

BACKGROUND DOCUMENT: UASA “BLOCK”

M B Beck and A Saltelli

TAUC’s essential purpose:

“Handling uncertainty associated with using models in a policy-focused context in the service of making better environmental decisions.”

Guiding principle for this Background Document:

“What are the needs of UASA at each stage of the *Life Cycle* of an environmental policy?”

Other than the fact that this document is very much a “work in progress” — indeed, “half-baked”, as some would say — the following guidance is offered. First, as the first co-author, I (MBB) take full responsibility for this text. I left insufficient time for my co-author to have a reasonable shot at incorporating his material into a genuinely co-authored draft. Second, I allowed myself to be drawn into more of a personal take on the “overview” of the entire narrative of TAUC, hence in part the excessive length of this document, hence too its migration away from its original focus (as set out above). Third, shortage of time has left me with no capacity even to list the references that are cited (too many Becks, doubtless), let alone do a much more decent job of reflecting what our field as a whole has achieved in respect of this subject. Fourth, nevertheless, like all the other documents prepared as the basis for discussion during our Workshop, the purpose of this document is to raise questions and stimulate debate. Last, to save embarrassment on a wider scale, I am asking that this document not be shared with anyone outside our Workshop participants, except by direct request to me (not that you would be thinking of sharing such material more widely anyway).

1 INTRODUCTION

We propose to cut a quite particular path across the vast field marked out at the intersection of models, uncertainty, and the making of environmental policy. It will be organized around what we are provisionally calling the life-cycle of that policy, rule, directive, or regulation. Assuming we are able to define this life-cycle, and that it is a useful and meaningful organizing principle, albeit differing as it may between practice in the USA and the European Union (EU), the goal of this Background Paper is to answer the question:

What are the needs for Uncertainty Analysis and Sensitivity Analysis (UASA) of models as they are developed, drawn into, and employed within the various stages of the life-cycle of an environmental policy?

In other words, our discussion of UASA will be defined by the “demand pull” of these particular needs for it, not according to the “technology push” of what is computationally feasible today. We stress, however, that this is a matter of organization, not of placing pragmatism on a pedestal to the detriment of research. Exploring what might be the future promising directions for research — in handling uncertainty associated with using models in a policy-focused context in the service of making better environmental decisions — is one of the priorities of the TransAtlantic Uncertainty Colloquium (TAUC). From that perspective, however, it should be noted that dealing above all with epistemic uncertainty, i.e., uncertainty and error in the underlying concepts or *structure* of the model, is a distinguishing feature of TAUC. We acknowledge the fundamental insecurity and inadequacies of the science bases on which models are built. We recognize that there can be substantial errors in what we have chosen to include in the given model, the {presumed known} as we shall label it herein, and even more uncertainty arising from the ocean of ignorance, i.e., the {acknowledged unknown}, surrounding this small island of the {presumed known} (Beck, 2002, 2005; Saltelli ...).

According to our provisional organizing principle, this draft of the Background Document is gathered around the following sequence of stages in the life-cycle of an environmental policy:

- (i) Cultivating early warnings of issues, including discriminating those issues requiring a policy from those that do not;
- (ii) Developing the provisional environmental rule/directive/policy, from the perspective of, i.e., from within, the regulatory agency;
- (iii) The political science/economy of dispute and negotiation over the provisional rule;
- (iv) Structuring the legal discourse in resolving disputes;
- (v) Operating in a post-rule world of enforcement, learning, and adaptation.

At each of these stages, our goal is to illuminate how models are used and, more particularly, how

UASA is deployed alongside those models.

It will be helpful, however, to preface our discussion with some brief remarks on how we view epistemic uncertainty, how it manifests itself in computational models, and the changing nature of the relationship between science and policy.

2 MODELS, SCIENCE, AND POLICY

2.1 Epistemic Uncertainty, Models, and Virtual Realities

To serve some purpose, an “industry” of experts and analysts is engaged in the enterprise of developing computational models capable of replicating the behavior of environmental systems. At any stage we are unable to include in these models all that we “know” — or believe we know — of the things that must be affecting the behavior of the system “in reality”. According to conventional wisdom, scientific (epistemic) uncertainty is reduced by including greater refinement of description in the model, that is, the inclusion of more state variables to represent both a greater diversity of chemical and biological species and the spatial distribution of these species in greater detail. Uncertainty in this sense should, in principle, be reduced by unfolding the coarser-scale parts of the model, specifically some of its parameters, or coefficients, into sets of more refined state variables, and inter-relationships among them containing parameters that are truly invariant at this finer scale of resolution. Parameters at one scale subsuming state variables at another, which states are not constant in time-space (because they are variables), are themselves capable of variation. At any given scale of representation, the model’s parameters are thus commonly used as interim, empirical expedients — temporary parking places, as it were, for knowledge of a too detailed or too speculative nature. These parameters are the ports of entry into the search for *invariance* — at a finer scale of resolution. And this search, the very essence of progress in understanding, may one day culminate in use alone of the irreducible set of the three universal constants of Nature (the velocity of light in empty space; the quantum constant of Max Planck; and the gravitational constant of Isaac Newton).

For as long as unit computational times and costs are decreasing, such uncertainty might continue to be resolved and reduced in this manner. Our models will expand, to become enormous vessels laden with all the theory that can be crammed into them. They will elide into virtual realities. And then, in our minds, we shall discreetly shed the limiting qualification of their being but “virtual”. Who has not watched an animated film of proteins chopping, changing, clipping, snipping, and transcribing bits and pieces of DNA, and been anything but mightily impressed? As lay observers of the scene (as the vast majority of us are) we perceive we could comprehend the previously incomprehensible — much as we have been enlightened by the powerful insight of circulation in the North Atlantic behaving as a conveyor belt (which we can all comprehend). Indeed, this insight has itself been liberated from the vast bulk of the requisite numerical oceanography, now at last realized on a sufficiently large computational platform (and fully appreciated by just a few professional *cognoscenti*).

And yet, those animated molecular graphics of protein-DNA behavior are still a *simulation*. What is more, this is not reality resolved down to the finest scale, of the most elementary subatomic particles. It is but the virtual reality of what happens within one cell, as part of some tissue, within one organ, within one organism, within a population of such organisms, interacting with each other, and with organisms of other species, in an ecological system, inter-connected with other ecosystems, in what constitutes our environments and they, collectively, our biosphere. Few models of environmental systems resolve the world down to a scale more refined than that of a population of micro-organisms, expressed as a single, aggregate state variable with uniform properties in time-space. We know of no environmental models populated only by the three universal constants (parameters) of Nature.

So while there is innate fascination in the march towards virtual realities — the allure of the powerful insight and discovery, and the satisfaction of expunging the inelegant expedient of parameterizing some of the microscopic parts — there is an inescapable illusion. A model cannot evolve towards becoming the real thing, by definition. The illusion, however, is perfectly in order. This is what draws Science on, in its unending quest. But the virtual reality is an ever receding horizon, for all of that. Thirty years on from what has been called the “youthful exuberance of systems ecology” of the 1970s (Beck, 2002), the need to issue such a reminder endures.

2.2 Homing in on the “Truth” of the Matter

As long ago as 1973, O’Neill (1973) observed that the error in the predictions from a model should decrease with a decreasing degree of model aggregation (in his case, aggregation of the behavior of several species of organisms into a single state variable). However, he also noted that precisely this increasing refinement of detail — more complex kinetic expressions, more state variables — would tend to increase the prediction errors resulting from the necessarily increasing number of model parameters with uncertain values.

Given no known environmental models populated only by the three universal constants (parameters) of Nature, all the parameters of a model may need to be estimated in principle, within the given structure of the model, through the process of reconciling the performance/behavior of the model with that observed of the field system. We shall refer to this process as system identification. In more familiar terms, calibration of the generic model to the observed properties and behavior of the specific, prototype environmental system has almost always been necessary. Uncertainty in the estimated values of the model’s parameters would ideally be reduced thereby, but never entirely, and not necessarily so, if surprising, seemingly anomalous, or confounding behavior happens to have been observed. In a Bayesian spirit, the prior uncertainty attaching to the model, when the model is reconciled with a set of uncertain field data, will result in a model with a posterior pattern of associated uncertainties, most appropriately expressed as the uncertainties attaching to the posterior estimates of the model’s parameters. These posterior uncertainties, together with uncertainties arising from other sources, will be propagated forward with any predictions derived from the model, for the purpose of evaluating policy options (see for example, Beck, 1987). In this light, calibration should be seen more as a matter of calibrating the uncertainty of the model, less a matter of finding

some uniquely best set of estimates for the model's parameters.

Development and application of the (modern) algorithms of calibration and system identification suffered from just as much youthful exuberance in the 1970s as did systems ecology. Suffice it to say, it has been far from possible in the majority of cases to arrive at a “uniquely best” set of estimates to be inserted into the model, for predictive purposes. Frequently it has been disturbing to discover both how uncertain and how absurd — relative to prior beliefs about the science base underpinning the model — those estimates might turn out to be in terms of generating a match between the model and the observations. When this has occurred — provided that the patterns of posterior parametric uncertainty have been computed (by no means often the case) — the strong inference to be drawn from such results is that the structure of the model is flawed, i.e., subject to significant structural error/uncertainty, subject, that is, to significant epistemic uncertainty. Furthermore, belief in the presumption that better field data and substantially improved algorithms of system identification would lead ultimately to conquest of these difficulties has, for some of us, been abandoned. Indeed, there is no longer (Beck, 2002) adherence to what was once (in the 1970s) a bold program of research, intended to employ the methods of system identification in successively homing in on the “truth” of the matter, i.e., in identifying *the* correct structure of the model (Beck, 1987). In this resides, of course, acknowledgment of the profound importance of epistemic uncertainty relative to aleatory uncertainty, roughly equivalent to structural uncertainty relative to parameter uncertainty, as accommodated within the framework of environmental modeling.

