often been shown to change with the state of the system, or over time. Uncertainties in the input or forcing data as well as data used in model calibration, and also competing model structures, are intrinsic to the modelling process (see e.g. Beven 2009, 2012a). The dominant sources of uncertainty are often epistemic (i.e. the result of a lack of knowledge) rather than aleatory (i.e. statistical or resulting from random natural variability) in nature.

This chapter essentially addresses the question of how to do science when using models in the face of such epistemic uncertainties. This is discussed in the context of Popper's falsificationist approach in Sect. 6.2 and how this might then be applied as a rejectionist methodology in model evaluation when all models are known to be false to some extent. Sections 6.3 and 6.4 discuss the concepts of verisimilitude and fitness-for-purpose in the context of how models that are false might be useful. In Sect. 6.5, some principles of model invalidation as a positive methodology for advancing the science are discussed and it is shown how rejection can be considered within a modified Bayesian framework that either allows a choice between model structures or applies some limits of acceptability. Section 6.6 discusses how epistemic uncertainties impact on a rejectionist framework, and Sect. 6.7 how to resolve the advocacy of models that might not be fit-for-purpose with making scientific progress.

There is a long and continuing debate in both science and philosophy about what constitutes, or what should constitute, a scientific method in different domains of science (e.g. Chalmers 1976; Howson 2000; Hackett 2013). The recent rise of simulation models as a methodology for doing science has been much discussed in this respect (e.g. Cartwright 1999; Winsberg 2003). Simulation models combine elements of deductive inference in arguing from premises based on established con-

cepts and theories, and inductive inference as a way of deriving functional relationships and parameterisations (see Young 2013 for a recent discussion in the field of hydrology), but also as a way of inferring values for the parameters of models by calibration against observational data. Perhaps reflecting the extent to which current scientific method has given primacy to observation, the parameter values required to make the model reproduce those data may either have no physical equivalent or vary from those that have been measured (see Lane et al. 2011; Lane 2012). They are "effective" in that the calibrated values are needed to make the model perform against observational data (Beven 1989, 2016). Nearly all, if not all, environmental simulation models incorporate "conceptual" or inductive elements of this type.

However, as Hume (1748) first pointed out, there is a problem with induction. It demands that the future will be the same as the past. There is then an implication that any theory or model of nature can never be verified on the basis of past observations because there is always a possibility that "*the course of nature may change*" (Hume 1748, Sect. IV.2). It has proven difficult to provide any philosophical resolution to Hume's problem of induction even though it challenges the fundamental belief that environmental simulation models are a means of getting a handle on events that have yet to happen. Indeed, Howson (2000) argues that Hume is correct, but that does not mean that we cannot use reasoned argument based on observations, as well as deductive argument, to modify our scientific understanding and beliefs about the future. Widespread use of the term "physically-based" in environmental modelling is the implicit manifestation of a faith in this form of deductive argument. The term "physically-based" suggests a simulation model is based upon assumed-to-be time-invariant "laws of nature" and so capable of better getting at the future than other kinds of approaches (e.g. belief systems; expert judgement). Of course, however, physically based a model might be, Hume's proposition means that we will sometimes get surprises as the future unfolds.

It would appear evident that the use of simulation models that involve inductive elements as either parameterisations or calibrated/effective parameter values, might be most susceptible to future surprise when the observational data used in the inference comes from the past. As modellers, we expect the future to be uncertain and past experience suggests that we should expect some element of surprise, if only because future boundary conditions cannot be known (see, for example the post-audit analysis of groundwater model simulations of Konikow and Bredehoeft 1992; Anderson and Woessner 1992, discussed later). One aim of simulation modelling is then to minimise the element of surprise by ensuring that any model used for predicting the future is fit-for-purpose, in so far as its current and past performance has been evaluated. This is the process of model evaluation or validation or, from another perspective, model invalidation or falsification of the theoretical or conceptual components of a simulation model.

6.2 The Falsification Framework of Karl Popper

We hold that the invalidation view of model evaluation is a useful alternative to model validation because of the critical role of falsification in the development of science. Popper (1959) argued that either inferring universal statements from singular or particular ones, or confirming a universal statement with particular statements could not be justified because no matter how many instances something was observed and used to justify a particular universal statement, there was always the possibility that one observation may falsify that statement. In this context, for a hypothesis or theory to be considered scientific it must be advanced a priori, be testable and have the capacity to be falsified in some way. Hypotheses or theories that cannot be falsified in this way consequently can be considered as only pseudo-scientific. Scientific method is then the process of developing hypotheses and confronting them with the available evidence. Successful hypotheses in this context are not more probably true, because obtaining more evidence does not necessarily change the probability that a hypothesis might be falsified. For this reason, Popper argued that it is better to talk of the corroboration of hypotheses, where a better corroborated hypothesis is one that has been tested more rigorously individually, widespread or for a longer period of time. Equally, the most rapid of scientific progress may be made when a long-established or well-corroborated hypothesis is shown to no longer hold.

There are distinct parallels with the notion of multiple working hypotheses (Chamberlain 1895) and the idea that it may be necessary to work with a set of potentially contradictory hypotheses. In a Popperian framework, in which hypotheses are subject to testing and potential rejection, a hypothesis is defined as admissible if it is testable. This concept can be applied to simulation models, in that any particular realisation of a model of a process or system can be considered as a working hypothesis of how that system functions (e.g. Herskowitz 1991; Beven 2002). While recognising that all models are idealisations and consequently necessarily false in some respects, models that are successful in making useful predictions over a period of time can be considered corroborated; models that not successful should be considered as invalid or not fit-for-purpose and revised by changing beliefs (Klein and Herskowitz 2007). Similar considerations apply to simulation models used for different purposes, either for testing scientific concepts or for practical applications. The criteria of invalidation might, however, be different for different types of purpose.

Popper's falsification approach to scientific inference has not been without its dissenters. Indeed, it has been suggested that falsification itself cannot be falsified and that, in many celebrated examples, theories have not been falsified, despite contradictory observational evidence being available, because of some other intrinsically attractive features (e.g. Chalmers 1976; Ladyman 2002). Certain theories cannot be rejected because it would be too costly to do so (Latour and Woolgar 1979). It has also been pointed out that experimental observations are often conditioned by the theoretical framework within which they are developed, allowing free parameters to be derived from the observations and leaving no possibility of falsification (the Duhem–Quine thesis, see Quine 1975; Chalmers 1976). It may take a change of

paradigm to evaluate a theory in a different way, causing it to be replaced (though in the past, this has sometimes happened even though the new paradigm has been initially less supported by the available observations, see Kuhn 1970; Feyerabend 1975; Lakatos 1978).

More recently a statistical version of the falsification method has been promoted by Mayo and her co-workers (see, for example, Mayo 1996 and the discussions in Mayo and Spanos 2010, cf. also Chap. 19 by Robinson in this volume). This approach recognises that all experimental methods are subject to observational and sampling uncertainties, so hypotheses and theories should be exposed to strong statistical testing in validation. Failure of such tests would then constitute falsification. An example is the "5σ" test used in particle physics, where σ represents the standard deviation of the observations, such as in the identification of the Higgs Boson in the Large Hadron Collider at CERN (e.g. CMS Collaboration 2013). This requires that the variability of the data are well described by a Gaussian distribution, but if this assumption is accepted, then 5σ represents a 1 in 3.5 m chance ($p = 0.0000003$) of making a Type I error, where a false theory is accepted as correct. If this test is passed, then the hypothesis or theory is not rejected and can be considered as corroborated by the evidence. Similarly, tests on discrepancies between the data and theoretical predictions can be used to suggest when a theory should be rejected, though interestingly there do not seem to be any equivalent accepted standards in such cases for the probability at which falsification is confirmed. This is almost certainly an effect of the general bias against the publication of failures (see, for example, Masicampo and Lalande 2012), even though the statistics of negative results might be an important consideration in risk management (e.g. Mayo 1991). More often, hypotheses and theories that (to a more or less extent) conflict with observations are modified or replaced rather than simply being discredited or falsified in the literature. We revise our beliefs and hence our theories (Quine 1969; Morton 1993; Klein and Herskowitz 2007) through the addition of auxiliary information (e.g. empirical parameterisations of momentum loss and secondary circulation in rivers) even though we know that the reason that makes this auxiliary information needed (depth-averaging of the full 3D Navier–Stokes equations) fundamentally invalidates the capacity of depth-averaged models to represent the nature of river flow.

