Empirical agreement in model validation

Julie Jebeile and Anouk Barberousse (Université Paris-Sorbonne)

Empirical agreement is often used as an important criterion in model validation (Oberkampf and Trucano 2002; Oberkampf, Trucano and Hirsch 2002; Trucano et al. 2005). However, it is by no means a sufficient criterion as a model can be so adjusted as to fit available data even though it is based on hypotheses whose plausibility is known to be questionable. Our aim in this paper is to investigate into the uses of empirical agreement within the process of model validation as it is performed in scientific practice.

In order to do so, we first present the main reason why empirical agreement is not a sufficient criterion for model validation, namely, Duhem problem of refutation and confirmation holism. What we here call "Duhem problem" is the model-oriented version of the Duhem-Quine thesis (Lenhard and Winsberg 2010, Winsberg 2010): When a model's outputs are not the expected ones, the modeler has usually no way to cut the model into pieces that could be confirmed or refutated isolatedly. As a result, she cannot identify which part is responsible for the failure. Conversely, when a model's outputs do agree with available empirical data, it is not easy to tell whether it is only due to adjustments or to the model's core hypotheses. The model faces the court of experimental data as a whole in such a way that it is not easy to determine the precise role of each of its components.

However, models do not all suffer in the same way from the Duhem problem. According to their goal and component hypotheses, it is more or less easy to overcome the Duhem problem. Accordingly, empirical agreement is endowed with different meanings in different modeling situations. In order to account for these differences, we put forward a typology of models in the following.

At last, we put forward a special type of models that illustrates another difficulty in interpreting empirical agreement. This new difficulty is perhaps even more troublesome for the use of models than is the Duhem problem.

1. The Duhem problem

Even though empirical agreement does play an important role in the activity of model validation, it cannot be considered a straightforward criterion for model validation. Here, we understand validity is a purpose-relative notion. A valid model is one that performs the task for which it has been designed, whether predictions, experiment planing, prototype construction, etc. Even though validity is so construed as to be purpose-relative, empirical agreement seems to play an important role it assessing it. Why isn't empirical agreement a simple criterion for model validity, though? Because when a model's outputs are consistent with data acquired by observation or measurement, it is usually not possible to assess to which element within the model this match is due. More precisely, it is not possible to tell whether it is due to adjustments in the model or to the fact that the model's hypotheses accurately represent the underlying processes accounting for the investigated phenomena. Adjustments are mainly of two sorts: a model is "adapted" to the phenomenon at hand either by calibrating it or by introducing *ad hoc* terms into it. Calibration is a (usually long) process consisting in tuning some parameters — i.e., numerical constants in the model — in order to progressively guarantee that the model

outputs better fit the database. Generally, the outputs to fit are associated with observable variables while the parameters to tune are neither known from observation and measurement nor from theories. This process has been described and analyzed by a number of people, including Winsberg (1999), Hitchcock and Sober (2004), Epstein and Forber (2013). The second sort of adjustment consists in introducing *ad hoc* terms into the model's equations in order to compensate initially omitted target features in the model. These terms are either correlations between variables of the model, which are not derived from theories, or measured values. (For instance, air friction is often neglected in order to describe a free falling body, but an *ad hoc* term to account for air friction can be added in order to study light objects). This process has been extensively described and analyzed by Cartwright (1983, Essay 6 "For phenomenological laws").

In order to better characterize the Duhem problem, we first give a short analysis of these complex objects called models by describing their usual components. This will allow us to present a more precise diagnosis of the Duhem problem.

1.1 Models and their components

A model is a representation of a class of phenomena that may have a variety of goals. According to the desired degree of generality of this representation and the modeler's goal, the importance of the various components may vary quite a lot. For instance, there are models whose inability to provide accurate predictions is not considered unacceptable, when their goal is mostly heuristic, whereas for others, precision is a major virtue. In order to shed some light on this diversity, we put forward a distinctive analysis of models based on the identification of the nature and function of their components.

In order to devise a fully general analysis, we both include models that are written and solved by hand, which we call "analytical models", and computational models, which are written for, and solved by computers. This amounts to saying that we do not focus on the purely conceptual aspect of models, that is, their instantiating theoretical principles. On the contrary, we want to include a large variety of representations that are used for various tasks, from prediction to experimental design or artifact production.

The first distinction we introduce is between the conceptual components of models and the components that transform them into *usable tools* (of investigation, prediction, design, etc.). This distinction is meant to capture the intuition that contrary to theories, models encompass elements that do not possess strict justification but are required for these models to be *applied* to concrete situations.

