Of data and models

Jean A. Cunge

ABSTRACT

Relationship between the data, such as direct observations of nature and recorded measurements, and the models is very complicated in the 'water domain'. It is not at all as clear and explicit as it is often presented by teachers to students, by consultants to clients, or by authors to readers of publications. A number of aspects of this relationship are discussed using examples to illustrate the author's views. Limitations of data-driven tools (correlations, Artificial Neuronal Networks, Genetic Algorithms, etc.) and data-mining, when applied without physical knowledge of the relevant phenomena, are discussed, as are those of deterministic models. The currently used 'good practice' paradigm in modelling (the model is to be set up, calibrated, validated and run) is rejected when deterministic models are concerned. They should not be calibrated. A new paradigm, a new 'code of good practice', is proposed instead. Strategic and tactical aspects of various available approaches to modelling of physical phenomena and data exploitation have practical engineering and financial consequences, most often immediate and sometimes very important: hence the significance of the subject that concerns the everyday occupations of modellers, their clients and end-users .

Key words | calibration of models, data-driven models, deterministic models, good practice modelling paradigm, limitations of models, validation of models

Jean A. Cunge
International Institute for Infrastructural, Hydraulic and Environmental Engineering (IHE),
P.O. Box 3015,
2601 DA Delft,
The Netherlands
E-mail: *jacunge@aol.com*

INTRODUCTION

All engineering-oriented activities of modelling practice and monitoring, as well as of collection, analysis and archiving of information, are concerned with the subject of data. Obviously one cannot expect to deal with this subject within a paper of a length acceptable to a technical journal. Thus only a few particular points and problems, chosen rather arbitrarily by the author, will be discussed in what follows. The author's intention is to stress some of them because he feels that they should be clear in the minds of those who are involved in modelling, especially those who model for engineering purposes. Indeed, there are various reasons for becoming involved in modelling activities. One important category is that of research: when a researcher hypothesises some feature of natural phenomena, by this very hypothesis he is creating a model in his mind, a concept.[1] 'He is selecting certain signs from the phenomena themselves and merging these to form an expressive sign that appears to be meaningful and true' (Abbott 2002*a*). Then, building a material model (analytical, numerical, laboratory scale, etc.) and comparing the results of the latter with observed data, he can validate or invalidate his hypothesis. The model provides a better understanding of nature in the sense that the comparison can either confirm the acceptability of the image of reality proposed by the researcher or, to the contrary, indicate that the image is not adequate and further developments are necessary. Another category of reasons for becoming so involved are engineering applications. In

[1]All terms employed here are those set up in Abbott (2002*a*) but employed more loosely and less formally than in that paper.

most cases, when an engineer is building a model, he does it for one of two reasons:

- either he wants to *predict* the consequences of situations that have not been observed so far: for example, the consequences of building hydraulic structures on a river (dams, dykes, etc.), transformation of an exceptional, never before observed rainfall episode into runoff and discharge, etc;
- or he aims at monitoring or controlling a process during which the variations are not exceptional but should be correctly forecast: for example, manufacturing processes, flood control and navigation.

As compared to a researcher, the engineer's involvement with numerical or scale modelling is more utilitarian, oriented towards construction, safety, design and analysis rather than any special understanding of the physics of the phenomena. This is because, in most situations, the engineer already has in mind his own model, based upon his own concept of the underlying reality. This model, and its basic concept, this knowledge composed of what he has learnt before his graduation and then during his professional life, is in most cases sufficient to use a specific modelling tool (numerical, reduced scale) and interpret its results for his *engineering* purposes. Such an application can in most cases be legitimely called 'state of the art' or 'best engineering knowledge'. In other words an engineer, rather than trying to improve or invalidate an *existing formulation* of real phenomena wll most often use an *acquired* knowledge of physics and modelling to set up a model and to ensure that the model results are sound and interpreted in a useful way. In hydrology and within the context of a more historical discussion of modelling, one can refer to Klemeš (1986) and his distinction between

- models for *use inside hydrology* as exploratory research tools (, *the*) purpose of which is to contribute to (*an*) understanding of hydrological processes, and
- models intended for *use outside hydrology*, in particular for planning, design, or operational decisions'.

There are of course overlaps between research and the engineering use of models, especially when the physical domain is not yet well explored or understood but when engineering intervention is required, even though we know that the risks of error are considerable. The climate change domain is possibly the widest known example of such a situation. In this area, the understanding of physics, the consequences of human dealings with nature, the value of models and the availability and the collection of data and their interpretation are all interwoven. In hydraulics, modelling of the sedimentation and geomorphology of rivers and estuaries provides an example of such a situation.

The models can be misused. A typical case of the misuse of models arises when engineers go beyond their 'best knowledge', or simply beyond their knowledge, and apply modelling tools in ignorance or in contempt of the limitations of these tools. Another occurs when researchers, who often ignore engineering art, are pushed by their political masters to get some 'pocket money' for their universities and therefore apply *research* tools for *engineering* purposes.

The relationship between data, such as direct observations of nature and recorded measurements, and models is very complicated. It is not at all as clear and explicit as it is often presented by teachers to students, by consultants to their laymen clients, or even by otherwise serious authors to readers of scientific and technical publications. In the present paper, a number of aspects of this relationship will be discussed using examples that are thought to illustrate the essence of the author's point of view concerning the subject. The author's views expressed here are, of course, controversial and their criticism by readers is most welcome because the subject allows for multiple approaches and the process of selection of any particular approach does not necessarily obey the rules of a Cartesian logic. Strategic and tactical aspects of various available approaches to modelling of physical phenomena and data exploitation have, however, practical engineering and financial consequences, most often immediate and sometimes very important.

The aspects discussed will be limited to the domain known to the author, i.e. the 'water domain' covering hydraulics, hydrology, surface and groundwater

modelling, flood simulation and forecasting. They will be considered in the context of 'data and models', models being here restricted to numerical models and data being considered as related, in one way or another, to modelling activities. These aspects are as follows:

- data and information, data interpretation;
- data and models: data-driven models and deterministic models;
- data mining and models: theory from data, data assimilation;
- data and modelling methodology: 'good practice' in modelling (calibration, simulation and validation);
- data and choice of modelling tools: truth, ambitions, illusions and engineering.

DATA MINING, DATA AND INFORMATION AND DATA INTERPRETATION

It is supposed here that data mining, in the restricted sense of the interpretation of data in order to produce information, cannot be expected to lead to findings useful for engineering if it is carried out in a 'blind' manner, even when the most recent and sophisticated mathematical tools and enabling methodologies are applied. This thesis is certainly difficult either to prove or to invalidate in any formal way. It is, however, supported by historical experience and the way that the hydraulic/hydrology theories and engineering tools, based on experimental data, have been developed in the past.

The significance of measured data

Consider *quantitative* data obtained through measurements. When we measure and record something (a 'variable'), such as the velocity of flowing water, or the head loss, or any other magnitude, we always measure the variation of one particular magnitude of one particular variable. The measure can be very accurate, but it is always limited in duration and space (it is a local measure). The first difficulty with which one has to deal is the choice, or rather definition, of *significant variable*(*s*), this choice depending upon the definition of the *purpose of measuring* that in turn defines the significance. The choice can be made only through a physical analysis, from knowledge of the physical phenomena concerned and through the intuition or 'feeling' of the practitioner.

When research experiments are concerned, the historical paradigm in physical science is that a measured phenomenon should be as much as possible isolated from others so that the measurement is not influenced, and in particular not distorted (today we say 'polluted'), by other phenomena. Such an approach allowed in the past for the formulation and verification of a number of physical laws, valid only for isolated systems but prevalent to the considered phenomena even when the real systems were not, and could not in fact, be isolated.

Within the context of modelling, this paradigm often does not answer the questions that are actually asked, and indeed it may well bring more problems to solve than it actually solves. Three crucial points are encountered in the engineering practice of modelling:

- the modelled phenomenon is characterised not by one measurable variable but by several of them or, often, by composite variables that cannot be derived rigorously from basic measurable ones;
- the 'state of the art' theories, techniques, methodologies and models use state variables that are impossible to measure directly but result from an interpretation (sometimes again through modelling) of other measurable variables;
- the physical parameters of the models, necessary for their calibration and validation, are measurable only at punctual locations over a limited duration, while they are supposed to represent large areas and times; this point brings in the problem of the representativity of measured parameters and of their spatial and temporal variability.

The first crucial point is that most often the phenomenon to be modelled, and hence also described by the measurements, is characterised not by one variable to be measured but by several of them. Thus a significant variable that characterises the phenomenon and might make it possible to draw general conclusions, or to explain the phenomenon, may be a composite variable. Or there may be

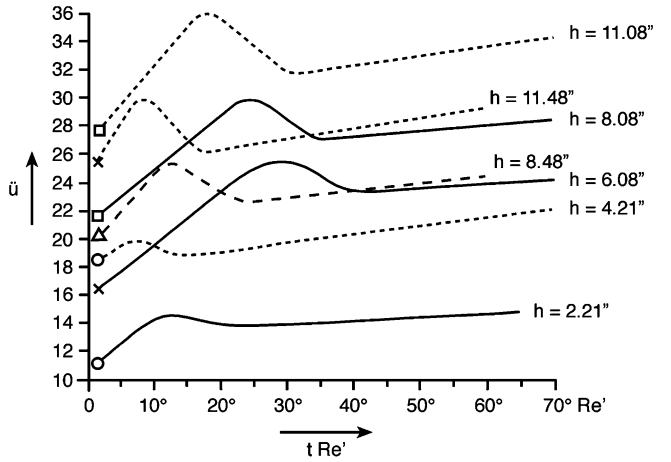


Figure 1 | Relation between the flow velocity $\bar{u}$ (in Rheinland inch per second) and the temperature (in degrees Réaumur) for various pipe diameters and heads h (in Rheinland inch). After Prandtl & Tietjens (1934).

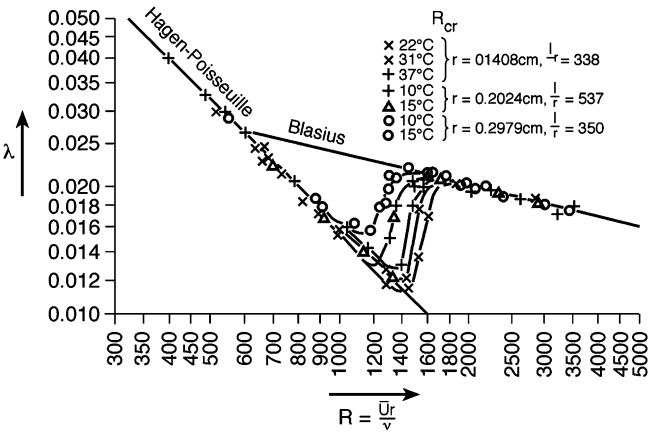


Figure 2 | Pressure-drop coefficient vs Reynolds' number, being Hagen's tests of Figure 1 replotted. Definition of Blasius law. After Prandtl & Tietjens (1934).

several of those. Such significant variables *may not be measurable directly at all* although their components are measurable.