2.3 Models as Tools

To continue thinking along the lines of models as formally codified, succinct expressions of our current best beliefs about the truth of the matter, however, is to be distracted from a most helpful insight (Ravetz, 1992): the conception of models as tools, including as truth-generating machines, if needs must, when making predictions of future behavior. Nowhere more so does this insight assist us than in escaping from the impasse of the “paradox of prediction” (Beck *et al*, 1997). We construct models, more often than not, for exploring our future possibilities. Of the greatest intrigue is this: what might occur that has not previously been observed, for example, in releasing into the environment an entirely novel xenobiotic substance or genetically modified organism? Given no historical record of past occurrences — by definition — on what other device should we rely than a computational model? If we place our trust in such a model, however, proscription of extrapolation beyond the span of the history already matched by the model, as advocated by Konikow and Bredehoeft (1992), is hardly an option. And looking back, we might suggest, these authors were so minded because we could not (and cannot yet) have models of the environment populated alone by the three universal constants of Nature. In short, when a model is most needed, for the purpose of radical extrapolation, there is the least basis for our investing trust in it on the grounds of its resemblance to the past empirically observed behavior of the prototype, given that we cannot be entirely secure in its companion theoretical foundations, as a consequence of ineluctable epistemic uncertainty; hence the paradox of prediction.

A model is a tool designed to fulfil a designated task, and that task may come in a variety of forms.

Besides intended as a truth-generating *machine*, the model might alternatively be constructed for the purposes of being, for instance: a succinct *archive* of current hypotheses about the behavior of the system; an *instrument* for forecasting and foresight generation in the making and evaluation of an environmental policy/decision; a *device* for communicating environmental science to an audience of scientifically lay stakeholders; or even a *vehicle* for discovering our ignorance — the {acknowledged unknown} — and at the earliest possible juncture (Beck, 2002a,b).

With this crucial insight, then, we can conceive of developing and evaluating models in a somewhat different, broader context. At the outset is the task the model is to fulfil, as elaborated by those holding a stake in the issue at hand, be they model developer, model user, and/or the parties affected by the consequences flowing from the purpose to which the model is to be put. Then comes the matter of designing the model against the goal(s) of this task description, subject to a variety of constraints. Like any other tool designed to fulfil a purpose — a hammer, a screwdriver, a computer program, and so on — there may be tradeoffs amongst these design constraints, which tradeoffs govern what is put into the model, i.e., which bits of the science bases are mobilized in its construction. More general notions of good/bad design of the tool can come into play, where what constitutes goodness/badness will lie in the eye of the beholder. In principle, a poor tool can result from a poor specification of the task, but the problem would then be to discover the poverty of the task description in the first place.

2.4 The Changing Relationship Between Science and Policy

In a recent paper Funtowicz and Strand (2006) chart a succession of conceptual “models” — *not* mathematical models — reflecting the changing relationship between Science and Policy (if not Society, in the sense dealt with by Gibbons (1999) and Nowotny *et al* (2001)). Taking as their point of departure what they call the “modern model”, wherein “Science informs policy by producing objective, valid and reliable knowledge”, they examine how various successor alternatives have attempted to overcome the essential challenges to which the modern model is subject: that most important real-life environmental and health issues display both complexity and scientific (here epistemic) uncertainty.

[To be continued, not least with a view to understanding how Funtowicz and Strand relates to our subsequent discussion in Section 4 and, perhaps more so, Section 5 (and 6?).]

3 CULTIVATING EARLY WARNINGS

Policies are oriented to the future; and to the future can attach the greatest uncertainty.

The essential questions for any foresight scheme for cultivating early warnings, especially in the prospective mode of facing the threats of the future (as opposed to the retrospective mode of curing the ills of the past), are these:

- How do issues emerge, to approach and confront us, from over the horizon? Is something of potential significance happening “out there”, such as nascent technologies with adverse consequences alongside their proposed benefits, or behavior in an environmental system moving barely perceptibly towards a cumulative tipping point (a nonlinear dislocation in behavior; a surprise)?
- How do we discriminate, if we can and do, issues requiring the formulation or adaptation of a policy from those that do not?
- What are the roles of models and UASA in responding to (answering) these questions?

3.1 Design for Discovery of Ignorance

In 1995 the Science Advisory Board of the US Environmental Protection Agency (EPA) produced a report entitled *Beyond the Horizon: Using Foresight to Protect the Environmental Future* (Science Advisory Board, 1995). In particular, the report observed that

... the [Environmental Protection] Agency has an obligation to search for the “weak signals” that portend future risk to human health and to ecosystems ...

and that any such search should be

... eclectic in its use of information sources ...

It recommended that

... EPA should establish an early-warning system to identify potential future environmental risks ...

and went on to discuss possible systems of enquiry for meeting this objective. The Executive Summary of a later report from the US National Research Council on *Grand Challenges in Environmental Sciences* says (NRC, 2001):

People want to be warned of major environmental changes and, if the environment is under threat, want to know how to respond.

The image of some kind of threat-probing radar is apt. What, then, might be such a “system of enquiry”, the means of generating foresight? There is, without belittling it, the obvious response: an area is to be identified in which scientific data are sparse and/or in conflict and the scientist conducting the enquiry is to submit — to the process of scientific peer review — an opinion on the interpretation of the data as portending some threat to the environment¹. In other words, the extant

¹ This appeared subsequently in a request for proposals for research on *FUTURES: Detecting the Early Signals* (1999 Science to Achieve Results (STAR) Program, National Center for Environmental Research and Quality Assurance,

historical record gathered within the paradigm of scientific enquiry is to be examined and interpreted by a practising scientist whose opinion will be judged by other practising scientists. Such judgement, one senses, would instinctively be urged on towards the “singularity” of consensus on the most plausible speculation/interpretation.

The current European Research Area (ERA) network project Scientific Knowledge for Environmental Protection (SKEP; www.skep.org) has a Work Package devoted to the problem of foresight. Its “systems of enquiry” are several, ranging from [revisit the SKEP website for Paris 06 presentations] ... through learning from the lessons of the rich history of Foresight in other areas (technology) ... to Schemes of horizon-scanning, embracing the basic principles of those advocated by the US EPA’s Science Advisory Board, are in use, but hugely expanded now in their reach across the world wide web, and substantially mechanized by web-crawling software devices, in ways barely imaginable a decade ago when the SAB published its *Foresight* report (Huggins, 2006; see also Pascual/Street, 2006; Rejeski, 200x). In generating syntheses from probing what lies in certain directions on the horizon, the practising scientist exercises his/her judgement over the relevance and trustworthiness of not just the “extant historical record gathered within the paradigm of scientific enquiry”, but also information from a host of various other sources — literally, all those actors able to post material to the web and have a “voice” there. This material, moreover, can include all manner of speculations about future threats and the future capacities, for better or for worse, of nascent and imagined technologies (Huggins, 2006).

In his book *State of Fear* (2004/5) — a novel set around the dispassionate, legal basis of a judgement for or against the occurrence of climate change — Michael Crichton argues thus. Social control is best managed through fear (according to the book’s character Professor Hoffman), but the fall of the Berlin Wall created a vacuum of fear, which had to be filled; and indeed was filled by threats of a toxic environment. Once it was the military-industrial complex that was the driver of Society. Now it is the political-legal-media complex: politicians need fears to control the population; lawyers need dangers to litigate (and make money); and the media need scare stories to capture an audience. Universities today are factories of fear; they produce a steady stream of new anxieties, dangers and social terrors to be used by the politicians, lawyers and reporters. We can readily appreciate Crichton’s point: the *National Geographic* television channel is replete with vivid and colorful narratives (tales) — made all the more vivid through highly sophisticated computer animation — of what might happen to our planet and our environment in the future; and they are overwhelmingly threats and fears, backed up by research scientists from, amongst other institutions, universities. But does each and every one of these threats, fears, and aspirations call for a policy?

Looking back, we can see just how “eclectic” has become the mix of information sources into which we are now tapping for our environmental foresight. In cryptic form, we might say that we began with recourse merely to {scientific empirical observations & scientific opinion} — in the SAB’s *Foresight* paper — and seem to be ending up with merely {stakeholder imagination}, some of which, for some stakeholders, may indeed subsume scientific opinions on the scientific

interpretation of scientific empirical observations (and include the use of models and UASA). So great has become the capacity of actors outside the customary scientific enterprise to manufacture surprises and disruption of “normal service”, in particular, for those bearing “government responsibility” for environmental stewardship, that scanning the horizon in order to contend with this stakeholder capacity itself has become as important as trying to detect those weak signals (portending future risk to human health and ecosystems) that are candidates worthy of a policy. This all echoes what we have said above (in Section 2.4, hopefully) of Post Normal Science (Funtowicz, Ravetz, Saltelli, Zidek), socially robust science (Gibbons, 1999; Nowotny *et al*, 2001), the changing relationship between expert and lay knowledge (Darier *et al*, 1999), and of the relationship between Cultural Theory and environmental stewardship (Thompson *et al*, 1999). In a sense, we are saying nothing more than that the activities (social solidarities?) of the lay public matter; and that they matter in highly significant — and highly creative — ways.

All of this gives no hint, however, of how models and UASA might be used in procedures of horizon scanning or foresight generation for cultivating early warnings. In none of the case studies from the computational era of the 1960s onwards in *Late Lessons From Early Warnings* [Check this!] was any use of a formal computational model cited as having been instrumental in prompting an early warning. In many ways, where we are today is summed up in the title of a recent book: *Environmental Foresight and Models: A Manifesto* (Beck, 2002b). Developments, if there are to be any, lie ahead of us.