An incorrect rejection would be a Type II or false negative error (rejecting a model as a hypothesis that should not be rejected). In any statistical test there is a trade-off between Type I and Type II errors so the lower the required probability of avoiding a Type I error (as in the 5σ case), the higher the probability of a Type II error. This probability can be reduced by adding more informative observations, when this is feasible. Mayo's response to the Duhem–Quine thesis is to suggest that strong statistical testing implies the testing of any auxiliary conditions related to the theory. "*A claim can only be said to be supported by experiment if the various ways in which the claim could be at fault have been investigated and eliminated*" (Mayo 1996, p. 199). This represents severe testing but is not always possible, particularly when we wish to test the implementation of theories, and complex, multi-component models based on theories, to situations where controlled experiments are impossible or difficult to justify economically. This is the case for the very many models of environmental

systems currently being used, where knowledge of parameter values and boundary conditions may be subject to significant epistemic uncertainties. However, it also emphasises the need to test not only the model per se, but the constituent hypotheses, theories, models or auxiliary relations that are contained within it. Testing model outputs may not be sufficient.

6.3 Simulation Models, Invalidation and Falsification

These difficulties become particularly apparent where theories about some aspect of reality are combined and implemented as a computer simulation model, and it is the outputs from the model that are compared with observations. In many cases, for applications of environmental models to real-world open systems, the models are based on theories that are not expected to represent fully the complexity of the real world. This may be because full knowledge of the processes relevant to that complexity is lacking; because the processes have had to be simplified, or even ignored, to make the model tractable; because knowledge about the boundary conditions, initial states and characteristics of the system is insufficient; or it may simply be because the currently available computational resource does not allow a closer degree of approximation. These are all sources of epistemic uncertainty that might result in complex and nonstationary structures in model residuals when simulation outputs are compared against observations.

In such cases, auxiliary rules are often introduced to represent the consequences of simplification of the system being modelled, whether of the hypotheses being used in the model, the boundary or initial conditions needed to apply the model or the spatio-temporal scale at which the model is applied. Such rules commonly invoke free parameters, difficult to estimate a priori given limited information about the complex system and thus they are often calibrated against available observations (e.g. Morton 1993; Beven 2002). For the modeller, such parameters may not simply be a consequence of model implementation (e.g. simplification, approximation) but a necessary element of being able to make a model perform through the process of model calibration (Lane 2012). For example, in river flood modelling, modellers have typically used a single empirical parameter to represent friction losses due to a range of different processes (e.g. dispersion effects due to secondary circulation, turbulence, friction at the stream bed and energy losses at the water surface). Lane (2014) reports that an attempt to improve the determination of one of these parameters (the Manning roughness coefficient) was largely rejected in practice, because it was needed as an adjustable effective parameter that allowed modellers to make their model perform against observations. The improved parameterisation was not and could not be adopted. The notion that a model is made to perform reminds us that this performance might achieve the right results but this is not necessarily for the right reasons (Beven 1989): a model can be forced to be empirically adequate (Oreskes et al. 1994) and in some sense acceptable by the calibration of effective values of its parameters; even if it might be falsified in terms of the validity of the auxiliary

relations that are used to make it acceptable. The question is then whether it will be equally fit-for-purpose in predicting future changed conditions.

There can also be issues about the commensurability of observables and model variables, due to differences in scale or meaning, even when both are given equivalent names in the theoretical context used. In environmental systems, for example it can often be the case that observations are made at a "point" in space and time, while a model predicts a variable of the same name at some larger space–time discretisation. When there is little information about the sub-discretisation heterogeneity of the observable, it can then be difficult to relate one to the other. In many circumstances, it can also be difficult to assess that heterogeneity. For example, Hills and Reynolds (1969) examined the variability of point soil moisture measurements in a field and concluded that more than 150 measurements were necessary to estimate the mean value to within ±5%. Even in research projects such a sampling density is rarely affordable and such a field might represent just a single model grid element. In this case, recent advances in measurement technology can help overcome this problem by sampling surface soil moisture at larger scales (e.g. the COSMOS method, Zreda et al. 2012). However, hydrologists are not only interested in the surface soil moisture, but also in the water stored in the full soil profile, which is even more difficult to observe experimentally (but see the recent study of Güntner et al. 2017, using microgravity as an indication of how new measurement techniques might help constrain uncertainties). Similar issues arise at larger scales for variables within global or earth system science models. Such commensurability issues represent a fundamental limitation for the validation or falsification of such models.

These issues underlie George Box's aphorism that "*all models are wrong but some are useful*" (Box 1979), or as expressed by Morton (1993, p. 662): "the modelling assumptions are generally false, **and known to be false**, relative to a standard governing theory" (emphasis added). There is thus an **expectation** that our models could be falsified, especially if we look at what they predict in close detail (even if this is not reflected in how those models are presented in the literature). In this situation, therefore, there is an issue of what degree of approximation to the observational data we are prepared to accept before we allow that our modelling assumptions are wrong, knowing that there are uncertainties associated with the boundary conditions and evaluation data for any model application. Effectively, this requires a definition of the point at which we accept that a model might be invalidated as not fit-for-purpose in making the predictions required of it, while making proper allowance for the epistemic and aleatory uncertainties in the modelling process. We can, therefore, differentiate between invalidation of a simulation model structure based on the outputs relevant to a particular purpose, and the falsification of any of the individual hypotheses or theoretical constructs that might be involved as components of that model based on more controlled experimental testing (see also the frameworks suggested by Bennett et al. 2013; Augusiak et al. 2014).

6.4 Fitness-for-Purpose, Verisimilitude and Likelihood

The question of fitness-for-purpose is analogous to, but somewhat different from, Popper's original discussion of the evaluation of the verisimilitude or truthlikeness of a theory about reality. Popper suggested that we should accept that ultimately we could never be sure to have found a correct theory; even if it has survived all tests to date, the next inference it makes might prove to be wrong. However, the very process of testing and rejecting in this way and consequently building new theories should, over time, increase the degree of verisimilitude of the theory being applied. Reasoned argument suggests that we should, in principle, prefer theories or models as hypotheses with a greater degree of verisimilitude than others. This requires a scale of verisimilitude in order to determine a ranking of the multiple working hypotheses under consideration. Popper made some specific suggestions about the nature of that scale: that for a hypothesis to have greater verisimilitude than some competing hypothesis, the truth content of the first should include that of the second; while the false content of the first should be a sub-set of that of the second (Popper 1976). This proposal was shown to be logically untenable by Miller (1974). Subsequently, a variety of other technical definitions of verisimilitude have been proposed to try and overcome this limitation (see the recent discussion of Niiniluoto 2017). It also led to Popper to suggest later that the concept of verisimilitude need not be considered an essential part of his theory (Introduction 1982, p. xxxvi, in Popper 1983).

However, as scientists we still tend to think that it is possible to move from hypotheses that are known to be false in some sense, towards hypotheses that are closer to a correct description of the real system, even if still false in some lesser sense, i.e. from a lower to a higher degree of verisimilitude. Watkins (1985) expresses this in the sense of trying to assess the relative merits of hypotheses when one might be more readily corroborated than another, even if both might be far from the truth. In his later writings Popper accepted that, even if corroboration could not be used as a scale of verisimilitude, it could be used as an indicator of verisimilitude. Thus: "*If two competing theories have been criticized and tested as thoroughly as we could manage, with the result that the degree of corroboration of one of them is greater than that of the other, we will, in general, have* **reason to believe** *that the first is a better approximation to the truth than the second*" (Popper 1983, p. 58). In this context, the aim of the method is to justify a *preference* for one hypothesis over another, as a closer approximation to the truth, based on the evidence available, and using reasoned argument (Deutsch 1997; Klein and Herskovitz 2007). This does not now imply, however, that such a preference will necessarily be equivalent to a greater degree of verisimilitude.

But such corroboration with the evidence can be considered as a form of induction (e.g. O'Hear 1975), at odds with Popper's aim of providing a hypothetico-deductive scientific method. This will be even more the case when the hypotheses are implemented as computer simulation models with free parameters that need to be calibrated for some specific application, especially in the case of models that become over-parameterised with respect to the information content of the available observations.

This inevitably invokes induction from the (uncertain) empirical observations used in calibration when making inferences about the future behaviour of the system under study. It also makes falsification and the assessment of degrees of verisimilitude more difficult. Many potential models might fit the available observations to some acceptable degree of error (e.g. Beven 2006; Chap. 33 in this volume); some might be more truth-like or fit-for-purpose than others, but how do we make such an assessment?

Howson (2000) has suggested that one solution to the problem of induction is to work within a Bayesian framework (see also Chap. 7 by Beisbart and Chap. 20 by Jiang et al. in this volume). When we may not be able to assess a degree of verisimilitude of a hypothesis, we might be able to assess how the evidence could change our degree of belief in that hypothesis (see Howson and Urbach 1993 and this volume, Chap. 19). In modern applications of Bayes, the degrees of belief are most commonly expressed as terms of probability and the degree of explanation is called the likelihood. As new evidence becomes available Bayes theorem can be applied recursively so that hypotheses that are successful in the sense of having higher likelihoods will gradually develop higher posterior probabilities or degrees of belief. At no point, however, is it necessary to invoke any measure of truthfulness or verisimilitude, which makes the framework evidently suitable for application to hypotheses implemented as models while accepting that all models are idealisations of reality (or to some greater or lesser extent false).

