Evaluating the Predictive and Explanatory Value of Atmospheric Numerical Models: Between Relativism and Objectivism

D.G. Steyn[*,1] and S. Galmarini[2]

[1]*Department of Earth and Ocean Sciences, University of British Columbia, Vancouver, B.C., Canada, V6T 1Z4*

[2]*European Commission-DG Joint Research Centre, Institute for Environment and Sustainability, Ispra, 21020, Italy*

Abstract: We discuss the ways in which atmospheric numerical models can be shown to have both predictive and explanatory value. This argument rests largely on the established predictive value of numerical weather forecasts, and can be extended to atmospheric models at larger- and smaller scales. We demonstrate that atmospheric modellers have not been particularly critical about the logical basis of model evaluation, and recommend strategic approaches to remedy this. Our suggestions involve the creation of a spatio-temporal scale-dependent and context-relative ordinal scale for model evaluation that must be applied in an *a-priori*, context-dependent fashion.

INTRODUCTION

The term model, as used in science, means an *abstract, analogue* representation of the prototype whose behaviour is being studied. In specific instances, the type of analogue, and level of abstraction can vary widely. Three types of analogue commonly used are: *analytical models* (in which variables in analytically tractable mathematical equations are taken as analogous to measurable properties of the world); *physical scale models* (in which the physical behaviour of measurable properties of the small scale model are taken as analogous to corresponding environmental properties at full scale) and *numerical models* (in which variables in a numerically solved system of equations are taken as analogous to measurable environmental quantities). Regardless of level of abstraction, or type of analogue, the ultimate objective of all modelling studies is to provide insight into the workings of a given system. In specific cases, the sought-after insight can be the attainment of a deeper understanding of processes governing the behaviour of a phenomenon (an *explanatory* use of the model) *or* to provide a prognosis of the evolution in time and structure in space of the phenomenon (a *predictive* use of the model). In what follows, we concentrate solely on numerical models, but believe that in principle, the ideas we explore are applicable to the evaluation of analytical and scale modeling as well. In particular, we examine approaches to evaluating what sort of insight can be attained from atmospheric numerical models. We do this because of the long history, and extensive use of numerical models of atmospheric phenomena. The depth and breadth of atmospheric numerical modelling work in both research and operational realms has resulted in a rich literature and wide ranging practises that inform questions concerning the utility of such models. While we concentrate on numerical models of the atmosphere, our conclusions will not be specific to models of atmospheric phenomena, but rather will be transferable to models of environmental phenomena and processes in general.

Associated with both explanatory and predictive uses of numerical models of the atmosphere is an unstated assumption that the behaviour of the model can be treated as if it at least closely matches, in some essential way, the behaviour of the real atmosphere. Since models are abstract analogues, rather than the real atmosphere, before model results can be interpreted as if they fairly represent real atmospheric behaviour, some test must be applied to both the model itself, and to its performance in simulating specific cases. Almost inevitably, such tests involve a comparison between model output and observations of the real atmosphere under specified conditions. These tests are often called model-evaluation, -verification or -validation [1], and are exercises designed to establish the extent to which the model captures behaviour of the modelled phenomenon. It is important to note that these evaluation exercises are also directed towards selected runs of the model. In order to avoid becoming ensnared in the semantic debates that often accompany this matter [3,4], we use the term "model evaluation"[1], to signify a process (whose details we do not specify here) in which model output is compared with atmospheric data so as to test the ability of a model to predict and explain atmospheric phenomena. Without such tests, the models can be no more than interesting intellectual tools.

The oft-cited work of Oreskes *et al.* [2] (hereafter called OSFB94) argues that "verification or validation of numerical models of natural systems is impossible", with the consequence that " Model(s) ….. predictive value is always open to question. The primary value of models is heuristic"[1]. The objectives of this paper are to examine model evaluation practices (but not techniques) as they exist in the field of atmospheric numerical modeling, in light of the OSFB94 views on numerical models of natural systems. Few atmospheric scientists will need convincing that numerical models of the atmosphere do have predictive value, both in terms of providing useful prognoses of the future state of the atmos-

*Address correspondence to this author at the Department of Earth and Ocean Sciences, University of British Columbia, Vancouver, B.C., Canada, V6T 1Z4; E-mail: dsteyn@eos.ubc.ca

[1] As used by OSFB, "heuristic" means "useful for guiding future study".

