Algorithm for Model Validation: Theory and Implementation

D. Sornette[1,2], A. B. Davis[3], K. Ide[1,4], K. R. Vixie[5], V. Pisarenko[6], and J. R. Kamm[7]

[1] Institute of Geophysics and Planetary Physics, University of California, Los Angeles, CA 90095, USA

[2] Department of Earth and Space Sciences, University of California, Los Angeles, CA 90095, USA
and Laboratoire de Physique de la Matière Condensée, CNRS UMR 6622,
Université de Nice-Sophia Antipolis, 06108 Nice Cedex 2, France

[3] Los Alamos National Laboratory, Space and Remote Sensing Sciences Group (ISR-2),
Los Alamos, NM 87545, USA

[4] Department of Atmospheric Sciences, University of California, Los Angeles, CA 90095, USA

[5] Los Alamos National Laboratory, Mathematical Modeling and Analysis Group (T-7),
Los Alamos, New Mexico 87545, USA

[6] International Institute of Earthquake Prediction Theory and Mathematical Geophysics,
Russian Academy of Sciences, Warshavskoye sh., 79, kor. 2, Moscow 113556, Russia

[7] Los Alamos National Laboratory, Materials Science Group (X-7),
Los Alamos, New Mexico 87545, USA

Abstract: Validation is often defined as the process of determining the degree to which a model is an accurate representation of the real world from the perspective of its intended uses. Validation is crucial as industries and governments depend increasingly on predictions by computer models to justify their decisions. We propose to formulate the validation of a given model/code as an iterative construction process that mimics the often implicit process occurring in the minds of scientists. We offer a formal representation of the progressive build-up of trust in the model. We thus replace static claims on the impossibility of validating a given model/code by a dynamic process of constructive approximation. This approach is better adapted to the fuzzy, coarse-grained nature of validation. Our procedure factors in the degree of redundancy versus novelty of the experiments used for validation as well as the degree to which the model predicts the observations. We illustrate the new methodology first with the maturation of Quantum Mechanics as the arguably best established physics theory and then with several concrete examples drawn from some of our primary scientific interests: a cellular automaton model for earthquakes, a multifractal random walk model for financial time series, an anomalous diffusion model for solar radiation transport in the cloudy atmosphere, and a computational fluid dynamics code for the Richtmyer-Meshkov instability.

Introduction: Model Construction and Validation

At the heart of the scientific endeavor, model building involves a slow and arduous selection process, which can be roughly represented as proceeding according to the following steps: (1) start from observations and/or experiments; (2) classify them according to regularities that they may exhibit [1]; (3) use inductive reasoning, intuition, analogies, and so on, to build hypotheses from which a model is constructed; (4) test the model with available observations, extract predictions that are tested against new observations or by developing dedicated experiments. The model is then rejected or refined by an iterative process, a loop going from (1) to (4). A given model is progressively validated by the accumulated confirmations of its predictions by repeated experimental and/or observational tests. Whereas *verification* deals with whether the simulation code correctly solves the model equations, *validation* carries an additional degree of belief in the value of the model vis-à-vis experiment and, therefore, may convince one to use its predictions to explore beyond known territories [2].

The validation of models is becoming a major issue as humans are increasingly faced with decisions involving complex tradeoffs in problems with large uncertainties, as for instance in attempts to control the growing anthropogenic burden on the planet [3] within a risk-cost framework [4] based on predictions of models. For policy decisions, federal, state, and local governments increasingly depend on computer models that are scrutinized by scientific agencies to attest to their legitimacy and reliability. Cognizance of this trend and its scientific implications is not lost on the engineering [5] and physics [6] communities.

How does one validate a model when it makes predictions on objects that are not fully replicated in the laboratory, either in the range of variables, of parameters or of scales? Indeed, a potentially far-reaching consequence of validation is to give the "green light" for extrapolating a body of knowledge, which is firmly established only in some limited ranges of variables, parameters and scales. Predictive capability is what enables us to go beyond this clearly defined domain into a more fuzzy area of unknown conditions and outcomes. This problem has repeatedly appeared in different guises in practically all scientific fields. A notable domain of application is risk assessment: how to quantify the potential for a catastrophic event (earthquake, tornado, hurricane, flood, huge solar mass ejection, large bolide, industrial plant explosion, ecological disaster, financial crash, economic collapse, etc.) of amplitude never yet sampled from the knowledge of past history and present understanding? This is crucial, for example, in the problem of scaling the physics of material and rock rupture tested in the laboratory to the scale of earthquakes. This is necessary for scaling the knowledge of hydrodynamical processes quantified in the laboratory to the length and time scales relevant to the atmospheric/oceanic weather and climate, not to mention astrophysical systems. Perhaps surprisingly, the same problem arises in the evaluation of electronic circuits; to quote Hefner [7]: "The problem is that there is no systematic way to determine the range of applicability of the models provided within circuit simulator component libraries." The example of validation of electronic circuits is particularly interesting because it identifies the origin of the difficulties inherent in validation: the fact that the dynamics are nonlinear and complex with threshold effects, that it does not allow for a simple-minded analytic approach consisting in testing a circuit component by component. This same difficulty is found in validating general circulation models of the Earth's climate or end-to-end computer simulations of complex engineering systems such as an aircraft or a nuclear weapon. The problem is fundamentally due to its systemic nature. The theory of systems, sometimes referred to as the theory of complex systems, is characterized by the expectation of surprises. The biggest one may be the phenomenon of "emergence" in which qualitatively new processes or structures appear in the collective behavior of the system, while they cannot be derived or guessed from the behavior of each element. The phenomenon of "emergence" is similar to the philosophical law on the "transfer of the quantity into the quality." A full control of the validation process therefore requires an attempt to account for this emergence phenomenon.

Impossibility Statements

For these reasons, the possibility to validate numerical models of natural phenomena, often endorsed either implicitly or identified as reachable goals by natural scientists in their daily work, has been challenged;

quoting Oreskes et al. [8]: "Verification and validation of numerical models of natural systems is impossible. This is because natural systems are never closed and because model results are always non-unique." According to this view, the impossibility of "verifying" or "validating" models is not limited to computer models and codes but to all theories that rely necessarily on imperfectly measured data and auxiliary hypotheses, as Sterman et al. [9] put it: "Any theory is underdetermined and thus unverifiable, whether it is embodied in a large-scale computer model or consists of the simplest equations." Accordingly, many uncertainties undermine the predictive reliability of any model of a complex natural system in advance of its actual use.

Such "impossibility" statements are reminiscent of other "impossibility theorems." Consider the mathematics of algorithmic complexity [10], which provides one approach to the study of complex systems. Following reasoning related to that underpinning Gödel's incompleteness theorem, most complex systems have been proved to be computationally irreducible, i.e., the only way to predict their evolution is to actually let them evolve in time. Accordingly, the future time evolution of most complex systems appears inherently unpredictable. Such sweeping statements turn out to have basically no practical value. This is because, in physics and other related sciences, one aims at predicting *coarse-grained* properties. Only by ignoring most of molecular detail, for example, did researchers ever develop the laws of thermodynamics, fluid dynamics and chemistry. Physics works and is not hampered by computational irreducibility because we only ask for approximate answers at some coarse-grained level [11]. By developing exact coarse-grained procedures on computationally irreducible cellular automata, Israeli and Goldenfeld [12] have demonstrated that prediction may simply depend on finding the right level for describing the system. Similarly, we propose that validation is possible, to some degree, as explained below.