The conceptual components of models are themselves of two sorts: first, the description of the target system's properties and second, a set of equations supposed to represent the behavior of the target phenomenon. The description of the system's properties consists in a selection of properties that are supposed to bear on the problem at hand; as models are representations whose scope and validity are determined by their purpose, this description is not supposed to hold absolutely, but only locally, that is, for the concrete situation at hand, and for a specific purpose. The same is true of the model's equations: they are not designed in the first place to hold for ever, but are only meant to help modelers solve the problem they face. Thus the validity of a model is assessed in terms of the accuracy of its predictions *or* the correctness of its core hypotheses *depending on* its specific purpose. For instance, some statistical models used in life insurances are considered valid only

because, in virtue of the adjustments they contain, they give accurate predictions; it is not expected from them to include any substantial hypotheses about subject matter[1].

Both the description of the system's properties and the equations are written by relying on already established modeling practice in the relevant domain: they do not come out of nowhere. However, they need not be exclusively grounded on available theories. Sometimes, the equations are just meant to catch a basic type of behavior, like the linear dependence of one variable of interest against another. It is also important to emphasize that in many cases, the assessment of the model's quality is not based on the quality of the representation *per se*, but only relative to the variables that have been identified as interesting ones for the purpose at hand. Purpose-relativity is a major component of the way empirical agreement is taken into account when performing the validation task.

As emphasized above, if one is willing to account for models as *usable tools*, it is necessary to include other components than the conceptual ones and to mention simplifying assumptions, idealizations, approximations, to which we come back below, but also algorithms and computational schemes that are essential parts of computational models. Let us take the ballistic equation as an example to illustrate the difference between our two types of components. The ballistic equation is:
(1) $d(v\cos\tau)/d\tau = c/g\ vF(v)$
where v is the projectile's velocity, τ is the angle between its direction and the horizontal, g is the gravitational constant, c and F express air resistance.
This equation can be derived from Newton's second law and is thus well justified. However, it only holds when the following idealizations are accepted: the projectile is a point mass, there is no wind, and the Earth is flat. For sure, in any concrete situation, at least the first two idealizations have to be dismissed and the equations transformed accordingly. But there is worse, as the ballistic equation is only integrable in very few cases. So there is still another reason why it has to be transformed in order to be applied to a concrete situation.

One may have the impression that the components allowing a model to be usable are inessential, because the genuine scientific content is carried by what we have called the conceptual components. However, this impression is erroneous. The characteristic feature of a model, as opposed to a theory, is precisely to include, as unremovable components, those elements that allow scientists to use it for whatever purpose they may have determined.

In order for models to be usable, model equations need to be (1) expressed in mathematical terms and (2) (analytically or numerically) tractable. Both requirements entail including simplifying assumptions. This both holds for models that are written and solved by hand and for computational models. Simplifying assumptions come in two sorts, approximations and idealizations. Approximations are modifications of the equations that are governed by tractability requirements: they are needed to find out the solutions to the equations (Redhead 1980; Ramsey 1990, 1992; Laymon 1989a, 1989b). For instance, equation (1) can only be integrated when F has special properties; otherwise it has to be replaced by e.g. polynomials. Why by polynomials? Simply because we know how to perform the integration. The choice of an approximation is determined by whatever mathematical or computational method is available. Idealizations are convenient omissions

[1] We are grateful to our anonymous reviewer for mentioning this example to us.

or deliberate deformations of the target system's features[2]. For instance, in the original representation, the target system's behavior is supposed to depend on temperature, but in the usable model, the target system is idealized as isolated from its environment, so that the temperature dependence is removed. For sure, the distinction between approximations and idealizations is not always easy to make, as some features of the target system are represented by mathematical objects or relationships; however, it is useful to keep it in mind in the analysis of Duhem problem because, unlike idealizations, approximations are often justified *a posteriori* and their role in the empirical (dis)agreement of the model is therefore more difficult to determine.

The approximations are designed differently in analytical and simulation models. While in analytical models the approximations aim at making the equations solvable by hand, in simulation models they aim at making the equations solvable by the computer. In other words, a *simulation model* — or computational model — is a digital translation of the conceptual model containing the required approximations for the resolution of the equations on the computer. In simulation models, the approximations are determined by the chosen *numerical method* (e.g. Runge-Kutta's method, the finite difference method, or the Monte Carlo method). For instance, one can apply either analytically or numerically the Navier-Stokes equations to the problem of the fluid past a cylinder. In the analytical model of Prandtl, the fluid is represented as having two interacting components which both contain approximations: The first component is the boundary layer which is around the obstacle and whose viscosity is not zero and velocity profile evolves linearly. Within the second component, the wake, the viscosity is zero and the viscosity is constant. In a discretization-based numerical approach, the Navier-Stokes equations are integrated following a finite element method. This method involves two main approximations: first it consists in discretizing the domain under study into finite elements, second the partial differentials of the equation variables are replaced by approximate values which obtain from the values of the variables at the nodes of each finite element.