1874-2823/08

phere, and in terms of their ability to provide a process level understanding of atmospheric phenomena. Atmospheric scientists use their models in ways that extend well beyond their being mere guides for future study. In spite of the assertions of OSFB94, a vast majority of citations to their work are made by modelers engaged in mode-evaluation exercises designed to show that their models do have predictive value. We engage this apparent contradiction by examining all citations of OSFB94 in works that involve numerical modeling of natural (including atmospheric) systems. Our reading of works citing OSFB94 is conducted in order to highlight weaknesses in model evaluation practices commonly employed by atmospheric modelers. On the basis of that reading, we assert that OSFB94 seems to underestimate the accepted utility of atmospheric models, and has had a mildly negative influence on approaches to model evaluation. One might have hoped that a work so clearly important as OSFB94 would have caused modellers to exercise particular rigour in evaluating their models, or to more carefully examine their conceptual framework for model evaluation. Alternatively, an even more salutary effect of a paper like OSFB94 might have been an examination of the strategic, logical and philosophical bases of model evaluation. We find no evidence that OSFB94 has had either of these effects on atmospheric modelers. Our analysis of the influence of OSFB94 on numerical modeling in a range of scientific fields leads us to believe there are several extremely important strategic questions surrounding model evaluation in the atmospheric sciences that need urgent consideration. We close this paper by providing a starting point for a consideration of the nature, practice and meaning of evaluation of atmospheric numerical models and their output. Our considerations of these questions lead us to recommend a set of generalized starting points for a reformed approach to model evaluation.

THE EVALUATION AND UTILITY OF ATMOSPHERIC MODELS

Numerical models of natural systems are characterized by OSFB94 as containing verifiable mathematical components, as well as requiring input parameters that are incompletely known. Atmospheric models conform well to this characterization. Indeed, they are based on constitutive equations (which are specific expressions of the conservation of mass, momentum and energy). In addition, atmospheric models at all scales employ parameterizations of processes at smaller, unresolved scales. These parameterizations are not arbitrary, but are representations of small-scale processes and phenomena built upon observations. These models are also implemented in numerical schemes which include many approximations and discretizations. Atmospheric models also require input data, in the form of boundary- and initial conditions (all derived from observations), which are incompletely known in the sense that they contain measurement errors. An added complication comes from the practice of data assimilation in which the model, while running, is forced towards newly available measurements of modeled variables. The output of atmospheric models is subsequently compared to observations in the process we have called "model evaluation". This process is technically[2] demanding because of the multi-dimensional and vector nature of phenomena being modelled; the fundamental incommensurability of model output (being averaged values on a regular grid) and observations (being point values on an irregular grid); the existence of landscape features not captured or incompletely resolved by models, but effecting observations [5-7].

We accept fully the statement of OSFB94 that the combination of a numerical model and the data against which it is to be evaluated constitutes a logically open system, and that any assertions about undiluted truth in such a system are impossible. By contrast with OSFB94, we argue, that atmospheric numerical models can provide valid insight in the sense of being demonstrated to have predictive value. Their conclusion is based on a strong reliance on rigid Popperian falsifiability [4], which has application only in the most limited and narrowly technical fields, and generally not in the natural sciences. We do not mean to imply that atmospheric numerical models have limitless predictive capability, but rather that they can have predictive and explanatory value that must (and can) be established by appropriately exercised model evaluation. Our assertion is thus an intermediate one, between the admitted extremes of outright rejection of the possible utility of atmospheric numerical models and the blind acceptance of atmospheric numerical models as capturing faithfully the behaviour of atmospheric phenomena. This latter approach leads to the (regrettably frequent) practice amongst atmospheric scientists of referring to model output as "data", and the rather more serious (but related) flaw of treating model results as if they were a complete and faithful representation of the real atmosphere, and therefore "true". These two extremes are also characterized by Kleindorfer *et al.* [8], who refer to them as "objectivist" versus "relativist" stances. We take it as self-evident that model evaluation is an essential part of model development and application, and that effective model evaluation is necessary to produce models that have meaningful predictive value and utility.

What we are advocating is evidently a much more difficult, intermediate position, in which it is recognized that imperfect models can still have tangible explanatory and/or predictive value. We believe that resorting to either of the extreme positions suppresses both the advance of atmospheric science and possible social benefits from the application of scientific results. The intermediate position we suggest is not new, and is argued forcibly by Beven [9], among others.

The Predictive Value of Numerical Atmospheric Models

As demonstrated by Tribbia and Anthes [10], numerical models are a central tool in predicting the evolution and structure of the atmosphere. Indeed, the use of computational resources for the preparation of daily weather forecasts by national weather services worldwide, and the running of general circulation models in climate research exceeds resource use for any other form of modeling of natural sys-

[2]While acknowledge the importance of technical and statistical difficulties inherent in the process of discovering or quantifying the utility in numerical models of the atmosphere, we will not tackle those difficulties here.

tems. The atmosphere is, of all natural systems, the most heavily, and frequently subjected to numerical modelling. Weather forecasts based on numerical model output are widely disseminated by radio, television and print media, and over the internet. Millions of people rely on these weather forecasts to make decisions about many aspects of their daily lives, and these decisions cover matters as diverse as: conduct of weather sensitive agricultural activities; choice of clothing to wear; involvement in outdoor sporting activities; scheduling of outdoor entertainment events; conduct of crop spraying programs and making route selection for transportation of persons and goods. Over longer time scales, seasonal weather forecasts are used as bases for insurance policies; investment decisions in the agricultural sector; energy supply decisions and water resource distribution. Many of these decisions involve substantial amounts of money, and often involve risks to human life and well-being. Clearly many people and organizations behave as if numerical models of the atmosphere do have predictive value.