Validation and Hypothesis Testing

We start by recognizing that validation is closely related to hypothesis testing and statistical significance tests of mathematical statistics [13]. In hypothesis testing, a null H_0 is compared with an alternative hypothesis H_1, in their ability to explain and fit data. The result of the test is either to "reject H_0 in favor of H_1" or "not reject H_0." One never concludes "reject H_1," or even "accept H_0 or H_1." If one concludes "do not reject H_0," this does not necessarily mean that the null hypothesis is true, it only suggests that there is not sufficient evidence against H_0 in favor of H_1; rejecting the null hypothesis may suggest but does not prove that the alternative hypothesis is true, only that it is better given the data. Thus, one can never prove that an hypothesis is true, only that it is less effective in explaining the data than another hypothesis. One can also conclude that an hypothesis H_1 is not necessary and the other more parsimonious hypothesis H_0 should be favored. The alternative hypothesis H_1 is not rejected, strictly speaking, but can be found unnecessary or redundant with respect to H_0. This is the situation when there are two (or several) alternative hypotheses H_0 and H_1, which can be composite, nested, or non-nested (the technical difficulties of hypothesis testing depends on these structures of the competing hypotheses [14]). This illuminates the status of code comparison in verification and validation [15]. Viewed in this way, it is clear why code comparison is not sufficient for validation since validation requires comparison with experiments and several other steps described below. The analogy with hypothesis testing makes clear that code comparison allows the selection of one code among several codes but does not help to conclude about the validity of a given code or model when considered as a unique entity independently of other codes or models.

In the theory of hypothesis testing, there is a second class of tests, called "tests of significance," in which one considers a unique hypothesis H_0 (model), and the alternative is "all the rest," i.e., all hypotheses that differ from H_0. In that case, the conclusion of a test can be the following: "this data sample does not contradict the hypothesis H_0," which is of course not the same as "the hypothesis H_0 is true." In other words, a test of significance cannot "accept" an hypothesis, it can only fail to reject it because the hypothesis is found sufficient at some confidence level for explaining the available data. Multiplying the tests will not help in accepting H_0.

Since validation must at least contain hypothesis testing, this shows that statements like "verification

and validation of numerical models of natural systems is impossible" [8] are best rephrased in the language of mathematical statistics [13]: the theory of statistical hypothesis testing has taught mathematical and applied statisticians for decades that one can never prove an hypothesis or a model to be true. One can only develop an increasing trust in it by subjecting it to more and more tests which "do not reject it." We attempt to formalize below how such trust can be built up to lead to validation viewed as an evolving process.

Validation as a Constructive Iterative Process

In a standard exercise of model validation, one performs an experiment and, in parallel, runs the calculations with the available model/code. Then, a comparison between the measurements of the experiment and the outputs of the model/code calculations is performed. This comparison uses some metrics controlled by experimental feasibility, i.e., what can actually be measured. One then iterates by refining the model/code until (admittedly subjective) satisfactory agreement is obtained. Then, another set of measurements is performed, which is compared with the corresponding predictions of the model/code. If the agreement is still satisfactory without modifying the model, this is considered progress in the validation of the model/code. Iterating with experiments testing different features of the model/code corresponds to mimicking the process of construction of a theory in Physics [16]. As the model/code is exposed to increasing scrutiny and testing, the testers develop a better understanding of the reliability (and limitations) of the model/code in predicting the outcome of new experimental and/or observational set-ups. This implies that "validation activity should be organized like a project, with goals and requirements, a plan, resources, a schedule, and a documented record" [6].

We thus propose to formulate the validation problem of a given model/code as an iterative construction that embodies the often implicit process occurring in the minds of scientists:

1. One starts with an a priori trust quantified by the value $V_{\rm prior}$ in the potential value of the model/code. This quantity captures the accumulated evidence thus far. If the model is new or the validation process is just starting, take $V_{\rm prior} = 1$. As we will soon see, the absolute value of $V_{\rm prior}$ is unimportant but its relative change is important.

2. An experiment is performed, the model/code is set-up to calculate what should be the outcome of the experiment, and the comparison between these predictions and the actual measurements is made either in model space or in observation space. The comparison requires a choice of metrics.

3. Ideally, the quality of the comparison between predictions and observations is formulated as a statistical test of significance in which an hypothesis (the model/code) is tested against the alternative, which is "all the rest." Then, the formulation of the comparison will be either "the model/code is rejected" (it is not compatible with the data) or "the model/code is compatible with the data." In order to implement this statistical test, one needs to attribute a likelihood $p(M|y_{\rm obs})$ or, more generally, a metric-based "grade" that quantifies the quality of the comparison between the predictions of the model M and observations $y_{\rm obs}$. This grade is compared with the reference likelihood q of "all the rest." Examples of implementations include the sign test and the tolerance interval methods [17]. In many cases, one does not have the luxury of a likelihood; one has then to resort to more empirical notations of how well the model explains crucial observations. In the most complex cases, these notations can be binary (accepted or rejected).

4. The posterior value of the model/code is obtained according to a formula of the type [18]

$$V_{\rm posterior}/V_{\rm prior} = F\left[p(M|y_{\rm obs}), q; c_{\rm novel}\right] \ . \tag{1}$$

   In this expression, $V_{\rm posterior}$ is the posterior potential, or coefficient, of trust in the value of the model/code after the comparison between the prediction of the model and the new observations

have been performed. By the action of $F(\cdots)$, $V_{\rm posterior}$ can be either larger or smaller than $V_{\rm prior}$: in the former case, the experimental test has increased our trust in the validity of the model/code; in the later case, the experimental test has signaled problems with the model/code. One could call $V_{\rm prior}$ and $V_{\rm posterior}$ the evolving "potential value of our trust" in the model/code or, paraphrasing the theory of decision making in economics, the "utility" of the model/code [19].

The transformation from the potential value $V_{\rm prior}$ of the model/code before the experimental test to $V_{\rm posterior}$ after the test is embodied into the multiplier F, which can be either larger than 1 (towards validation) or smaller than 1 (towards invalidation). We postulate that F depends on the grade $p(M|y_{\rm obs})$, to be interpreted as proportional to the probability of the model M given the data $y_{\rm obs}$. It is natural to compare this probability with the reference likelihood q that one or more of all other conceivable models is compatible with the same data.

The factor F depends also on a parameter $c_{\rm novel}$ that quantifies the importance of the test. In other words, $c_{\rm novel}$ is a measure of the impact of the experiment or of the observation, that is, how well the new observation explores novel "dimensions" of the parameter and variable spaces of both the process and the model that can reveal potential flaws. For instance, repeating the same observation twice does not teach more on the model, except for the statistical improvement by noise reduction in the observational data obtained by the increased sample size, which allows one to refine the grade p attributed to the model. In this case, $c_{\rm novel} \to 0$ and $V_{\rm posterior} \to V_{\rm prior}$ irrespective of the value of p, since the new experiment is a repetition of a past experiment and does not validate further the model/code. A fundamental challenge is that the determination of $c_{\rm novel}$ requires, in some sense, a pre-existing understanding of the physical processes so that the value of a new experiment can be fully appreciated. In concrete situations, one has only a limited understanding of the physical processes and the value of a new observation is only assessed after a long learning process, after comparison with other observations and experiments. Providing an a priori value for $c_{\rm novel}$, as required in expression (1), remains a difficult and key step in the validation process. This difficulty is similar to specifying the utility function in decision making [19].

One experimental test corresponds to a entire loop $1-4$ transforming a given $V_{\rm prior}$ to a $V_{\rm posterior}$ according to (1). This $V_{\rm posterior}$ becomes the new $V_{\rm prior}$ for the next test, which will transform it into another $V_{\rm posterior}$ and so on, according to the following iteration process:

$$V^{(1)}_{\rm prior} \to V^{(1)}_{\rm posterior} = V^{(2)}_{\rm prior} \to V^{(2)}_{\rm posterior} = V^{(3)}_{\rm prior} \to \ldots \to V^{(n)}_{\rm posterior}\,. \tag{2}$$

After n validation loops, we end up with a posterior trust in the validation of the model given by

$$V^{(n)}_{\rm posterior}/V^{(1)}_{\rm prior} = F\left[p^{(1)}(M|y^{(1)})_{\rm obs}, q^{(1)}; c^{(1)}_{\rm novel}\right] \;\ldots\; F\left[p^{(n)}(M|y^{(n)}_{\rm obs}), q^{(n)}; c^{(n)}_{\rm novel}\right]\,, \tag{3}$$

where the product is time-ordered since the sequence of values for $c^{(j)}_{\rm novel}$ depend on preceding tests. Validation can be said to be asymptotically satisfied when the number of steps n and the final value $V^{(n)}_{\rm posterior}$ are sufficiently high. This construction makes clear that there is no absolute validation, only a process that may or may not converge. This product (3) expresses the assumption that successive observations give independent multipliers. This assumption keeps the procedure simple because determining the dependence between different tests with respect to validation would be highly undetermined. We propose that it is more convenient to measure the dependence through the single parameter $c^{(j)}_{\rm novel}$ quantifying the novelty of the jth test with respect to those preceding it. In full generality, each new F factor should be a function of all previous tests.