Let us recall the words we are using here: “signals”, “radar”, “horizon-scanning”, and so forth. They evoke the image of a satellite or observatory, which we know are either already in place (GEOSS; Gary Foley) or under consideration, in the proposed Environmental Observatories (EOs) of the US National Science Foundation (NSF; ORION (2005), NEON (2006), CUAHSI (2006), CLEANER (2006)). In short, there is the ambition to instrument our environment, from the oceans to the lands and to their ecologies, watersheds, and cities, to a massive extent, thereby to open up the final frontier in observing the behavior of those environments — in *real time*, that is — substantially automated, and with signals flowing from the observatory to the cyber-infrastructure, and *vice versa*. Not the least of the significance of this derives from the conceptual novelty of the much closer and more intimate juxtaposition of historically odd disciplinary bed-fellows: between Geophysics and Ecology on the one hand, and Control Theory on the other (see also Eigbe *et al*, 1998). Mathematical filtering theory (Jazwinski, 1970) and its more recent progeny, often referred to as algorithms of data assimilation (notably in the Ocean Sciences), deal *par excellence* with signal processing in a real-time manner and the detection of faults, failures, and anomalies in the behavior of systems.

Historically, control and filtering theory have been geared almost exclusively to applications in supervising the performance of man-made systems, where epistemic uncertainty — or at least undue sensitivity to it — might have been “engineered out” of the system’s design and construction. Components in engineered systems can fail to perform as expected, however. Models of the system’s behavior — formal mathematical representations of the {presumed known} — can thus be embedded in a filtering-like algorithm, itself an embodiment of the Bayesian account of accommodating the analysis of uncertainty, and trained on detecting, at the earliest possible juncture,

any hint of incipient failure, i.e., departure from the expected behavior of the reference model. In the case of environmental systems, we are not afforded the luxuries of either deliberate design for discrimination against epistemic uncertainty or confinement of the source of failure to the {presumed known} alone. We now know well, given appropriate technologies for environmental observation, such as the relatively modest network of sensors for water quality strung across the Lagoon of Venice, Italy, that we have equally appropriate (filtering) algorithms for detecting incipient instrument failure, conditional upon the {presumed known} of the structure of the embodied model being utterly correct (Ciavatta *et al*, 2004)². But it may well not be “utterly correct”. For the structure of the {presumed known} may be flawed. It may be in error, i.e., subject to significant epistemic uncertainty, so that the fault, or anomalous behavior, may reside either in the observing system or in the behavior of the observed environment (as noted towards the close of Ciavatta *et al*, 2004). More complicated yet, something of significance to the way in which the system functions (and may function in the future) — something technically not ascribable to the actions of pure chance (or to the aleatory uncertainty attaching to the model’s parameters) — may alternatively have been detected in the vast ocean of ignorance and epistemic uncertainty lumped under the {acknowledged unknown}.

From this line of argument, we can now appreciate epistemic uncertainty, as accommodated within a model, in more subtle and variegated terms: as manifesting itself both as errors/flaws in the {presumed known} of the model’s structure and as something of emergent, palpable, non-random, deterministic significance in the {acknowledged unknown}, i.e., residing in that which has been left out of the model’s structure. Hence, we can contemplate the prospect of deploying models and UASA for cultivating early warnings, by conceiving of the purpose of the model as a tool *designed expressly* for discovery of our ignorance (as we have already signaled in section 2.3). How one might realize such a design remains a matter for research, well ahead of us, perhaps, beyond its first adumbrations in Beck (2002b), Ciavatta *et al* (2004), Beck (2005), and Lin and Beck (2006) [and hopefully elsewhere!].

3.2 Reachable Futures: Key Unknowns in Policy, Technology, and Science

To have discerned a future threat in the accumulating “extant historical record”, in particular, through deploying a formal, scientific model, is one thing. But the second question we set ourselves for discussion in this section requires us to divine whether the apprehended threat merits the development of a policy to pre-empt or counter it. Again, our need is to discern possible roles for models and UASA in this. The issue is *not* of divining what future behavior may be like (a matter of prediction/forecasting). Rather, the question is now: is the apprehended threat, or the host of

² The parameters appearing in the structure of the model are nevertheless considered fully subject to uncertainty and, indeed, this is actually vital to the subsequent discussion, although a technicality to which we do not need to descend in detail herein (see, for example, Beck, 2005; Lin and Beck, 2006)

threats deriving from the {stakeholder imagination}³ fermenting under Michael Crichton's prevailing *State of Fear*, reachable or plausible as an image of target future behavior? Clearly the intent is that a feared (desired) future should have a low (high) probability of being reached; and that policy formulation, along with appropriate, nascent technologies, should have a reasonable prospect of changing these probabilities (in desirable directions).

Being beset by a host of threats deriving from the creativity of {stakeholder imagination} is no bad thing. We need all the richness of perspective we can get. What happens in real life is often surprising; what might come about, at the periphery of our imagination, ought therefore to be within the purview of our analysis; consensus on the long-term outlook seems insufficient on its own. After all, to be forewarned is to be forearmed. The consensus might well merit the forearming of a policy; the peripheral might warrant merely a watching brief. But which is the more plausible, the more reachable, in scientific (and policy) terms? For these outcomes of the {stakeholder imagination} need anchoring back into the peer-reviewed science base, all its epistemic uncertainty notwithstanding.

We have argued elsewhere (Beck, 2002b) a case in favor of generating such foresight through a procedure of tapping into {scientific models and stakeholder imagination}, as a compliment of the more customary {scientific empirical data & scientific opinion} of the SAB's *Foresight* document. We have also demonstrated our case in a prototype study of Lake Lanier, just to the north of metropolitan Atlanta, Georgia, USA, in respect of deploying an aquatic foodweb model and the methods of UASA to assess the reachability of the best hopes and worst fears for the future of the lake, as imagined by a group of generally scientifically lay stakeholders (Beck *et al.*, 2002; Osidele and Beck, 2004). Think of the procedure thus. We are familiar with having a definition of past behavior of the system, its extant historical record. And we are equally familiar with the notion of reconciling (calibrating) an uncertain model with that uncertain history. Here instead we have imagined future behaviors — each grossly uncertain in their characterization, of course — against which the model is to be reconciled. In the Lanier study, the results sought were broadly threefold: (i) relatively speaking, how more (less) reachable were the feared/hoped-for futures, one to another; (ii) which bits of the uncertain science base (in fact, parameters in the model) were key in governing whether target future was attained from those that were redundant in discriminating attainment from non-attainment; and (iii) were the same bits of the uncertain science base thus key with respect to both the feared and the hoped-for futures? Given a limited budget for future research, with which to conduct scientific enquiries into reducing the epistemic uncertainty surrounding the many possible scientific unknowns, it was clearly desirable to identify just the handful of *key* unknowns — and even more so should they be apparently key to the reachability of a range of possible futures of concern.

We here are interested centrally in policy, however, not just the purchase of more science. In principle, the challenge is technically no greater, since the elements of policy, and any associated

³ We take professional scientists and engineers to number amongst the stakeholders and will accept descriptions of target futures derived from whatever sources, by whatever means, through constructing scenarios (Schwarz) or the more formal construction and manipulation of Belief Networks (Varis, 2002), for instance.

technologies, can be “parameterized” in the model, just as are the bits of the science base. Hence opens up the prospect of being able to answer questions such as: is avoidance of a future threat “policy-feasible” or “policy-infeasible” at present; which elements of policy, in concert with what types of nascent technology, and which current scientific unknowns, appear key to making a desired future more reachable and a feared future more avoidable; and so on? The study of such questions, using models and UASA, remains largely to be done (although see Osidele *et al*, 2003).

4 DEVELOPING THE PROVISIONAL RULE/POLICY

In the life cycle of a policy, the potential future environmental risk has been apprehended, and identified as both in need of a policy and capable of being averted, in principle, by wielding the current portfolio of policy instruments. Putting ourselves in the place of that single, particular stakeholder charged with developing the provisional rule/policy, i.e., the governmental regulatory agency, its tasks are now broadly twofold: to choose a promising policy from amongst several options, through assessment of respective future consequences for the environment at risk (and the economy); and to be prudent in anticipating the nature of the reception of the provisional policy amongst the agency’s various external stakeholders and affected parties.

In other words, we too here must have an eye on what follows in Sections 5 and 6 in this document, supposing these external “audiences”, amongst whom the “pilot product” of the provisional policy is to be “test marketed”, will have a variety of attitudes towards risk — typically those of risk-averse, risk-seeking, and risk-managing (Thompson, 1989). The present section is lengthy, in part because it will set out the ground rules for the terms of the debate and disputation amongst the various stakeholders, but also because relatively deeply embedded technical matters — in particular, those of formally enumerating aspects of both aleatory and epistemic uncertainty — must be brought to the surface and exposed for wider inspection and appreciation. The present section is the *terra cognita*, if one so wishes, preceding the *terra infirma* of Sections 5 and 6. What is seen of the subject of our paper primarily from the perspective of a single, particular stakeholder in this section must be viewed from the multiple perspectives of the several stakeholder groups in Sections 5 and 6.