This Bayesian framework, however, has been criticised for its subjectivity in both the prior assessments of degree of belief and in the choice of likelihood measure. The latter subjectivity has been addressed by statisticians in developing formal likelihood measures (or objective functions) that follow from specific assumptions about model errors (see, for example, Box and Taio 1992; Bernado and Smith 2000; Fernandez and Steele 1998; Beven 2009; Schoups and Vrugt 2010; Rougier 2007) but in applications to complex open systems it may be difficult to justify those assumptions. In such cases, the use of a formal statistical likelihood can lead to overconfidence in model evaluation when a large number of observations are available, for example, when time series are used in model evaluation (e.g. Beven 2012b, 2016; Beven and Smith 2015). This is because of the way in which the contributions of individual model residuals are combined multiplicatively, which may lead to models that have nearly equal error variance simultaneously having orders of magnitude differences in likelihood (even when bias and autocorrelation of model residuals are included in the likelihood function, see Beven 2016). Alternative subjective definitions of likelihood, that allow for the fact that model errors may not be simply stochastic, can avoid this stretching of the likelihood surface but do not have the same formal theoretical foundation.

There are other aspects of the formal Bayesian framework as based on probabilities that are relevant to the current discussion. The first results from the fact that the probability and statistical likelihood distribution functions that are commonly used have infinite tails (e.g. Bernado and Smith 2000). This means that no hypothesis that has a finite prior probability will be given a posterior probability of zero. The posterior probability might become very small for those models that do not perform well relative to the observations, but never zero. Consequently there is no falsification within this framework, unless some other, more subjective, threshold of

incompatibility with the evidence is imposed such that the likelihood can be set to zero. Falsification is then a limiting case of updating, but is outside the framework of formal statistical likelihood theory.

Another aspect of the formal Bayesian framework is that the hypothesis or model with the highest posterior likelihood will not necessarily be good enough to be useful for its intended purpose (let alone approach a truth-like representation of the real system). A further, related, point is that the approach normally takes no account of the fact that the probabilities might be incomplete: the approach is normally applied without taking any account of the fact that there might be other competing hypotheses (and consequent model structures) that have not been included.

6.5 If All Models May Be False, When Can They Be Considered Useful?

In some sense, we are all Bayesians because we have an expectation that additional evidence should lead to a refinement in our hypotheses and models about how the real-world system works. The question, therefore, is whether we have sufficient information to differentiate between hypotheses given the uncertainties associated with the modelling process. This in the Bayesian context equates to how best to define a likelihood to condition our degree of belief in a particular hypothesis, and to determine when the likelihood should be set to zero in cases where we infer that not only is the model false, but we have no belief that it will be useful for the purpose for which it is intended to be used.

This represents a challenge for four reasons, that apply to all models in the environmental and ecological sciences, including those that claim to be based on physical principles (Cartwright 1999; Beven 2002, 2012a, 2016). First, repeated runs of the computer simulation program using Monte Carlo techniques to make many different realisations using the same model structure, but different parameter sets and (sometimes) boundary and initial conditions, will often reveal a spectrum of responses from the best models found to those that clearly do not represent the observed behaviour well at all. Very different values of the same parameter (or even models with very different structures) may lead to equally "good" evaluation (or likelihood) measures; this is the equifinality thesis of von Bertalanffy (1968) and Beven (1993, 2006, Chap. 33 in this volume). Second, the evaluation or likelihood measures may reveal different things about what constitutes a good model performance. There may be Pareto trade-offs between the rankings of different models when evaluated against different criteria. Different periods of evaluation data can also change the rank ordering. Third, some of the data available to drive a model and to evaluate the outcomes of a model run might be disinformative in respect of whether a model performs well or not (Beven and Smith 2015; Beven 2016). Fourth, fitness-for-purpose implies more than just an epistemological concern as to when a model cannot be rejected against certain statistical criteria but also a series of wider concerns that relate to the way in

which the model sits within both wider scientific communities and decision-making processes.

In terms of scientific communities, models may continue to be used, even when it can be shown that alternative model structures can give better performance or even when the fundamental bases of the model (e.g. an auxiliary relation, as in the case of effective roughness parameters noted above) are not correct. For example simple empirical models of climate seem to provide better predictions to recent periods of historical data than general circulation models of climate (GCMs) climate models, even for global mean temperature (Fildes and Kourentzes 2011; Suckling and Smith 2013; Young 2018; for validation of climate simulations see Chap. 30 by Rood in this volume). However, GCMs continue to be used on the basis of the argument that their theoretical physical basis allows a greater degree of belief in their projections for the future (Shackley et al. 1998; Knutti 2018), whereas we cannot be sure that data-based models developed from historical observations will continue to be valid into the future. This is an argument for fitness-for-purpose based on the physical bases of process representations (Knutti 2018). Yet, GCMs involve empirical or conceptual elements in many process representations, and may be just as "empirical" as simpler models in terms of their dependence upon observational data to parameterise them (Shackley et al. 1998; Parker 2018). GCMs may also contain significant epistemic uncertainties, notably because of unknown boundary conditions (e.g. future decisions on fossil fuel use), which is why GCMs are run with different scenarios of future emissions. Yet, the number of such runs into the far future is often small because of the computational expense involved in resolving finer and finer detail in the atmospheric and oceanic circulations with each generation of model. Given these issues, Shackley et al. (1998) argue that GCMs remain dominant because they have mutually reinforced relations between GCM scientists, policy communities, climate impact communities and surrounding scientists, such that they have developed "*a wider symbolic significance than implied by their scientific credentials alone*" (Shackley et al. 1998, p. 188; see Winsberg 2003, for a wider discussion). The resilience of these relations to being challenged may explain why the question of fitness-for-purpose has rather rarely been questioned within the climate modelling community (though see Collins et al. 2012; Hargreaves and Annan 2014; and comments in Parker 2009, 2018 on the adequacy for purpose of climate models).

The above points emphasise that it is necessary to decide on what constitutes fitness-for-purpose and that such a decision may not be one that is only defined by scientific communities and past performance. What constitutes being fit-for-purpose, in general, will be highly context dependent (e.g. Barraque 2002; Wimsatt 2007; Knutti 2018) even where models are not developed with a pragmatic purpose in mind, but more because "*we are intrigued by the possibility of assembling our knowledge into a neat package to show that we do, after all, understand our science and its complex interrelated phenomena*" (Kohler 1969). For this purpose it is sufficient to be able to justify giving a likelihood of greater than zero in model evaluation, i.e. to have some degree of belief that the model mimics the functioning of the real system in some measurable sense. As Beven (2002) suggests, most modellers are pragmatic realists in this context. They would like to be able to equate the variables

in their computer models with quantities and fluxes in the real system, but they are pragmatic in recognising that there are real limitations as to how far that is possible. As with GCMs, however, past performance may not be the only factor in deciding on that degree of belief: there may be strong prior beliefs about the nature of the assumptions that underlie a model, beliefs that might vary between research groupings as well as being subject to strong influence by those who wish to use model results. For this purpose model evaluations are made not only with respect to demonstrable performance, but also in terms of what is considered acceptable within a research programme in terms of assumptions and degrees of uncertainty or error in the predictions, as well as the suitability of the predictions for the purpose to which they are to be put; the "*antecedently established credentials of the model building techniques developed over an extended tradition of employment*" (Winsberg 2003, p. 122). Thus, the evaluation could be against the *opinions* of experts or users, as conditioned on expectations about sources of uncertainty in the modelling process, as much as against any kinds of observable variables used to test a model. Similar considerations will apply to experts as referees on scientific papers and research reports, with their own experiences and impressions of what might be considered as acceptable.

Given the subjectivity implicit to the above argument, it might be expected that the faith in models as a contribution to decision-making might be undermined by the eventual realisation that those models were not fit-for-purpose when viewed after the fact. But, modellers are protected to some extent from being judged as to whether past predictions were fit-for-purpose because model predictions are generally constructed as scenarios or projections. With GCMs, for instance, the most recent Intergovernmental Panel for Climate Change report (IPCC 2013, p. 21) estimates a range in globally averaged warming by 2100 (as compared to 1986–2005) of between +0.4 °C and +5.5 °C according to the combination of scenario and aleatory uncertainty chosen. It is not generally expected that any of the assumed scenarios regarding future boundary conditions will actually prove to be correct. The simulations are projections not predictions. These projections are intended as the best available simulations **conditional on** the assumed emissions scenarios and other assumptions (and therefore not expected to occur in the future). In this way they are deemed to be useful, despite the better performance of data-based models on decadal time scales noted earlier.