The loop $1-4$ together with expression (1) are offered as an attempt to quantify the progression of the validation process, so that eventually, when several approximately independent tests exploring different features of the model/code and of the process have been performed, $V_{\rm posterior}$ has grown to a level at which most experts will be satisfied and will believe in the validity of the model/code. This formulation has the

advantage of viewing the validation process as a convergence or divergence built on a succession of steps, mimicking the construction of a theory of reality. Expression (3) embodies the progressive build-up of trust in a model or theory. This formulation provides a formal setting for discussing the difficulties that underlay the so-called impossibilities [8, 9] in validating a given model/code. Here, these difficulties are not only partitioned but quantified:

- in the definition of "new" non-redundant experiments (parameter $c_{\rm novel}$),
- in choosing the metrics and the corresponding statistical tests quantifying the comparison between the model and the measurements of this experiment (leading to the likelihood ratio p/q), and
- in iterating the procedure so that the product of the gain/loss factors $V_{\rm posterior}/V_{\rm prior}$ obtained after each test eventually leads to a clear-cut conclusion after several tests.

This formulation makes clear why and how one is never fully convinced that validation has been obtained: it is a matter of degree, of confidence level, of decision making, as in statistical testing. But, this formulation helps in quantifying what new confidence (or distrust) is gained in a given model/code. It emphasizes that validation is an ongoing process, similar to the never-ending construction of a theory of reality.

The general formulation proposed here in terms of iterated validation loops is intimately linked with decision theory based on limited knowledge: the decision to "go ahead" and use the model is fundamentally a decision problem based on the accumulated confidence embodied in $V_{\rm posterior}$. The "go/no-go" decision must take into account conflicting requirements and compromise between different objectives. Decision theory, created by the statistician Abraham Wald in the late forties, is based ultimately on game theory [19, 20]. Wald [21] used the term *loss function*, which is the standard terminology used in mathematical statistics. In mathematical economics, the opposite of the loss (or cost) function gives the concept of the *utility function*, which quantifies (in a specific functional form) what is considered important and robust in the fit of the model to the data.

Properties of the Multiplier of the Elementary Validation Step

The multiplier $F\left[p(M|y_{\rm obs}), q; c_{\rm novel}\right]$ should have the following properties:

1. If the statistical test(s) performed on the given observations is (are) passed at the reference level q, then the posterior potential value is larger than the prior potential value: $F > 1$ (resp. $F \leq 1$) for $p > q$ (resp. $p \leq q$), which can be written succinctly as $\ln F / \ln(p/q) > 0$.

2. The larger the statistical significance of the passed test, the larger the posterior value. Hence

$$\frac{\partial F}{\partial p} > 0 \,, \tag{4}$$

   for a given q. There could be a saturation of the growth of F for large p/q, which can be either that $F < \infty$ as $p/q \to \infty$ or of the form of a concavity requirement $\partial^2 F/\partial p^2 < 0$ for large p/q: obtaining a quality of fit beyond the noise level should not be over-interpreted.

3. The larger the statistical level at which the test(s) performed on the given observations is (are) passed, the larger the impact of a "novel" experiment on the multiplier enhancing the prior into the posterior potential value of the model/code: $\partial F/\partial c_{\rm novel} > 0$ (resp. ≤ 0), for $p > q$ (resp. $p \leq q$).

The simplest form obeying these properties (not including the saturation of the growth of F) is

$$F\left[p(M|y_{\rm obs}), q; c_{\rm novel}\right] = \left(\frac{p}{q}\right)^{c_{\rm novel}} \,. \tag{5}$$

This form provides an intuitive interpretation of the meaning of the experiment impact parameter c_{novel}. A bland evaluation of the novelty of a test would be $c_{\text{novel}} = 1$, thus $F = p/q$ and the chain (3) reduces to a product of normalized likelihoods, as in standard statistical tests. A value $c_{\text{novel}} > 1$ (resp. < 1) for a given experiment describes a nonlinear rapid (resp. slow) updating of our trust V as a function of the grade p/q of the model with respect to the observations. In particular, a large value of c_{novel} corresponds to the case of "critical" tests. A famous example is the Michelson-Morley experiment for the Theory of Special Relativity. For the Theory of General Relativity, it was the observation during the 1919 solar eclipse of the bending of light rays from distant stars by the Sun's mass and the anomalous precession of the perihelion of Mercury's orbit.

The alternative multiplier,

$$F\left[p(M|y_{\text{obs}}), q; c_{\text{novel}}\right] = \left[\frac{\tanh\left(\frac{p}{q} + \frac{1}{c_{\text{novel}}}\right)}{\tanh\left(1 + \frac{1}{c_{\text{novel}}}\right)}\right]^4 , \tag{6}$$

is plotted in Fig. 1 as a function of p/q and c_{novel}. It emphasizes that F saturates as a function of p/q and c_{novel} as either one or both of them grows large. A completely new experiment corresponds to $c_{\text{novel}} \to \infty$ so that $1/c_{\text{novel}} = 0$ and thus F tends to $[\tanh(p/q)/\tanh(1)]^4$, i.e., $V_{\text{posterior}}/V_{\text{prior}}$ is only determined by the quality of the "fit" of the data by the model quantified by p/q. A finite c_{novel} implies that one already takes a restrained view on the usefulness of the experiment since one limits the amplitude of the $\text{gain} = V_{\text{posterior}}/V_{\text{prior}}$, whatever the quality of the fit of the data by the model. The exponent 4 in (6) has been chosen so that the maximum confidence gain F is equal to $1/(\tanh(1))^4 \approx 3$ in the best possible situation of a completely new experiment ($c_{\text{novel}} = \infty$) and perfect fit ($p/q \to \infty$). In contrast, the multiplier F can be arbitrarily small as $p/q \to 0$ even if the novelty of the test is high ($c_{\text{novel}} \to \infty$). For a finite novelty c_{novel}, a test that fails the model miserably ($p/q \approx 0$) does not necessarily reject the model completely: unlike with the expression in (5), F remains > 0. Indeed, if the novelty c_{novel} is small, the worst-case multiplier (attained for $p/q = 0$) is $[\tanh(1/c_{\text{novel}}) / \tanh(1 + (1/c_{\text{novel}}))]^4 \approx 1 - 6.9\,\mathrm{e}^{-2/c_{\text{novel}}}$, which is only slightly less than unity if $c_{\text{novel}} \ll 1$. In short, this formulation does not heavily weight unimportant tests.

In the framework of decision theory, (1) with one of the specific expressions in (5) or (6) provides a parametric form for the utility or decision "function" of the decision maker. It is clear that many other forms of the utility function can be used, however, with the constraint of keeping the salient features of expression (1) with (5) or (6), in terms of the impact of a new test given past tests, and the quality of the comparison between the model predictions and the data.

Practical Guidelines for Determining p and c_{novel}

These two crucial elements of a validation step are conditioned by four basic problems, on which one can exert at least partial control:

1. *How to model?* This addresses model construction and involves the structure of the elementary contributions, the hierarchical organization of the routines, and requires dealing with uncertainties and fuzziness.