The *code*, written in a computer language (e.g., C++ or fortran), translates the algorithm which allows the computer to process calculations based on the simulation model. The expression "computer simulation" is both used to designate these calculations and the simulated behavior of the targeted system on the computer screen; we avoid the second use in this paper. The numerical data are the solutions resulting from the computational process. When they come from differential equations, they are generally converted into visual representations, i.e., tables, graphs, pictures, or films. Once converted, numerical data can be interpreted as simulation outputs. These outputs can then be directly compared with empirical data.

1.2 Validation

Now that we have presented our conception of a model's components, let us turn to the notion of validation of a model. This notion has mostly been addressed in specific circumstances of complex simulation models such as climate models (e.g. Parker 2009; Lenhard and Winsberg 2010) but here we are treating simple as well as complex models equally in order to frame a general explication of the notion of validation. First, we emphasize that it is only what we have called the *usable* model that can be validated, that

[2] There is presently no commonly agreed upon definition of "idealization" in a scientific model (see e.g. Jones 2005, Godfrey-Smith 2009, and Weisberg 2007 for various definitions).

is, the representation that has been built up for a specific purpose. Accordingly, as emphasized above, validation can only be meaningful with respect to this purpose.

Through the validation process, the modeler assesses whether the model fulfills its purpose: whether it provides her with precise predictions, with a practical experimental design, etc. For sure, checking whether the model's outputs are consistent with available empirical data is an important step within the validation process, but it cannot be the only one because validation is holistic, as we make clear in the following.

What is to be validated? It is not only the conceptual components, supposed to be the only bearers of scientific content, but the representation *as a whole*. What does it mean to be valid for a model? This means that the model can be used for a specific purpose, like prediction, further investigation, building up experimental setup, … A valid model is one on which one can rely in the task it was designed for.

The validation process includes two types of comparison: comparison with available data and comparison with models which are already well-accepted and well-used in virtue of the highly confirmed underlying theories or in virtue of their past predictive successes. The comparison with available data is not a one-shot act but is carried only through multiple iterations. Thus, empirical agreement is not given once and for all, but is the result of a series of actions on the part of the modeler. In the same way, the comparison with current models that are accepted and used in comparable cases includes plausibility assessments as there is no algorithm for validation. Validation cannot be judged on any absolute basis from any single element that would be detached from the other components of the model.

From the above description of the validation process, it is easy to infer why empirical agreement cannot be a unique criterion of validation. Indeed, the question rather seems to be to what extent empirical agreement could indicate something other than the effect of model adjustment. How can it be determined that when a model agrees with available data, it is *not* because it is too well adjusted to fit those data? If this is the case, it is not possible to infer from empirical agreement that the original description and equations are correct or even good enough for the purpose at hand. Let us suppose that a model has been adjusted and model parameters have been tuned or *ad hoc* terms have been introduced into the equations. Then, when empirical agreement is obtained, it is impossible to isolate the set of elements that can be said responsible for this achievement. This is Duhem problem of confirmation and refutation holism.

Duhem problem is the very reason why “over-fitting” or “over-tuning” is a genuine risk for modelers (see Epstein and Forber 2013, Hitchcock and Sober 2004). Over-fitting results from the same process as calibration, but it can be described as twisted calibration for it ends up with an undesirable product as the over-fitted model is less falsifiable than the correctly calibrated model. A model is over-fitted to the data when a great number of parameters are tuned. In this case, “[t]he problem with tuning is that it artificially prevents a model from producing a bad result” (Randall and Wielicki 1997, p. 404). Here, “ “tuning” a model, or tweaking the micro-parameters whenever we get results we do not like, can amount to slapping an ad hoc bandaid on a broken model, insulating the model from any empirical risk.” (Epstein and Forber 2013, p. 204). The chance that a model will be invalidated is low or non-existent when it is a precondition for successfully sanctioning a model.

As validation is thoroughly purpose-relative, it is now time to introduce a classification of models according to their purpose that will allow us to refine our analysis of Duhem

problem. We shall then have a better grip on the role empirical agreement can play within the validation process according to the type of model at hand.

2. Differential sensitivity to Duhem problem

In section 1, we have made clear why empirical agreement cannot be the only criterion for a model's validity. Confirmation and refutation holism precludes modelers to consider empirical agreement as a reliable guide to the validity of each component of a model when taken in isolation from the others. However, models of different types are not equal in front of holism. According to their purpose and components, they can use different weapons to face this problem. In this section, we first present a basic typology of models and then we analyze how they can cope with confirmation/refutation holism. This typology, albeit conceptual, actually reflects the way modelers classify the models depending on their components, and therefore the way they build their judgments about the models' validity in practice.

2.1 Different types of models

Basically, models can be generated by following two different routes: first, by starting from available data; second, by starting from available theoretical knowledge. Accordingly, we can identify four types of models, according (i) to the way they were generated and (ii) to the distance they are from their origin.