There has been an interesting discussion in the simulation modelling community about the value of such projections in terms of the robustness of simulating future outcomes (e.g. Weisberg 2006; Lloyd 2010, 2018). This debate has recognised that individual models might be deficient in their predictions in the past, but that across an ensemble of models, some features of the projections might be robust to the specification of parameterisations and auxiliary conditions in individual models (Oldenbaugh 2018). The general trends in global warming in response to specific emission scenarios in the CMIP5 ensemble of GCMs is an example. It has also been pointed out, however, that in the climate model case the different model projections are not independent, but share common histories of development and prioritisation of added components over time (Oreskes 2018). It is also the case that robustness of projec-

tions across the ensemble, in terms of a commonality of outcomes, does not imply that any of the models is fit-for-purpose, but is simply corroboration of one model by another (Parker 2018). To get round this, it has been suggested that robustness should only be inferred when all the models in the ensemble have been empirically validated against past observations (Lloyd 2010), but clearly GCMs have limitations in reproducing past observations in detail (Parker 2009, 2018). The CMIP5 ensemble is currently the set of best available models; it remains unclear as to whether they are fit-for-purpose when they require bias corrections and flux corrections when used for evaluating the impacts of future climate changes on societies.

In other areas, where (rarely) post hoc assessments of modelled futures have been carried out more formally, the results have not been good. Examples are provided in the post hoc assessments of groundwater models reported in Konikow and Bredehoeft (1992) and Anderson and Woessner (1992). In some of the cases considered the conceptual model of the groundwater system proved to be inadequate; in others, the conceptual model was adequate but the estimation of future boundary conditions proved to be totally inadequate. Groundwater modelling is an example of where modelling technology has developed rapidly in the more than two decades since those papers were published and the four decades since the original modelling studies have been done. But, in most groundwater modelling applications, we still have limited knowledge of the subsurface geological characteristics and parameters, particularly in fractured rock systems, and future boundary conditions (climate, recharge, well development, pumping rates, etc.) are necessarily uncertain. Similar issues will arise in all areas of the inexact natural sciences. It is likely to be an even greater impediment for the social sciences even though that has not stopped attempts to model the joint development of natural and social systems into the future (e.g. in sociohydrology, see Viglione et al. 2014; Elshafei et al. 2014; Jeong and Adamowski 2016; Pande and Savenije 2016).

6.6 Defining Fitness-for-Purpose and Model Invalidation

The above argument is predicated upon the idea that a simulation model should be shown to be fit-for-purpose, that is corroborated against some kind of observation or judgment, even if there are few rules about precisely what constitutes "fit" and "purpose", such that its use can be justified. For both the purpose of understanding our science and informing decisions, the question that arises is how good is good enough to be useful, given the uncertainties in the modelling process. This can be posed as a problem of showing that a simulation model is invalid for the purpose intended, while taking proper account of those uncertainties. No modeller wants to present a model that is invalid of course: within research programmes considerable efforts are put into ensuring that the assumptions on which the model is based are justifiable; that the equations derived from those assumptions are correctly formulated; that the coded version of those equations is debugged and numerically accurate; that the parameter values used within the model are suitable; and that the model produces presentable

results against some evaluation observations. However, we wish to argue here for the importance of seeing model *invalidation* as a good thing, perhaps the ultimate goal of model use in science, in contrast with the simple use of the best models available in applications to society, when the best models (or ensemble of best models) might not be fit-for-purpose.

From a scientific perspective, model rejection is a positive outcome; it implies that we need to do better, either in defining better model structures or in generating better observations to drive and evaluate models. Of course, when modelling is used in practice, and uncertainties in the modelling process are recognised, there can be substantial constraints upon the capacity for a model to be shown to be false or invalid. The limited research that has traced the transition of model development into model adoption has revealed how social and economic constraints determine the extent to which a scientifically rejected model leads to the evolution of modelling practice (e.g. see Lane et al. 2013, for the case of flood inundation models and the discussion of GCMs above). Such constraints emphasise the difficulty that can exist in rejecting a model formulation as false. The philosopher of science, Isabelle Stengers (2013)[1] argues for a resistance to the constraints upon scientific practice related to both socio-economic limits as well as scientists' own institutional and community settings. She argues that being "scientific" requires us to recover our own capacity to be wrong and, in so doing, to raise different questions to those which we are being forced to ask. In 2005 she wrote: "*How can we present a proposal intended not to say what is, or what ought to be, but to provoke thought, a proposal that requires no other verification than the way in which it is able to 'slow down' reasoning and create an opportunity to arouse a slightly different awareness of the problems and situations mobilising us?*" (Stengers 2005, p. 994). Stengers' position here is interesting because it is in marked contrast to one of the traditional raison d'être of models which is to speed up time, to allow the future to become present today, such that society can invest now to make the future that becomes manifest more palatable. We develop Stengers' ideas more specifically below.

There is a very strong parallel here between the notion of model rejection or invalidation and the Popperian concept of falsification. By allowing for models to be invalidated, we may be able to move towards truer theories and models in an evolutionary way (e.g. Popper 1969; Dolby 1996; Deutsch 1997; Wimsatt 2007). Popper also made this point in saying that a falsificationist would "*prefer to solve an interesting problem by a bold conjecture, even (and especially) if it turns out to be false, to any recital of a sequence of irrelevant truisms*" (1969, p. 231). Learning from our mistakes should bring us further to a realistic representation of a system of interest, even if only an approximation to reality is attainable. The **nature** of the rejection can then provide valuable information about the assumptions on which a model is based, or the data needed to apply and evaluate the model, provided we allow it to do so. The question that then arises is twofold. First, how do we define criteria to invalidate a model as fit for its intended purpose? This is a problem analogous to defining a measure of verisimilitude in the Popperian framework, albeit that fitness-

[1] This is written in French. See Lane (2017) for an English interpretation.

for-purpose is a lesser requirement than truthlikeness. The second question, addressed in part below, is how to reconcile the self-interest of model advocates who want to present predictions as acceptable and useful, with the fundamental scientific progress that comes from accumulating our (posterior) beliefs that a model is no longer fit-for-purpose.

Wimsatt (2007, pp. 100–106) provides an analysis of 7 ways in which models might be wrong, and 12 ways of learning from models that are wrong (and sometimes designed to be wrong as a way of illuminating system processes). He suggests that the ways in which models are modified over time as a result of testing and thoughtful reasoning is the way in which much of science is normally practiced (similarly arguments are made by Koen (2003) in a discussion of engineering practice, and Klein and Herkowitz 2007, from a simulation philosophy perspective). This is a rather instrumentalist view of scientific method, in that all the time that theoretical tools and models provide some utility, they will not be rejected; and when they appear to be wrong, we learn from how they appear to be wrong. However, it is very similar to Quine's (1969) notion of "belief revision". Mayo (1996, Chap. 1) also considers learning from mistakes, but firmly within a falsificationist approach, with a heavy use of error statistics within a statistical theoretical approach. Such an approach depends, of course, on making strong aleatory assumptions in statistical testing, which may be difficult in the applications of models to open systems with epistemic uncertainties that are characteristic of the environmental sciences.

The discussion of the previous sections and past experience suggests some principles on which to base any assessment of model invalidation.

a. Within the feasible model space (of model structures and parameter sets) it should be accepted that model outputs often show a wide spectrum of goodness-of-fit from the best models found to those that are far from any evaluation data or evidence.
b. Fitness-for-purpose is concerned with the best simulation models found, but these may be localised in a high dimensional model space and may not be easy to find.
c. The best simulation models found will depend on the criteria of evaluation used, and also on the set of forcing and evaluation data used. The criteria used should therefore, as far as possible, reflect the framing of the purpose intended.
d. Uncertainty in the input or forcing data is important—by analogy with statistical hypothesis testing we do not want to accept a "false" model or reject a "useful" model just because of uncertainties or disinformation in the forcing and boundary condition data (or other auxiliary conditions).
e. The structure of a simulation model should add value; we should not accept a simulation model that is not significantly better than a parsimonious non-parametric data-based model for the variable of interest. The data-based model might be overfit, but so could the simulation model when used with the same forcing data.
f. Fitness-for-purpose should be defined prior to running any model simulations, taking account of understanding of uncertainties in the modelling process; we do not want to compensate poor performance simply by an error model with large variance.

g. A simulation model that is deemed fit-for-purpose should not be expected to necessarily remain fit-for-purpose if the assimilation of further evidence suggests the model fails in some important respect.

There are a variety of methods for model choice available. These include methods based on Bayesian inference, statistical implausibility measures and methods based on tolerance thresholds or limits of acceptability. As discussed earlier Bayesian inference is based on defining a measure of likelihood together with any estimates of prior probability for model formulations that might be based on past applications or testing. The definition of a likelihood measure is now commonly based on more or less complex statistical assumptions about the nature of the model residuals (e.g. Bernado and Smith 2000).