2. *What to measure?* This relates to the nature of c_{novel}: ideally, following Palmer et al. [22], one should target adaptively the observations to "sensitive" parts of the system. Targeting observations could be directed by the desire to access the most "relevant" information as well as to get information that is the most reliable, i.e., which is contaminated by the smallest errors. This is also the stance of Oberkampf and Trucano [23]: "A validation experiment is conducted for the primary purpose of determining the validity, or predictive accuracy, of a computational modeling and simulation capability. In other words, a validation experiment is designed, executed, and analyzed for the

purpose of quantitatively determining the ability of a mathematical model and its embodiment in a computer code to simulate a well-characterized physical process." In practice, $c_{\rm novel}$ is chosen to represent the best-guess estimate of the importance of the new observation and the degree of "surprise" it brings to the validation step. [24]

3. *How to measure?* For given measurements or experiments, the problem is to find the "optimal" metric or cost function (involved in the quality-of-fit measure p) for the intended use of the model. The notion of optimality needs to be defined. It could capture a compromise between fitting best the "important" features of the data (what is "important" may be decided on the basis of previous studies and understanding or other processes, or programmatic concerns), and minimizing the extraction of spurious information from noise. This requires one to have a precise idea of the statistical properties of the noise. If such knowledge is not available, the cost function should be chosen accordingly. The choice of the "cost function" involves the choice of how to look at the data. For instance, one may want to expand the measurements at multiple scales using wavelet decompositions and compare the prediction and observations scale by scale, or in terms of multifractal spectra of the physical fields estimated from these wavelet decompositions [25] or from other methods. The general idea here is that, given complex observation fields, it is appropriate to unfold the data on a variety of "metrics," which can then be used in the comparison between observations and model predictions: the question is then how well is the model/code able to reproduce the salient multiscale and multifractal properties derived from the observations? The physics of turbulent fields and of complex systems have offered many such new tools with which to unfold complex fields according to different statistics. Each of these statistics offers a metric to compare observations with model predictions and is associated with a cost function focusing on a particular feature of the process. Since these metrics are derived from the understanding that turbulent fields can be analyzed using these metrics that reveal strong constraints in their organization, these metrics can justifiably be called "physics-based." In practice, p, and eventually p/q, has to be inferred as an estimate of the degree of matching between the model/code output and the observation. This can be done following the concept of fuzzy logic in which one replaces the yes/no pass test by a more gradual quantification of matching [26]. We thus concur with Ref. [27], while our general methodology goes beyond.

4. *How to interpret the results?* This question relates to defining the test and the reference probability level q that any other model (than the one under scrutiny) can explain the data. The interpretation of the results should aim at detecting the "dimensions" that are missing, mispecified or erroneous in the model. What tests can be used to betray the existence of hidden degrees of freedom and/or dimensions? This is the hardest problem. It can sometimes find an elegant solution when a given model is embedded in a more general one. Then, the limitation of the "smaller" model becomes clear from the vantage of the more general model.

We now illustrate our algorithmic approach to model validation using the historical development of quantum mechanics and four examples based on the authors' research activities. In these examples, we will use the form (6) and consider three finite values: $c_{\rm novel} = 1$ (marginally useful new test), $c_{\rm novel} = 10$ (substantially new test), and $c_{\rm novel} = 100$ (important new test). When a likelihood test is not available, we propose to use three possible marks: $p/q = 0.1$ (poor fit), $p/q = 1$ (marginally good fit), and $p/q = 10$ (good fit). Extreme values ($c_{\rm novel}$ or p/q are 0 or ∞) have already been discussed. Due to limited experience with this approach, we propose these ad hoc values in the following examples of its application.

Quantum Mechanics

Quantum mechanics (QM) offer a vivid incarnation of how a model can turn progressively into a theory held "true" by almost all physicists. Since its birth, QM has been tested again and again because it presents a view of "reality" that is shockingly different from the classical view experienced at the macroscopic scale.

QM prescriptions and predictions often go against classical intuition. Nevertheless, we can state that, by a long and thorough process of verified predictions of QM in experiments, fueled by the imaginative set-up of paradoxes, QM has been validated as a correct description of nature. It is fair to say that the overwhelming majority of physicists have developed a strong trust in the validity of QM. That is, if someone comes up with a new test based on a new paradox, for instance, most physicists would bet that QM will come up with the right answer with a very high probability. It is thus by the on-going testing and the compatibility of the prediction of QM with the observations that QM has been validated. As a consequence, one can use it with strong confidence to make predictions in novel directions. This is ideally the situation one would like to attain for the problem of validation of models and of codes discussed below. We now give a very partial list of selected tests that established the trust of physicists in Quantum Mechanics.

1. Pauli's exclusion principle states that no two identical fermions (particles with non-interger values of spin) may occupy the same quantum state simultaneously [28]. It is one of the most important principles in physics, primarily because the three types of particle from which ordinary matter is made, electrons, protons, and neutrons, are all subject to it. With $c_{\text{novel}} = 100$ and perfect verification in numerous experiments ($p/q = \infty$), this leads to $F^{(1)} = 2.9$.

2. The EPR paradox [29] was a thought experiment designed to prove that quantum mechanics was hopelessly flawed: according to QM, a measurement performed on one part of a quantum system can have an instantaneous effect on the result of a measurement performed on another part, regardless of the distance separating the two parts. Bell's theorem [30] showed that quantum mechanics predicted stronger statistical correlations between entangled particles than the so-called local realistic theory with hidden variables. The importance of this prediction requires $c_{\text{novel}} = 100$ at the very minimum. The QM prediction turned out to be correct, winning over the hidden-variables theories [31, 32] ($p/q = \infty$), leading again to $F^{(2)} = 2.9$.

3. The Aharonov-Bohm effect predicts that a magnetic field can influence an electron that, strictly speaking, is located completely beyond the field's range, again an impossibility according to non-quantum theories ($c_{\text{novel}} = 100$), The Aharonov-Bohm oscillations were observed in ordinary (i.e., nonsuperconducting) metallic rings, showing that electrons can maintain quantum mechanical phase coherence in ordinary materials [33]. This yields $p/q = \infty$ and thus $F^{(3)} = 2.9$ yet again.

4. The Josephson effect provides a macroscopic incarnation of quantum effects in which two superconductors are predicted to preserve their long-range order across an insulating barrier, for instance, leading to rapid alternating currents when a steady voltage is applied accross the superconductors. The novelty of this effect again warrants $c_{\text{novel}} = 100$ and the numerous verifications and applications (for instance in SQUIDs: Superconducting QUantum Interference Devices) argues for $p/q = \infty$ and thus $F^{(4)} = 2.9$, as usual.

5. The prediction of possible collapse of a gas of atoms at low temperature into a single quantum state is known as Bose-Einstein (BE) condensation, again so much against classical intuition ($c_{\text{novel}} = 100$). Atoms are indeed bosons (particles with integer values of spin) which are *not* subjected to the Pauli exclusion principle evoked in above test #1 of QM. The first such BE condensate was produced using a gas of rubidium atoms cooled to $1.7 \cdot 10^{-7}$ K [34] ($p/q = \infty$), leading once more to $F^{(4)} = 2.9$.

6. There have been several attempts to develop a paradox-free nonlinear QM theory, in the hope of eliminating Schrödinger's cat paradox, among other embarrassments. The nonlinear QM predictions diverge from those of orthodox quantum physics, albeit subtly. For instance, if a neutron impinges on two slits, an interference pattern appears, which should, however, disappear if the measurement is made far enough away ($c_{\text{novel}} = 100$). Experiment tests of the neutron prediction rejected the nonlinear version in favor of the standard QM [35] ($p/q = \infty$), leading to $F^{(6)} = 2.9$.

7. In addition, measurements at the National Bureau of Standards in Boulder, CO, on frequency standards have been shown to set limits of order 10^{-21} on the fraction of the energy of the rf transition in ^{9}Be ions that could be due to nonlinear corrections to quantum mechanics [36]. We assign $c_{\rm novel} = 10$, with $p/q = 10$), to this result, leading to $F^{(7)} = 2.4$. Although less than $F^{(1-6)}$ this is still meant to be an impressive score.