A classic example is provided by one of the famous experiments of Hagen (1854) aiming at determining the relationship between the velocity of water flow in pressurised pipes, the water temperature, the diameter of the pipe and the pressure head. Hagen's results are shown in Figure 1. Hagen had a very good conception of laminar and turbulent flow but his approach lacked a physical analysis allowing for a unifying principle (essentially, a model) governing the flow. Simple transcriptions of carefully measured and recorded data gave diagrams (Figure 1) that were awkward to use and certainly did not lead to a unifying law or principles.

As is well known the credit for finding such a law belongs to Osborne Reynolds (1883). He demonstrated by means of *dimensional analysis* that, if flow in a circular pipe is mainly influenced by inertia and the viscosity of the liquid, and by the pipe diameter, then the resistance to the flow (pressure drop) should be a function of a dimensionless expression

$$\mathrm{Re} = \frac{R_h \bar{u}}{\nu} = \frac{D\bar{u}}{4\nu} \tag{1}$$

This expression, in which u is the flow velocity, R_h is the hydraulic radius, $D = 2r$ is the pipe diameter and ν is the kinematic viscosity, is of course the Reynolds number, Re. Consider then a dimensionless head loss coefficient λ:

$$\lambda = \frac{\Delta p}{\rho \overline{u^2}/2}\frac{r}{L} = \frac{hg}{\bar{u}^2/2}\frac{r}{L} = \frac{hg}{\bar{u}^2}\frac{D}{L} \tag{2}$$

where L is the length of the conduit and h is the difference in elevation between its upstream and downstream ends. This formula expresses simply the proportionality between the pressure drop gradient I and square of the velocity $\bar{u}$:

$$I = \frac{\Delta p}{L} \propto \bar{u}^2 \tag{3}$$

Then Hagen's data of Figure 1 could be plotted in a new coordinate system (Re, λ). The well known result is shown, after Prandtl & Tietjens (1934), in Figure 2: it allows not only for explicit definitions of flow laws in smooth pipes but also shows clearly the difference between laminar (Hagen–Poiseuille) and turbulent flows, with Blasius' (1913) law applicable to the latter:

$$\lambda = \frac{0.316}{(\mathrm{Re})^{0.25}} \tag{4}$$

It should be observed, however, that Blasius' law is not valid for rough pipes, for which one cannot leave out

of the analysis the roughness height k_s of the wall, so that

$$\lambda = f\left(\mathrm{Re}, \frac{k_s}{D}\right) \tag{5}$$

At this point we would like to stress that the limitation of validity of Blasius' law cannot be deduced at all from the diagram in Figure 2. It was only the experimental data of Nikuradse (1933) that showed that, for turbulent flows, there are again two domains, one within which Blasius' law is valid and another, for pipes with rough surfaces, for which it is not. From Nikuradse's data the implicit Colebrook–White formula was derived for λ in rough turbulent flow:

$$\frac{1}{\sqrt{\lambda}} = -2\log_{10}\left(\frac{k_s}{3.7D} + \frac{2.51}{\mathrm{Re}\sqrt{\lambda}}\right) \tag{6}$$

This formula is widely used to define the roughness and, hence, the resistance to pressurised flow in pipes. Complicated diagrams, such as the generalised Moody diagram shown in Figure 3, have been used for some 70 years to define λ as a function of k_s, D and Re. We shall come back later to this formula, well known to all hydraulic engineers.

Replacing in the Reynolds Number formula the flow velocity by the shear velocity u^* and the hydraulic radius by the roughness height k_s, Prandtl (1933) plotted the data in new coordinates:

$$\text{Abscissa:} \quad \log_{10}\mathrm{Re}^* = \log_{10}\left(\frac{k_s u^*}{\nu}\right) = \log_{10}\left(\mathrm{Re}\sqrt{\frac{\lambda}{8}}\right) - \log_{10}\frac{D}{k_s} \tag{7}$$

$$\text{Ordinate:} \quad \frac{1}{\sqrt{\lambda}} - 2\log_{10}\frac{D}{2k_s}$$

The resulting diagram, shown in Figure 4, reproduced after Jaeger (1954) gives a single-valued curve (which actually can serve as a model of the flow) for the whole domain of turbulent flow, with the smooth-pipe domain being represented by one straight line and the rough-pipe domain by another, the two being related through a transition zone. Note also that, beyond the value of $\mathrm{Re}^* = 100$, the coefficient λ does not depend any longer on the value of the Reynolds number, but is constant for a given ratio D/k_s—a clear clue to the existence of a similarity law for all rough turbulent pressurised pipe flows. This curve takes into account as well the measured

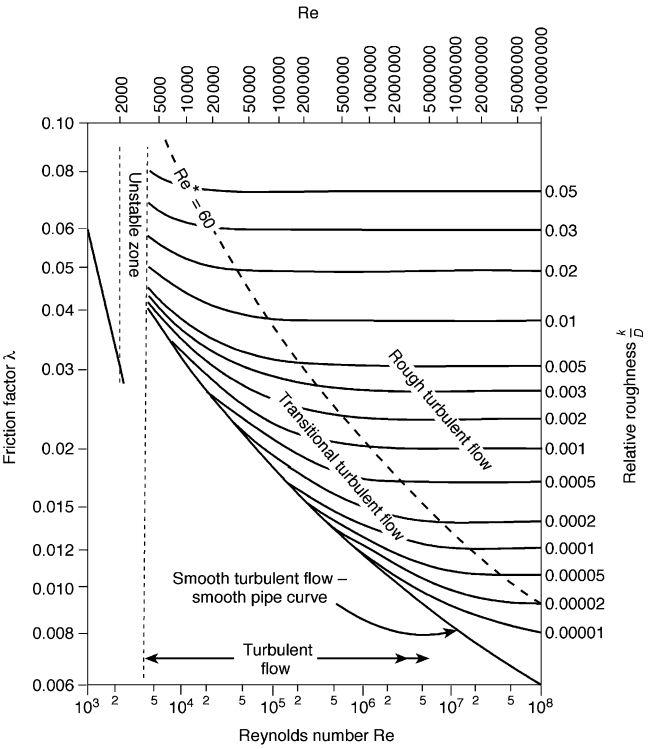


Figure 3 | Typical textbook generalised Moody diagram showing the variation of λ with Re and k_s/D. After Hamil (1995).

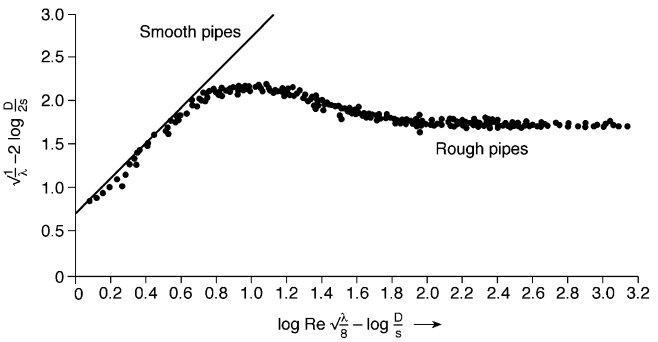


Figure 4 | Roughness function (after Prandtl and Nikuradse) λ as a function of Re*. After Jaeger (1954).

values of λ as the velocities (i.e. corresponding to von Karman's theory) and, as remarked by Eck (1941) and Bakhmeteff (1941), it solves completely the general problem of friction in pipes.

The purpose of the above historic reminiscence is to show that the presence of raw experimental data has not been sufficient to allow any general conclusions to be drawn. Reasoning and analyses based on physics (such as dimensional analysis, the introduction of shear velocity as a variable, etc.) have been necessary to formulate useful laws giving real insight into physical problems, and measured data has confirmed the theory, including its formulation. Another purpose here is then to show how this process was carried out.

The second crucial point that is encountered in engineering applications appears as an important methodological difficulty: the variables or parameters considered as significant may not only be composite (such as the Reynolds number in the above examples) but, often, they are impossible or very difficult to measure. And yet it is just these that are used for all analysis, decisions and design. Thus, in hydrology, the main variable to which traditionally hydrologists refer is discharge. This magnitude, typically needed to define the frequency of occurrence of events, necessary as a boundary condition for one-dimensional hydrodynamic numerical models of river flow, and so on, is, however, very difficult to measure accurately. It is even very often impossible (for practical or budgetary reasons) to measure, such as during high flood conditions in streams (which is exactly when the measurement would be most useful). Incidentally it is worth noting that discharge data in rivers are the results of measurements of local velocities across the river section. Thus what is really measured (when it is measured at all) are local velocities during defined time intervals and, of course, wetted cross-sectional areas. The integration of their product gives 'instantaneous' discharge. In hydraulic practice the most often measured data are not discharges but rather water stages because their measurements are generally feasible and, at a given location of a sensor, can be made accurately. But in unsteady flow there is no accurate way to transform measured water stages into discharges unless a *reliable numerical simulation model* is used. As a result, most of the hydrology of rivers is based on the information (discharges) that is not directly *measured* but rather *estimated from water stages* with the help of models of various kinds and with widely varying reliabilities.

Another illustrative case is groundwater flow. While it is feasible and not too costly to measure piezometric heads in an aquifer, it is impossible to relate them to the discharges except through the use of a more or less reliable numerical simulation model. The latter is based on such parameters and variables as the hydraulic conductivities and the diffusivities or infiltration coefficients, none of which can be measured except at a precisely given location and during a limited time, i.e. in practice in laboratory or laboratory-like conditions.

One example showing how the data and modelling are interwoven and how 'obvious' solutions may well not be feasible at all is the problem of open (sea) boundary conditions for two-dimensional estuarine models. Free surface elevations and water velocities within such models usually depend closely upon the imposed open seaward tidal boundary condition (at least for the area influenced by the tide). However, to measure with sufficient temporal and vertical space accuracy the free surface elevation at a number of points along an open boundary is a task that is next to impossible, even if very expensive and elaborate measurement systems are introduced. Indeed, very small errors in elevation measurements or time shifts (of the order of millimetres and seconds) along the open sea frontier of a model commonly create transverse artificial slopes and result in complete changes in the pattern of currents inside the model. Accordingly the *reliable* models of this kind are calibrated not by the variation of roughness coefficient within the modelled domain (that must be assessed using engineering experience of the modeller) but rather by the variation of in-flowing and out-flowing discharge hydrographs and their distribution at the open seaward boundary (Abbott & Cunge 1975). The reason is that the bottom roughness inside the modelled domain does not influence significantly the velocity fields and free surface elevations within the model while the boundary distribution of discharges and tidal elevations at the boundary do. The concept of 'automatic' calibration through approaching mathematically some 'objective' function by varying roughness coefficients inside a model is doomed

to fail in such situations. The failure in each such case is usually of a twofold nature: totally unphysical values of the roughness coefficients necessary to 'calibrate' the free surface elevations are introduced, while the incorrect velocity fields are still present. It is necessary for the modeller to understand the physics of the modelled tidal areas very well in order to follow the adequate approach to this difficulty.