National meteorological services worldwide continue to operate at considerable cost, a large fraction of which is involved with the running of numerical models and interpretation of model output. Clearly the cost is deemed justified, a judgement which is based on the generally accepted value of numerical models upon which weather forecasts are based. This value is demonstrated forcibly by Katz and Murphy [11] and the growing literature on the subject [12] reinforces this view. This seems in direct contrast to the assertions of OSFB94. A more direct argument is that, at a rather crude level, models used in numerical weather prediction (NWP) can capture a major fraction of atmospheric behaviour. For example, NWP models simulate very successfully the progression of mid-latitude cyclones, and based on the structure and evolution of these weather systems, they forecast conditions such as temperature, precipitation and wind with remarkable precision, for up to a few days in the future [10]. It is well recognized that the atmosphere is not limitlessly predictable, due to well-known (but poorly understood) chaotic processes. Thus, efforts to improve prediction involving both continuous data assimilation and ensemble forecasts may push back the limits to predictability, but can never eliminate them completely. Nevertheless, NWP models do have predictive value, which is captured, quantified and monitored by a variety of "skill scores" in which the forecast is compared with actual weather in an after-the-fact evaluation exercise. The continuous evaluation of forecast model performance based on these skill scores constitutes an essential part of NWP model evaluation, and is used in the continual improvement of both NWP models themselves, and the forecast system that relies on their results [10, 13]. This is done in recognition that model evaluation is an essential part of modelling. While acknowledging limits to their accuracy, we believe the assertion that the value of weather forecasts (being specific instances of predictions produced by numerical models of natural systems) is "open to question" is unreasonably critical.

Aside from their use in the operational meteorological realm, numerical models of the atmosphere are used regularly in the analysis of regional to continental scale air pollution, and in the production of forecasts of regional air pollution. Several national air pollution regulatory agencies worldwide provide operational pollution forecasts [14-16]. The analyses are often mandated by regulatory agencies, which are used as a basis for the operation of air quality management plans. The most demanding task to which these models are applied is the determination of the relationship between emissions and ambient air quality [17]. Again, the models accurately capture atmospheric features and phenomena of importance to the dispersion of air pollutants, such as mean wind in the lower atmosphere, boundary layer depth, rate of dispersion of pollutants within the boundary layer, rate of deposition of pollutants to surface receptors and chemical transformation of pollutants. Often, such models and their output are subjected to very rigorous evaluation against data collected during air pollution field measurement campaigns, as well as against data from operational air quality monitoring networks. The purpose of the evaluation is to gauge the level to which the model simulates observed phenomena and thus to determine its utility in air quality management or abatement strategies. Again, there are limits to the accuracy and precision of these models, but they do have applied predictive value in the sense of being able to reliably capture at least part of the essential behaviour of complex natural systems.

At a larger scale, a wide range of climate modeling studies are directed at understanding the climatic consequences of the accumulation of greenhouse gases in the atmosphere and are reported in the various IPCC document. These are examples of the predictive use of numerical models of the atmosphere. Of course this application is made particularly difficult by the complexity of the models, the unavailability of data on the future behaviour of Earth's atmosphere, and uncertainty in data on future anthropogenic emissions of greenhouse gases into the atmosphere. These difficulties are at the core of the debate regarding the predictive value of this particular class of models [18]. In any case, the massive amount of modeling research conducted in this field is considered sufficiently sound to become an important element of policy making (e.g. the Kyoto protocol).

The Explanatory Value of Numerical Models of the Atmosphere

In addition to having predictive value, atmospheric models can have explanatory value, in the sense that dynamic-, or diagnostic studies utilizing model results can be used to further our understanding of atmospheric phenomena. In this type of application, dynamic or thermodynamic processes are studied by extracting (from the numerical model) values of terms in the underlying conservation equations. An analysis of the relative magnitude of the terms is then used to judge what underlying processes, and which combination of processes are responsible for the evolution of the phenomenon being studied. This kind of analysis can be applied at any scale, as done by Reed *et al.* (1988 [19]) (at the synoptic scale), Steyn and Kallos (1992) [20] (at the mesoscale), and Moeng and Wyngaard [21] (at the boundary-layer scale). Before undertaking the dynamic or diagnostic analysis, it is essential that the model be evaluated by comparison with

field data. The purpose of this evaluation is to ensure that essential features of the dynamics, kinematics or thermodynamics are adequately simulated by the model thus establishing its utility as an explanatory tool.

These kinds of studies are clearly explanatory in nature as their objective is an enhanced understanding of the workings of the phenomena being studied. We thus argue that numerical models of the atmosphere have both predictive and explanatory value.