6.6.1 Using Bayes Ratios to Differentiate Between Models

Model comparisons can be made in terms of the posterior marginal probability distributions for different model structures, expressed as Bayes factors or ratios. The Bayes ratio can be defined as

$$K_B = \frac{\int [P_o(M_1\{\theta_1\})L(O \vee M_1\{\theta_1\})]d\theta_1}{\int [P_o(M_2\{\theta_2\})L(O \vee M_2\{\theta_2\})]d\theta_2} \tag{6.1}$$

where M_1 and M_2, with parameter vectors θ_1 and θ_2, are two different model structures under consideration; P_o is the prior probability for each model and L is the likelihood when model predictions are evaluated against the observations O. Since the ratio is defined in terms of probability integrals, it will not give a crisp differentiation between valid and invalid models. Some rules of thumb have been suggested for model choice using the Bayes ratio. Thus, for ratios of >20 we should have a strong preference for M_1 over M_2; and for ratios >150 we should have a very strong preference for M_1 over M_2. (e.g. Kass and Rafferty 1995). Note, however, that to be directly comparable the likelihood definition used in evaluation of each model should be directly comparable. Where this is based on statistical assumptions about the nature of the model residuals it requires the same structural assumptions. This may, or may not, be appropriate for the different error model structures and is an assumption that should be checked in good practice. Experience suggests that such ratios can be sensitive to such assumptions and can vary dramatically (by tens of orders of magnitude) depending on what periods of data are used in the evaluation (see the discussion in Beven 2016).

For cases where it is difficult to define an explicit likelihood measure, the Bayes ratio can be approximated using Approximate Bayesian Computation (ABC e.g. Robert et al. 2011). Interestingly the ABC methodology depends on defining some tolerance level for model acceptance. This is sometimes refined as the search within the model space (or spaces in the case of multiple model structures) proceeds. We

know of no cases, however, where it has been defined on the basis of fitness-for-purpose, rather than ensuring a sufficient sample of acceptable models.

Note also that the integral for each model in Eq. 6.1 integrates over all plausible model parameter sets; it does not focus on the best performance for each model structure. In evaluating fitness-for-purpose it might therefore be better to consider only the maximum likelihood associated with each model in which case [3] reduces to a likelihood ratio test that involves only a single parameter set in each model structure. Again under the proviso that a similar error model assumption is appropriate for each of the models considered, the likelihood ratio can be used to evaluate whether one model is more acceptable than another, but not necessarily whether either is fit-for-purpose.

6.6.2 *Use of Implausibility Measures to Differentiate Between Models*

A somewhat different statistical approach has been suggested by Vernon et al. (2010) for cases where it is difficult to specify a likelihood measure based on residual error characteristics. Rather than use a likelihood measure, they propose the use of an implausibility scaling of the following form:

$$I^2(x_i) = \frac{\{O_i - M(x_i;\theta)\}^2}{\left\{Var\left(e_{M,i}\right) + Var\left(e_{O,i}\right)\right\}} \tag{6.2}$$

where x_i is the ith model output variable, $M(x_i; \theta)$ is the model prediction of x_i given a parameter set θ; O_i is the equivalent observed variable, $e_{M,i}$ is an estimate of model uncertainty (arising from allowable model discrepancy or from stochastic forcing) and $e_{O,i}$ is an estimate of the observation uncertainty for the ith variable. Separate implausibility measures can be calculated for all available observation–prediction matching couples, and combined into a total measure of implausibility. The measure can be updated as new information becomes available. Implausibility, as defined in this way, is similar to the Bayes ratio, in that it represents a continuous relative scale with no sharp cut-off. Again some rule of thumb is required to decide where the limit of plausibility lies on that scale. In Vernon et al. (2010) and Woodhouse et al. (2015) the plausible model space is defined by a threshold of $I < 3$, based on the 3σ rule, implying that the plausible region contains the most plausible model, allowing for both model and observational uncertainty, with probability greater than 95%. Other forms of plausibility measure are discussed in Halpern (2005).

6.6.3 *Use of Limits of Acceptability to Define Behavioural Models*

Both the Bayesian and implausibility measure approaches depend on the magnitude of the model residuals evaluated after each model run. They do not require any decision to be made about some threshold of acceptability before making a model run. An alternative method that does require a prior definition of acceptability is the Limits of Acceptability implementation of the Generalised Likelihood Uncertainty Estimation (GLUE) methodology as outlined by Beven (2006). GLUE is based on Monte Carlo sampling of the model space to identify an ensemble of acceptable or "behavioural" models that will be used in prediction. Simulations that do not pass the limits of acceptability test are rejected as non-behavioural or invalidated, i.e. they are not considered to be fit-for-purpose. The approach is general in that it can be applied to parameter sets and uncertain boundary conditions for one or more model structures, with likelihood measures defined and combined in different ways (Beven and Binley 1992, 2014). Statistical likelihood functions and combining likelihoods using Bayes equation represent a special case within GLUE, where the necessary assumptions can be justified. Different search algorithms can be used to explore the model space (e.g. Beven and Binley 1992, 2014; Blasone et al. 2008; Vrugt 2016; Vrugt and Beven 2018).

Within this framework, the ensemble of behavioural models can be used to produce likelihood weighted predictions, but it also allows for the possibility that none of the sampled models reach the level of performance required for a particular purpose. Thus, in GLUE, the choice of a behavioural threshold assumes a particular importance, but allows the consideration of fitness-for-purpose for a given application in doing so. In the past GLUE has been criticized for the subjectivity in making such a choice so Beven (2006) suggested that the choice should be made more objective by considering what is known about the data that is used to drive and evaluate the model, as well as what level of performance is needed for the predictions to be considered useful. The use of limits of acceptability in this way is analogous to the tolerance limits used in ABC (e.g. Nott et al. 2012; Sadegh and Vrugt 2013), or applying a limit to an implausibility measure, except in that the limits should be defined before making any model runs.

In doing so, limits of acceptability can be applied to predictions of either individual observations (e.g. Liu et al. 2009), or of summary statistics relevant to the purpose (e.g. Westerberg et al. 2011; Westerberg and McMillan 2015). It is, therefore, possible that (harking back to Popperian falsification) a model could be rejected on the basis of the failure to simulate a single observation within the limits of acceptability, if that observation is considered sufficiently important. Popper notes, however, that "*a few stray basic statements contradicting a theory will hardly induce us to reject it as falsified. We shall take it as falsified only if we discover a* **reproducible effect** *which refutes the theory. In other words, we only accept the falsification if a low-level empirical hypothesis which describes such an effect is proposed and corroborated. This kind of hypothesis may be called a* **falsifying hypothesis**" (1959,

p. 86). On the other hand we should, perhaps, be rather wary of generalising this idea of reproducibility to a form of simple statistical 3σ/95% threshold, since it is quite possible that the remaining 5% might be those observations that are of most interest to the purpose for which the model is being used (e.g. the hydrograph peaks in a hydrological model application, see Beven 2016). However, in considering single observations the limits of acceptability should reflect the impact of input errors on how well a model might be expected to perform. This is important so as to avoid the Type II error of rejecting a good model that would be fit-for-purpose just because of errors in the inputs or forcing data (models are very much subject to the "garbage in garbage out" phenomenon).

The advantage of an approach based on model invalidation is that it encourages honesty in the modelling process, including about just how well we might expect a model to perform given the understanding of how a particular system works, and the data available with which to drive and evaluate the model performance (see also Smith and Stern 2011). It also allows for the possibility that all the models tried might be invalidated as not fit-for-purpose (see for example Brazier et al. 2000; Choi and Beven 2007; Dean et al. 2009; Mitchell et al. 2011; Hollaway et al. 2017). Where this happens, the model structure has been effectively invalidated, at least for that application. Commonly, however, it will survive in other applications, perhaps with less constrictive evaluation measures or limits of acceptability, rather than being reconsidered and modified. We would surely learn more from trying to understand the reasons for such rejections (Beven 2018).

6.7 Epistemic Uncertainties and Model Invalidation

The types of open system models that have been discussed in the last section are commonly subject to epistemic uncertainties or knowledge gaps. In such cases, the use of strong statistical assumptions about the sources of uncertainty might lead to overconfidence in inference because they result in a stretching of the likelihood surface, such that model and parameter uncertainty tends to be underestimated, and the residual error variance will expand to compensate. Where there are time series of data, with large numbers of observations, this stretching can be extreme and unrealistic (see Beven 2016).

Clearly, other forms of likelihood measure can be used (as, for example, in the GLUE methodology), but at the expense of losing the formal probabilistic interpretation embodied in a formal statistical likelihood function that follows from specific distributional assumptions about the model residuals. However, for good epistemic reasons, it will remain difficult to capture the nature of perceived epistemic uncertainties in the form of a statistical likelihood measure. This is particularly true for input data that might be subject to epistemic uncertainties because such uncertainties will be propagated through the (generally nonlinear) dynamic structure of the system model, interacting with any model structural error to produce complex output error structures. Even if input errors could be defined simply (e.g. as Gaussian distribu-

tions with homoscedastic variance) the output errors would then be nonstationary in bias, variance and autocorrelation, depending on the sequence of events. But the input errors are more likely to be epistemically nonstationary in complex ways, compounding the problem of how to represent the uncertainty in model evaluation. In extreme cases, the available input and output data might, at least in part, be physically inconsistent and therefore not informative about whether a model is fit-for-purpose. Where this can be identified, it can also be taken into account in setting limits of acceptability and making predictions (e.g. Beven and Smith 2015).