Finally, evoking the third crucial point, we have mentioned the problem of data measured locally and the difficulty of interpretation of its representativity and significance for larger modelled domains. This point is immediately obvious when we think of rain water infiltrating vertically through the unsaturated zone in, say, a 100 m × 100 m column (e.g. in a hydrological distributed model, with a computational grid resolution of 100 m) represented by a Richards equation with soil parameters measured locally. The characteristics of such measured parameters most likely are valid only within the immediate neighbourhood of a few metres around the measurement point. Here again interpretation through a model is necessary. The Richards equation has been developed through reasoning on infinitely small volumes. If we have to deal with finite volumes, would the same law be valid? If yes, can there be conceived a model of spatial variability of locally measurable parameters leading to averaged parameters with physical meaning? And, finally, what is a minimum sample of locally measured parameters allowing for the estimation of the spatially averaged ones? The same type of question can be asked about Strickler–Manning coefficients when a river stretch or an area of inundated plain are concerned. This time we do not talk about measurements of parameters, but rather about estimates of coefficients that are local, but otherwise the questions are of the same kind.

An obvious conclusion from the above is that the data must be interpreted in the light of our understanding of the physical phenomena and those theories that describe these phenomena to our satisfaction; otherwise they are not very useful.

Qualitative data and information

The problem of *qualitative* data, such as descriptions of physical events made in common, current or 'narrative' languages, is another difficulty that cannot be put aside. Some people may object that such narrative language descriptions are not equivalent to scientific data, but 'just a description'. But, if we stick only to two fields, hydrology and climatology, many data on past floods or rainfalls are only descriptive. When we try to recover information on the past going backwards beyond the 19th century, the information is nearly exclusively qualitative. And even today, consider a phone call of a citizen living next to a mountain torrent and saying that 'the rain is extremely heavy and I can hear a horrible noise made by the stream upstream in the hills'. If and how a real-time flood forecasting/warning system can integrate such data is a crucial and difficult engineering problem. At this point it may be useful to introduce for the time of reading of this paper a convention on making the distinction between information and data, especially because qualitative data clearly convey information and it is rather difficult to make a distinction between these two. Let us call the information a description of the phenomena or of the situation as a whole, while the quantitative data will refer to the measurements of specific individual magnitudes of physical variables or parameters. In this context we understand that qualitative data refers to qualitative description or estimate of individual parameters of interest (see, further, Abbott 2002*a*).

Incidentally one may observe that one of the main difficulties we face when applying information society technologies is to extract *user intelligible information* from measured *quantitative* data and collected *qualitative* data. How far and in what conditions is a reversible approach possible: to quantify qualitative data and information? Or how can a fusion of both types of data be achieved? And how do these problems influence modelling practices? (See, further to this, Jonoski 2002.)

DATA AND MODELS: DATA-DRIVEN MODELS AND DETERMINISTIC MODELS

Models, or at least the category of models useful for engineering applications in the hydrology/hydraulic field, are representations of one part of reality, of real

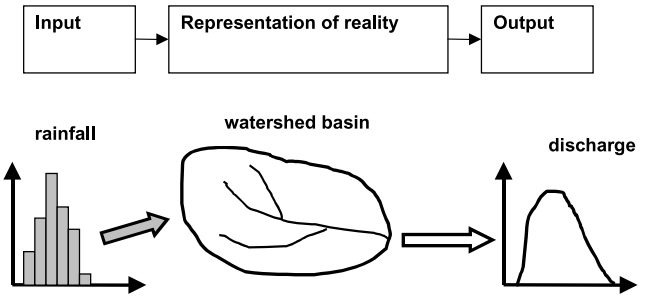
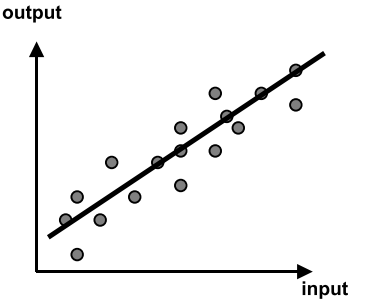


Figure 5 | Symbolic representation of a watershed basin as a system to be modelled.



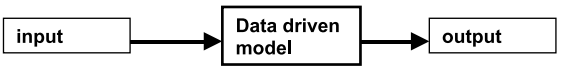


Figure 6 | Data-driven modelling—symbolic representation.

phenomena. In nearly all cases the aspect of reality in which we are interested and that we try to model goes far beyond the magnitudes of the isolated parameters that can be, and customarily are, measured. Consider as an example Figure 5 which shows schematically the object modelled (a catchment basin) with input (rainfall) and output (outflow discharge hydrograph). We may need to model such a catchment for various reasons (land use modifications, hydraulic structures to be implemented, to obtain outflow hydrographs for exceptional rainfalls, etc.).

The complexity of the catchment's physical characteristics (parameters), as well as the complexity of physical phenomena of such a rainfall/runoff transformation illustrates well the difficulty involved in assessing anything from knowledge of a few measured variables. A model is a consequence of a synthetic view of reality; it simplifies reality, but it encompasses much more than the sum of a limited number of measurable parameters, or parameters which it is possible to isolate. In the context of hydroinformatics, a model with its input/output/interpretation environment can be considered as a tool transforming data into information. Hence the modelling problem is always a cognitive problem and the data, quantitative or qualitative, including the model results, are rarely useful in their raw forms. There is an obvious need for their interpretation: 'and it is only through this interpretation that we are truly in the presence of a model—or ''the model comes to presence'' ' (Abbott 2002*b*).

In hydraulics/hydrology engineering practice there are essentially two classes of models to represent reality, namely data-driven models and deterministic models. It is useful to introduce here a few definitions of the vocabulary we use. We talk essentially about mathematical representations (formulations) of what we think describes some aspect of reality. In our case, each such formulation is more or less related to a physical reality: not at all or little (the typical situation for data-driven modelling) or very strongly (in deterministic modelling). A *modelling tool* is an informatics code for the numerical computerised solution of mathematical formulations. A *model* is related to a given physical situation: data-driven modelling tool, the components and parameters of which have been trained on past observed data of a catchment basin, is a data-driven model of this catchment. A deterministic model of unsteady flow in a river reach, if based on De Saint-Venant's equations, is a modelling tool that contains all topographic, geometric and hydraulic characteristics of the modelled reach.

Data-driven models

A data-driven model is a transfer or correlative function generator, the parameters, coefficients and components of which have been fitted (trained, calibrated) using input/output data recorded in the past. The simplest model of this class is of course a linear input/output correlation, as depicted in symbolic form in Figure 6.

The symbolism of Figure 6 shows clearly the basic conception of data-driven modelling tools: they are not based on any explicit integration of physical laws. This

class of models comprises, however, a large variety of sophisticated methods such as (from simple to complex) multicorrelations, ARMA methods, transfer function identification, Artificial Neural Networks (ANNs), Genetic Algorithms (GAs), Genetic Programming (GPs) and methods based on chaos theory.

The main advantages of data-driven models are:

- rapid and nearly always automatic training and validation using past recorded data;
- once the model is defined, the calculations of input/output transformation are extremely rapid and, hence, well adapted to real-time applications.

Data driven models have essentially two predictivity limitations:

- their validity is limited by the training sample of inputs/outputs;
- they are not predictive when the system is modified.

At this point it is useful to define what might be called the *predictivity* of a model. First of all consider Figure 6, assuming that a data-driven model has been trained using observed recorded daily input (rainfall)/output (discharge) data for the last 25 years and that during this period the period or occurrence of the most important observed rainfall episode has been 50 years. Try now to compute with this model the discharge corresponding to an exceptional rainfall, the frequency of occurrence of which is 200 years. The data point is definitely out of range of the sample and one must be a very strong believer in the model's virtue to maintain that it would predict the output correctly. Then consider Figure 7(a), i.e. a free-flowing river reach for which there is a large sample of observed inflows and outflows and for which a data-driven model has been trained.

Suppose now that a dam project is proposed for this river as in Figure 7(b) and the same model is used in an attempt to compute expected outflows for given inflows. This time the outflows are influenced by the dam operations rules and obviously the 'trained' parameters obtained from the original input/output sample are not adequate any more and the model cannot predict the outflows because it does not take into account the modification in the system itself.

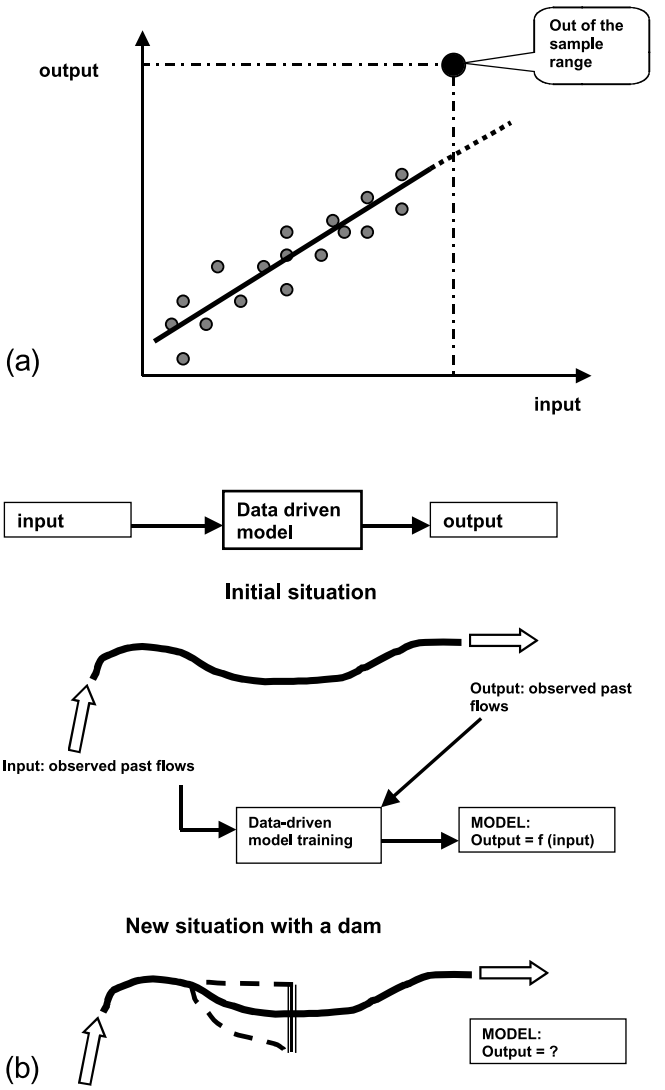


Figure 7 | (a) Predictivity limitations of data-driven models, (b) symbolic representation.