While the foregoing discussion has drawn its material solely from atmospheric numerical modelling, the conclusion that those models can have both predictive and explanatory value is not limited to atmospheric numerical models. By analogy, process-based numerical models of any environmental system or phenomenon could have predictive and explanatory value.

INFLUENCE OF THE ORESKES *ET AL.* PAPER ON ATMOSPHERIC MODELLING

A search for citations of OSFB94 using the *ISI Web of Science* facility reveals more than 599 citations since publication, with citations appearing in papers published in a wide range of journals, including most of the major environmental, atmospheric science, management science, hydrological and ecological journals. The mean citation rate of 46 papers per year appears to continue without decrease. Clearly, scientists employing numerical models of natural systems see the need to take into consideration conceptual criteria surrounding the evaluation of those models, and this is expressed in their frequent citation of OSFB94. Numerical models are widely used in atmospheric science, and substantial effort is put into the evaluation of such models. Because OSFB94 deals with numerical models of natural systems, and is widely quoted, it seems worthwhile to ask how OSFB94 has influenced the way atmospheric numerical models have been evaluated.

An examination of all papers reporting on atmospheric science, and citing OSFB94 reveals a rather disturbing feature among a subset of these works. Papers in this category are: Alapaty *et al.* (1995) [22], Dennis *et al.* (1996) [23], Grassi *et al.* (2002) [24], Hanna *et al.* (1996) [25], Huebert *et al.* (2001) [26], Kambezidis and Psiloglou (1995) [27], Katz (2002) [28], Pinty *et al.* (2001) [29] and Roselle and Schere [30]. In these papers, the impossibility of validating a model serves to inhibit a full model evaluation. In most of these papers, reference to OSFB94 is made in passing, without referring specifically to their ideas. An example of this sort of citation is: "As a recent review critically demonstrates [2], a numerical model *cannot*, by its very essence, be validated or verified. So we will use here the more acceptable term "performance assessment".". Rather more reassuring is that some atmospheric modelers [1, 17, 31-33] have taken seriously the cautions offered by OSFB94, and have incorporated their cautionary ideas into work on atmospheric modeling. It is notable that these works are generally non-technical in the sense that they offer analyses of the practice of model evaluation, rather than explicitly modifying their approach to model evaluation in response to OSFB94. More encouraging is the substantial body of work by non-atmospheric scientists who take seriously the cautions offered by OSFB94, and engage in often far-reaching examination of modeling, model evaluation, and the utility of models in their respective sciences. This is particularly evident in ecology and hydrology. Papers in this category are: [9, 34 -42]. Of particular note is the searching analysis of Rykiel [4], who concludes that models can be accepted as having utility, as a consequence of successfully passing predetermined evaluation criteria, but that the criteria, and therefore acceptance are relative to the context in which models are to be used. This matter will be returned to later in this paper.

A third category of papers referring to OSFB94 is those written by social scientists, philosophers of science and analysts of scientific practice. These range in position from fairly strong relativist analyses which present arguments that numerical models of natural systems have very limited predictive value, to arguments that numerical models of natural systems can have utility if treated appropriately. Papers in this category are: [8, 44 -48]. Most interesting is the work of Konikow and Bredehoeft [43] which, while predating OSFB94, provides very carefully considered arguments that models cannot be validated rigorously, but do have considerable explanatory value. They take a strongly practical approach to the use of models for predictive purposes, arguing that while models can have predictive value, caution must be exercised in using them in this way. Particularly interesting are Kleindorfer *et al.* (1998) [8] who argue that OSFB94 contains elements of an intermediate position, between the two extremes characterized earlier. They argue that both extremes allow modellers to avoid their responsibility for model evaluation, and urge that modellers not be allowed to escape this requirement. A rather smaller group of papers referring to OSFB94 takes strong exception to the idea that numerical models have limited predictive value - clear examples of the realist position. Papers in this category are: Roache (1998) [3], Spear (1997 [49]), Oberkampf and Trucano (2002) [45] and Saltelli and Scott (1997) [50].

From the foregoing, it is clear that, in spite of being frequently cited, OSFB94 has had relatively little positive effect on the practice of numerical model evaluation in the atmospheric sciences. In a few cases, it appears to have inhibited a thorough model evaluation, or to have resulted in a less than critical model evaluation. These outcomes have happened in spite of warnings put forward by OSFB94. We believe this development occurred chiefly because a cursory reading of OSFB94 will give the impression that, since models cannot be validated, all model evaluation exercises must fail, and therefore need not be pursued with great diligence. Sadly, citation of OSFB94 in work that evaluates environmental models of natural systems has become almost obligatory, but of no identifiable positive effect. Atmospheric modellers, because of their frequent use of numerical models often use OSFB94 as a "throw-away" citation. By contrast, many writers in non-atmospheric sciences have been inspired by OSFB94 to engage in careful and often important considerations of the meaning and evaluation of models. We believe that atmospheric modellers should be similarly inspired.