That is one reason why such limits should be defined a priori, before running a model, to avoid rejecting periods of data just because they are not well fitted by the model. The question is then how to do so, if we expect that there will be a significant impact of epistemic input errors on model predictions and consequently the appropriate limits of acceptability in assessing fitness-for-purpose. This is analogous to the problem of defining the term $e_{M,I}$ in the implausibility framework, but without knowing how to define the stochastic input variation. This remains a problem to be resolved, including for cases where interaction with stakeholders and decision makers might introduce more qualitative evaluation of models (see, for example Landström et al. 2011; Haasnoot et al. 2014).

6.8 The Model Advocacy Problem

We want to finish this Chapter with some thoughts on what we call the "model advocacy problem": how is it that we can move from advocating our models as somehow useful to seeing scientific progress as arising when we realise from our accumulated (posterior) beliefs that a model is no longer fit-for-purpose? The relevance of this question has been touched upon at a number of points throughout this Chapter, in relation to Global Climate Models and flood inundation models, for instance. It is an important concern because it has been shown (e.g. Landström et al. 2011) that "*[A]ccustomed to living in their entrenched fields, researchers end up with eyes only for the problems which are born in their laboratories*" (Callon et al. 2009, pp. 94–95). Research that has followed the evolution of modelling as a practice has shown that models can become bound into an assemblage that resists attempts (e.g. new knowledge) that might break it apart. In relation to flood inundation modelling, the Manning's *n* roughness parameter was too valuable as a model parameterisation tool that attempts to improve its measurement and representation failed (Lane 2014). If models can develop resistance to their own invalidation through the assemblage of people (scientists, consultants, policy-makers), technologies and places of which they come a part, what are the conditions that may break down that resistance, that make model invalidation possible?

One response is a fundamentally scientific one, to be empirical in the very broadest sense of the term. How is it that we can establish practices that allow the world "to speak back" to the modeller, to challenge the way the world is being represented (Baker 2017; Lane 2017; Beven and Alcock 2012; Beven 2018) by the model. This

is not always straightforward because of the assembled network of constraints that serve to protect the model's (and modeller's) status as it has become (e.g. Lane et al. 2013). Stengers (2013) argues that one way of doing this is through finding ways that make a scientist turn away from their normal communities of practice (as scientists) and the abstraction of their investigation out of the milieu of which it is normally a part (see also Baker 2017, in relation to hydrology; Landström et al. 2011, in relation to flood modelling). For Stengers, this should be done through the "*enrolment of phenomena*" (trans. p. 127) that don't dictate how they should be described but rather are given the "*capacity to evaluate the relevance of the way they are being described*" (trans. p. 68). Stengers' argument points to the need to focus less on a model's goodness-of-fit and more on those points that don't fit the model and, as a result, cause us to slow down our reasoning to the point at which other kinds of hypotheses and simulation models might be deemed suitable or other, quite different, kinds of approaches meaningful (Lane 2017).

It is right, then, to admit that our models can be wrong (see Beven 2016, 2018), in that this implies that further improvements to either input data or modelling hypotheses need to be made. How this might be done in practice is not, however, evident. We can perhaps distinguish between model use in relation to applied questions, where a model might be a tool that assists with decision-making, and model use in scientific research where progress will be made when a model is found to be invalid. When the latter is the case, it implies that the model might not be fit-for-purpose for applied uses, but it is clearly evident that for applied use there is so much investment and vested interests in the development of modelling packages that any invalidation will tend to be hidden within the improvements associated with new version releases. A new version will be developed when it is found that modifying parameters or auxiliary conditions within a modelling framework is not sufficient to match the observational data to a degree acceptable to the client (or a critical bug in the code is found), but there may still be significant resistance to the invalidation of the fundamental concepts on which a modelling package is based. The question of when to use model invalidation is then intrinsically embedded in the communities of practice within which model applications are situated, and dependent on critical feedback from those communities.

Stengers suggests that model advocacy works against thoughtful scientific progress. There is also the issue as to whether models that can be considered invalidated with respect to the science can be considered useful when providing predictions for applied decision-making. We suggest therefore that a new way of appreciating a problem is required that allows invalidation to be pursued more widely and more thoughtfully. One way of doing so might be to give the concept of fitness-for-purpose more prominence in both the scientific and applied use of models.

6.9 Conclusions

This paper has discussed a number of aspects of invalidation of models as not fit-for-purpose for a particular application, drawing an analogy with the Popperian idea of falsification of hypotheses and theories. It has been shown that as soon as epistemic uncertainties in observational data and boundary conditions are acknowledged invalidation loses some objectivity. The original Popperian concept of falsification as a way of resolving Hume's problem of induction then becomes less tenable, in favour of a Bayesian framework of corroboration that contains elements of induction, particularly when evaluation allows the modification of prior estimates of boundary or auxiliary conditions and parameter values in model calibration.

This is particularly the case of models of open systems that are subject to epistemic uncertainties such that there is an expectation of models being (more or less) false when examined in detail and where it can be difficult to represent model error in terms of well-defined probabilistic structures. This means that it can be difficult to justify the strong assumptions of formal definitions of likelihood within a Bayesian conditioning framework. In addition, a Bayesian framework based on statistical likelihood functions does not explicitly allow for model invalidation, only evaluation of relative likelihoods of different model formulations (and that only under the assumption that the same statistical error structure is appropriate). Some rules of thumb for Bayes ratios have been proposed in comparing different model representations, but where the integral likelihoods are used to define the ratio, the approach does not explicitly evaluate whether the maximum likelihood models are fit-for-purpose. Other approaches based on implausibility measures and the prior definition of limits of acceptability are discussed, both of which can be applied to the evaluation of simulated individual observations for different variables and which attempt to allow for input and observational error, either as variances or in terms of support for the limits of acceptability. The limits of acceptability approach also focuses attention on how good a performance is required for a model to be fit-for-purpose in a particular application, whether that is to demonstrate scientific understanding or to inform a decision-making process. Some problems remain in applying these techniques, particularly in assessing the role of input uncertainties on fitness-for-purpose, but the approach allows for a more thoughtful and reflective consideration of model invalidation as a positive way of making progress in the science.

Acknowledgements The discussions on which this paper is based were initiated while KB was supported by the Fondation Herbette as visiting professor at the University of Lausanne. We thank Claus Beisbart, Nicole Saam and an anonymous referee for their comments on an earlier draft of this chapter.

References

Anderson, M. P., & Woessner, W. W. (1992). The role of the postaudit in model validation. *Advances in Water Resources, 15*(3), 167–173.

Augusiak, J., van den Brink, P. J., & Grimm, V. (2014). Merging validation and evaluation of ecological models to'evaludation': A review of terminology and a practical approach. *Ecological Modelling, 280,* 117–128.

Baker, V. R. (2017). Debates— Hypothesis testing in hydrology: Pursuing certainty versus pursuing uberty. *Water Resources Research, 53,* 1770–1778.

Barraque, B. (2002). Modélisation et gestion de l'environnement. In P. Nouvel (Ed.), *Enquète sur le concept de modèle* (pp. 121–141). Paris: Presses Universitaires de France.

Bennett, N. D., Croke, B. F., Guariso, G., Guillaume, J. H., Hamilton, S. H., Jakeman, A. J., et al. (2013). Characterising performance of environmental models. *Environmental Modelling and Software, 40,* 1–20.

Bernado, J. M., & Smith, A. F. M. (2000). *Bayesian theory*. Chichester: Wiley. ISBN 978-0-471-49464-5.

Beven, K. J. (1989). Changing ideas in hydrology: The case of physically-based models. *Journal of Hydrology, 105,* 157–172.

Beven, K. J. (1993). Prophecy, reality and uncertainty in distributed hydrological modelling. *Advances in Water Resources, 16,* 41–51.

Beven, K. J. (2002). Towards a coherent philosophy for environmental modelling. *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences, 458,* 2465–2484.

Beven, K. J. (2006). A manifesto for the equifinality thesis. *Journal of Hydrology, 320,* 18–36.

Beven, K. J. (2009). *Environmental modelling: An uncertain future?* Routledge: London.

Beven, K. J. (2012a). *Rainfall-runoff modelling: The primer* (2nd ed.). Chichester: Wiley-Blackwell.

Beven, K. J. (2012b). Causal models as multiple working hypotheses about environmental processes. *Comptes Rendus Geoscience, Académie de Sciences, Paris, 344,* 77–88. https://doi.org/10.1016/j.crte.2012.01.005.

Beven, K. J. (2016). EGU Leonardo Lecture: Facets of hydrology—epistemic error, non-stationarity, likelihood, hypothesis testing, and communication. *Hydrological Sciences Journal, 61*(9), 1652–1665. https://doi.org/10.1080/02626667.2015.1031761.