What we can say is that enabling techniques for data-driven models are efficient and, in certain engineering domains, extremely useful. Their application in physical domains, however, must not be blind, but must be related to physical interpretations of phenomena. It is vain to expect that putting together, on the one hand, a mass of measured data and, on the other, mathematical data processing methodologies that ignore any physical understanding of the phenomena concerned will produce useful results, unless the old story of a couple of hundred monkeys typing randomly on keyboards and eventually

producing Shakespeare's complete works is supposed to be believed.

It is very important that not only the users of modelling tools but also the end-users of the results that these tools produce know what are the premises of the quality of predictivity of these tools. For example, it can be shown that the ANN method provides a universal function approximator. This means, following Babovic *et al*. (2001) that 'given a *sufficiently long* training time, a *sufficiently complex* network architecture, and a *sufficiently representative* data set, an artificial neural network can approximate any function to an arbitrary degree of accuracy' (italics mine). The so-called 'fast operators' in the modelling field retain only the last part ('. . . an artificial neural network can approximate any function to an arbitrary degree of accuracy') of the quoted statement. The more careful reader will realise that the approximation of one function by another means that the difference between the two can be reduced as much as we wish, but at the cost of more and more effort and unlimited data availability, the latter being related to the word *sufficiently*, used several times in the quotation. And that the approximation is valid within some interval only—did the reader ever try to extrapolate with a spline approximation beyond the considered interval?

As a basic conclusion to this paragraph the author proposes another citation from Babovic *et al*. (2001): '. . . we strongly believe that the most appropriate way forward is to combine the best of two approaches—theory-driven (*i.e. deterministic—JC*), understanding-rich, with data-driven modelling processes'. The author would like to stress with all necessary emphasis the corollary to the above citation: there is a danger in entrusting modelling activities carried out with data-driven tools to persons who do not have sufficient understanding of the physics of the modelled phenomena, whatever their qualifications concerning the methods for setting up and training of such models may be. The methods and algorithms used in data-driven models may be the same for hydraulic problems, for supermarket inventories and for stock exchange market predictions but this does not mean that a financial analyst is the best person to model hydraulic phenomena, even if he is an expert on data-driven models.

The following (real-life) example illustrates the problem. In coastal areas one needs the short time (from, say, 2–12 hours of lead time) forecast of sea surface elevations, for shipping as well as for warning purposes. When meteorological factors are 'normal', the elevations depend only upon the tide. The situation is very different when wind and pressure variations create storm-surge conditions—the elevation will then be different from that predicted from tides alone and may handicap near-coast navigation. Hence the necessity to forecast such conditions and issue adequate warnings for navigation. It is tempting to set up a forecasting system using a data-driven model that is based on free-surface elevations observed in the past at a number of coast-located gauging stations. In the forecasting stage, such a model would use as input the free-surface elevations measured at these stations during a few hours (say up to 6–12 hours) before the time of issuing the forecast. It will produce as output the forecasted free-surface elevations, at the location of a desired forecast, in 2, 4 and 6 hours. The actual method at the base of such a data-driven model is not of concern here. The tool can be a simple one or it may use very sophisticated mathematical techniques, such as chaos theory. The principle is the same: the coefficients and parameters of the model are 'trained' upon the data observed during many years in the past.

One way to train such a model is to use the raw data of free-surface elevations. The model is then expected to forecast directly the free-surface elevations. On the other hand, in the search for higher accuracy, the data used to train the model could be limited to the data related to the surge only, i.e. to the elevation due only to the meteorological effects, above the 'normal' tidal level. Then the model would forecast only the surge-induced elevations that are then superimposed over the forecast 'normal' tidal level. Both approaches are based on the idea that one can, in some sophisticated way, extrapolate a sequence of locally observed elevations at times $t-1$, $t-2$, . . ., $t-n$ into the future and obtain forecasts for times $t+1$, $t+2$, . . ., $t+m$.

In one case known to the author, a very sophisticated and advanced (mathematically) data-driven model has been used to develop a system for forecasting navigation channel levels, the forecast being based solely on past

observed free-surface water elevations available at a number of coast-located stations.

The objection here is that the model assumes that all possible surge-tide events in this area have some kind of common behaviour that can be defined by free-surface elevations observed during a short period prior to the forecast time. Further, it assumes its own capacity to extract from past observed local values these common characteristics. This can be correct in restricted areas, but it could also be widely incorrect in more extended seas. For example, to forecast storm surge levels at North Sea coastal areas one should take into account the meteorological *forecast* (pressure, winds, etc) at stations located very far from the place for which the forecast has to be issued. Only a physical knowledge of the genesis of storm surges in the area can indicate where the data and forecasted input should be taken.

The second approach (forecast of the surge alone as superimposed on the 'normal' tide) introduces another difficulty: how, for past records, can one possibly separate the surge from 'normal' tidal elevations and how, for the purposes of the forecast, can one possibly define such 'normal' elevations upon which the forecast surge should be 'superimposed'? The astronomical tides in the oceans can be computed by adding their harmonic components, so it seems simple. Unfortunately, near coasts the non-linearity of tides becomes more and more accentuated and the deformation of tidal waves by friction, bathymetry and coastal topography more and more important. The actual free-surface elevation at any point is the result of a series of nonlinear interactions of all these influences with the forcing tidal wave *and* with the meteorological forces. Thus one cannot reasonably expect that the storm surge effect can be considered as simply superimposed upon a tide, and certainly not upon an astronomical tide in a coastal zone—they are likely to be interactive, and the stronger the tide and the surge, the stronger this interactivity will be. While the usefulness of the forecast depends upon its reliability and accuracy, it is likely that the only really satisfactory approach is to obtain the data necessary to train a data-driven model from a deterministic model. Such a deterministic model should be designed to simulate both tide and storm surge and then can create the data upon which a data-driven model can be 'trained'. But that means that there is no 'short cut' through data-driven models to such problems of storm surge surface elevation forecasting. It is true that the existence of an established (trained) data-driven model representing the relationship between recently observed and near-future forecast free-surface elevation values allows for rapid and simple real-time applications to such problems as those of aiding navigation. But the value of such a data-driven model will depend upon the physical situations, the analysis and understanding of the physical phenomena and upon the value of the deterministic model!

Deterministic models

A deterministic modelling tool is based on a number of physical laws that have been formulated in suitable mathematical form (usually in the form of equations), the latter being solved numerically through the application of appropriate algorithms. The laws that provide the foundations of the tool define its domain of application and the user may choose one or another tool as a function of the adequacy of the laws upon which the tool is built, together with the user's requirements and needs.

As an illustration, consider two examples of very different physical problems and very different engineering applications.

Reservoir model of rainfall-runoff transformation over a catchment

This conception, schematised in Figure 8, is based upon the formulation of the law of conservation of volume in the form:

$$\frac{\mathrm{d}S}{\mathrm{d}t} = I(t) - Q(t) \tag{8}$$

where $S(t)$ is the volume stored in the reservoir, $I(t)$ is the input (rainfall) and $Q(t)$ is the output (outflow). The purpose of using such a formulation is to be able to define, for a catchment, the outflow $Q(t)$ when the rainfall $I(t)$ is known.

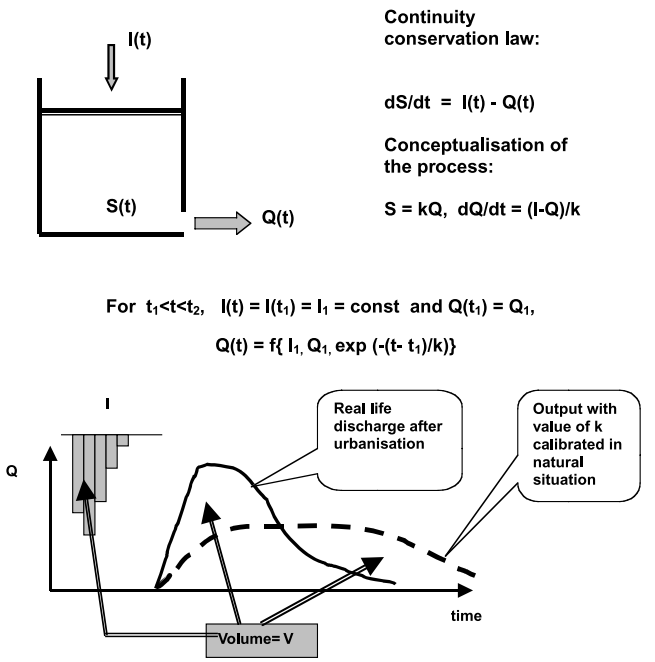


Figure 8 | One-parameter reservoir model of a watershed basin.

To make (8) operational for such a purpose one should add to it a relationship between storage S and state variables $I(t)$ and $Q(t)$. The simplest and most often used relationship is

$$S = kQ \tag{9}$$

where k is a supposedly constant coefficient of proportionality. Note that we have already introduced into the system two different categories of concepts. Equation (8) is a physical law stating that the volume of rainfall must be equal to the volume of outflow. Coefficient k is a parameter of unknown but constant value that is supposed to represent all the dynamics of transfer, including all the complex physical phenomena of propagation across the catchment. The introduction of (9) corresponds to a simple *conceptualisation* of an infinitely more complex reality.