A GENERALIZED APPROACH TO MODEL EVALUATION

We are in full accord with the contention of OSFB94 that numerical models of natural phenomena cannot be validated, in the sense of being proved true in the most rigorous scientific sense. It is also evident that, even if the output of a model corresponds exactly to the behaviour of the phenomenon being modeled, we can never know with certainty which terms correspond to which particular process and why. However, we assert that such models can have predictive- as well as explanatory value and that this value can and must be demonstrated in model evaluation exercises comparing model results and/or output with observations. In this regard, several extremely important strategic questions surrounding atmospheric model evaluation need urgent consideration and are outlined in the following sections.

Establishing Model Utility

The utility (either predictive or explanatory) of a model and its output cannot be established on an all-or-nothing basis. That is to say, the aforementioned properties of atmospheric numerical models as being essentially imperfect automatically imply the existence of a range of utility values. It is possible to establish that a model has some utility, and that the degree of utility can be measured by a continuous (ratio) variable. However, it also seems unlikely that the resolution in model utility afforded by a continuous variable is relevant, and that an interval (or even ordinal) measure of model utility is more appropriate. Atmospheric modellers implicitly recognize the appropriateness of an ordinal scale when they perform model evaluation based on statistical measures of model agreement with data, and then declare the agreement between model and observations to be "good", or "acceptable", or "adequate", or sometimes "poor". If agreement is "unacceptable", or "poor", researchers understand that either further work must be done on the model in order to improve the agreement or the model must be abandoned in favour of a better one. Statements using these value-laden terms are not particularly useful since the terms are too flexible to be meaningful. Having accepted that models can have utility, and that an all-or-nothing acceptance criterion is meaningless, we have no choice but to develop a formally defined, interval (or ordinal) scale of model utility. The purpose of model evaluation exercises should be to determine the position of models and their output on this scale. Numerical models of the Environmnetal phenomena and processes exist for a wide range of scales (micro-, meso-, synoptic- or global-scale), and are used in widely differing applications (from purely scientific investigations to public policy making). Since most processes in the atmosphere are scale-dependent, it would be very surprising if criteria for evaluating atmospheric models were not also scale dependent. It seems inevitable therefore that the utility scale must be relative to the dynamical/physical scale of the model. Both Hogrefe *et al.* (2001) [51] and Beven (2002) [9] provide strong arguments that this should be the case.

Establishing Model Success

It seems obvious that before a model is run for a particular application, the criterion for success (position on the utility scale) should be established as a prior condition. *A-priori* setting of the criterion (or criteria) is necessary since in the absence of the pre-established criterion, model evaluation might proceed in a gradualist way until, out of exhaustion (of funds, human energy or computer resources), the model is declared satisfactory. Once this is done, the model will be treated as valid, an attitude strongly at variance with arguments put forth by OSFB94, and with our ideas. If the criteria for success, on an established interval or ordinal scale are known beforehand, no gradualism can be tolerated. If successful, the model success will always be judged relative to the position within the scale it was required to achieve. If the model cannot meet the specified criteria, it will not be judged invalid, but merely inappropriate for the purposes which resulted in the specified criteria being selected in the first instance.

Different Evaluations for Different Models

While we have viewed models used for explanatory or predictive purposes as quite separate, a more careful consideration leads to the conclusion that they are quite closely linked, as demonstrated by Tribbia and Anthes (1987) [10]. It does however seem reasonable that evaluation requirements for explanatory uses of numerical models will be different than those for predictive models. This should be so since the explanatory use of models requires that the relationships between process parameters in the model match those in observations, thus allowing dynamic or thermodynamic analyses. By contrast, a predictive model will be judged to have utility if the temporal evolution of chosen variables (the forecast variables) corresponds to those that actually occur. That this judgment can only be made after the fact does not alter the nature of model evaluation, which supports it. That a conceptual model may be wrong (and known to be so), while still being sufficiently realistic for certain restricted purposes is argued by Beven (2002) [9]. It is thus entirely possible that an atmospheric model may be demonstrably unacceptable at a process (explanatory) level, but still produce acceptably precise forecasts of a property such as surface temperature. In a similar way, but by different logic, a model that performs unacceptably in a forecasting context may have utility in a policy making realm. These and related considerations led Funtowicz and Ravetz [52] and Haag and Kaupenjohann [53] to define a *post-normal* science, in which issue-driven problems in the realm of policy-related research force us to use numerical models of natural systems in a context characterized by high uncertainty, disputed values, high stakes and urgent decisions. What we argue therefore is that model evaluation criteria and possibly also model utility scale are most likely to be dependent on the context in which they are to be applied.