Combining the multipliers according to (3) leads to $V^{(8)}_{\rm posterior}/V^{(1)}_{\rm prior} \simeq 1400$, which is of course only a lower limit given the many other validation tests not mentioned here. Tests of QM are ongoing [37].

Four Further Examples Drawn from the Authors' Research Activities

The Olami-Feder-Christensen (OFC) sand-pile model of earthquakes. This is perhaps the simplest sand-pile model of self-organized criticality, which exhibits a phenomenology resembling real seismicity [38]. To validate it, we examine the properties and prediction of the model that can be compared with real seismicity, together with our assessment of their $c_{\rm novel}$ and quality-of-fit. We are careful to state these properties in an ordered way, as specified in the above sequences (2)–(3).

1. The statistical physics community recognized as an important step in the development of a theory of earthquakes the discovery of the OFC model: without a conservation law but nevertheless exhibiting a power law distribution of avalanche sizes resembling the Gutenberg-Richter (GR) law [38]. On the other hand, many other models with different mechanisms can explain observed power law distributions [39]. We thus attribute only $c_{\rm novel} = 10$ to this evidence. Because the power law distribution obtained by the model is of excellent quality for a certain parameter value ($\alpha \approx 0.2$), we formally take $p/q = \infty$ (perfect fit). Expression (6) then gives $F^{(1)} = 2.4$.

2. Now turning to foreshocks and aftershocks, exponents for the inverse and direct Omori laws are smaller than for real seismicity [40]. There are two aspects in this prediction of the model: (i) the finding of foreshocks and aftershocks with similar qualitative properties, and (ii) their inverse and direct Omori rates. The first aspect, deserves a large $c_{\rm novel} = 100$ as the observation of foreshocks and aftershocks came as a rather big surprise in such sand-pile models [41]. The clustering in time and space of the foreshocks and aftershocks are qualitatively similar to real seismicity [40], which warrants $p/q = 10$, and thus $F^{(2a)} = 2.9$. The second aspect is secondary compared with the first one ($c_{\rm novel} = 1$). Since the exponents are only qualitatively reproduced (but with no formal likelihood test available), we take $p/q = 0.1$. This leads to $F^{(2b)} = 0.47$.

3. Scaling of the number of aftershocks with the main shock size (productivity law) [40]: $c_{\rm novel} = 10$ as this observation is rather new but not completely independent of the Omori law. The fit is good so we grant a grade $p/q = 10$ leading to $F^{(3)} = 2.4$.

4. Power law increase of the number of foreshocks with the mainshock size [40]: this is not observed in real seismicity, probably because this property is absent or perhaps due to a lack of quality data. This test is therefore not very selective ($c_{\rm novel} = 1$) and the large uncertainties suggest a grade $p/q = 1$ (to reflect the different viewpoints on the absence of effect in real data) leading to $F^{(4)} = 1$ (neutral test).

5. Most aftershocks are found to nucleate at "asperities" located on the mainshock rupture plane or on the boundary of the avalanche, in agreement with observations [40]: $c_{\rm novel} = 10$ and $p/q = 10$ leading to $F^{(5)} = 2.4$.

6. Earthquakes cluster on spatially localized geometrical structures, known as faults. This property is arguably central to seismicity physics ($c_{\rm novel} = 100$), but absolutely not reproduced by the OFC model ($p/q = 0.1$). This leads to $F^{(6)} = 4 \cdot 10^{-4}$.

Combining the multipliers according to (3) up to test #5 leads to $V^{(6)}_{\rm posterior}/V^{(1)}_{\rm prior} = 18.8$, suggesting that the OFC model is validated as a useful model of the statistical properties of seismic catalogs, at least with respect to the properties which have been examined in these first six tests. Adding the crucial last test strongly fails the model since $V^{(7)}_{\rm posterior}/V^{(1)}_{\rm prior} = 0.0075$. The model can not be used as a realistic predictor of seismicity. It can nevertheless be useful for the study of certain statistical properties.

The multifractal random walk (MRW) as a model of financial returns. We now consider the MRW model introduced as a random walk with stochastic "volatility" endowed with exact multifractal properties [42], which has been proposed as a model of financial time series. Among the documented facts about financial time series, we have the absence of correlation between lagged returns, the long-range correlation of lagged volatilities, and the observed multifractality. These can not be taken as validation tests of the model since they are the observations that motivated the introduction of the MRW. These observations thus constitute references or benchmarks against which new tests must be compared. The new properties and prediction of the MRW model that can be compared with real financial return time series are the following.

1. The probability density distributions (PDF) of returns at different time scales: the MRW exhibits the remarkable property of accounting quantitatively for the transition from fatter-than-exponential PDFs at small time scales to approximately Gaussian PDFs at large time scales. But, because the MRW is intrinsically a model developed as the continuous limit of a cascade across scales, this is perhaps not very surprising. We thus rate the novelty of this observation with $c_{\rm novel} = 10$. In absence of formal likelihood tests on the PDFs, we take $p/q = 10$ to reflect the apparent excellent fits of the data at multiple scales, leading to $F^{(1)} = 2.4$.

2. Different response functions of the price volatility to large external shocks compared with endogeneous shocks, which are well-confirmed quantitatively by observations on a hierarchy of volatility shocks [43]. This prediction has been verified to hold with remarkable accuracy without any adjustable parameters (i.e., the parameters were adjusted previously and fixed before the new test). We thus rate the novelty of this test with a high $c_{\rm novel} = 100$ and the agreement is quantified by $p/q = 10$, leading to $F^{(2)} = 2.9$.

3. The sharp-peak/flat-trough pattern of price peaks [44] as well as accelerated speculative bubbles preceding crashes [45] is not captured by the MRW. In view of the debated importance of such patterns, we rate these observations with $c_{\rm novel} = 1$ and $p/q = 0.1$, leading to $F^{(3)} = 0.47$.

4. The leverage effect and volatility dependence on past volatility and returns (see [46] and references therein). These features are not captured by the MRW at all. We rate $c_{\rm novel} = 10$ and the lack of agreement is quantified by $p/q = 0.1$, leading to $F^{(4)} = 0.0037$.

Combining the multipliers according to (3) leads to $V^{(5)}_{\rm posterior}/V^{(1)}_{\rm prior} = 0.012$, rejecting the model. But if we stop the validation steps at $V^{(3)}_{\rm posterior}/V^{(1)}_{\rm prior} = 7$, we obtain a clear validation signal. The two additional tests fail the MRW because the observed effects involve mechanisms that are absent in it. Here, we should conclude that the MRW is a useful model that is validated with respect to certain properties on the memory of volatility but is not validated for a fully faithful description of the stock market returns. These mechanisms can be actually incorporated into extensions of the MRW, corresponding to the addition of new dimensions lacking in the MRW. If we had used the long-range correlation of lagged volatilities and the observed multifractality (each with parameters $c_{\rm novel} = 10$ and $p/q = 10$) as tests #-1 and #0, F would have gained a factor $2.4^2 = 5.9$, changing $V^{(5)}_{\rm posterior}/V^{(1)}_{\rm prior} = 0.012$ into $V^{(5)}_{\rm posterior}/V^{(-1)}_{\rm prior} = 0.07$, still far from sufficient to validate the model

An anomalous diffusion model for solar photons in cloudy atmospheres. To properly model climate dynamics, it is important to narrow the significant uncertainty associated with clouds. In particular, estimation

of the radiation budget in the presence of clouds needs to be improved since current operational models for the most part ignore all variability below the scale of the climate model's grid (a few 100 km). So a considerable effort has been expended to derive more realistic mean-field radiative transfer models [47], mostly by considering only the one-point variability of clouds (that is, irrespective of their actual structure). However, it has been widely recognized that the Earth's cloudiness is fractal over a wide range of scales [48]. This is the motivation for modeling the paths of solar photons at non-absorbing wavelengths in the cloudy atmosphere as convoluted Lévy walks [39], which are characterized by frequent small steps (inside clouds) and occasional large jumps (between clouds). These paths start downward at the top of the highest clouds and end in escape to space or in absorption at the surface. In sharp contrast with most other mean-field models for solar radiative transfer, this diffusion model with anomalous scaling can be subjected to a battery of observational tests.