Eliminating $S(t)$ between (8) and (9) leads to a simple ordinary differential equation, (10), that can be solved analytically:

$$\frac{dQ}{dt} = \frac{1}{k}(I - Q) \tag{10}$$

As an example, consider an inflow hydrograph $I(t)$ divided into time intervals of t. Suppose that during one time interval, $t_1 < t < t_2$, $\Delta t = t_1 - t_2$, the rainfall $I(t)$ can be considered constant: $I(t) = I(t_1) = I_1$. Then, during this interval the solution for the outflow is

$$Q(t) = Q(t_1)\exp\left(\frac{t - t_1}{k}\right) + I_1 k\left[1 - \exp\left(\frac{t - t_1}{k}\right)\right] \tag{11}$$

All this is generally known and trivial *except for its implicit meaning*. Indeed, the modelling concept represented by (11) is a deterministic one in as much as volume continuity is concerned. There is no way to say something *a priori* about the value of parameter k, however. Usually the value of k results from some kind of calibration. Consider a suburban catchment area which is mostly occupied by parks and playgrounds and for which long periods of data (i.e. of synchronous $I(t)$ and $Q(t)$ hydrographs) have been recorded. Then the value of k can be calibrated as, say, $k = k_1 =$ constant by a trial and error procedure reproducing past events. Now we have a model and, if there is a new input $I_A(t)$, we can compute a corresponding real-life output $Q_A(t)$ such as shown in Figure 8. Note what happens, however, if the catchment is urbanised and parks and playgrounds are transformed into parking areas, streets and houses. Then the input (rainfall) $I_A(t)$ will result in real life in the outflow shown in Figure 8 as $Q_B(t)$. But our model, with parameter value k_1, will still give the same outflow hydrograph $Q_A(t)$ as before. Thus we can conclude that such a model is deterministic and predictive in the sense that the water volume is conserved. Its predictive capacity is, however, very limited because the parameter k corresponds to the whole set of physical processes and to the whole set of characteristics of the catchment (e.g. soil occupation) and, hence, has no explicit physical meaning. This model is *not* predictive as far as engineering purposes (i.e. modifications of the catchment or consequences of exceptional input) are concerned.

One-dimensional unsteady open channel flow

The formulation taken as an illustrative example is the well known De Saint-Venant set of two equations:

- water volume continuity

$$\frac{\partial A}{\partial t} + \frac{\partial Q}{\partial x} = 0 \tag{12}$$

- energy gradient conservation

$$\frac{1}{g}\frac{\partial u}{\partial t} + \frac{\partial (u^2/2g)}{\partial x} + \frac{\partial y}{\partial x} + S_f = 0\,, \qquad S_f = \left(\frac{Q}{K}\right)^2,$$
$$K = \frac{A}{n} R_h^{2/3} \tag{13}$$

where

- $y(x,t)$ = free surface elevation and $u(x,t)$ = water velocity are dependent unknown variables;
- $A(y)$ = wetted area, $R_h(y)$ = hydraulic radius and $b(y)$ = free-surface width are fixed geometric data depending explicitly upon the free-surface elevation y; these data are considered as known from surveys, digital terrain models (DTMs), etc;
- n = Manning coefficient representing roughness of the river bed.

Equations (12) and (13) are physically sound conservation laws of water volume and energy balance of the flow. It means that, whatever happens, models built with tools based on these laws and constructed in appropriate conservation form will *always conserve these two variables* (volume and energy). A model may be inadequate or wrong if the geometric or hydraulic data are inadequate or wrong. It will not predict the future situations due, for example, to dyking of a river if the dykes are not introduced into the model as topographical modifications. But it *is* predictive in the sense that the two basic laws will always be satisfied.

Note that in (12) and (13) the only parameter that is not measurable directly is Manning's coefficient n: it is a proportionality coefficient between velocity and the square root of the slope in the Manning formula. At the scale of the model it does not replace or represent any set of physical processes. At each point of the river bed or inundated plain one can attribute one well defined value of the Manning coefficient after a visual inspection of the site. The values of this coefficient may vary within a narrow range, but hydraulic experience over the last 90 years allows us to link them to very well defined characteristics, such as the extent of silt, sand, pebbles, shrubs, prairies, sinuous streams, etc (see Chow (1959) and Barnes (1967)). The error in the estimate is normally within a few per cent of the value of the parameter.

Adequate modelling tools derived from this formulation also allow, besides the numerical solution of equations, for the input of all geometric characteristics, and for the simulation of the hydraulic effects of engineering structures, such as dykes, dams, gates, weirs, other singular head losses, etc.

Thus the models built and run with tools based on (12) and (13) are fully predictive in the engineering sense of the word because:

- Two physical laws essential for the description of one-dimensional open channel flows are always satisfied, even beyond the range of past-observed situations of the system existing in nature. They are valid even if the system has been modified by human intervention changing soil occupation, introducing structures, etc.
- Hydraulic characteristics are represented by Manning coefficients or singular head loss coefficients that can be estimated with sufficient accuracy, or at least within the well defined uncertainty range on the basis of engineering experience.
- Topographic data and man-made structures, including the singular head losses they produce, can be inserted into the model while conserving their dimensions, specific characteristics, etc; thus these features can be reproduced if they exist or can be implemented in the model when they are planned in order to study their influence.

The model/data relationship is here of a very different kind as compared to the one related to the data-driven model. A classical example is the case of modelling flood wave propagation between two stations along a river reach.

– A data-driven model allowing for prediction of the output (downstream hydrograph) when input is known can be built if a considerable number of upstream and downstream hydrographs have been observed in the past ('considerable' means statistically meaningful, which results in tens or scores of years of records). Such a lumped model is a 'passive' one: it does not tell us what happens along the reach and cannot be used beyond the range of the input/output data samples observed in the past.
– A physically based deterministic model built for the same purpose asks for topography and hydraulics characteristics *within* the reach and for synchronous upstream/downstream flow conditions, as well as, if possible, some intermediate, water stage hydrographs of a few (1–3) past observed floods. The hydrographs are needed to validate the model, to make sure that all essential features of the reach have been simulated. This model does not ask for long periods of records, it is predictive and can produce by itself the data that can be used in turn for building a data-driven model. Indeed, as much as a deterministic model is predictive, one may introduce as input the hydrographs that are exceptional, or introduce the modifications to structures, etc, and produce through simulation corresponding output hydrographs.

Note: we may observe already here that, in sharp distinction from the data-driven models, or the not fully described deterministic models (such as the reservoir model discussed above), *in principle* there is no need for any kind of calibrating procedure for models built with tools based on equations such as (12) and (13). We shall come back to this essential point below.

DATA MINING AND MODELS: THEORY FROM DATA, DATA ASSIMILATION

Theories and data

A theory is both a description and, in a sense, an *explanation* of physical situations or processes, as far as the proposer of the theory understands them. A theory may be based on hypotheses but it explains and defines causal links between the relevant inputs and outputs and the description variables. An input/output correlation, even one obtained through the most highly sophisticated methods, is not a theory. It simply contains a hypothesis that the output is correlated to the input and that there is a certain probability of the existence of a causal link between them.

Is it possible to obtain theories from the data through data-mining techniques and methods? If the question is narrowed to ask if it is reasonably possible to deduce physical theories (equations) from data-driven trained models in a 'blind way', our answer is highly sceptical. A 'blind way' means here that the deduction is based only on the data, without previous knowledge and without using physical hypotheses, and even without taking account of possible 'suggestions' for such models.

Theories in their formulation can be construed as models, i.e. they are capable of inducing a human perception of reality. Thus what is said in the following about theories applies also to models in the sense of physically based deterministic simulation models. Physical laws, represented by equations (as formulations of theories) might well be validated by the observed, measured values. The example given by (7) and Figure 2 shows this clearly. But their deduction, even when triggered by acquired data, asks for a creative mental process based on physical knowledge. It is certainly premature, and probably dangerous, to affirm that, because we master data-mining techniques, we can derive such theories automatically from the data.

We are interested in theories that are useful to engineers, that can be used by engineers for projection of the present into the future, to predict the evolution of current situations, influenced or not by human intervention. A theory is not useful or, if the readers prefer, accepted by engineers and scientists, just because it has been published, even if it is formulated and supported by some experimental evidence.

We like to judge the correctness of a theory by making measurements of what happens in nature and compare these with predictions provided by theory, but in practice this may be not possible, or at least not fully possible. Thus

we must have other criteria besides in order to judge our theories. A theory formulated to provide a model may be judged by whether it *explains* the otherwise perceived aspects of reality. But, to qualify as an explanation and thus to serve truly as a model, the theory, as pointed out by a physics' Nobel Prize '. . . . has to be simple–not necessarily a few short equations but equations that are based on a simple physical principle. . . . And the theory has to be *compelling*–it has to give us the feeling that it could scarcely be different from what it is' (Weinberg 2000). When we have a simple (i.e. clear but not necessarily short), compelling, mathematically consistent concept and its formulation that *explains* all the essential perceived aspects of reality in which we are *interested*, then we can expect to build a simulation tool that is satisfactory for the prediction of the evolution of such aspects. Such a theory, such a model, is not simply compelling–it is also *appealing*.

Consider the ancient avatars of the theory of open channel flow. Brahms (1754) seems to have been the first to realise that the slope of a uniform flow is determined by the resistance of the river bed and by the gravity force (acceleration) ΔL in a unitary length slice of the stream:

$$W = \gamma A(\Delta L) i_0 \tag{14}$$

where A = cross-sectional area, i_0 = bed slope and γ = specific weight.

P.-S. Girard (1813) states that in 1775, Chèzy was the first to assume that the resistance force is proportional to the square of mean velocity v_m, to the wetted perimeter P and to the specific weight of the fluid γ:

$$W = \frac{\gamma}{C^2}(v_m{}^2) P(\Delta L),$$

(C^{-2} being a proportionality factor) (15)

Hence, from comparison of (14) and (15), we draw the famous Chèzy formula:

$$v_m = C\sqrt{R_h i_0} \tag{16}$$

Experience showed that C, the Chèzy coefficient, is not constant and depends not only upon bed roughness but also upon water depth. This led to the Strickler (1924) formula that uses a coefficient depending solely upon the bed roughness:

$$v_m = k(R_h^{2/3})\sqrt{i_0}, \qquad k = \frac{23.7}{\sqrt[6]{k_s}} \tag{17}$$

Manning (1891), many years before Strickler, proposed an analogous formula that is not related, however, to the roughness elements but employs, rather, an empirical coefficient n allowing for a qualitative definition of the hydraulic resistance characteristics:

$$v_m = \frac{1}{n}(R_h^{2/3})\sqrt{i_0} \tag{18}$$

For river engineers, formulae such as those of Strickler/Manning or Chèzy are simple, easy to memorise and easy to apply as rules of thumb. The order of magnitude of the velocities are, in the great majority of cases, accurate enough for the first assessments of results, or, for that matter, to ascertain roughly the correctness of the output of simulation numerical models. These theories show clearly the shape of the relationship between the variables: water flow velocity, depth and slope.

A further step was the De Saint-Venant system of (12) and (13) based on the two clearest and simply expressed principles concerning a control volume of flowing water: the continuity of the mass and the equilibrium between the pressure and the inertia forces, on the one hand, and the resistance to these forces on the other. Again, these expressions are *appealing, simple* and *compelling* to use, explaining the form of the relationship between variables. At our present state of knowledge it is doubtful if such a system could be derived by any existing data-mining technology without the input of the concepts that (12) and (13) embody. It is interesting to note that the De Saint-Venant theory has been accepted since 1881 without systematic validation! Although isolated punctual studies (Favre 1935; Cunge 1966) have been carried out, it is only later (Brutsaert 1971) that a systematic validation through comparison with experiments took place. The last study is not very well known because its interest was mostly formal and academic: everybody had been sure that

De Saint-Venant equations do represent a certain reality that is of interest in one-dimensional nearly horizontal flows.