Model Evaluation as a Continuous Process

Many of our ideas can be summarized in a simple diagram (Fig. **1**), which depicts either the development of a particular model, or the development of modeling capability for a particular atmospheric phenomenon. The figure contains an extension of the idea (captured in Fig. (**1a**)) that data and theory are inextricably bound in a circular relationship. All theories are based on data, and no data are collected without a theoretical antecedent. In Fig. (**1b**), we depict this circular process as occurring in an evolutionary way, and in order to bring it into the present context, replace "data" from Fig. (**1a**) with "observations", and "theory" with "model". We recognize that, in a formal sense, a model is not a theory, but that the relationship between observation and model approximates that between data and theory. Each arrow depicts a half cycle of either model runs or observational campaigns. In early stages, there is a relatively large variance between observations and model, represented by the horizontal separation between O and M symbols. The horizontal dimension can be thought of as simply one component of a multi component (multi variable) comparison between model and observation. The number of components will in general depend on the richness of the observational data, and the realism of the model. As the process of model evaluation and data collection advances, represented by distance downward in the vertical direction, model-observation variance, ideally, decreases. Of course this is not necessarily always the case, and it may be that subsequent cycles result in no such decrease. It is also possible that additional observations will reveal substantial flaws in a model, thereby increasing the variance between model and observations. For simplicity we do not include these possibilities in the figure, which shows a monotonic decrease in variance between model and observations. What is most important is that on the long term, model and observations do converge, but never coincide completely, in recognition of the existence of measurement error in observations, of incomplete representivity (in both space and time), of observations and of conceptual, discretization, roundoff and parameterization errors in models.

The asymptote towards which model and observations tend is generally assumed to be not greatly distant from "truth", though the distance can, in principle, never be known, as already pointed out. This is a reflection of our assertion that models can have both explanatory and predictive value, but can never be treated as encompassing the entire truth of a system. The five shaded vertical bands are intended to represent the various criteria for model success described above. The multiplicity of such criteria is intended to indicate that model utility is to be judged relative to the context within which the model is being applied. It is then presumed that when the model-observation variance lies within a given band, the model can be judged to have utility for purposes associated with that band, in spite of residual model-observation variance.

The stage of evolution (roughly vertical distance down the figure) needed to achieve utility within each context is indicated by the vertical position at which the dashed lines cross the bands. We will not explicitly label the bands, but suggest that, working from the inner to outer bands, they represent: Explanatory uses within a scientific context; Predictive uses within a scientific context; Predictive uses within an engineering context; Predictive uses within a policy context. It appears that evaluation practices for models used in the policy realm are relatively relaxed, and that relatively simple models are acceptable in this realm because model uncertainty is presumed to be small compared to uncertainty about societal and economic matters.

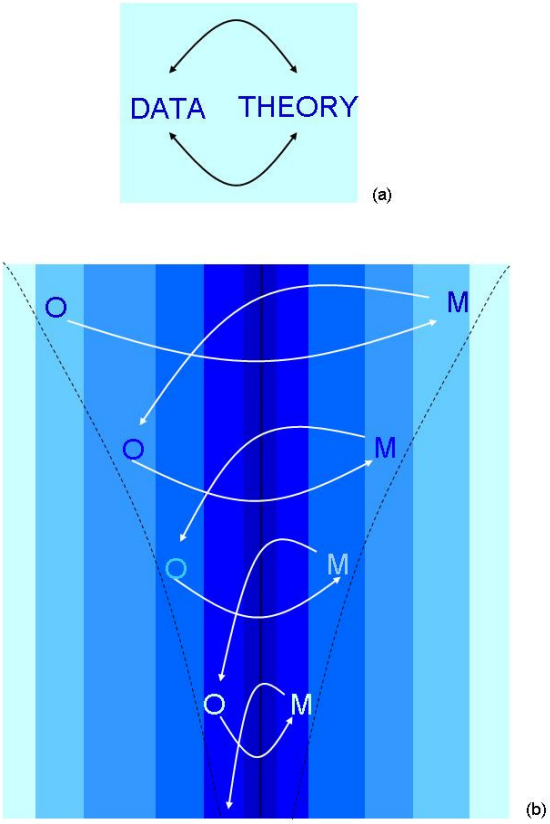

Fig. (1). A diagrammatic representation of (**a**) The circular relationship between data and theory, and (**b**) of the evaluation of model (M) against data (D). The dashed line represents the narrowing envelope of model (M) - observation (O) variance, and the shaded zones represent various criteria for model success. Darker shades indicate more stringent criteria.

What we suggest here is a set of starting points for devising a reformed approach to model evaluation. Many operational, technical and statistical details remain to be established. We believe that an approach such as the one we suggest here will place model evaluation practices in a logically more defensible position, and thereby make models more valuable tools in both scientific and public policy-making realms, as well as result in that value achieving proper recognition.

CONCLUSION

We review the nature of atmospheric models and conclude that such models can be used in both predictive and explanatory modes, and that they can be shown to have utility in the predictive mode. From our consideration of the evaluation of atmospheric numerical models, we conclude

that the perspective on numerical models of natural systems provided by OSFB94 has had no particularly positive influence on the practice of evaluation, and that, in some cases, has had a mildly negative influence.