Our greatest concern, once more, is to review the present and future challenges in deploying models and the methods of UASA across this stage in the life cycle of a policy. Two things are immediately striking: first, the sharp difference between the US and EU in the extent of using models in developing a provisional policy [See TAUC Legal Background Document; Fisher *et al*, 2006]; and second, the long tradition in the US EPA of doing just this, i.e., formally employing models, if not UASA, in rule-making — witness the extensive role of EPA’s SAB in this context, as well as, more recently, creation of EPA’s Council on Regulatory Environmental Modeling (CREM), the Council’s issue of its *Draft Guidance* document (Pascual *et al*, 2003), the (2005) review of this document by the SAB, and the associated work of the current National Research Council (NRC) Committee on

*Models in the Regulatory Decision Process*⁴. To summarize, much of this substantial effort is directed at the issue of “model evaluation in the regulatory setting”. And model evaluation, together with the deployment of models in regulatory impact assessment and the policy-screening process, are the two focal topics of the present section.

In contrast to the foregoing discussion of foresight generation and horizon scanning in Section 3, the questions are not ones of whether models are to be used, but those of the extent to which the methods of UASA are applied and of how epistemic uncertainty in models is treated, if it is. But before we look across this field, extensive in its own right, in order to highlight the past and possible future roles of UASA, two strategic changes of the past 15 years or so must be noted.

The first of these changes relates to the perception of models as tools, as already noted. This is especially important at the present stage in the life cycle of a rule, where models are to be employed as instruments of forecasting in the making and evaluation of an environmental policy, not least because it was in this setting that the notion itself first arose (Ravetz, 1992), subsequently to be more fully elaborated, but by no means sufficiently, in Beck *et al* (1997) and Chen and Beck (2000). Indeed, as we shall see, the ramifications of these earlier developments are still evolving, including as a function of writing this paper. Much could be said on the topic of model evaluation — formerly known as model “validation” — and a very great deal has been. Suffice it to say that some of this extensive debate has been entrained already into our previous discussions of the subject, without in any way suggesting these have been either definitive or exhaustive statements thereof (Beck *et al*, 1997; Ford *et al*, 1999; Beck and Chen, 2000; Beck, 2002). And yet more will appear in the outcome of the current NRC Committee on *Models in the Regulatory Decision Process*. The important consideration herein is this: conceiving of models as tools implies conceiving too of a “design space” for models, with then significant implications for applying the methods of UASA.

The second strategic change has to do with the nature of policy-making for stewardship of the environment, where the “style” of decision-making has moved from that of a command-and-control technocracy to something of a more participatory, more open democracy (Darier *et al*, 1999). We must address the changing perception of what it takes, therefore, to trust a model when employed in environmental policy-making at the Science-Society interface, and who will be involved in having a legitimate right to be asked whether they can indeed accord the model, and the decisions emanating from its application, such trust.

4.1 Strategic Changes: Trusting the Design of the Tool

An Essential Matter of Trust

On 29 August, 2003, the Office of Management and Budget of the US Federal Government issued

⁴ One of whose charges is to emulate the success of the NRC’s earlier “red book” on risk assessment (*Risk Assessment in the Federal Government: Managing the Process*, 1983).

a “Proposed Bulletin on Peer Review and Information Quality”. [See Jasanoff e-mail comment of 16 December, 2003] The purpose of the Bulletin was to ensure “meaningful peer review” of science pertaining to regulation, as part of an “ongoing effort to improve the quality, objectivity, utility, and integrity of information disseminated by the federal government”. Responding to the manner in which the Bulletin was proposing to meet this intent, Jasanoff argued that, in short, making progress may depend more on getting stakeholders — the public, the regulators, the scientists, and so on — to agree in advance on appropriate methodologies and investigative protocols, than on subsequent scientific peer review, at least in *regulatory* science. Establishing, and demonstrating, the reliability and credibility of the peer review process itself — we would say, evaluation process — are every bit as crucial as the conventional challenge of establishing the reliability and credibility of the information to be reviewed in the process, including that from models. In what would be Jasanoff’s preferred form of “extended peer review”, it is the process, not the product, that matters; and the scientifically lay public, as legitimate stakeholders, should be engaged therein from the very beginning.

These issues of trust and reliability trouble all of us holding a stake in the development of a model and the consequences that might flow from its predictions. Take those who develop the models as their profession. What concerns them is this: does the peer group of scientists approve of the constituent elements of theory mobilized in the model; and can the model’s articulation of theory be reconciled with the empirical experience of observed behavior — in particular, in a manner minimally tainted by what some may fear are the arbitrary parametric adjustments of model calibration? Then there are the decision-makers and policy makers, the decision-owners, that is, and the wider community of model users. What, they wonder, are the risks of making the wrong decision, as a result of using the model? What is the risk of subsequent litigation, either way — of actions leading to overly burdensome constraints on the economy or inadequate stewardship of the environment? And what bothers the people, the community of scientifically lay stakeholders, upon whom will be visited the consequences of the actions born of models and their predictions? It is this: can we trust those developing and using the models to respect and protect our interests, as the most palpably affected parties; can we see through the model; and can our worst fears be avoided, while our greatest hopes are yet attained?

Design Space of the Tool

Keeping in mind the image of the model as a tool, we can conceive of criteria for designing the tool against a specified task, i.e., we can imagine a design space for the model, from which can follow the idea of judging the quality — good or bad — of the model’s design, where we recall further that “quality” may lie in the eye of the beholder, amongst our various groups of stakeholders.

This is key, not only because it encapsulates the two strategic changes in style and context from the past decade and more, but because it confronts those who build models with the notion of designing a tool, whose purposes may be much more varied than that of the previously predominant, somewhat singular, pursuit of models as truth-generating machines. In other words, it should arrest thinking along the single line of models as a secondary science, following merely in the wake of the primary

field sciences of the environment (physics, chemistry, biology, ecology, and so on) — and of trying solely to emulate them. That is but one legitimate purpose to which a model might be turned. Our focus must instead be upon a model designed as an *instrument* of prediction⁵ in support of formulating a policy. For example, the model might have to be designed to predict high-end exposures of a subject population to a given hazardous substance (Task A we might say). In this rudimentary sense, requiring the model to serve the purpose of predicting low-end exposures would constitute an alternative description of the task (Task B); and different designs of models might be preferred for the two different tasks.

There are at least three criteria against which to judge the quality of the design of a model to meet a specified purpose, as judged from as many varieties of perspective as there are stakeholders (and here, in particular, we must acknowledge this text as a “work in progress”):

- *Fidelity*, in the sense of faithfully reflecting as much as possible of what scientists consider (currently) to be the essence of the system’s workings, based on all past experience of those workings;
- *Relevance*, in the sense that the model is, in principle, capable of generating the task-relevant projections of future behavior — as a consequence (herein) of the contemplated policy actions — with a trustworthy accounting for the propagation of uncertainty;
- *Transparency*, in the sense that an informed, but scientifically lay, stakeholder can comprehend the essential workings of the model.

It is helpful to think of these as marking the vertices of a triangular design space (Figure 1), where most models designed for most tasks, i.e., as feasible solutions to the design of the tool, will require pragmatic trade-offs to be made amongst the three, for instance, relinquishing some Fidelity in acquiring greater Transparency. The three are not mutually exclusive; nor is the design necessarily some form of “zero-sum game”. For example, if the task were to have a model as an archive, its design might place it very close to the Fidelity vertex, with that design scoring highly on the questions of the approved materials in the composition of the model’s structure and the resemblance of its behavior to that of the real system. At least, this would be the kind of scoring the group of model builders, as stakeholders, would wish to have. On the other hand, to meet the task of the model as a communication device, a good design of model might be located close to the Transparency vertex, with emphasis on the word “might”. For it is not difficult to imagine that under conditions of high decision-stakes, disputed facts, and great uncertainty surrounding the science base, some affected parties *might* prefer opacity as a virtue of design in their model, for the purpose of the politics of rule negotiation (a topic reserved for discussion below in Section 5). Of perhaps greatest interest herein, where, if anything, the task of the model is primarily to serve as an

⁵ “Prediction” is understood here in rather broad terms, such as, principally, the exploration of future consequences, statements about future behavior in the system, or the projection of current understanding into the future. Specifically, the word is not used in the sense of “the future state of nature will be this at some specified point in time and space”.

instrument of prediction, might be to err on the side of Relevance (as opposed to Fidelity or Transparency). But even in this, we can readily appreciate that so to err would be to have adopted, in effect, the stance of the decision-owner, who may nevertheless want high Fidelity and Transparency, not least as a basis for negotiation with the other groups of stakeholders.

4.2 Model Evaluation: The Role of UASA

Let us now look in somewhat greater detail at each of the vertices of Figure 1, thus to tease out a past and possible future role for applying the methods of UASA in evaluating the model, *prior* to its actual deployment in screening the provisional policy alternatives. As an insight on where research in the future might be directed, we can observe from the outset that much is to be said on Fidelity, little on Relevance, and less still on Transparency.

Beginning then with Fidelity, there are two facets to this notion: (i) peer-group approval (or otherwise) of the theoretical underpinnings of the bits of the science mobilized expressly in the model, which we shall label Fidelity (Theory) for short; and (ii) evidence of the extent to which the model and its (parameterized) constituent hypotheses have been reconciled with observations of past behavior, i.e., Fidelity (Observation).

Fidelity (Theory)

In their book *Uncertainty and Quality in Science for Policy*, Funtowicz and Ravetz (1990) introduce the idea of a *pedigree* of a domain of science, a word expressing something about the history — and the quality of the provenance — of the concepts and theories behind the model and, possibly more appropriately, each of its constituent parts. Table 1 reproduces their research-pedigree matrix. Over the years, the materials to be employed in constructing the model have been consolidated and refined, to produce — at maturity — a product with a fine pedigree. To draw upon the words of Table 1, an embryonic field of study, such as modeling of lake eutrophication, passed thus through the adolescence of competing schools of thought (Vollenweider, 1968), to the gathering of consensus around a single scientific outlook (disputed only by the sub-discipline’s “rebels”), thence to the adulthood of the fully consolidated outlook, contested, if at all, only by those considered “cranks” by the overwhelming majority (a history recounted, up to a point, in Schertzer and Lam (2002)). The status of the model’s pedigree, we note, should change over time, with the strong implication of *ever* improving quality. In a qualitative manner, progression up through the various gradations of the “colleague consensus” column in Table 1 — and down, in the event of confounding/surprising countervailing evidence — mimics the customary view on how epistemic uncertainty is reduced (as already discussed in Section 2.1).

