DATA INTEGRATION INTO ENVIRONMENTAL MODELS AND SENSITIVITY TO INPUT DATA

San José R., Salas I. and López J.
Environmental Software and Modeling Group, Computer Science School, Technical University of Madrid, Madrid, Spain

González R.M.
Department of Geophysics and Meteorology, Complutense University of Madrid, Madrid, Spain

Keywords: Mathematical Modeling, Environmental Modeling, Air Quality Modeling, Deposition Modeling, Mesoscale

Contents

Summary

We have presented a brief initial discussion of the concept of the environmental modeling system and laid particular emphasis on the relation between the data which is used to run such models and the mathematical model itself. We have also discussed the relations between the data and the output of the models on the one hand, and the relations of the data with the sensitivity of the environmental model on the other hand. We have discussed specifically the mesoscale air quality models and the different modules by making a distinction between peripheral modules (DATA) and the mathematical model itself. We have presented a specific example of the application of a mesoscale air quality model (OPANA) to a domain (Madrid, Spain) and the impact of using remote sensing data or field experiment data for parameterising the canopy resistance in deposition modeling. Finally, we have discussed the foreseen future evolution of air quality modeling (and particularly the mesoscale air quality modeling) by using tools such as GIS or satellite data and statistical packages together with the future more powerful parallel computer platforms.

1. Introduction: Mathematical Modeling

Substantial research has been devoted to mathematical modeling of environmental systems. Most of the work has been published in a wide range of journals and conference proceedings, each venue usually being specific to a particular, narrow range of disciplines or problems. Models are usually tools which provide integrated results of complex problems. In the case of environmental matters, the models essentially

contribute to the understanding of the complex relations between the different areas in ecological systems such as atmosphere, biosphere, etc. Models, and particularly mathematical models over computer platforms, are useful in simulating complex relations, and are providing answers to many different types of problems. Models are particularly useful when interdisciplinary subjects must be managed to understand a fixed problem or matter. Environmental processes are usually excellent examples of where an application of models is suitable. Environmental processes are also excellent examples of interdisciplinary applications which usually lead to solutions that are only understandable when all objects or modules are interconnected in the proper way, trying to simulate real processes in nature. These processes are characterized, in general terms, by being highly non-linear, so that modeling simulations are probably the only method to understand the complexity of the processes and, furthermore, the answers.

Mathematical modeling (as a set of differential equations) is, as a consequence, in many cases the essential tool for environmental engineering to simulate processes in the atmosphere, water or any ecosystem. In this contribution, we will focus particularly on atmospheric models. (For information on Measurement Tools for Atmospheric Systems, see *Measurement Tools: Atmospheric Systems*). However, many of the concepts can be easily transferred to other areas of environmental studies, such as water, waste, energy, etc. The model represents a tool to reproduce reality by simulating processes on a computer platform. In broad terms there are two reasons for constructing a mathematical model. From a *pragmatic* point of view, decisions regarding restoration and protection of the environment must be made. From a more *philosophical* perspective, a mathematical model may be the only means of representing our understanding of the complex behavior of an environmental system; such a model may be the most appropriate vehicle for interpreting observations of this system's past behavior.

In the context of classical decision analysis, however, a decision may take one of two forms: of either an action - figuratively, the pulling of a lever of policy in order to drive the system in a desired direction; or, the collection of further observations - for identification of those parts of the system that are not well understood, yet crucial to the success of knowing which lever of policy might subsequently best be pulled. We know that governments may delay taking (expensive) action while (much cheaper) research is undertaken in order to reduce uncertainty. Such research is often equated largely with field work and experimentation. Yet, it has not been the tradition for these processes of monitoring the environment to be guided by use of a model.

In a more refined sense, therefore, there are three objectives of modeling:

1. Prediction of future behavior under various courses of action, i.e.,, in the service of informing a decision.
2. Identification of those constituent mechanisms of behavior that are crucial to the generation of a given pattern of future behavior but insufficiently secure in their theoretical or empirical basis, i.e., in designing the collection of further observations.
3. Reconciliation of the observations of past behavior with the set of concepts embodied in the model, i.e., in the modification of theory and in explaining why a

particular input disturbance of the system gave rise to a particular output response.

Developments in hardware and software of digital computing, as the platform on which our models are made, cannot be separated from the way in which these models are conceived. Precisely which of these conceptual frameworks, and therefore computational representations, would be best suited to detecting and predicting environmental change is an entirely open question. We have barely begun to define the problem, let alone establish the means of its solution. Most of the contributions to environmental modeling, in general, presume the use of classical differential calculus as their conceptual framework, for which purposes we shall need to be concerned with the reliability of a numerical solution scheme. Yet there is equally a need to be aware of the possibility of better, alternative conceptual frameworks.