Based on these considerations we recommend that atmospheric modellers develop a formally defined, semi-quantitative scale of model utility in order to express the level of predictive value inherent in their models. In any model application, we argue that the level of acceptable agreement between model and observations (on the scale of model utility) must be determined as the starting point of a model evaluation exercise. We argue that the scale of model utility will be contextually relative, and as a starting point, suggest there be different evaluation scales for explanatory and predictive purposes, and that model applications in scientific and policy realms also be evaluated on different scales. By extension, we assert that these ideas apply to process-based numerical models of all environmental systems, though acknowledge that few such models are as richly developed as atmospheric numerical models.

ACKNOWLEDGEMENTS

This work was conducted while DGS was appointed as visiting research scientist at Joint Research Center, (Ispra, I). His work is supported by grants from Natural Science and Engineering Research Council of Canada and Canadian Foundation for Climate and Atmospheric Science. Bruce Ainslie, Anton Beljaars, Paul Bovis, Peter Builtjes, Alison Munro, Lee Gass, Frank Raes and Silvio Funtowicz commented on an early version of the manuscript.

REFERENCES

[1] Canepa E, Irwin JS. Evaluation of Air Pollution Models. In Zannetti P (Ed.). Air quality modeling: theories, methodologies, computational techniques, and available databases and software. Fundamentals. Air & Waste Management Association 2002. Chapter 17, Volume I.

[2] Oreskes N, Schrader-Frechette K, Belitz K. Verification, Validation and confirmation of numerical models in the earth sciences. Science 1994; 263: 641-646.

[3] Roache PJ. Verification of codes and calculations. AIAA J 1998; 36: 696-702.

[4] Rykiel EJ. Testing ecological models: The meaning of validation. Ecol Model 1996; 90: 229-244.

[5] Willmott CJ. Some comments on the evaluation of model performance. Bull Am Meteorol Soc 1982; 63: 1309-1313.

[6] Gates WL , Boyle J, Covey C, *et al*. An Overview of the Results of the Atmospheric Model Intercomparison Project (AMIP I). Bull Am Meterol Soc 1998; 73: 1962-1970.

[7] Fuentes M, Guttorp P, Challenor P. Statistical assessment of numerical models. Technical report NRCSE, University of Washington 2003.

[8] Kleindorfer GB, O'Neill L, Ganeshan R. Validation in simulation: Various positions in the philosophy of science. Manage Sci 1998; 44: 1087-1099.

[9] Beven K. Towards a coherent philosophy for modelling the environment. Proc R Soc London Ser A Math Phys Eng Sci 2002; 458: 2465-2484.

[10] Tribbia JJ, Anthes RA. Scientific basis of modern weather prediction. Science 1987; 237: 493-499.

[11] Katz RW, Murphy AH. Economic value of weather and climate forecasts. Eds., Cambridge University Press, 1997; pp 222.

[12] Zhu YZ, Toth R, Wobus D, Richardson, Mylne K. The economic value of ensemble-based weather forecasts. Bull Am Meteorol Soc 2002; 83: 73-83.

[13] Simmons AJ, Hollingsworth A. Some aspects of the improvement in skill of numerical weather prediction. Quart J Roy Meteorol Soc 2002; 128: 647-677.

[14] United Kingdom Meteorological Office, 2003: http://www.metoffice.gov.uk/environment/boxurb/. Accessed December 2007.

[15] United States Environmental Protection Agency, 2003: http://www.epa.gov/cgi-bin/airnow.cgi?MapDisplay=FOREMAP, Accessed December 2007.

[16] Meteorological Service of Canada, 2003, http://www.msc-smc.ec.gc.ca/aq_smog/index_e.cfm Accessed December 2007.

[17] National Research Council Rethinking the ozone problem in urban and regional air pollution. Washington, National Academy Press 1992; p 267.

[18] Petersen AC. Philosophy of climate science. Bull Am Meteorol Soc 2000; 81: 265-271.

[19] Reed, Richard J, Albright, *et al.* Per The Role of Latent Heat Release in Explosive Cyclogenesis: Three Examples Based on ECMWF Operational Forecasts. Weather Forecast 1988; 3: 217-229.

[20] Steyn DG, Kallos G. A Dynamical Study of Hodograph Rotation in the Sea Breeze of Attica, Greece, Boundary-Layer. Meteorology 1992; 58: 215-228.

[21] Moeng CH, Wyngaard JC. Evaluation of turbulent transport and dissipation closures in second-order modeling. J Atmos Sci 1989; 46: 2311-2330.

[22] Alapaty K, Olerud DT, Schere KL, Hanna AF. Sensitivity of regional oxidant model predictions to prognostic and diagnostic meteorological fields. J Appl Meteorol 1995; 34: 1787-1801.

[23] Dennis RL, Byun DW, Novak JH, Galluppi KJ, Coats CJ, Vouk MA. The next generation of integrated air quality modeling: EPA's Models-3. Atmos Environ 1996; 30: 1925-1938.