These levels of approval in Table 1, however, are being offered exclusively from the perspective of the *model-builder* alone. The entire matrix, as Funtowicz and Ravetz (1990) fully intended at the time, is confined to conventional matters of the scientific peer review of laboratory-anchored science, which will guide the choices the model-builder makes regarding the bits of that science base

to be incorporated into a computational model. The other stakeholders, the decision-owners and the affected parties, ought arguably *not* to be concerned so much about approving of the materials employed in the construction of the model — if we accept Jasanoff’s recommendations on extended peer review — but instead approving of the scientists and model-builders chosen to conduct the evaluation by conventional peer review. Yet how exactly the gradations of Table 1 translate into computational accounts of epistemic uncertainty, for instance, in applying subsequently the methods of UASA in screening the policy options, remains somewhat unclear (although we note that some preliminary steps in this direction have recently been taken; Refsgaard *et al*, 2006).

Let us now turn to the second facet of Fidelity in the design space of Figure 1, namely that of judging the extent to which the model matches the observed record of history.

Fidelity (Observation)

In the entirely hypothetical context of a model populated only by the three universal parameters of Nature (our introductory discussion of Section 2), the significance of needing to be convinced by any such match of the model with the real thing might be vanishingly small. Peer approval of the *internal* constituent materials and parts of which the model has been constructed would be at its maximum; the pedigree of the model might be deemed impeccable. Successful juxtaposition of the empirical, measured record with the model’s *external* responses, i.e., the outward, observable nature of its behavior in specific settings, would be tantamount to gilding the lily. At most, the content of any residual errors of mismatch between the model and the prototype could be consigned to the consequences of purely random chance — the consequences of imperfect observation at most, vagaries incapable of any more deterministic interpretation (and essentially free of epistemic uncertainty). This state, where the (internal) pedigree of the model is so impeccable as to render assessment of its (external) match with history almost redundant, is not going to obtain. Its expression marks an unattainable benchmark, against which all lesser levels of performance can be graded — along a metric now to be outlined.

The American pragmatic philosopher Lewis held that knowledge accumulated through “acts”, whereby prior “concepts (theory)” could be seen to have been reconciled with the observed “given data” (MacFarlane, 1990). These “acts”, which Lewis argued were essential to progress in science (to reiterate our introductory discussion of Section 2.2), amount to what is called system identification in control theory, in effect, calibration and “validation” (and more) when placed here in this more detailed discussion of evaluating models (Beck, 1987). A common, less high-minded view of model calibration holds that the more parameters there are to be “tuned”, the more degrees of freedom the analyst has to fit the curve to the data, and the less will be the quality of the exercise in reconciling theory with observation. This freedom can be curtailed, either by restricting the number of model parameters to be adjusted in the process of calibration, or by bounding the permissible ranges of values the parameters may assume. Both prior choices are a function of the pedigrees of the model’s various parts (here “prior” in the sense of before calibrating the model in the context of the current task specification). They are guided by the model-builders’ judgements on the foregoing peer-reviewed literature, in which values for the parameters have been quoted. In

the event of obtaining an acceptable matching of the model's outputs with the data, the status of this "acceptability" would be elevated by the return of calibrated (posterior) values for the parameters well away from their (prior) upper and lower bounds of acceptability, as judged, that is, by the peer group of colleagues with previous experience of working with models embodying these parameters. Who — rebel or crank — would risk choosing a parameter value at the edges of the tolerance of that collective wisdom, or even beyond, just to get that somewhat better fit? Acceptability and quality would tend to be diminished by a match of the model with the observations achieved through the assignment of absurd values to some of the model's parameters.

Many exercises in calibrating environmental models to observed history have been conducted over the years. The record of this experience, with very few exceptions, can be stated thus. For anything but the simplest of models, which is not the majority of models, there will be an inescapable lack of model (and parameter) identifiability: many combinations of parameter values may enable the model to fit the data more or less equally well (Beck, 1987, Beven, 1996). So which combination should one eventually use for extrapolation of behavior into the future? It is safe to conclude this lack of identifiability is inescapable in calibrating today's environmental models. To achieve the single line of the model's response passing through the dots of the observations is something; it is better than not doing so. In the absence of any appreciation of a lack of identifiability of the model, the community of stakeholders might move on to the generation of predictions and the making of decisions. We can see now how such ignorance would be risky indeed. On the outward surface the curve can be fitted — ostensibly there *is* resemblance to behavior of the real thing — but at the cost of a host of ambiguities, absurdities, and distortions in the internal workings of the model. In this there is a further echo of our introductory comments on abandoning pursuit of the singular truth of the matter (in Section 2.2), now with growing pragmatic consequences.

Based on these kinds of arguments, and employing now (in its proper context) the narrower understanding of "validation" in the common vocabulary of model builders, a metric of the quality of the resemblance of the behavior of the model to the observed behavior of the real thing can be expressed as shown in Table 2 — to accompany the grades of pedigree in Table 1. Put cryptically, the strength of the evaluation of the constructed model will weaken progressively as one descends from validation, through calibration, to model-model inter-comparisons. This decline is simply a function of the quantity and quality of the empirical observations, whose essential purpose is to provide a source of "experience" of the behavior of the environmental system maximally independent of the "experience" deriving from theory (recalling Lewis's perspective on the growth of knowledge). In more refined terms, understanding the details of the composition of Table 2 requires discussion of matters of model uncertainty, to which we now turn.

Because so much attention has been focused historically on identifying singular — uniquely best — values to be assigned to a model's parameters, the concept of calibration as a means of transforming a prior image of model-parameter uncertainty into a posterior image has not been prominent. In others words, in a Bayesian sense, reconciling the constructed model with its prior uncertainty against any one set of uncertain data will, in principle, allow computation of the posterior uncertainty attaching to the model. Like the curate's egg, the model may be revealed to be: good, i.e., less uncertain, in parts; bad, more uncertain, in others. Again, in a Bayesian sense, were

a second set of data available, a second act of reconciling model and data would facilitate further updating of the image/enumeration of model-parameter uncertainty. The traditional interpretations of calibration and validation, with their uniquely best, singular sets of parameter estimates and single curves passing through data, yield no such insights about the quality and reliability of the model's internal parts. Any flaws, distortions, and ambiguities of calibration remain latent. But by the same token, the act of taking *exactly* the same single set of model parameter estimates from calibration and applying it unchanged in a test of the model's capacity to match a second set of data, i.e., "validation" in Table 2, can be seen to be vital. The consequences of any latent flaws in calibrating the model ought to be maximally apparent, within the limitations of such a test.

Armed with this broader interpretation of calibration, we need to unpack the meaning of "model-parameter uncertainty" in order to proceed to a more complete appreciation of the detail of Table 2. For there can be uncertainty not only about the values of the parameters in a given *structure* of the model, i.e., particular arrangement of the interactions amongst its input, state, and output variables, but also in that structure itself. So there is an important distinction to be made between model structural error/uncertainty and parameter uncertainty. Conceptually, the two are distinct. Computationally, they are correlated, in the sense that estimating values of the parameters, i.e., calibration, of an incorrect model structure will lead typically to seemingly absurd parameter estimates, for example. Structural error/uncertainty, however, is (a) hard to detect at the stage of reconciling the constructed model with the empirical data, (b) yet harder still to rectify and resolve, and (c) hard enough to enumerate, as a *distinctly* separate source of uncertainty from that of parameter uncertainty. In fact, identifying and accounting for the consequences of model structural error/uncertainty has only recently become the subject of more sustained and systematic research (Beck, 1987, 2005; Beven, 2005; Refsgaard *et al*, 2006).

For these reasons the content of Table 2 is phrased in the terms of model parameter uncertainty alone, under the presumption — predominantly the case in practice — of just a single, candidate structure for the constructed model. That the residual errors of mismatch between the simulated outputs of the model and the observed data should be seen to have the statistical properties of white noise (Table 2) implies that the two have been reconciled to the fullest extent, with no further explanation for the cause of any mismatch being ascribable to anything but pure, inexplicable chance. In the absence of high-volume, high quality data, however, such may not be readily demonstrable or computable. Elsewhere (in Table 2), the preference for low variances in the estimation (errors) of the parameters signals broadly a consistency between the function of the parameter within the model and the observed effects of its workings in the empirical evidence. It is also broadly indicative of the absence of a lack of model identifiability, which tends to be revealed, in addition, through higher covariances across the estimation (errors) of pairwise combinations of model parameters. For example, in any model embracing biomass activity, it could well be that high (low) growth rates and high (low) death rates of the organisms generate equally good fits of the model to the data, because it is the value of the difference between the two that is the key to the workings of the model in respect of the behavior encapsulated in the data, not the value of the one rate parameter chosen independently of the other.

To summarize, evaluation of the model up through the grades of model-model inter-comparison,

calibration, and then “validation” is mostly implemented without reference to the uncertainty in the model (without application of the methods of UASA). This might not only be a rather risk-prone strategy, when it comes to evaluating the model in terms of its Relevance (and reliability) as an instrument of prediction (Figure 1), it also permits no recourse to features of model performance, such as the “map” of parameter uncertainty, that are both revealing and highly relevant in evaluating the constructed model (provided these more detailed diagnostics are reasonably easily computable). This is why these features have an elevated status in Table 2 (within the span of calibration, but equally so in “validation”), suggesting greater strength in the attaching evidence of evaluation as a result.