For example, it is obvious that the movement of a substance (pollutant) through a medium (air, water, solid) and its fate in that environment are of fundamental importance. In the purely formal terms of solving numerically the differential equations for characterizing these features, the differential operator may be split into the three components: advection, dispersion, and biochemical reaction. In spite of much progress in the use of *operator splitting* schemes, a proper simulation of the advective component remains problematic. There will always be benefits to be gained from improved schemes of numerical solution. For instance, Somlyody and Varis (1992) have argued that better schemes of operator splitting are needed for improved identification of the non-hydraulic terms in models of water pollution, i.e., those elements associated with the biochemical operator. In other words, if the hydraulic basis upon which our biochemical assumptions are founded is made more secure, we ought to be more confident of correctly identifying from the field data aspects of behavior attributable to these assumptions. This presumes, of course, that the concepts underlying the theory of advection and dispersion, to which the numerical operators more faithfully approximate, are themselves correct. Yet, in spite of the longevity of their study, the debate over how one perceives of what is meant by advection and dispersion has not diminished.

On the other hand, given the receding image of the fine-grained "truth" that motivates the enquiry, and given the grid (of either discrete points or volumes) that will result from any numerical scheme of solution, there will always be phenomena operating at finer scales of resolution than that of the model's numerical grid, and they must therefore be excluded from the model. These phenomena lie beyond the resolving power of the model; they may influence the behavior of the system as reflected in the more macroscopic terms of the model (i.e., in the values of its state variables); such microscopic influences must be described by expressions that are functions of these relatively macroscopic state variables; and, the formulation (or parameterization) of these expressions is problematic. In short, how are we to quantify the effects of that which must be excluded from the model in terms of that which can be included? This problem - of sub-grid scale variability and its relevance to modeling change in environmental systems - is already familiar to us from the discussion in the opening session. It is most often conceived in respect of characterizing spatial variability, and, for this reason, it is almost always intertwined with the technicalities of a numerical solution. But this spatial dimension has its counterpart in the differentiation among - or conversely, agglomeration of - chemical and biological species (as best illustrated in the

now almost forgotten works on models utilizing the concept of trophic length). The problem of sub-grid scale variability stands, therefore, above the mere technicalities of numerical solution. At this time in the development of the model, concern must necessarily shift from construction to evaluation. To be concise, in the following, let us assume, without great loss of generality, that the model of the environmental system can be defined by the following representation of the state variable dynamics:

$$\dot{x}(t) = f(x, u, \alpha; t) + \varphi(t) \qquad (1)$$

with observations of the state of the system sampled discretely in time as:

$$y(t_k) = h(x, \alpha; t_k) + \mu(t_k) \qquad (2)$$

Here x is the vector of state variables (such as pollutant concentrations in a defined volume of water), u is a vector of measured input disturbances (precipitation, solar radiation, effluent characteristics, and so on), y is a vector of output responses, α is a vector of model parameters (such as, for example, dispersion coefficients, growth-rate constants), φ is a vector of disturbances of the state variable dynamics that are not observable (the system *noise*), μ is a vector of (output) observation errors (the measurement *noise*), f and h are vectors of nonlinear functions, t is continuous time, t_k is the kth discrete instant in time, and the dot notation in $\dot{x}$ denotes differentiation with respect to time t.

Spatial variability of the state of the system can be assumed to be accounted for by, for example, the use of several state variables for the same quantity at several discretely defined locations (or within several discretely defined volumes). Typically, the outputs (y) are simply the error-corrupted values of the states, although in element-cycle models it is common to find that a single output variable, such as the observed concentration of total phosphorous, may refer to the aggregate sum of this element distributed in the system among several chemical and biological species, each denoted as a separate state variable.

We need now to prepare our discussion of the procedural steps of system identification with some observations on the objectives of analysis and the possible reorientation of some parts of the procedure for the purpose of detecting change.

MacFarlane (1990) has presented a three-element characterization of knowledge. According to the American philosopher Lewis, these three elements are (as reported by MacFarlane):

- The given data
- A set of concepts
- Acts which interpret the data in terms of concepts.