Coming back to such theories as those Chezy or Manning/Strickler, if somebody tries to obtain formulae for velocity from raw measured data without the intervention of the modeller who presumes the 'shape' of the results, only a most improbable chance would lead him to the exponent of velocity equal to 2. And then comes the problem of the definition of proportionality factor or, preferably, the description of the roughness. The Manning/Strickler, or even the Chèzy, approach, essentially based on the experience of the hydraulic engineer, is popular and again mostly used because it is *compelling*, *appealing*, *simple* and *related to the physics*. This approach is also used for pressurised pipe flows in the rough turbulent range. An example of a less compelling theory is the Colebrook–White formula (White & Colebrook 1937), (6), introduced to find a more accurate relationship between the velocity and the head loss friction coefficient λ in pipes. It is implicit, difficult to apply and memorise and impossible to use as a rule of thumb or for quick reasoning:

$$\frac{1}{\sqrt{\lambda}} = 2\log_{10}\left(\frac{k_s}{3.7D} + \frac{2.51}{\mathrm{Re}\sqrt{\lambda}}\right) \tag{6}$$

where k_s = roughness height, D = pipe diameter and Re = Reynolds number.

The Colebrook–White formula has been accepted as an implicit expression of a physical law even if it is not really compelling or appealing. At least it points out in an implicit way that there is a causal relation between the variables λ and Re and the parameters k_s and D. The simplest way to find the value of λ from (6) as a function of Re, k_s and D is to solve (6) numerically using Newton's iteration method, preferably with a quadratic rate of convergence. The results with 0.01% error are obtained after 4–6 iterations in this case. Equation (6) is widely used for pipe pressure flows with the help of some rather complicated nomographs. The version of this formulation proposed for river flows is ignored by most engineers because the Chezy, Manning and Strickler formulae are better accepted.

Consider, on the other hand, a formulation aiming at an explicit expression for λ as a function of Re, k_s and D that is 'simpler' than (6) as obtained through the use of the GP (Genetic Programming) method by Davidson *et al.* (1999). The GP data-driven model was trained on a large sample of solutions of (6), values of Re, k_s and D being considered as inputs and the corresponding calculated values of λ as outputs. The result is a variety of polynomial expressions, each expression corresponding to a certain degree of accuracy. In order to obtain λ values with a 0.01% error a polynomial expression (19) of 14 terms obtained by the GP method is to be used:

$$\begin{aligned} y = {} & 1.222\,995\,307(10^{-5})x_1^6 - 2.242\,748\,136(10^{-5})x_1^5x_2^5 \\ & - 2.482\,162\,347(10^{-4})x_1^5 \\ & + 9.286\,977\,109(10^{-6})x_1^3x_2^3 + 3.645\,038\,671(10^{-2})x_1^3 \\ & - 1.180\,446\,94(10^{-3})x_1^2x_2^2 \\ & - 0.384\,932\,3423x_1^2 + 6.598\,4017\,65(10^{-2})x_1x_2 \\ & + 2.522\,401\,137x_1 \\ & + 6.471\,827\,292(10^{-4})x_2^4 - 1.776\,888\,826(10^{-2})x_2^3 \\ & + 0.182\,981\,6121x_2^2 \\ & - 0.936\,953\,0943x_2 - 0.369\,821\,4152 \end{aligned} \tag{19}$$

where

$$x_1 = 1000k_s/\mathrm{D}, \qquad x_2 = \mathrm{Re}/10^5$$

$$y = 10\,\frac{\lambda - 0.019\,945}{0.038\,5035 - 0.019\,9435}$$

While this formulation may be useful for encapsulation of the Colebrook–White formula, it is neither compelling nor appealing to any practising engineer—the word *appalling* seems to be more appropriate. We observe that, since it does not express anything, or have any evident meaning, it does not constitute a model.

Data assimilation within the context of forecast and deterministic modelling

There is no room here for an in-depth analysis of questions related to the problems of data assimilation, even within such a limited context as forecasting and deterministic modelling. Only a few main points can be roughly sketched. Data assimilation in such contexts concerns

methodologies that can extract information from observations and combine it with, or assimilate it into, numerical models, thus supplying more accurate or more reliable results than the models by themselves. The way of presentation used here is borrowed from Babovic *et al*. (2001), who based themselves on WMO (1992) and Refsgaard (1997): the interested reader will find more details and a development of these ideas, as well as applications, in these references. Babovic *et al*. (2001) list four methodologies of data assimilation in the domain of forecast: this is thought sufficient to show the link between the data and the use of models in this field.

The list of methodologies will be illustrated using an example of a flood forecasting system based on a one-dimensional deterministic model of unsteady flow propagation along a river reach. Assume that all necessary data, observed during a sufficient number of past flood events, are available and that the events have also been simulated by a one-dimensional deterministic model which can be considered as reliable. In real time, during a flood, at every forecast time, the deterministic model is run in order to forecast, at the forecast times T_i ($i = 1, 2, \ldots n$) the output (water stages and discharges) at a number of locations along the river during the forecast period. The model runs can be used in conjunction with data assimilation, the data being recorded either before the current flood, during previous events, and/or observed during a short period ($T_i - k\Delta t$) preceding the forecast time T_i. The four methodologies, all starting with the difference between observed and forecast values during the preceding forecast period (T_{i-1}, T_i), are as follows:

- Updating input parameters. For example, the inflows of tributaries and upstream incoming discharges can be increased if the water stages forecast during the period (T_{i-1}, T_i) turn out to be lower than observed. The underlying idea is that the inflow estimates are the most uncertain.
- Updating of state variables. The variables are water stages and discharges computed by the deterministic model. The methodology would consist of an adjustment of the initial state of these variables at the forecast time and then all through the forecast period, using such techniques as Kalman filtering.
- Updating of parameters. In our example this would correspond to varying such parameters as the Manning coefficients in the model (the coefficients that had been considered as reliable), the modification criterion being the minimisation of differences between forecast and observation during the last (T_{i-1}, T_i) period.
- Updating the output variables. This methodology is based on the existence of a *model of errors*. Such a model can be built using the deviations between observed past events (floods) and the forecast, repeatedly simulated at a number of times T_i during the same events with the same deterministic model. Note that this is always a data-driven model. Then the actual real-time forecast is carried out in two stages: one using the deterministic model that supplies a first estimate of forecast stages and discharges; then the second, using the model of errors that supplies corrections to the forecast.

It is the author's opinion that only the fourth methodology seems to limit possible 'mechanical' intervention. The first three open the doors to purely formal corrections and adjustments, some of them without any reasonable physical basis and, hence, it is impossible to interpret or assess their value. As for the fourth (most often used in practice) it is founded on data-driven error modelling and hence subject to everything that has been written above about such models. The author would like to stress that this methodology is extremely useful and should certainly be used whenever possible in real-time forecast systems. It needs, however:

- sufficient data recorded during past-events,
- extensive simulation of the forecast during these past events using exactly the same modelling tool as the one used in real time, this in order to create sufficient data samples for a reliable data-driven model of errors;
- correct physical analysis of the phenomena involved and of differences between the simulated forecast and the recorded data.

DATA AND MODELLING METHODOLOGY: 'GOOD PRACTICE' IN QUESTION

'Good practice' is really the state of the art at a given time (historically speaking). In engineering it is nothing other than a certain way to apply norms, carry out calculations, etc, that is recognised as 'acceptable and correct' by the profession. It is presumed that, if an engineer follows 'good practice' rules, his project will not fail, the bridge will not collapse and the dam will not break. And, if a catastrophe occurs, notwithstanding following 'good practice', it is usually, quasi-automatically, considered as an 'Act of God', or at any rate not the engineer's fault. So much so that, when there are doubts expressed concerning such and such an engineering action or proposal and questions are asked, the answers such as 'according to the best engineering knowledge' or 'following the state of the art' are considered as definitive. If we analyse these two expressions, however, we find them meaningless and irrelevant. They simply mean that the person who used them learnt at school to follow such and such a procedure, or that she/he does what everybody else does, without necessarily knowing why or even knowing if the procedure is applicable to the situation that he or she has at hand.

What the author would like to express is that 'good practice' may be an important safety net for professionals, but it can prove inefficient or, worse, be a cause of a disaster. It can be a barrier of obscurantism resisting progress to new approaches and technologies, because the latter, of course, are not encompassed by such 'good practice', which is essentially based on past experience. Just for the sake of being relevant, let me evoke:

- Tacoma Narrows Bridge, built according to 'good practice', but which did not integrate already well-established experience of the dynamic behaviour of such structures;
- Freyssinet's pre-stressed concrete that eventually revolutionised the reinforced concrete industry but did not correspond to the usual reinforced concrete code of 'good practice';
- the idea that 'when the data is scarce, a black box model is more appropriate than a deterministic physically based one, the latter being too complicated and demanding too many data'.

The last statement is still (even today) very much considered as 'good practice' in modelling. Even more so because it is based on a 'scientific' argument: since the data are poor, let us be homogeneous and use the model that is coarse. Such an approach adds the certainty of a poor quality of modelling to the uncertainty of the data and the synergy of the two is likely to result in a very poor, if not absurd, representation of reality. Thus this apparently 'logical' argument proves to be pseudo-scientific, when the uncertainty of the final result is analysed.

In true 'good practice', a lack of adequate data necessitates the use of the most advanced and reliable modelling tools over the greatest range of uncertainty in the data in order to compute–assess the uncertainty in the results of the coarse model. And thus the cost is *increased* by inadequate data.

Current modelling paradigm

In modelling practice we have a well-established 'good practice' principle taught at universities, engineering schools and in general assumed as a dogma. It says that model application is to be carried out in four stages:

(i) *Instantiation* or set-up or 'construction'. This consists in defining such features and parameters as discretisation, computational grid, limits and boundary conditions; an introduction of topography, soil occupation, structures, initially assessed values of roughness coefficients, etc.

(ii) *Calibration*, which consists in executing a number of simulations of past observed events and in varying the parameters of the model until an acceptable (to the modeller) coincidence between observations and computations is obtained.

(iii) *Validation*, which consists in executing with a calibrated model a number of simulations of past observed events (different from those used for calibration) and checking to see if the simulated results are sufficiently close to observation.

(iv) *Exploitation* runs (studies) with the model recognised as a validated tool.