Relevance

Fidelity is essentially retrospective in its outlook. Evidence from the cradle-to-cradle process of evaluation has been accumulating up to this point: from the essentially retrospective orientations of evaluating the conceptual and constructed model, in terms of pedigree from the *past* (Table 1; Fidelity (Theory)), and resemblance to *past* behavior of the real thing (Table 2; Fidelity (Observation)). Our thinking must now be cast forwards. Formulating rules and policy for regulation and stewardship — the essential task — is necessarily futures-oriented.

So what more is there to be asked? It is this, of course: does the model fulfil its intended purpose, as an “instrument of prediction in support of making a decision or formulating a policy”, as far as we can tell? For in being prepared to undertake exercises in extrapolation, in exploring imagined future possibilities, there must always be an element of the “unknowable” in prospect. This then is the essence of the paradox of prediction (as discussed earlier in Section 2); this then is why focusing on the notion of the model as a tool has become so important. Is the model well, or ill, suited to — well or ill designed for — its task? How might it perform when set to work on its task, of screening the policy options and in performing a regulatory impact assessment? Can we, at bottom, trust it, or not?

Let us invoke an analogy. Suppose the task had been specified as “achieve flight”, the novel design of aircraft conceived, and the prototype constructed. In evaluating now the model relative to the task description, it as though the prototype aircraft is at the end of the runway, with all holding a stake in this endeavor contemplating whether, at the other end of the runway, flight *will* be achieved. Stakeholders will then be pondering the ultimate question of evaluation, of whether to place their trust in the model. But before that stage, we ought to be able to put our model through the equivalent of the wind-tunnel experiments employed in designing the prototype aircraft. The outcomes of these “predictive exercises” — a kind of shadow-boxing, one might say — will tell us something about how well, or ill, suited the model is to serving its intended forward-looking purpose⁶, but not that this purpose is achieved, with some predicted consequences. The wind-tunnel experiments with the

⁶ Such exercises would not be needed for all of the tasks of which we could conceive, such as, for example, employing the model as an “archive”.

prototype aircraft design — indeed a (physical) model of the real thing — tell us not that “flight *is* achieved” but that, when evaluated under the circumstances likely to be encountered by the real thing in the future, eventual flight seems probable/improbable with this particular design of the device. In evaluating the model, the decision before the regulatory agency (as the author of the device) is that of whether to accept, as legitimate, use of the given model for the given task. It is akin to the decision of those holding a stake in the eventual success of the prototype aircraft, of whether to embark, or not, as the craft is readied at the end of the runway prior to its attempt at maiden flight.

To turn the analogy around to suit our needs here, not least as we anticipate test-marketing of the provisional policy (in Section 5): who, amongst the stakeholders, would climb aboard the prototype aircraft at the end of the runway; how did they arrive at the summary judgement prompting them to embark; on the basis of what evidence and through what reasoning; and displaying thereby what attitudes towards risk? We cannot foresee what might be in the minds of stakeholders, including their attitude towards risk (as we look ahead to Section 5), any more than we can speculate on the politics of resolving conflicts amongst legitimate competing interests, except that the stakeholders should agree on the process of conflict-resolution — Jasanoff’s “appropriate methodologies” and “investigative protocols” — before it is entered into (including into the legal process; Section 6). All of which gaming, posturing, and negotiation will result in the particular design of the model eventually chosen to inform — at last — the *final* policy/regulatory decision. It is clearly prudent, therefore, for our notional stakeholder, the regulatory agency, to anticipate something of the debate and disputation — in its decision at this stage — to place its trust in a particular model in order to formulate a *provisional* policy.

Consider, therefore, the stylized format of classical decision analysis (Figure 2). In the present (now), we assume there are choices among $i = 1, 2, \dots, m$ actions, and a state of nature in the future with $j = 1, 2, \dots, n$ possibilities, the features of some or all of which may be conditioned by the particular choice of action, some generic threshold standard (S) to be satisfied by the decision-state (i, j) combination, and the costs and benefits attaching to each such combination (and to whom these accrue). Someone — cancer specialist, field ecologist, community activist — or some group of stakeholders, supported by some scientific and model-derived data, will have had to have thought through the content and categories of the n future states/domains of nature relative to standard S ⁷. They could be as simple as suggested earlier, in discussing the tasks of a predictive exposure assessment: here, in effect, that one domain of future behavior is above standard S , while the other outcome is below, i.e., just two branches ($n = 2$) emanating from the circular nodes in Figure 2. To put this another way, we can look out towards the imagined future from the (square-node) decision point of Figure 2, but such remains in the imagination, no matter how well supported by the model. The decision, however, must be made in the here and now.

Let us recall, once more, the notion of a model as a tool designed to fulfil a (predictive) purpose and introduce the thought of calibrating that tool against such a task specification. Suppose the future

⁷ Crichton might argue that under his political-legal-media complex there would be no risk of n being small on many occasions.

states of nature ($j = 1, 2, \dots, n$) in Figure 2 can be conceived of, and represented as, a set of bounded domains of behavior in the model's state space: in familiar, elementary terms, "above S " and "below S "; or, to emphasize extrapolation into the unknowable, "essentially similar to current and past behavior" and "radically different from what we have observed hitherto". When first discussed above, calibration was referred to as the matter of reconciling one form of "experience" of the system's behavior and functioning, namely theory, with another, maximally independent, form of "experience", i.e., *past* observation. It is but a short, logical extension of this to suggest the possibility of reconciling "experience" from theory against "experience" expressed in terms of imagined aspirations for the *future*. Indeed, this is nothing more than the same logical extension already set out in Section 3.2 above, when discussing the analysis of reachable futures in generating foresight and horizon-scanning.

To reiterate what was said in Section 3.2., but without going into technical details (not least because such forms of analysis are the subject of current research; Beck and Chen, 2000), there are methods — known as Regionalized Sensitivity Analysis (RSA; Hornberger and Spear, 1978; Spear and Hornberger, 1978) — allowing identification of those parameters of the model that are key to discriminating whether or not a domain of behavior is generated by the model, and those that are redundant. Possibilities for judging the quality of the design of the model, as a tool for fulfilling a specified predictive task, are thereby revealed. Is a good design of the model one with minimally few redundant parameters? Further, conjecturing that a good design of model will be one in which key parameters are known with little uncertainty, is a poor design that in which the key parameters are highly uncertain? Would such a model, found to be poorly designed when calibrated thus against its intended purpose, be considered of low Relevance and reliability (in Figure 1), all its Fidelity notwithstanding, of impeccable pedigree and great resemblance to the (past of the) real thing?

These are open questions, which we do not intend to answer herein: in part for technical reasons; in part because judging the quality of the design of any object, not to mention attitudes towards risk, is in the eye of the beholder. The point, however, is that we can discern a means of acquiring evidence of evaluation of a rather different kind from that obtainable from the predictive exercises of worst-case determinism and pragmatic, uncertain, ensemble propagation. Unlike those foregoing exercises, assessment of performance is here reflected back from information cast in the *external* behavior space of the future states of nature, to information expressed in terms of the *internal* workings of the model (its constituent hypotheses; the materials of its construction). In principle, we can be told something — from the wind-tunnel experiments of our analogy — of what it may take, regarding the design and construction of the model aircraft (its internal struts and structural members), to achieve eventual flight.

Transparency

[All the thinking here remains to be developed, but the bottom line of this lengthy discussion should be something along the following lines]

In particular, in the context of evaluating models as part of the regulatory, decision-making process,

we might even consider phrasing a summary judgement on the legitimacy of the model thus: we accept Model A primarily because it is well designed in respect of serving its intended purpose (Task Z), looking out to the future; secondarily, we make this choice because the behavior of Model A also bears some resemblance to the past observed behavior of the real thing. All analogies have their limitations, however. In judging the quality of the design of a model as a tool, new territory is being encountered; for a tool does not have to resemble the real thing, whereas it is desirable for a model so to do — and the more so, so very much the better.

4.3 Handling Uncertainty in the Screening Process/Regulatory Impact Assessment

The decision — to trust a particular model as the basis for fashioning a provisional environmental policy — has thus been taken. In terms of our analogy, the regulatory agency, as stakeholder, has committed itself to boarding the prototype aircraft, readied at the end of the runway for its attempt at maiden flight. Supposing a sequence of test flights now to be embarked upon — the subject of the present sub-section — the regulatory agency must look forward, to the public’s reception of the provisional policy, the test marketing of its pilot product. But in implementing these tests, the agency cannot divorce itself from the foregoing discussion of evaluation and its pivotal decision of trust a particular model. It may indeed be obliged to revisit them, as a function of the outcome of the current screening exercises, as we shall see.

We begin by examining the consequences of screening exercises conducted *without* formal considerations of uncertainty. If the natural environment is presumed to behave in as risk-prone a manner as possible, “conservative” values can be assigned to the model’s parameters, possibly towards the bounds of current collective wisdom. Thereafter, pure determinism may prevail in the analysis of success/failure relative to S . Alternatively, making no such detailed presumptions about fragility or vulnerability in the behavior of the environment, standard S may simply be tightened by a safety factor/margin. In either case, success or failure can occur with but probability one or zero, for any given policy (i). Only the one branch will emanate from either of the circular nodes for the future state of nature in Figure 2. Such predictive exercises will be composed of *singular* statements alone about future behavior: that it falls either above or below S . We would accordingly be falling back, in effect, on the custom of “sound science”— with soundness vested predominantly the pedigree of the model, Fidelity (Theory) — sufficing to carry us through into trustworthy model-based statements about the unknowable elements of the future. If the stakes held in the decision by the scientifically lay members of the community are high, the design of the model might need to err well towards the vertex of Transparency in Figure 1. All its scientific Fidelity notwithstanding, they might wonder, was assignment of the model’s parameter values “conservative” enough; do the builders and intellectual owners of the model really know what they are doing; can we trust *their* “current collective wisdom”?