Beven, K. J. (2018). On hypothesis testing in hydrology: Why falsification of models is still a really good idea. *WIRES Water*. https://doi.org/10.1002/wat2.1278.

Beven, K. J., & Alcock, R. (2012). Modelling everything everywhere: A new approach to decision making for water management under uncertainty. *Freshwater Biology, 56,* 124–132. https://doi.org/10.1111/j.1365-2427.2011.02592.x.

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

Beven, K., & Binley, A. (2014). GLUE: 20 years on. *Hydrological Processes, 28*(24), 5897–5918.

Beven, K. J., & Smith, P. J. (2015). Concepts of Information content and likelihood in parameter calibration for hydrological simulation models. *ASCE Journal of Hydrologic Engineering*. https://doi.org/10.1061/(asce)he.1943-5584.0000991.

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

Box, G. E. P. (1979). Robustness in the strategy of scientific model building. In R. L. Launer & G. N. Wilkinson (Eds.), *Robustness in statistics* (pp. 201–236). Academic Press.

Box, G. E. P., & Tiao, G. C. (1992). *Bayesian inference in statistical analysis*. New York: Wiley.

Brazier, R. E., Beven, K. J., Freer, J., & Rowan, J. S. (2000). Equifinality and uncertainty in physically-based soil erosion models: Application of the GLUE methodology to WEPP, the Water Erosion Prediction Project–for sites in the UK and USA. *Earth Surface Processes and Landforms, 25,* 825–845.

Callon, M., Lascoumes, P., & Barthe, Y. (2009). *Acting in an uncertain world. An essay on technical democracy*. Cambridge, MA: MIT Press.
Cartwright, N. (1999). *The dappled world. A study of the boundaries of science*. Cambridge: Cambridge University Press.
Chalmers, A. (1976). *What is this thing called science?* St Lucia, Queensland: University of Queensland Press.
Chamberlin, T. C. (1895). The method of multiple working hypotheses. *Science*, *15*(old series), 92–96.
Choi, H. T., & Beven, K. J. (2007). Multi-period and multi-criteria model conditioning to reduce prediction uncertainty in distributed rainfall-runoff modelling within GLUE framework. *Journal of Hydrology, 332*(3–4), 316–336.
CMS Collaboration. (2013). Observation of a new boson with mass near 125 GeV in pp collisions at $\sqrt{s} = 7$ and 8 TeV. *Journal of High Energy Physics*, *6,* 81.
Collins, M., Chandler, R. E., Cox, P. M., Huthnance, J. M., Rougier, J. C., & Stephenson, D. B. (2012). Quantifying future climate change. *Nature Climate Change, 2,* 403–409.
Dean, S., Freer, J. E., Beven, K. J., Wade, A. J., & Butterfield, D. (2009). Uncertainty assessment of a process-based integrated catchment model of phosphorus (INCA-P). *Stochastic Environmental Research and Risk Assessment, 2009*(23), 991–1010. https://doi.org/10.1007/s00477-008-0273-z.
Deutsch, D. (1997). *The fabric of reality*. London: Allen Lane.
Dolby, R. G. H. (1996). *Uncertain knowledge*. Cambridge: Cambridge University Press.
Elshafei, Y., Sivapalan, M., Tonts, M., & Hipsey, M. R. (2014). A prototype framework for models of socio-hydrology: Identification of key feedback loops and parameterisation approach. *Hydrology and Earth System Sciences, 18*(6), 2141–2166.
Fernandez, C., & Steel, M. J. F. (1998). On Bayesian modeling of fat tails and skewness. *Journal of American Statistical Association, 93,* 359–371.
Feyerabend, P. (1975). *Against method*. New York: Verso Books.
Fildes, R., & Kourentzes, N. (2011). Validation and forecasting accuracy in models of climate change. *International Journal of Forecasting, 27*(4), 968–995.
Güntner, A., Reich, M., Mikolaj, M., Creutzfeldt, B., Schroeder, S., & Wziontek, H. (2017). Landscape-scale water balance monitoring with an iGrav superconducting gravimeter in a field enclosure. *Hydrology and Earth System Sciences*, *21*, 3167–3182. https://doi.org/10.5194/hess-21-3167-2017.
Haasnoot, M., Van Deursen, W. P. A., Guillaume, J. H., Kwakkel, J. H., van Beek, E., & Middelkoop, H. (2014). Fit for purpose? Building and evaluating a fast, integrated model for exploring water policy pathways. *Environmental Modelling & Software, 60*, 99–120.
Hackett, J., & Zalta, E. N. (Eds.) (2013). *Roger bacon.* Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/archives/spr2015/entries/roger-bacon/.
Halpern, J. Y. (2005). *Reasoning about uncertainty*. Cambridge, MA: MIT Press.
Hargreaves, J. C., & Annan, J. D. (2014). Can we trust climate models? *WIREs Climate Change, 5,* 435–440. https://doi.org/10.1002/wcc.288.
Herskovitz, P. J. (1991). A theoretical framework for simulation validation: Popper's falsfication ism. *International Journal of Modelling and Simulation, 11,* 56–58.
Hills, R. C., & Reynolds, S. G. (1969). Illustrations of soil moisture variability in selected areas and plots of different sizes. *Journal of Hydrology, 8,* 27–47.
Hollaway, M. et al. (2017). The challenges of modelling phosphorus in a headwater catchment: Applying a 'limits of acceptability' uncertainty framework to a water quality model. Under review.
Howson, C. (2000). *Hume's problem: Induction and the justification of belief*. Oxford: Oxford University Press, Clarendon Press.
Howson, C., & Urbach, P. (1993). *Scientific reasoning: The Bayesian approach* (2nd ed.). Chicago, IL: Open Court.
Hume, D. (1748). *Philosophical essays concerning human understanding*. London: A. Millar.

IPCC. (2013). Summary for policymakers. In T. F. Stocker, D. Qin, G. -K. Plattner, M. Tignor, S. K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex & P. M. Midgley (Eds.), *Climate change 2013: The physical science basis. Contribution of working Group I to the fifth assessment report of the intergovernmental panel on climate change*. Cambridge: Cambridge University Press.

Jeong, H., & Adamowski, J. (2016). A system dynamics based socio-hydrological model for agricultural wastewater reuse at the watershed scale. *Agricultural Water Management, 171,* 89–107.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association., 90*(430), 791. https://doi.org/10.2307/2291091.

Klein, E. E., & Herskovitz, P. J. (2007). Philosophy of science underpinnings of prototype validation: Popper vs Quine. *Information Systems Journal, 17*(1), 111–132.

Knutti, R. (2018). Climate model confirmation: From philosophy to predicting climate in the real world. In E. A. Lloyd & E. Winsberg (Eds.), Climate modelling: Philosophical and conceptual issues. Palgrave Macmillan (Chap. 11).

Koen, B. V. (2003). *Discussion of the method: Conducting the engineer's approach to problem solving*. New York: Oxford University Press.

Kohler, M. A. (1969). Keynote address, in *Hydrological Forecasting*, WMO Technical Note No. 92, pp. X1–XV1, WMO, Geneva.

Konikow, L. F., & Bredehoeft, J. D. (1992). Ground-water models cannot be validated. *Advances in Water Resources, 15*(1), 75–83.

Kuhn, T. (1970). *The structure of scientific revolutions* (2nd ed.). Chicago, IL: University of Chicago Press.

Ladyman, J. (2002). *Understanding philosophy of science*. London: Routledge.

Lakatos, I. (1978). Philosophical papers. In J. Worrell & G. Curry (Eds.), *The methodology of scientific research programmes* (Vol. 1). Cambridge University Press.

Landström, C., Whatmore, S. J., Lane, S. N., Odoni, N., Ward, N., & Bradley, S. (2011). Coproducing flood risk knowledge: Redistributing expertise in critical 'participatory modelling'. *Environment and Planning A, 43*(7), 1617–1633.

Lane, S. N. (2012). Making mathematical models perform in geographical space(s). In J. Agnew & D. Livingstone (Eds.), *Handbook of geographical knowledge*. Sage, London (Chap. 17).

Lane, S. N. (2014). Acting, predicting and intervening in a socio-hydrological world. *Hydrology and Earth System Sciences, 18,* 927–952.

Lane, S. N. (2017). Slow science, the geographical expedition, and critical physical geography. *The Canadian Geographer, 61,* 84–101.

Lane, S. N., Landstrom, C., & Whatmore, S. J. (2011). Imagining flood futures: Risk assessment and management in practice. *Philosophical Transactions of the Royal Society, A, 369,* 1784–1806.

Lane, S. N., November, V., Landström, C., & Whatmore, S. J. (2013). Explaining rapid transitions in the practice of flood risk management. *Annals of the Association of American Geographers, 103,* 330–342.

Latour, B., & Woolgar, S. (1979). *Laboratory life: The social construction of scientific facts*.