These four stages historically come from hydrological correlative or black box modelling practice. They are a natural and indisputable approach when data-driven 'models' are concerned. This approach is the very essence of such 'models',[2] parameters of which in most cases have no physical meaning. Such 'models' cannot explain what is going on within the 'modelled' system: they do not describe the interaction of processes within the system. What counts is the 'training', i.e. the calibration of these parameters in such a way that, for given inputs, computed outputs correspond to observed ones.

Many true models are in reality composed of both deterministic and data-driven or black box correlative parts. Some of the processes within the model or modelling tool may well be described by physical laws and equations but others, within the same model, may not. The MIKE SHE modelling tool provides a deterministic approach for all processes of water transfer through a catchment, but it allows the user to replace some of the deterministic components by others that are of a data-driven type. This option is often used when there is no need for some components to be predictive. A similar situation arises for other applications:

- Modelling river floods often involves a model that is a combination of a deterministic component of river flow simulation and of a data-driven rainfall/runoff component (e.g. MIKE11 and NAM).
- Modelling pollutant fate in groundwater flow often involves deterministic components, such as the water flow itself and the advection–diffusion of the pollutant, but also data-driven ones, such as the adsorption of the pollutant in unsaturated zones.
- etc.

Clearly, in all such cases the data-driven components have to be calibrated ('trained') as explained above. And it is important to realise that they must be calibrated *separately* from the overall tool and, hence, there is the need for data that concerns only these components, as well as the methodology to calibrate them as specific components. It is not easy to calibrate some components as data-driven models and then consider others as deterministic ones! However, we still have to do so with a model in the sense that we can always trace an expressive property or meaning to the productions of the tool concerned.

When we consider, however, deterministic modelling (based on physical laws describing simulated processes and their interactions), this four-stage paradigm is not only illusory as a way of increasing accuracy but it may also lead to dubious and unreliable results. Hence it should be abandoned and a *modified paradigm is to be applied when physically based deterministic models are concerned*. More precisely, the calibration stage should be eliminated from the paradigm while the validation stage, as compared to current practice, should be carried out in a different way and in a different spirit. We shall illustrate this position in the following and by examples of current practice.

If to calibrate, what is to be calibrated?

When the calibration of *deterministic* models (or deterministic components of models) is considered, one may ask oneself: *what is to be calibrated*? What is modified during the calibration procedure? An obvious principle (often violated in practice) is that the calibration must be limited to the model parameters that are *invariant* between the instantiation and exploitation stages, unless the purpose is to study the sensitivity of the model to modifications in its parameters. To calibrate parameters that will subsequently be modified during the exploitation runs used for simulating the impact of future projects is most often a useless, as well as costly, exercise. It is better by far to let a model be truly deterministic, i.e. a model without 'inner black boxes' describing physical processes, the only parameters of which are empirical or experimental coefficients related to physical characteristics. For example, in open channel flows in rivers, such parameters are roughness Manning/Strickler/Chezy coefficients, singular head-loss coefficients and discharge coefficients of structures (weirs, gates, culverts, etc). But certainly not

[2]We place the word 'model' between quotes here because, having no expressive sign or meaning, these are not models in any standard dictionary sense.

topography, dyke elevations, operations rules for structures, etc. Roughness coefficients are *invariant* parameters between calibration and exploitation stages, unless the exploitation concerns projects that could modify them.

Cleaning up and dredging a badly maintained river stretch invaded by vegetation will necessitate the modification of the roughness coefficients. In this situation, roughness coefficients cannot be invariant. In order to study the impact of cleaning, one may wish to calibrate the current coefficients (before cleaning) in order to ascertain that the model reproduces the present conditions accurately. But to assess the impact of the change one has to modify the coefficients for the future situation and there is no way to calibrate these new values: they are defined through an engineering assessment. Any calibration of 'global' head-loss coefficients along a stretch including features, the characteristics of which vary between a calibrated situation and exploitation runs (structures, sills, narrowing or widening of the river bed, etc), may lead to a nonpredictive black box model.

Is a meaningful calibration feasible?

Another question to be asked and considered: in practice, is a *meaningful calibration possible*? It should be clear that the answer is negative, at least for most cases, because of the lack of appropriate data, or the cost of their acquisition. It can be shown in examples how 'obvious' applications of a paradigm including calibration may well lead to serious errors because of the belief that calibration is meaningful or because of a wrong choice of calibrated parameter. One problem area is that of open (sea) boundary conditions for two-dimensional tidal estuarine models, as introduced earlier. There are still modellers who impose tidal free-surface elevations at open boundaries and 'calibrate' roughness coefficients within the model using stage hydrographs recorded at a few shore stations. However, as explained earlier in this paper, water elevations within the modelled area, both in reality and in the model, follow the variations of the boundary elevations, which means that even large variations in roughness coefficients of the model have little influence on elevations that are compared with observed values. As mentioned before, the main parameters that are calibrated to make models of this kind *reliable* (Abbott & Cunge 1975) are the variation of in-flowing and out-flowing discharge hydrographs and their distribution at the open seaward boundary. Roughness coefficients within the domain must be assessed using the engineering experience of the modeller and then they can only be modified rarely and only locally. It is legitimate to calibrate the discharge distribution at an open sea boundary of a tidal model for the 'project tide' because the same distribution and the same 'project tide' would be applied to exploitation runs: this distribution is *invariant* for the model. It is futile to do so if projects proposed for the estuary could influence the discharge distribution at the boundary.

Another classic example concerns one-dimensional river modelling. One-dimensional models of rivers have a typical resolution of computational grids between 100–1,000 m, with distances between gauging stations where the water stages are recorded being of the order of 10 000 m. Thus along a 50 km channel there might be four calibration sections (boundary conditions excluded). In open channel flow engineers can evaluate values of roughness and head-loss coefficients by inspection, within a narrow range of error. If a visual inspection of the river stretch suggests a Manning coefficient of 0.03, it is easy to accept that the actual value of the coefficient may vary between, say, 0.025 and 0.035. If, however, the calibration of roughness (coincidence between computed and observed water stages at gauges at a distance of some 10 000 m) leads for this reach to values such as 0.04 or 0.05, this is unacceptable. Indeed, such a river bed would be, according to the Strickler formula, covered with equivalent roughness elements of diameters 0.78 or 3.00 m high! The only possible conclusion in such a case is that the model does not reproduce reality and that the calibration is meaningless. Obviously, when instantiating the model for this case, something has been forgotten: a bridge, a singular head loss, river shape-induced head losses, the appropriate representation of an inundated plain, etc. Another possibility is that the river geometric characteristics are not correct in the model and calibration gives absurd values because it compensates for narrows or for sills that influence more the surface elevations than does the roughness. Or, worst of all, the model is based on equations that do not describe

adequately the physical process, such as fixed-bed equations applied to alluvial bed rivers, or a diffusive wave equation model applied to downstream-influenced or inertia-dominated flows. At any rate, from the point of view of predictivity and future exploitation, the calibration effort is futile and useless.

Other typical examples of meaningless 'calibration' (should we say mindless 'fitting'?) of parameters until a coincidence between computed and observed free-surface elevations is reached are:

- Two-dimensional modelling of inundated plains. Indeed, the only past-observed data concerning the unsteady evolution of water stages that can be found on inundated plains are those rare marks of the highest elevations attained during historical floods. The only one known to the author—and a never repeated historical case—where the records were adequate for calibration purposes of such a situation was in the case of the Mekong Delta Model (Zanobetti *et al*. 1968; Cunge 1975). There were 350 computational points and three consecutive floods (1963, 1964 and 1965) were recorded at 300 gauges located over the modelled area. The cost of the modelling and measurement campaigns was over US$1 million (at 1963 values: this would be ten times more in 2002). This number alone shows that this approach would not be repeated today.
- Two-dimensional modelling of tidal coastal areas, for the same reason: the records are available at only a very few stations, generally located at the coast where conditions are specific and the tide is deformed. Thus the calibration of large areas through fitting computed and observed results may well be meaningless because the calibration criteria for large domains are really dependent upon the local effects of features located near stations.
- Groundwater flow modelling provides several flagrant examples, of which there is only space here for one. Following the current 'good practice' paradigm the permeability coefficients are systematically calibrated to obtain a coincidence between observed and computed evolutions of piezometric levels. In practice, the resolution of the network of piezometric recording wells, as compared to the resolution of computational points over the modelled domain, is always very small. It is rarely dense enough to make it certain that it 'captures' the possible main variations of geological conditions and of permeability parameters over the domain. Consider, on the other hand, the duration of the measurement period (at best 10 years) as compared to the time needed for the flow to cross the modelled area (often 50–100 years). The chances are great indeed that the modeller, by calibrating the parameters on the basis of 10 year records, 'twists the arm' of reality, and that the resulting 'calibrated' model is not a reliable image of reality at all. Once more the calibration is meaningless.

Why then, in the current paradigm, is calibration considered as a necessary step?

There are several reasons, such as:

- The end user's or client's measure of satisfaction. How does she or he know that the model is correct? Answering that the calibration reproduced past-observed results accurately makes him or her happy because, in the absence of any understanding of the inner mechanisms of the modelling he or she is reassured by the image of two coinciding curves.
- The end user (or client) might have spent a lot of money on collecting data, carrying out measurements, etc, and does not like the idea that this money was spent for nothing. It makes him or her happy to show that this data is being proved useful for the project: 'these data were essential for a calibration that in turn is essential for proving the accuracy of the model . . .'.
- Suppose problems arise with a project built on the basis of modelling studies where nature did not behave as the model predicted. Then a modeller has arguments in his favour if he followed 'good practice' and if the calibration was carried out which led to the coincidence of past-observed and computed results.

In most specific cases, as developed above, it can be shown that these reasons are fallacious. To take a very crude intellectual shortcut, one may attempt to say that the calibration is still a common practice because it makes both sides happy: the modeller (who may estimate his or her intellectual effort as finished when the model is 'calibrated') and the end-user/client who feels that his or her duty of control and supervising has been done. Neither realises that their satisfaction is so often related to a formal coincidence and not to any understanding of the physical problems, with this last criterion as the most important point for projects and future developments, and the very reason for commissioning the model at all. This is to say that the technology in such a case is not directed to an understanding of the underlying phenomena, but only to persuading an end-user or client that something of value has been done. It thus corresponds to the technologies of persuasion in their most negative sense (Abbott 2002*a*).

A proposed modified paradigm

The modified paradigm identifies the following stages in the modelling process:

- *Instantiation* or set-up or 'construction' of the model; definition of the methodology necessary to define the range of uncertainty in the results of the computations.
- *Validation*, which consists of executing a number of simulations of past-observed events with the model, computing or otherwise finding the range of uncertainty for the results and analysing and finding physically logical reasons for differences between the simulated and observed results. After this, analysing the impact of the differences as well as of the uncertainties upon the exploitation results.
- *Exploitation runs (studies)*: supplying the results and impacts *and* their range of uncertainty to the end-user or client in a comprehensible form.