Instead of eschewing any explicit account of uncertainty, by making it implicit in a *negotiated* worst-case determinism, it may of course be brought to center-stage. Success or failure will now be a matter of probabilities, lying anywhere between 0 and 1. The logic runs as follows (Beven, 19xx; Beck, 2002, 2005; Beven, 2006). We have assembled uncertain prior hypotheses into what may be

several candidate model structures, acknowledging epistemic uncertainty thereby, and necessarily parameterized each with its own particular set of mathematical expressions for the interactions amongst the input, state, and output variables — and embracing quantities other than just the three universal parameters of Nature. We have uncertain observations of the real thing, against which to assess the hypothetical content of the model. Whatever myriad combinations of parameter values achieve acceptable matches of the models’ behaviors with observed history, so be it. For these will be employed here to generate ensembles of the uncertain projected consequences of the alternative policy options. Calibration and “validation” of a candidate model will have served to elaborate its respective parametric uncertainties summarizing the distortions wrought in the process of calibration, whose consequences should now be accounted for in any predictions obtained from the model. We shall thereby have taken out the prudent insurance policy of covering our predictive screening and impact-assessment exercises in uncertainty. At least we shall have a faithful accounting for any inevitable lack of resemblance of the model to the (past of the) real thing. And if the uncertainty renders impotent our capacity to discriminate one amongst the several policy options (*i*) as preferred or acceptable, both the task specification and model could be returned to an earlier stage of evaluation, for re-development and adaptation.

[The intention here is to provide a platform on which Andrea can develop a much more extensive discussion of this topic.]

5 DISPUTATION AND NEGOTIATION

Emerging from such a lengthy discussion of how the regulatory agency might employ models and the methods of UASA in developing a provisional environmental policy, not least a discussion intended to provide us with the “ground rules for the terms of the debate and disputation amongst the various stakeholders”, we must ask whether this might prove to be so. Do the key elements of Section 4, i.e., Figures 1 and 2 and Tables 1 and 2, constitute a useful framework within which to conceive of resolving the issues of the present and next section (6) — issues harking back to Jasanoff’s concern for all the stakeholders to agree ahead of time on the “appropriate methodologies and investigative protocols” for the extended peer review of regulatory science (and the models embedded therein)?

In fact, we stand now on the *terra infirma* promised some time ago, when introducing Section 4. And our subsequent discussion will accordingly be unusually brief.

5.1 From Singularity to Plurality of Perspective

To begin with, as a working hypothesis, we offer a second version of the pragmatist position on handling uncertainty in models for regulatory/policy science, i.e., that: (i) the model is trusted by the affected parties/stakeholders (first and foremost), which has to do with the Transparency vertex of Figure 1; (ii) it fulfils its designated task, of leading to better stewardship of environments (hardly anything other than crucially important), which attaches to the Relevance of the model’s design; and

(iii), as a tertiary consideration, the model has a fine provenance/pedigree and matches history closely, all of which is a matter of the Fidelity in its design. That, however, would seem to have the pragmatist flying directly in the face of using “sound science”, above all else, in the service of policy formulation. S/he would also be balking at adopting the tried and tested “modern model” of science for policy (as discussed by Funtowicz and Strand, 2006).

But we can detect just such “balking”, oddly enough, in the search for a label for all the voluminous discussion of the preceding section, i.e., “evaluation”, following Oreskes (1998). And that is precisely because earlier terms used for describing this process of evaluating model performance have provoked rather vigorous debate, within which the word “validation” was first to be replaced by “history matching” (Konikow and Bredehoeft, 1992), to which was then added the phrase “quality assurance” (Beck *et al*, 1997; Beck and Chen, 2000), but from which debate “evaluation” has emerged, as we say, as the most appropriate descriptor. The difficulty in finding a label for the process is this. Validation and assurance prejudice expectations of the outcome of the procedure towards only the positive — the model *is* valid or its quality *is* assured — whereas evaluation is neutral in what might be expected of the outcome. Because models of environmental systems have become so widespread in serving purposes affecting a substantially more aware and engaged audience of (scientifically) lay stakeholders, words used within the scientific enterprise can have meanings that are misleading in contexts outside the confines of the laboratory world. The public knows well that supposedly authoritative scientists can have diametrically opposed views on the benefits of proposed measures to protect the environment. When there is great uncertainty surrounding the science base of an issue, groups of stakeholders within society can take this as a license to assert utter confidence in their respective versions of the science, each of which contradicts those of the other groups. Great uncertainty can lead paradoxically to a situation of “contradictory certainties” (Thompson *et al*, 1986), or at least to a plurality of legitimate perspectives on the given issue, with each such perspective buttressed by a model proclaimed as *valid*. Those developing the models have found this disquieting (Bredehoeft and Konikow, 1993), for scientists conventionally recoil from the idea of aiding and abetting the survival (even prosperity) of a plurality of truths. It matters greatly how Science and Society communicate with each other (Nowotny *et al*, 2001); hence, in part, shunning of the word “validation”.

It is this “plurality of legitimate perspectives on the given issue” that is here vital. At the foregoing stage of formulating the provisional policy, our caricature of the regulatory agency was that it would presume the choice of just a *single* model on which to base a policy and evaluate its potential consequences. Further, it would presume there is a consensus on what *the* single model is and but *one* view of the man-environment relationship (*one* world view), hence a broadly *single* view on what the future might be like with/without the policy. Here, in contrast, the focus of the discussion of this section must be on how model evaluation, the use of models, and application of the methods of UASA might change — if they do — when there is a *plurality* of alternative, competing, possibly contradictory claims to know (with certainty) the “truth of the matter”, i.e., alternative models, radically different outlooks (world views) on the man-environment relationship, and therefore very different visions of the future with/without the provisional policy, in respect of both what is desired and what feared.

Amongst those legitimately holding a stake in the outcome of the process of evaluation, not everyone will share the same formulation of the policy problem as that of Figure 2. Nor, given widely differing attitudes towards risk, will all come to the same conclusion/judgement, even under an identical formulation, especially when one thinks in terms of the metaphor of whether or not to board the prototype aircraft about to attempt its maiden flight. For example, only when the prototype aircraft is predicted to fail with certainty under a worst-case determinism, might those dare-devil thrill-seekers amongst the stakeholders be persuaded to come on board. Stakeholders, we know, can be risk takers, risk averse, and risk managing, to name but three classes of perspective (Thompson, 1989). Attitudes towards risk, furthermore, will be modulated according to judgements about where the current task specification lies along the routine-exceptional continuum.

But let us turn back to look at how uncertainty, in its widest possible interpretation, including as a set of contradictory certainties can affect the construction and pattern of Figure 2. All decision-making is conducted under uncertainty, since it must necessarily involve information attaching to the future; the degrees to which such uncertainty is formally acknowledged, however, can differ dramatically. As an unattainable platform from which to depart, consider that of (worst-case) determinism, under which but the one branch would emanate from the circular nodes representing the outcomes of a future state of nature accompanying each course of policy action. When the outcomes are plural (n), but it is known what the nature of each can be, and each can occur with a known probability, Figure 2 is as it is, its customary form for the analysis of risk. Under this more complete account of uncertainty, relative to that of worst- or best-case determinism, the additional information that can be brought to bear on making a decision can be significant. For instance, a course of action (i) leading to a higher expected level of satisfaction of standard S relative to its counterpart deterministic policy (i), but with an accompanying non-zero probability of massively failing to satisfy S , could alter entirely a stakeholder's support for such a course of action.

Acknowledging now progressively more uncertain problem settings, the full range of the n outcomes in Figure 2 may be known, but not the probability of occurrence of each. Indeed, going beyond this, neither these probabilities nor the span of outcomes is known, i.e., the number of branches (n) for the future state of nature is not known (Kreyer von Krauss *et al*, 2006). Thus, in a sense, things can seem to have come full circle. Under conditions of such gross, radical uncertainty, each solidarity amongst the community of stakeholders could construct its own version of Figure 2, asserting possession of knowledge sufficient to restrict n to 1 alone, and insisting that such singularity is *the* correct understanding of the problem, which in turn doubtless contradicts the positions of the other social solidarities — the paradoxical situation of uncertainty sufficient to encourage the emergence of a plurality of contradictory certainties buttressed by a plurality of models. Thus, to each version of Figure 2 would attach the considerations of Figure 1, the design space in which variously the several models would fall, each with their own locations in the metrics of Fidelity (Theory) (Table 1) and Fidelity (Observation) (Table 2), as well as Relevance and Transparency. Rule negotiation would somehow have to accommodate all of this, in coming to a summary judgement, not only on the trustworthiness of each model, *ergo* position, but on how models (plural) evaluation endorses selection and promulgation of a (singular) particular course of action.

5.2 Essential Questions for Further Consideration?

It is unclear at present whether imagining this situation of conflict and dispute over the proposed policy raises any novel questions on how the uncertainty of the situation is formally accommodated and accounted for within models and UASA. Presumably, what we are searching for here is an appreciation of how models and UASA can be deployed in this disputatious situation — which has yet to be referred to a court of law — to identify policies that are potentially robust in the face of this plurality of stakeholder positions. Here we might interpret “robustness” as “Doing well by the environment, while allowing a healthy plurality of plausible future visions (from within/across the community of stakeholders) and while promoting learning (about the environment), and at the earliest juncture possible”. To what extent will the discussion hinge not merely on whether the models being used by each party in the dispute are to be trusted, but more on whether those parties are imagining/supposing broadly “tenable/untenable” visions of desired/feared target futures?