Liu, Y., Freer, J. E., Beven, K. J., & Matgen, P. (2009). Towards a limits of acceptability approach to the calibration of hydrological models: Extending observation error. *Journal of Hydrology, 367,* 93–103. https://doi.org/10.1016/j.jhydrol.2009.01.016.

Lloyd, E. A. (2010). Confirmation and robustness of climate models. *Philosophy of Science, 77*(5), 971–984.

Lloyd, E. A. (2018). The role of "complex" empiricism in the debates about satellite data and climate models. In E. A. Lloyd & E. Winsberg (Eds.), *Climate modelling: Philosophical and conceptual issues*. Palgrave Macmillan (Chap. 6).

Masicampo, E. J., & Lalande, D. (2012). A peculiar prevalence of p values just below .05. *The Quarterly Journal of Experimental Psychology*.

Mayo, D. (1991). Sociological versus meta-scientific views of risk management. In D. G. Mayo & R. D. Hollander (Eds.), *Acceptable evidence: Science and values in risk management* (pp. 249–279). Oxford: Oxford University Press.

Mayo, D. G. (1996). *Error and the growth of experimental knowledge*. Chicago, IL: University of Chicago Press.
Mayo, D. G., & Spanos, A. (Eds.). (2010). *Error and inference*. Cambridge: Cambridge University Press.
Miller, D. (1974). Popper's qualitative concept of verisimilitude. *The British Journal for the Philosophy of Science, 23,* 166–177.
Mitchell, S., Beven, K. J., Freer, J., & Law, B. (2011). Processes influencing model-data mismatch in drought-stressed, fire-disturbed, eddy flux sites. *JGR-Biosciences, 116.* https://doi.org/10.1029/2009jg001146.
Morton, A. (1993). Mathematical models: Questions of trustworthiness. *British Journal for the Philosophy of Science, 44,* 659–674.
Niiniluoto, I. (2017). Verisimilititude: Why and how? In N. Ber-Am & S. Gattei (Eds.), *Encouraging openness: Essays for Joseph Agassi*. Springer. ISBN: 978-3-319-57669-5.
Nott, D. J., Marshall, L., & Brown, J. (2012). Generalized likelihood uncertainty estimation (GLUE) and approximate Bayesian computation: What's the connection? *Water Resources Research, 48*(12), W12602. https://doi.org/10.1029/2011wr011128.
O'Hear, A. (1975). Rationality of action and theory-testing in Popper. *Mind, 84*(334), 273–276.
Oldenbaugh, J. (2018). Building trust, removing doubt? Robustness analysis and climate modeling. In E. A. Lloyd & E. Winsberg (Eds.), *Climate modelling: Philosophical and conceptual issues*. Palgrave Macmillan (Chap. 10).
Oreskes, N. (2018). The scientific consensus on climate change: How do we know we're not wrong? In E. A. Lloyd & E. Winsberg (Eds.), *Climate modelling: Philosophical and conceptual issues*. Palgrave Macmillan (Chap. 2).
Oreskes, N., Shrader-Frechette, K., & Berlitz, K. (1994). Verification, validation and confirmation of numerical models in the earth sciences. *Science, 263,* 641–646.
Pande, S., & Savenije, H. H. (2016). A sociohydrological model for smallholder farmers in Maharashtra, India. *Water Resources Research, 52*(3), 1923–1947.
Parker, W. S. (2009). Confirmation and adequacy-for-purpose in climate modelling. *Aristotelian Society Supplementary Volume., 83,* 233–249.
Parker, W. S. (2018). The significance of robust climate projections. In E. A. Lloyd & E. Winsberg (Eds.), *Climate modelling: Philosophical and conceptual issues*. Palgrave Macmillan (Chap. 9).
Popper, K. R. (1959). *The logic of scientific discovery*. London: Hutchinson.
Popper, K. R. (1969). *Conjectures and refutations: The growth of scientific knowledge*. London: Routledge.
Popper, K. R. (1976). A note on verisimilitude. *British Journal for the Philosophy of Science, 27,* 147–159.
Popper, K. (1983). *Realism and the aim of science*. London: Hutchinson.
Popper, K. R. (1994). *The myth of framework: In defence of science and rationality*. London: Routledge.
Quine, W. V. (1969). *Ontological relativity and other essays*. New York: Columbia University Press.
Quine, W. V. (1975). On empirically equivalent systems of the world. *Erkenntnis, 9,* 317–328.
Robert, C. P., Cornuet, J., Marin, J., & Pillai, N. S. (2011). Lack of confidence in approximate Bayesian computation model choice. *Proceedings of the National Academy of Sciences, 108*(37), 15112–15117.
Rougier, J. C. (2007). Probabilistic inference for future climate using an ensemble of climate model evaluations. *Climatic Change, 81,* 247–264.
Sadegh, M., & Vrugt, J. A. (2013). Bridging the gap between GLUE and formal statistical approaches: Approximate Bayesian computation. *Hydrology and Earth System Sciences, 17*(12), 4831–4850.
Schoups, G., & Vrugt, J. A. (2010). A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic, and non-Gaussian errors. *Water Resources Research, 46*(10), W10531. https://doi.org/10.1029/2009wr008933.

Shackley, S., Young, P., Parkinson, S., & Wynne, B. (1998). Uncertainty, complexity and concepts of good science in climate change modelling: Are GCMs the best tools? *Climatic Change, 38,* 159–205.

Smith, L. A., & Stern, N. (2011). Uncertainty in science and its role in climate policy. *Philosophical Transactions of the Royal Society*, *369*(1956), 4818–4841 (Handling Uncertainty in Science).

Stengers, I. (2005). The cosmopolitical proposal. In B. Latour & P. Weibel (Eds.), *Making things public* (pp. 994–1003) Cambridge, MA: MIT Press.

Stengers, I. (2013). *Une autre science est possible*! Paris: La Découverte.

Suckling, E. B., & Smith, L. A. (2013). An evaluation of decadal probability forecasts from state-of-the-art climate models. *Journal of Climate, 26*(23), 9334–9347.

Vernon, I., Goldstein, M., & Bower, R. G. (2010). Galaxy formation: A Bayesian uncertainty analysis. *Bayesian Analysis, 5*(4), 619–669. https://doi.org/10.1214/10-ba524.

Viglione, A., Di Baldassarre, G., Brandimarte, L., Kuil, L., Carr, G., Salinas, J. L., et al. (2014). Insights from socio-hydrology modelling on dealing with flood risk–roles of collective memory, risk-taking attitude and trust. *Journal of Hydrology, 518,* 71–82.

Von Bertalanffy, L. (1968). *General systems theory*. New York: Braziller.

Vrugt, J. A. (2016). Markov chain Monte Carlo simulation using the DREAM software package: Theory, concepts, and MATLAB implementation. *Environmental Modelling and Software, 75,* 273–316.

Vrugt, J. A., & Beven, K. J. (2018). Embracing equifinality with efficiency: Limits of acceptability sampling using the DREAM (LOA) algorithm. *Journal of Hydrology, 559,* 954–971.

Watkins, J. (1985). *Science and scepticism*. Princeton: Princeton University Press.

Weisberg, Michael. (2006). Robustness analysis. *Philosophy of Science, 73*(5), 730–742.

Westerberg, I. K., & McMillan, H. K. (2015). Uncertainty in hydrological signatures. *Hydrology and Earth System Sciences, 19*(9), 3951–3968.

Westerberg, I. K., Guerrero, J.-L., Younger, P. M., Beven, K. J., Seibert, J., Halldin, S., et al. (2011). Calibration of hydrological models using flow-duration curves. *Hydrology and Earth System Sciences*, *15*, 2205–2227. https://doi.org/10.5194/hess-15-2205-2011.

Wimsatt, W. C. (2007). *Re-engineering philosophy for limited beings*. Cambridge: Harvard University Press.

Winsberg, E. (2003). Simulated experiments: Methodology for a virtual world. *Philosophy of Science, 70,* 105–125.

Woodhouse, M. J., Hogg, A. J., Phillips, J. C., & Rougier, J. C. (2015). Uncertainty analysis of a model of wind-blown volcanic plumes. *Bulletin of Volcanology, 77*(10), 83. https://doi.org/10.1007/s00445-015-0959-2.

Young, P. C. (2013). Hypothetico-inductive data-based mechanistic modeling of hydrological systems. *Water Resources Research, 49*(2), 915–935.

Young, P. C. (2018). Data-based mechanistic modelling and forecasting globally averaged surface temperature. *International Journal of Forecasting, 34*(2), 314–335. https://doi.org/10.1016/j.ijforecast.2017.10.002.

Zreda, M., Shuttleworth, W. J., Zeng, X., Zweck, C., Desilets, D., Franz, T., et al. (2012). COSMOS: The cosmic-ray soil moisture observing system. *Hydrology and Earth System Sciences, 16*(11), 4079–4099.