The modified paradigm for deterministic modelling as proposed above eliminates the calibration stage as such. The validation stage is not only maintained, but reinforced. In a way it incorporates the calibration stage. The past measured data will, of course, be as useful and as necessary as for calibration under the currently admitted paradigm but they will be used for a *validation analysis of the computed results*. It is claimed that a deterministic model, with values of parameters defined by inspection on the basis of engineering practice, should simulate reality correctly and its results should be close to past observed results *without* calibration in its irrational sense. 'Close' does not mean an immediately satisfactory coincidence. But making computed results nearer and nearer the observed ones must not be carried out through a calibration process as it is currently understood and applied. Indeed:

- If the differences between the computed and the observed lie within an acceptable interval of uncertainty, or can be explained by physical reasons, and if the consequences of differences upon exploiting the model as it is are analysed and acceptable, then there is no reason to go any further with the modification of parameters.
- If the differences are greater than the uncertainty interval, than they must be explained. The reasons must be found and analysed, taking into account, once more, the consequence of using the model as it is or amending it. Most often the findings lead to modifications of originally erroneous data, such as topography, hydraulics characteristics or boundary conditions, and have not much to do with parameters. Sometimes there are factually important errors in values of parameters assessed during a visual inspection. But, sometimes, one may find that the modelling tool is not adequate: such often occurs when using 1D models where only 2D can simulate the real flows.

The new paradigm insists on the fact that this modified approach is *not* a calibration under a new name. The new 'good practice' asks for the collection and analysis of data for the purpose of validation, and validation is not just a check that computed values are not very far from observed ones: it is a study of the *reasons* why there is a difference between the two! Also it must, of course, be substantiated by a report leading to an understanding how such an analysis was carried out and how the conclusions were reached.

Paramount to the successful acceptance of this modified paradigm by the engineering profession as 'good practice' are the following conditions:

- A broad understanding of the difference, with respect to calibration, between deterministic and data-driven or black box models. A clear understanding of the difference, within a model, between an empirical parameter and a black box replacing a process.
- An acceptance that the simulation results of past events do not reproduce exactly the on-site measurements. A requirement for an engineering analysis of the differences.
- An acceptance, and even a requirement, that the results of models should be presented as uncertain and with a sound evaluation of the uncertainty range or interval.

The need for this new approach, and for this modified paradigm, is not fully recognised and not yet instituted as good practice, but the first descriptions of and information on applications and studies that follow these ideas are beginning to become available. One such study is the modelling of flood propagation across an inundated plain in South-Western France (Sauvaget *et al*. 2000). Interested readers are encouraged to read this reference. The case is interesting precisely because a calibration, in the usual sense, would be meaningless. The link between the requirements put on the reliability of the results, on the one hand, and the means necessary to analyse and ensure an accessible degree of reliability, on the other hand, is described. It is worth noting that, even in this case the authors thought it useful to insert one paragraph titled 'Model calibration', although there was no calibration in the traditional sense. Was this the result of 'professional pressure' to be 'politically correct' and to follow the 'good practice' paradigm?

CONCLUSIONS

During a single decade (from the advent of numerical deterministic modelling at the end of the 1950s up to the end of the 1960s) the water engineering profession witnessed a veritable war between those who believed in the superiority of the reduced scale modelling technology and those who thought that numerical simulation would eradicate laboratories. Only towards the end of the 1970s did most practitioners understand that these technologies were complementary, and it was up to the hydraulic engineer to use one or other in his or her own design and needs. The corollary was that the engineer had to know the qualities and limitations of both, since both were now available tools and since the appropriate choice of available tools is an attribute of a good professional. Today we have to avoid a war between data-driven and data-mining technologies, on the one hand, and the deterministic modelling approach, on the other hand.

The leitmotif of the present paper is also its conclusion: the purpose of exploitation (or, if readers prefer, of mining) of the data, as well as that of building and running models, is to produce information that is *meaningful* to the end-users. *Hence the need for a new paradigm in deterministic modelling based on validation and not on calibration.* Whatever the data are, whatever are the models used, the value of the results can only be measured in terms of physical, engineering, environmental or other criteria meaningful for the end-user, and *not* in terms of the sophistication of the enabling technologies and methods. Clearly the developers must know their techniques, but this is only one part of the story. The value of the tools, the value of the data and eventually *the value of information* depends upon understanding both by those who develop and by those who apply the tools of physics, represented by the data and models, on the one hand, and the requirements of the end-users, on the other hand. There is a lot of truth in what was said during the 1914–18 war by the then French Prime Minister, Georges Clemenceau: 'War is too serious a matter to be left to the generals'. Engineering conclusions and decisions based on modelling are too serious a matter to be left to the developers of algorithms and software alone.

ACKNOWLEDGEMENTS

The author is deeply grateful to his friend and colleague, M. B. Abbott, for reviewing and criticising the text. He

made a number of suggestions that made the content of the manuscript valuable. Last, but not least, his considerable effort in improving the English of the text is acknowledged and highly appreciated.

REFERENCES

Abbott, M. B. 2002*a* On definitions. *J Hydroinformatics* **4** (2), electronic version.

Abbott, M. B. 2002*b* Private communication.

Abbott, M. B. & Cunge, J. A. 1975 Two-dimensional modeling of tidaldeltas and estuaries, case study: River Seine Estuary Coarse Grid Model. In *Unsteady Flow in Open Channels* (ed. K. Mahmood & V. Yevjevich) Chapter 18.8, pp. 795–799. Water Res. Publications.

Babovic, V., Canizares, R., Jensen, H. R. & Klinting, A. 2001 Neural networks as routine for error updating of numerical models. *J. Hydraul. Engng. ASCE* **127** (3), 181–193.

Bakhmeteff, B. 1941 *Mécanique de l'écoulement turbulent des fluides*. Dunod, Paris.

Barnes, H. H. Jr 1967 Roughness characteristics of natural channels. US Geological Survey Water Supply Paper 1849. US Government Printing Office, Washington, DC.

Blasius, H. 1913 Das Ahnlichkeitsgesetz bei Reibungsvorgängen in Flussigkeiten. *VDI-Forschung*. No 131.

Brahms, A. 1754 *AnPangs- Gründe der Deiche- und Wasserbaukunst*. Aurich, Hanover.

Brutsaert, W. 1971 De Saint-Venant equations experimentally verified. *J. Hydrol. Div., ASCE* v.97, HY9, September, pp. 1387–1401

Chow, V. T. 1959 *Open-Channel Hydraulics*. McGraw-Hill, New York.

Cunge, J. A. 1966 Comparaison des résultats d'essais d'intumescences effectués sur le modèle réduit et sur le modèle mathématique du canal Oraison-Manosque (Comparison of physical and mathematical model test results on translation waves in the Oraison-Manosque power canal). *La Houille Blanche* No. 1, 55–70.

Cunge, J. A. 1975 Two-dimensional modeling of flood plains. In *Unsteady Flow in Open Channels* (ed. K. Mahmood & V. Yevjevich), Chapter 17.9, pp. 731–750. Water Res. Publications.

Davidson, J. W., Savic, D. & Walters, G. A. 1999 Method for the identification of explicit polynomial formulae for the friction in turbulent pipe flow. *J. Hydroinformatics* **1** (2).

Eck, B. 1941 *Technische Strömungslehre*. Berlin, p. 100.

Favre, H. 1935 *Ondes de translation dans les canaux découverts*. Dunod, Paris.

Girard, P.-S. 1813 *Mémoire de la Classe de sciences de l'Institut de Paris*.

Hagen, G. 1854 Ueber den Einfluss der Temperatur auf die Bewegung des Wassers in Röhren (On the influence of temperature on the movement of water through pipes). *Math. Abh. Akad. Wiss. Berlin* 17.

Hamil, L. 1995 *Understanding Hydraulics*. Macmillan, London.

Jaeger, C. 1954 *Hydraulique Technique*. Dunod, Paris. (Original in German: 1949 *Technische Hydraulik*. Verlag Birkhaüser AG, Basel.)

Jonoski, A. 2002 Hydroinformatics as sociotechnology: promoting individual stakeholder participation by using network distributed decision support systems. *PhD Dissertation*, IHE, Delft. Swets & Zeitlinger BV, Lisse, The Netherlands.

Klemes, V. 1986 Operational testing of hydrological simulation models. *Hydrol. Sci. J. Sci. Hydrol.* **31** (1), 13–24.

Manning, R. 1891 On the flow of water in open channels and pipes. *Trans. Inst. Civil Engrs. Ireland* **20**, 161–207; 1895 *Suppl.* **24**, 179–207.

Nikuradse, J. 1933 Strömungsgesetze in rauhen Rohren, *VDI-Forschung.* No 361.

Prandtl, L. 1933 Neuere Erbgnisse der Turbulenzforshcung. *Z.V.D.I.* **12**.

Prandtl, L. & Tietjens, O. G. 1934 *Applied Hydro- and Aeromechanics* (by Tietjens, based on lectures of Prandtl). United Engineering Trustees Inc. (Unabridged reproduction by Dover Publications, New York, 1957).

Refsgaard, J. C. 1997 Validation and intercomparison of different updating procedures for real-time forecasting. *Nordic Hydro.* **28**, 65–84.

Reynolds, O. 1883 An experimental investigation of the circumstances which determine whether the motion of water will be direct or sinuous, and of the law of resistance in parallel channels, *Phil. Trans. R. Soc. London* 1883, or *Sci. Papers* vol 2, p.51.

Sauvaget, P., David, E., Demmerle, D. & Lefort, P. 2000 Optimum design of large flood relief culverts under the A89 motorway in the Dordogne-Isle confluence plain, *Hydrol. Process.* **14**, 2311–2329.

Strickler, A. 1923 Beiträge zur Frage der Geschwindigkeitsformel und der Rauhigketiszalen für Ströme, kanäle und geschlossen Leitungen. Eidg. *Wasserwirtschaftsamt Bern*, Mitteilung 1923, SB, **83** (23).

Weinberg, S. 2000 Will we have a final theory of everything? *Time* **155** (14), April 10, 52.

White, C. M. & Colebrook, C. F. 1937 Fluid friction in roughened pipes. *Proc. R. Soc. A* **161**, 367.

World Meteorological Organisation (WMO) 1992 Simulated real-time intercomparison of hydrological models. *Operational Hydrology Report* No 38, Geneva.

Zanobetti, D., Lorgéré, H., Preissmann, A. & Cunge, J. A. 1968 Le modèle mathématique du Delta du Mékong (Mathematical model of Mekong Delta). *La Houille Blanche* Nos 1, 4 and 5.