1. The original goal of this phenomenological model, which accounts for the clustering of cloud water droplets into broken and/or multi-layered cloudiness, was to predict the increase in steady-state flux transmitted to the surface compared to what would filter through that same amount of water in a single unbroken cloud layer [49]. This property is common to all mean-field photon transport models that do anything at all about unresolved variability [47], so we can only assign $c_{\rm novel} = 10$ to this test and, given that all models in this class are successful, we have to take $p/q = 1$, hence $F^{(1)} = 1$. The outcome of this first test is neutral.

2. The first real test for this model occurred when it became possible to accurately estimate the mean total path of solar photons that reach the surface. This breakthrough was enabled by access to spectroscopy at medium (high) resolution of oxygen bands (lines) [50, 51]. Along with simultaneous estimation of cloud optical depth (basically, column-integrated water [kg/m^2] times the average scattering cross-section per kg), the observed trends were explained only by the new model in spite of the relatively large instrumental error bars. So we assign $c_{\rm novel} = 100$ to this highly discriminating test and $p/q = 10$ (even though the other models were not in a position to compete), hence $F^{(2)} = 2.9$.

3. Another test was proposed using time-dependent photon transport with a source near the surface (cloud-to-ground lightning) and a detector in space (the DOE FORTÉ satellite) [52]. The quantity of interest is the observed delay of the light pulse (due to multiple scattering in the cloud system) with respect to the radio-frequency pulse (which travels in a straight line). There was no simultaneous estimate of cloud optical depth, so assumptions had to be made, informed by the fact that storm clouds are at once thick and dense. Because of this lack of an independent measurement, we assign only $c_{\rm novel} = 10$ to the observation and $p/q = 1$ to the model performance since this is only about the finite horizontal extent of the cloud (one could exclude only uniform "plane-parallel" clouds). So, again we obtain $F^{(3)} = 1$ for an interesting but presently neutral test that needs to be refined.

4. Min et al. [53] developed an oxygen-line spectrometer with sufficient resolution to estimate not just the *mean* path but also its *root-mean-square* (RMS) value. They found the prediction by Davis and Marshak [54] for normal diffusion to be an extreme (envelop) case for the empirical scatter plot of mean vs. RMS path, and this is indicative that the anomalous diffusion model will cover the bulk of the data. Because of some overlap with a previous item, we assign $c_{\rm novel} = 10$ and $p/q = 10$ for the model performance (since the anomalous diffusion model had not yet made a prediction for the RMS path, but the other models have yet to make one for the mean path). We therefore obtain $F^{(4)} = 2.4$.

5. Using similar data but a different normalization than Min et al.'s, more amenable to model testing, Scholl et al. [55] observed that the RMS-to-mean ratio for solar photon path is essentially constant whether the diffusion is normal or anomalous. This is a remarkable empirical finding to which we

assign $c_{\rm novel} = 100$. The new mean- and RMS-path data was explained by Scholl et al. by creating an ad hoc hybrid between the normal diffusion theory (which made a prediction for the RMS path) and the anomalous theory (which did not). This significant modification of the basic model means that we are in principle back to validation step #1 with the new model. However, this exercise uncovered something quite telling about the original anomalous diffusion model, namely, that its simple asymptotic (large optical depth) form used in all the above tests is inappropriate: for typical cloud covers, the pre-asymptotic terms computed explicitly for the normal diffusion case prove to be important irrespective of whether the diffusion is normal or not. Consequently, in its original form (a simple scaling law for the mean path with respect to cloud thickness and optical depth), the anomalous diffusion model fails to reproduce the new data even for the mean path. (This means that previous fits yielded "effective" anomaly parameters and were misleading if taken literally.) So we assign $p/q = 0.1$ at best for the original model, hence $F^{(5)} = 0.0004$.

Thus, $V^{(6)}_{\rm posterior}/V^{(1)}_{\rm prior} = 0.003$, a fatal blow for the anomalous diffusion in its simple asymptotic form, even though $V^{(5)}_{\rm posterior}/V^{(1)}_{\rm prior} = 7.0$ which would have been interpreted as close to a convincing validation. This is of course not the end of the story. The original model has already spawned Scholl et al.'s empirical hybrid and there is a formalism based on integral (in fact, pseudo-differential) operators that extends the anomalous *diffusion* model to pre-asymptotic regimes [56]. More recently, a model for anomalous *transport* (i.e., where angular details matter) has been proposed that fits all of the new oxygen spectroscopy results [57].

In summary, the first and simplest incarnation of the anomalous diffusion model for solar photon transport ran its course and demonstrated the power of oxygen-line spectroscopy as a test for the perfomance of solar radiative transfer models required in climate modeling for large-scale average properties. Eventually, new and interesting tests will become feasible when we obtain dedicated oxygen-line spectroscopy from space (with NASA's Orbiting Carbon Observatory mission planned for launch in 2007). Indeed, we already know that the asymptotic scaling for reflected photon paths [58] is different from their transmitted counterparts [54] in both mean and RMS.

A computational fluid dynamics (CFD) model for shock-induced mixing and shock-tube tests. So far, our examples of models for complex phenomena have hailed from quantum and statistical physics. In the latter case, they are stochastic models composed of: (1) simple code (hence rather trivial verification procedures) to generate realizations, and (2) analytical expressions for the ensemble-average properties (that are used in the above validation exercises). We now turn to gas dynamics codes which have a broad range of applications, from astrophysical and geophysical flow simulation to the design and performance analysis of engineering systems. Specifically, we discuss the validation of the (extensively verified) "Cuervo" code developed at Los Alamos National Laboratory, implementing compressible Euler equations [59]. A standard test case involves the Richtmyer-Meshkov (RM) instability [60, 61], which arises when a density gradient in a fluid is subjected to an impulsive acceleration, e.g., due to passage of a shock wave. Evolution of the RM instability is nonlinear and hydrodynamically complex and hence defines an excellent problem-space to assess CFD code performance.

In the series of shock-tube experiments described in [62], RM dynamics are realized by preparing one or more cylinders with approximately identical axisymmetric Gaussian concentration profiles of dense sulfur hexaflouride (SF_6) in air. This (or these) vertical "gas cylinder(s)" is (are) subjected to a weak shock —Mach number $\approx$1.2— propagating horizontally. The ensuing dynamics are largely governed by the mismatch of the density gradient between the gases (with the density of SF_6 approximately five times that of air) and the pressure gradient through the shock wave; this mismatch acts as the source for baroclinic vorticity generation. The visualization of the density field is obtained using a planar laser-induced fluorescence (PLIF) technique, which provides high-resolution quantitative concentration measurements. The velocity field is diagnosed using particle image velocimetry (PIV), based on correlation measurements of

small-scale particles that are lightly seeded in the initial flow field. Careful post-processing of images from 130 μs to 1000 μs after shock passage yields planar concentration and velocity with error bars.

1. The RM flow is dominated at early times by a vortex pair (per gas cylinder). Later, secondary instabilities rapidly transition the flow to a mixed state. We rate $c_{\rm novel} = 10$ for the observations of these two instabilities. The Cuervo code correctly captures these two instabilities, best observed and modeled with a single cylinder. At this qualitative level, we rate $p/q = 10$ (good fit), which leads to $F^{(1)} = 2.4$.

2. Older data for two-cylinder experiments acquired with a fog-based technique (rather than PLIF) showed two separated spirals associated with the primary instability, but the Cuervo code predicted the existence of a material bridge. This previously unobserved connection was experimentally diagnosed with the improved observational technique. Using $c_{\rm novel} = 10$ and $p/q = 10$ yields $F^{(2)} = 2.4$.

3. The evolution of the total power as a function of time offers another useful metric. The numerical simulation quantitatively accounts for the exponential growth of the power with time, within the experimental error bars. Using $c_{\rm novel} = 10$ and $p/q = 10$ yields $F^{(3)} = 2.4$.