[24] Grassi B, Redaelli G, Visconti G. A three-dimensional Chemical Transport Model of the stratosphere. Ann Geophys 2002; 20: 847-862.

[25] Hanna SR, Moore GE, Fernau ME. Evaluation of photochemical grid models (UAM-IV, UAM-V, and the ROM/UAM-IV couple) using data from the Lake Michigan Ozone Study (LMOS). Atmos Environ 1996; 30: 3265-3279.

[26] Huebert BJ, Phillips CA, Zhuang L, *et al*. Long-term measurements of free-tropospheric sulfate at Mauna Loa: Comparison with global model simulations. J Geophys Res Atmos 2001; 106: 5479- 5492.

[27] Kambezidis HD, Psiloglou BE. Measurements and models for total solar irradiance on inclined surface in Athens Greece-reply. Solar Energy 1995; 54: 443-445.

[28] Katz RW. Techniques for estimating uncertainty in climate change scenarios and impact studies. Clim Res 2002; 20: 167-185.

[29] Pinty B, Gobron N, Widlowski JL, *et al.* Radiation transfer model intercomparison (RAMI) exercise. J Geophys Res 2001; 106: 11937-11956.

[30] Roselle SJ, Schere KL. Modeled response of photochemical oxidants to systematic reductions in anthropogenic volatile organic-compound and NOx emissions. Geophys Res Atmos 1995; 100: 22929-22941.

[31] Randall DA, Wielicki BA. Measurements, models, and hypotheses in the atmospheric sciences. Bull Am Meteorol Soc 1997; 78: 399-406.

[32] Russell A, Dennis R. NARSTO critical review of photochemical models and modelling. Atmos Environ 2000; 34: 2283-2324.

[33] Schlünzen KH. On the validation of high-resolution atmospheric models. J Wind Eng Ind Aerodyn 1997; 67 & 68: 479-492.

[34] Rykiel EJ. The meaning of models. Science 1994; 264: 330.

[35] Alewell C, Manderscheid B. Use of objective criteria for the assessment of biogeochemical ecosystem models. Ecol Model 1998; 107: 213-224.

[36] Baumann M. On nature, models, and simplicity. Conserv Ecol 2000; 4.

[37] Ferguson RI, Church M, Weatherly H. Fluvial aggradation in Vedder River: Testing a one-dimensional sedimentation model. Water Resour Res 2001; 37: 3331-3347.

[38] Holzbecher E. Testing ecological models: the meaning of validation Remarks. Ecol Model 1997; 102: 375-377.

[39] Kirchner JW, Hooper RP, Kendall C, Neal C, Leavesley G. Testing and validating environmental models. Sci Total Environ 1996; 183: 33-47.

[40] Ludwig D, Mangel M, Haddad B. Ecology, conservation, and public policy. Annu Rev Ecol Syst 2001; 32: 481-517.

[41] Pescatore C. Validation: The eluding definition. Radioact Waste Manage Environ Restor 1995: 20: 13-22.
[42] Pielke RA. Jr. Room for doubt. Nature 2001; 410: 151.
[43] Konikow LF, Bredehoeft JD. Groundwater models cannot be validated. Adv Water Resour 1992; 15: 75-83.
[44] van Asselt MBA, Rotmans J. Uncertainty in integrated assessment modelling-From positivism to pluralism. Clim Change 2002; 54: 75-105.
[45] Oberkampf WL, Trucano TG. Verification and validation in computational fluid dynamics. Prog Aerosp Sci 2002; 38: 209-272.
[46] Sarewitz D, Pielke R, Byerly R. Prediction-A process, not a product. Geotimes 1999; 44: 29-31.
[47] Shackley S, Young P, Parkinson S, Wynne B. Uncertainty, complexity and concepts of good science in climate change modelling: Are GCMs the best tools? Clim Change 1998; 38: 159-205.
[48] Stedman JD. The meaning of models. Science 1994; 264: 329-330.
[49] Spear RC. Large simulation models: calibration, uniqueness and goodness of fit. Environ Modell Softw 1997; 12: 219-228.
[50] Saltelli A, Scott M. Guest editorial: The role of sensitivity analysis in the corroboration of models and its link to model structural and parametric uncertainty. Reliab Eng Syst Saf 1997; 57: 1-4.
[51] Hogrefe C, Rao ST, Kasibhatla P, *et al.* Evaluating model performance of regional-scale photochemical modelling systems-meteorological predictions. Atmos Environ 2001; 35: 4159-4174.
[52] Funtowicz SO, Ravetz JR. Post-Normal Science: A new science for new times. Sci Eur 1998; 169: 20-22.
[53] Haag D, Kaupenjohann M. Parameters, prediction, post-normal science and the precautionary principle-a roadmap for modelling for decision-making. Ecol Model 2001; 144: 45-60.

Received: December 23, 2007 Revised: January 18, 2008 Accepted: March 18, 2008