It is readily apparent that the problem of system identification is covered exactly by the third of these three elements. For, in the formal terms of the model of equation (1) above, we have the following counterparts:

- The observed input-output data (u,y) constitute the external description of the system's behavior and are the *given data.*
- The states and the parameters of the model (x,α) constitute the internal description of the system's behavior and are therefore associated with a formal realization of the *set of concepts* (through (f,h)), and
- Identification of the model is the 'act which interprets the data in terms of the set of concepts' (or alternatively, brings about reconciliation of the model with the data).

It is equally obvious that the distinction between the internal (x,α) and external (u,y) descriptions of the system's behavior is crucial to an appreciation of what will be possible as an outcome of implementing the subsequent procedure of analysis. The power of the classical experiments of laboratory science lay presumably in promoting the possibility of 'acts which interpret the data in terms of concepts' by reducing the set of concepts under scrutiny to as small a set as possible and by maximizing the scope for acquiring a large volume of the necessary data. The possibility of progress in the identification of a model should be enhanced when the order of (u,y) is very much greater than the order of (x,α), or $O(x,y)>>O(x,\alpha)$, for brevity. This can rarely be the case in the analysis of environmental systems; indeed, quite the opposite is the norm, i.e., $O(u,y) << O(x,\alpha)$.

Moreover, the art of the possible in system identification will be heavily circumscribed by what we may broadly label *the balance of uncertainties* between the set of concepts included in the model and the given data. On both sides of the dichotomy there are continua, from the almost certain to the highly uncertain. In contrast to what we would hope for, the inputs and outputs may well not have been observed with barely any error and with a high sampling frequency in space and time. The given data may frequently amount to little more than a qualitative expert opinion based on casual observation in the field. A similar spectrum of 'solidity *vs*. fluidity' is apparent in the constituent hypothesis by which the model is composed and - perhaps more importantly - those of which it is *not* composed. It is well known that some of the model's constituent hypotheses, and, therefore, its constituent parameters, are believed *a priori* to be more secure, or less uncertain, than others (as in the foregoing discussion of operator splitting schemes). This too can be accounted for.

What is achievable, then, in reconciling the set of concepts (the model) with the given data will be geared to the relative positions of the identification problem along the two continua, of uncertainty in (x,α), and uncertainty in (u,y). Thus, for example, to identify which of the model's constituent mechanisms are key, and which redundant, to a matching of observed behavior is a less sophisticated question to answer than attempting to establish which of these mechanisms are 'correct' and which 'incorrect'. To ask whether each constituent hypothesis is correctly expressed in the model is rather more sophisticated; indeed, to search for a single, uniquely best set of values for the parameters of those expressions will be even more demanding of the confidence that must attach to the set of concepts and given data. In a Bayesian spirit, and with a view to making decisions, the process of system identification is designed to bring about changes in the posterior probability of the model's parameters relative to the prior distributions (assumed before the model has been confronted with a given set of data). Ideally, the posterior distributions should reflect less uncertainty in the model than

before identification took place. In practice, they will also tend to reflect the inevitable distortions of an inadequate model structure and the ambiguities of an insufficiently incisive and extensive set of field data. It has been conventional to use (and to seek) merely a uniquely *optimal* set of model parameters for subsequent prediction. Contemporary studies in identification are more pragmatic, searching somewhat less strenuously for either several sets of relatively *good* or simply just *acceptable* candidate parameterizations of the model.

See: *Nonlinear systems, mathematical modeling, decision analysis, state variable dynamics, reliability, operator splitting, parameterization, noise, given data, set of concepts, "the balance of uncertainties".*

-
-
-

Bibliography

Schmugge T.J. and Jean-Claude A. (eds.). (1991) *Land Surface Evaporation.* Ed. Springer-Verlag. [An excellent overview of.]

Krishnamurti T.N. and Bounoua L. (1996) *An Introduction to Numerical Weather Prediction Techniques.* CRC Press Inc. [A book describing numerical weather prediction techniques.]

San José R. *Measuring and Modeling Investigation of Environmental Processes* (1999). WIT Press, Computational Mechanics Publications. [A pioneer book on combining measurements and modeling in environmental processes.]

McGuffie K., and Henderson-Sellers (1997) *A Climate Modeling Premier.* WILEY. [An excellent book with a CD containing climate models.]

Fränzle O. (1993) *Contaminants in Terrestrial Environments* Springer-Verlag. [An excellent overview of.]

Gyr A., and Rys Franz-S (1995) *Diffusion and Transport of Pollutants In Atmospheric Mesoscale Flow Fields.* Kluwer Academic Publishers. [An excellent book on flow dynamics.]