The essential questions to be answered in this Section might be as follows:

- Is model evaluation, as summarized in the foregoing Section 4, an issue relevant to the changed context of this section?
- If so, how does “plurality” change, if at all, the essential questions and procedure of model evaluation and application of the methods of UASA, as set out above in Section 4?
- Can we/should we seek to go forward with formulating and implementing policy under a plurality of constructions of the “truth of the matter”? For after all, a robust policy does not require a single version of the truth?

Building upon this last point — thus to summarize how this section differs in its posture from that of Sections 4 and 6 — we have:

- A {*plurality* of models coupled with a *plurality* of world views and, therefore, a *plurality* of visions of the future conditioned by a *plurality* of stakeholder solidarities}; and, if anything
- A focus of attention on finding *robust* policies, where robustness implies that none of the stakeholders perceive themselves as “losers” in the disputation and negotiation phase of the life-cycle — and not least because they perceive they could become “losers”, if the dispute were to be referred to a court of law (see next section).

6 LEGAL DISCOURSE

If Section 5 is a relevant stage in the life-cycle, the next critical question is that of whether, when the informal dispute (of Section 5) is referred to a court of law, this has any significant implications

for the manner in which models and UASA — in underpinning the formulation of policy — must be assessed and applied.

Does referral of the dispute to a court of law imply that there can only be one “winner” amongst the contesting parties, i.e., we must insist on there being but a *single* truth of the matter? If so, then presumably we shall need procedures of model evaluation and UASA capable of ranking — as more plausible/tenable/legitimate — *one* particular model, *one* world view(???), and *one* set of visions/aspirations (hopes/fears) for the future? Is this likely to draw, therefore, upon the notions of using models, under gross uncertainty, to analyze the reachability of target futures, as set out in the Beck and Chen (2000), thus to rank the reachabilities of those respective futures and, moreover, to rank the “fitnesses”, or levels of Relevance in Figure 1, of the various (plural) models in fulfilling the given policy-focused task???

In short, relative to what has been said of the posture of Sections 4 and 5 above, the posture of the present Section would appear to be:

- Imposing the *singularity* of just the one trusted, acceptable, fit-for-purpose model upon the *plurality* of possibilities present in Section 5, with an attaching *single* vision of the future, reflecting the predominance of a *single* group of stakeholders, namely the winning group (in the legal discourse)?

7 LEARNING IN A POST-RULE WORLD

The essential question here is: How is an environmental policy to be designed and implemented in a manner designed to promote learning — learning, that is, about the nature of the physical environment to which the policy relates, the nature of the man-environment interaction, and the manner in which the various social solidarities within the community of concerned stakeholders undergo learning about themselves and their community?

And then, of course, the key — if this is a relevant question — is what are the roles of models and UASA in answering such a question? A subsidiary question is whether models and UASA can have any role to play in promoting “smart legislation”, if such a concept has any merit?

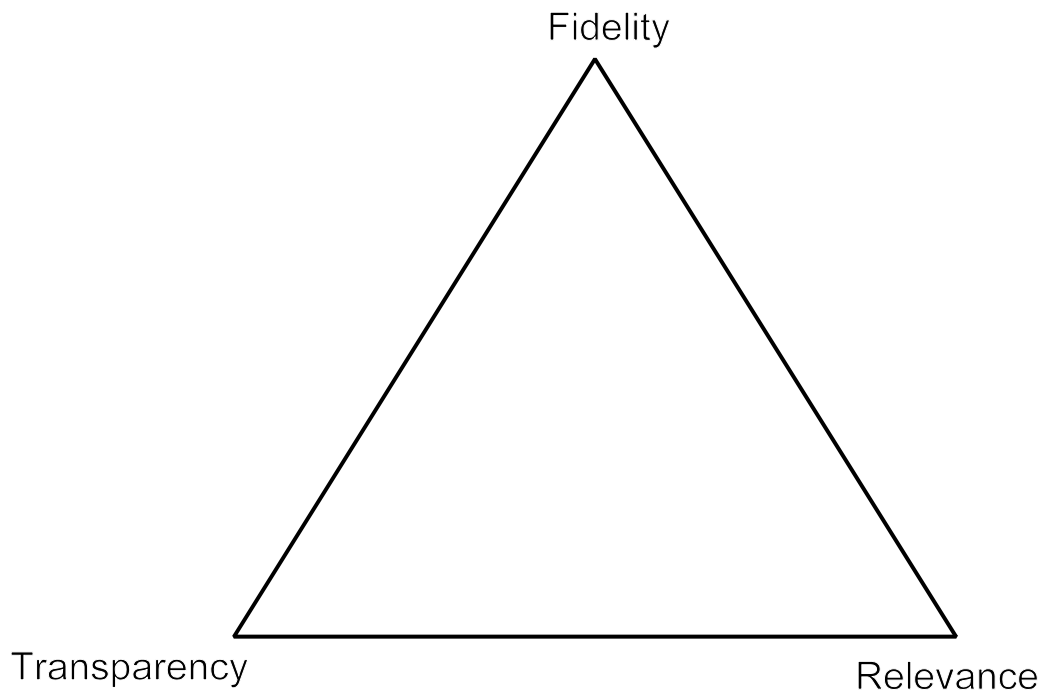


Figure 1. Design space of the model

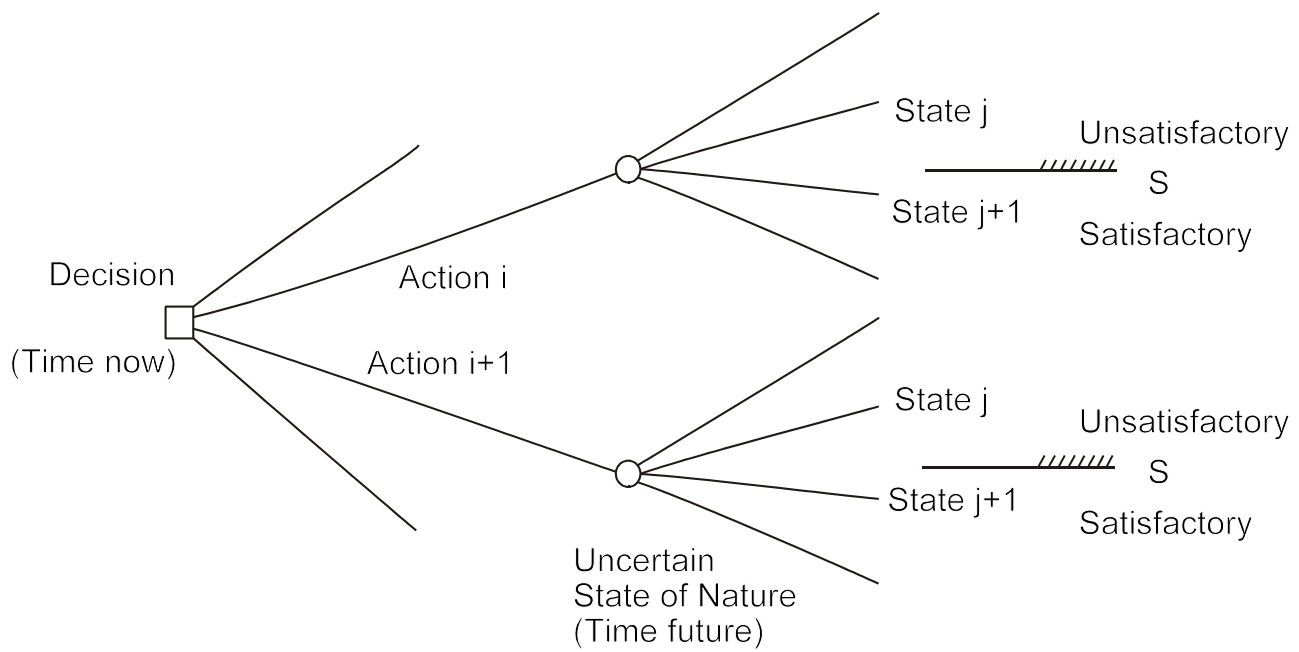


Figure 2. Stylized format of classical decision analysis in a regulatory context.

Theoretical Structures	Data-input	Peer-acceptance	Colleague consensus
Established theory	Experimental data	Total	All but cranks
Theoretically-based model	Historic/field data	High	All but rebels
Computational model	Calculated data	Medium	Competing schools
Statistical processing	Educated guesses	Low	Embryonic field
Definitions	Uneducated guesses	None	No opinion

Table 1. The research-pedigree matrix of Funtowicz and Ravetz (1990)

“Validation”	<p>(+) More than two independent sets of field data similarly so matched</p> <ul style="list-style-type: none"> ● Second set of independent data well matched, without any changes in the values assigned to the model’s parameters during calibration <p>(-) Match to second set of data bought at the expense of re-adjusting model’s parameter values</p>
Calibration (verification)	<ul style="list-style-type: none"> ● Low covariances (correlated errors) amongst estimated model parameter values ● Low variances (estimation errors) for model parameter values ● Minimal number of model parameter estimates at bounds of (peer-group) acceptability ● Residual errors of mismatch between model and observed outputs approximate zero-mean, low variance, white-noise sequences
Model-model inter-comparision	

Table 2. Tiers of the metric for gauging the extent to which the model’s behavior matches observed behavior, with progressively increasing strength from the bottom of table to top.