4. The concentration power spectrum as a function of wavenumber for different times provides another way (in the Fourier domain) to present the information of the hierarchy of structures already visualized in physical space ($c_{\rm novel} = 1$). The Cuervo code correctly accounts for the low wavenumber part of the spectrum but underestimates the high wavenumber part (beyond the deterministic-stochastic transition wavenumber) by a factor 2 to 5. We capture this by setting $p/q = 0.1$, which yields $F^{(4)} = 0.47$.

Combining the multipliers according to (3) leads to $V^{(5)}_{\rm posterior}/V^{(1)}_{\rm prior} = 6.5$, a significant gain, but still not sufficient to fully validate the Cuervo code. Intricate experiments with three gas cylinders have been performed [63] and others are currently under way to further stress CFD models.

These examples illustrate the utility of representing the validation process as a succession of steps, each of them characterized by the two parameters $c_{\rm novel}$ and p/q. The determination of $c_{\rm novel}$ requires expert judgment and that of p/q a careful statistical analysis, which is beyond the scope of the present report (see Ref. [27] for a detailed case study). The parameter q is ideally imposed as a confidence level, say 95% or 99% as in standard statistical tests. In practice, it may depend on the experimental test and requires a case-by-case examination.

The uncertainties of $c_{\rm novel}$ and of p/q need to be assessed. Indeed, different statistical estimations or metrics may yield different p/q's and different experts will likely rate differently the novelty $c_{\rm novel}$ of a new test. As a result, the trust gain $V^{(n+1)}_{\rm posterior}/V^{(1)}_{\rm prior}$ after n tests necessarily has a range of possible values that grows geometrically with n. In certain cases, a drastic difference can be obtained by a change of $c_{\rm novel}$: for instance, if instead of attributing $c_{\rm novel} = 100$ to the sixth OFC test, we put $c_{\rm novel} = 10$ (resp. 1) while keeping $p/q = 0.1$, $F^{(6)}$ is changed from $4 \cdot 10^{-4}$ to $4 \cdot 10^{-3}$ (resp. 0.47). The trust gain then becomes $V^{(7)}_{\rm posterior}/V^{(1)}_{\rm prior} = 0.07$ (resp. $\simeq 9$). For the sixth OFC test, $c_{\rm novel} = 1$ is arguably unrealistic, given the importance of faults in seismology. The two possible choices $c_{\rm novel} = 100$ and $c_{\rm novel} = 10$ then give similar conclusions on the invalidation of the OFC model. In our examples, $V^{(n+1)}_{\rm posterior}/V^{(1)}_{\rm prior}$ provides a qualitatively robust measure of the gain in trust after n steps; this robustness has been built-in by imposing a coarse-grained quality to p/q and $c_{\rm novel}$.

Finally, we remark that the proposed form for the multiplier (6) contains an important asymmetry between gains and losses: the failure to a single test with strong novelty and significance (as, e.g., for the localized seismicity on faults in the case of the OFC model and for the leverage effect in the case of the MRW model) cannot be compensated by the success of all the other tests combined. In other words,

a single test is enough to reject a model. This embodies the common lore that reputation gain is a slow process requiring constancy and tenacity, while its loss can occur suddenly with one single failure and is difficult to re-establish. We believe that the same applies to the build-up of trust in and, thus, validation of a model.

Summary
The validation of numerical simulations continues to become more important as computational power grows and the complexity of modeled systems increases. We have proposed an iterative, constructive approach to validation using quantitative measures and expert knowledge to assess the relative state of validation of a model instantiated in a computer code. In this approach, the increase/decrease in validation is mediated through a function that incorporates the results of the model vis-à-vis the experiment together with a measure of the impact of that experiment on the validation process. While this function is not uniquely specified, it is not arbitrary: certain asymptotic trends, consistent with heuristically plausible behavior, must be observed. In five fundamentally different examples, we have illustrated how this approach might apply to a validation process for physics or engineering models. We believe that the multiplicative decomposition of trust gains or losses (given in Eq. 3), using a suitable functional prescription (such as Eq. 6), provides a reasoned and principled description of the key elements —and fundamental limitations— of validation. It should be equally applicable to biological and social sciences, especially since it is built upon the decision-making processes of the latter.

Acknowledgments: This work was supported by the LDRD 20030037DR project "Physics-Based Analysis of Dynamic Experimentation and Simulation." We acknowledge stimulating interactions and discussions with the other members of the project, including Bob Benjamin, Mike Cannon, Karen Fisher, Andy Fraser, Sanjay Kumar, Kathy Prestridge, Bill Rider, Chris Tomkins, Peter Vorobieff, and Cindy Zoldi.

References

[1] This can be done using cluster analysis and/or pattern recognition methods in complex situations, as needed, but often simply by visual inspection.

[2] P.J. Roache, *Verification and Validation in Computational Science and Engineering*, Hermosa Publishers, Albuquerque, NM (1998).

[3] Living Planet Report 2004, WWF International, Published in October 2004 by WWF–World Wide Fund For Nature (formerly World Wildlife Fund), Gland, Switzerland. ISBN: 2-88085-265-X, A BANSON Production.

[4] R. Costanza, et al., Nature **387** (May 15), 253 (1997); S. Pimm, Nature **387** (May 15), 23 (1997).

[5] I. Babuska and J.T. Oden, Comput. Methods Appl. Mech. Engrg. **193**, 4057 (2004).

[6] D.E. Post and L.G. Votta, Phys. Today **57** (1), 35 (2005).

[7] A.R. Hefner, Working Group on Model Validation, http://ray.eeel.nist.gov/modval.html

[8] N. Oreskes, K. Shrader-Frechette, and K. Belitz, Science **263**, 641 (1994); N. Oreskes, Environmental Health Perspective Supplements **106** (S6), 1453 (1998).

[9] J.D. Sterman, E.J. Rykiel Jr., N. Oreskes, K. Belitz, and K. Shrader-Frechette, Science **264**, 329 (1994).

[10] G.J. Chaitin, *Algorithmic Information Theory*, Cambridge University Press, New York, NY (1987).

[11] M. Buchanan, Revealing order in the chaos, New Scientist **2488**, 26 February (2005); N. Goldenfeld, cited therein.

[12] N. Israeli and N. Goldenfeld, Phys. Rev. Lett. **92**, 074105 (2004).

[13] A.A. Borovkov, *Mathematical Statistics*, A. Moullagaliev (Translator), 1st ed., Taylor & Francis, Amsterdam, The Netherlands (1998).

[14] C. Gourieroux and A. Monfort, Testing non-nested hypotheses, Chapter 44 in *Handbook of Econometrics, Volume IV*, R.F. Engle and D.L. McFadden (Eds.), Elsevier Science B.V. (1994).

[15] T.G. Trucano, M. Pilch, and W.L. Oberkampf, *On the Role of Code Comparisons in Verification and Validation*, SAND2003-2752, Sandia National Laboratory, Albuquerque, NM (2003).

[16] F. Dyson, *Infinite in All Directions*, Penguin, London (1988).

[17] L. Pal and M. Makai [Remarks on statistical aspects of safety analysis of complex systems, http://arXiv.org/abs/physics/0308086] have used the mathematical statistics of hypothesis testing as a way to validate the correctness of a code simulating the operation of a system with respect to a level of confidence for safety problems. The main conclusion is that the testing of the input variables separately may lead to incorrect safety related decisions with unforeseen consequences. For this, they have used two statistical methods: the sign test and the tolerance interval methods for testing more than one mutually dependent output variables. We propose to use these and similar tests delivering a probability level p which can then be compared with a pre-defined likelihood level q.

[18] Although this formula is reminiscent of a Bayesian analysis, we are *not* dealing with probabilities here.

[19] J. von Neumann and O. Morgenstein, *Theory of Games and Economic Behavior*, Princeton University Press (1944).

[20] S. Kotz and N. Johnson, *Breakthroughs in Statistics (Foundations and Theory)*, vol. 1, Springer-Verlag (1993).

[21] A. Wald, *Statistical Decision Functions*, Wiley, New York, NY (1950).

[22] T.N. Palmer, R. Gelaro, J. Barkmeijer, and R. Buizza, J. Atmos. Sci. **55**, 633 (1998).

[23] W.L. Oberkampf and T.G. Trucano, Progress in Aerospace Sciences **38** (3), 209 (2002).

[24] In this approach, the quantification by $V_{\rm posterior}$ of a validation step described by the loop $1-4$ in **Valdiation as a Constructive Iterations Process** is nothing more than a "projection" operation P of the model space onto the observation space. Since projections are idempotent, i.e., remain identical when iterated on themselves: $P \circ P = P$, this means that an observation "projects" the model/code in a certain subset of all its potential domain of applications and tests the model/code in that restricted subset. Hence, successive observational tests should be performed to optimize the validation process: the projections P_1 and P_2 associated with two successive experiments should be as "orthogonal" as possible. If the space of experiments can be endowed with a scalar product, the coefficient $c_{\rm novel}$ associated with a given experiment can be assigned less arbitrarily. In particular, it should taken to increase with the inverse of the product of all scalar products of the normal to a new hyperspace (which is the image by the projection operator P of the model onto the new experiment)

and all the normals of the hyperspaces associated with the projection operators of all previously performed experimental tests.

[25] J.-F. Muzy, E. Bacry, and A. Arnéodo, Int. J. of Bifurcation and Chaos **4**, 245 (1994).

[26] L.A. Zadeh, Management Science **17** (4), B141 (1970); Knowledge representation in fuzzy logic, in *An Introduction to Fuzzy Logic Applications in Intelligent Systems*, R.R. Yager and L.A. Zadeh (Eds.), Kluwer Academic, Boston, Ma, pp. 2-25 (1992).

[27] W.L. Oberkampf and M.F. Barone, *Measures of Agreement Between Computation and Experiment: Validation Metrics*, SAND2005-4302, Sandia National Laboratory, Albuquerque, NM (2005).

[28] M. Massimi, *Pauli's Exclusion Principle, The Origin and Validation of a Scientific Principle*, University of Cambridge Press (2005).

[29] A. Einstein, B. Podolsky, and N. Rosen, Phys. Rev. **47**, 777 (1935).

[30] J. Bell, Physics **1**, 195 (1964).

[31] A. Aspect, C. Imbert, and G. Roger, Optics Comm. **34**, 46 (1980).

[32] N.J. Rarity and J.P. Tapster, Phys. Rev. Lett. **64**, 2495 (1990).

[33] R. Webb, S. Washburn, C. Umbach, and R. Laibowitz, Phys. Rev. Lett. **54**, 2696 (1985); B. Schwarzschild, Phys. Today **39** (1), 17 (1986).

[34] M.H. Anderson, J.R. Ensher, M.R. Matthews, C.E. Wieman, and E.A. Cornell, Science **269**, 198 (1995).

[35] R. Gähler, A.G. Klein, and A. Zeilinger Phys. Rev. A **23**, 1611 (1981).

[36] S. Weinberg, Phys. Rev. Lett. **62**, 485 (1989).

[37] A.J. Leggett, J. Phys. Condens. Matter **14**, R415 (2002).

[38] Z. Olami, H. J. S. Feder, and K. Christensen, Phys. Rev. Lett. **68**, 1244 (1992).

[39] D. Sornette, *Critical Phenomena in Natural Sciences*, 2nd ed., Springer Series in Synergetics, Springer-Verlag, Heidelberg, Germany (2004).

[40] A. Helmstetter, S. Hergarten, and D. Sornette, Phys. Rev. E **70**, 046120 (2004).

[41] S. Hergarten and H.J. Neugebauer, Phys. Rev. Lett. **88**, 238501 (2002).

[42] J.-F. Muzy, J. Delour, and E. Bacry, The European Physical Journal B **17**, 537 (2000); E. Bacry, J. Delour, and J.-F. Muzy, Phys. Rev. E **64**, 026103 (2001).

[43] D. Sornette, Y. Malevergne, and J.-F. Muzy, Risk **16** (2), 67 (2003). [http://arXiv.org/abs/cond-mat/0204626]

[44] B.M. Roehner and D. Sornette, Eur. Phys. J. B **4**, 387 (1998).

[45] A. Johansen, D. Sornette, and O. Ledoit, J. of Risk **1**, 5 (1999).

[46] L. Borland, J.-P. Bouchaud, J.-F. Muzy, and G. Zumback, The dynamics of financial markets – Mandelbrot's multifractal cascades, and beyond [preprint http://arXiv.org/abs/cond-mat/0501292] (2005).

[47] H.W. Barker and A.B. Davis, Approximation methods in atmospheric 3D radiative transfer, Part 2: Unresolved variability and climate applications, in *3D Radiative Transfer in Cloudy Atmospheres*, A. Marshak and A.B. Davis (eds.), Springer-Verlag, Heidelberg, pp. 343-383 (2005).

[48] S. Lovejoy, Science **216**, 185 (1982).

[49] A. Davis and A. Marshak, Lévy kinetics in slab geometry: Scaling of transmission probability, in *Fractal Frontiers*, M.M. Novak and T.G. Dewey (Eds.), World Scientific, Singapore, pp. 63-72 (1997)

[50] K. Pfeilsticker, J. Geophys. Res. **104**, 4101 (1999).

[51] Q.-L. Min, L.C. Harrison, and E.E. Clothiaux, J. Geophys. Res. **106**, 7375 (2001)

[52] A.B. Davis, D.M. Suszcynski, and A. Marshak, Shortwave transport in the cloudy atmosphere by anomalous/Lévy photon diffusion: New diagnostics using FORTÉ lightning data, in *Proceedings of 10th Atmospheric Radiation Measurement (ARM) Science Team Meeting*, 03/13-17, 2000, San Antonio (Tx), US DOE, http://www.arm.gov/docs/documents/technical/conf_0003/davis-ab.pdf (2000).

[53] Q.-L. Min, L.C. Harrison, P. Kiedron, J. Berndt, and E. Joseph, J. Geophys. Res. **109**, D02202 (2004).

[54] A.B. Davis and A. Marshak, J. Atmos. Sci. **59**, 2714 (2002).

[55] T. Scholl, K. Pfeilsticker, A.B. Davis, H.K. Baltink, S. Crewell, U. Löhnert, C. Simmer, J. Meywerk, and M. Quante, J. Geophys. Res., accepted for publication (2005).

[56] S.V. Buldyrev, M. Gitterman, S. Havlin, A.Ya. Kazakov, M.G.E. da Luz, E.P. Raposo, H.E. Stanley, and G.M. Viswanathan, Physica A **302**, 148 (2001).

[57] A.B. Davis, Effective propagation kernels in structured media with broad spatial correlations, Illustration with large-scale transport of solar photons through cloudy atmospheres, in *Computational Transport Theory - Granlibakken 2004*, F. Graziani (ed.), Springer-Verlag, New York (NY), pp. 85-140 (2006) [to appear].

[58] A.B. Davis, R.F. Cahalan, J.D. Spinhirne, M.J. McGill, and S.P. Love, Phys. Chem. Earth (B) **24**, 177 (Erratum 757) (1999).

[59] W.J. Rider, J.A. Greenough, and J.R. Kamm, Int J. Num. Meth. Fluids **47**, 1253 (2005).

[60] R.D. Richtmyer, Commun. Pure Appl. Math. **8**, 297 (1960).

[61] E. Meshkov, Izv. Akad. Nauk SSSR, Mekh. Zhidk. Gaza **4**, 151 (1969).

[62] R.F. Benjamin, Comput. Sci. Engng. **6**, 40 (2004).

[63] S. Kumar, G. Orlicz, C. Thomkins, C. Goodenough, K. Prestridge, P. Vorobieef, and R.F. Benjamin, Phys. Fluids **17**, 1 (2005).

Figure 1: The multiplier defined by (6) is plotted as a function of p/q and c_{novel}.