Route 1: Statistical models
The starting point of statistical models is a (usually large) set of data among which regularities are searched for. These regularities are then used to build up both the description of the target system and the equations governing its behavior. The latter are looked for automatically, without reliance on any previously available theoretical knowledge. This is why statistical models are both theoretically blind and closer from their origin, i.e., sets of data.

Route 1: Phenomenological models
Phenomenological models are not entirely generated by the identification of patterns within data sets but their aim is to reproduce these patterns from postulated mechanisms or processes that are known not to be the right ones. Empirical agreement is thus the primary aim of the building up of phenomenological models. However, they are a little further from their origin in data sets because they encompass some hypothesis.

In order to better capture the difference between statistical and phenomenological models, let us take linear regression as an example[3]. When a linear regression model is obtained from a data set by purely automatic procedures based on e.g. learning algorithm, it is statistical. However, when the modeler has a previous idea of the relevant correlations and looks for them within the data said, this is a phenomenological model because some form of (however informally formulated) hypothesis has been included into it.

Route 2: Theoretical models
Theoretical models aim at representing the physical processes underlying the phenomenon under study. They have a broader scope than the restricted domain of the

[3] We are grateful to our anonymous reviewer for mentioning this example to us.

phenomenon being studied. They contain theoretical hypotheses, that is, hypotheses based on currently accepted theories. Within the analysis we are putting forward, theoretical models are defined by their reliance on existing knowledge. For instance, hydrodynamics models are based on fluid mechanics, models of galaxies on Newtonian physics. Contrary to Route 1 models, theoretical models draw their content from their connection with surrounding theories and models. Nevertheless, this is not to say that their validity is always *justified* by this connection, because it may obtain as a result of calibration or the addition of ad hoc terms when the involved theoretical hypotheses do not apply to the target phenomenon.

Route 2: Formal models
Formal models are very idealized. They are not used to recover data but to identify asymptotic behaviors of the target system. Examples of formal models are prey-predator models and various types of ecological models. These models play a heuristic role but are not based on already established knowledge. As such, they are further from their origin than theoretical models.

It is clear from the above list that empirical agreement cannot play the same role in the validation of these different types of models. On the one hand, empirical agreement can only play a role in the validation of generalizations or extensions of statistical models as statistical models are empirically valid for their target by construction. On the other hand, empirical disagreement cannot be a criterion of rejection for formal models which do not aim at accurate predictions. In the next sections, we focus on phenomenological and theoretical models and show how it is sometimes possible to circumvent the problems raised by confirmation and refutation holism. Throughout this section, we want to make clear that validation and refutation holism is a well-known obstacle to model appraisal. As such, it is taken care of by several methods, none of which can provide any final solution.

2.2 Empirical agreement - phenomenological models

As emphasized above, models are representations whose validity is purpose-relative. Therefore, the role empirical agreement in the validation of models is different for models with different purposes. We first focus on phenomenological models, which aim at empirical agreement. In order to illustrate this point, we present two examples. The first one is a model of the animation of the sea and the second one is a traffic flow model.

(i) Interactive Phenomenological Animation of the Sea (IPAS) (Parenthoën 2006) is a system of virtual sea whose purpose is to help mariners to decide on sailing strategies. The IPAS model is based on a mathematical description of sea patterns in terms of wavelets. "Wavelets" here are mathematical entities; the behavior of the virtual sea is not based on any physical principle, like energy, mass or momentum conservation. The wavelet tool allows the modeler to reproduce the relevant aspects of the sea as experienced by the mariners. The purpose of IPAS is to produce a substitute to the motion of waves as they are perceived by the mariners but this substitute does not need to be produced from the known underlying, physical laws.

It is clear that the aim of IPAS is *not* to explain the phenomenon of swell for example, or any other sea motion. Empirical agreement is no indication of the validity of any physical hypothesis; it is a necessary condition to the use of the model as a substitute of actual sea to test craft shells. Moreover, the validity of this model, that is, its capacity to help mariners, is strictly restricted to the concrete situation at hand.

(ii) Nagel and Schreckenberg 1992 traffic flow model is used to predict traffic jams. The rules of the model govern the behavior of each car-driver depending on the distance to the next car ahead: if the distance is larger than a fixed size, then the car-driver accelerates (limited by a maximal value though); in the opposite case, she reduces the speed of her vehicle. The rules of the model are not supposed to be false *strictu sensu* but the model is phenomenological nevertheless because first it is only based on these basic rules and not on other pieces of knowledge which could make the model more general (according to gender, age, weather conditions, period of the day, of the year, etc.), and second these basic rules contravene the best knowledge we have from sociological studies on the behavior of car-drivers.

In the case of the traffic flow model, empirical agreement is usually not considered as implying the superiority of the basic rules over more complex pieces of knowledge about the behavior of car-drivers. It is required for the use of the model in concrete, local predictions setups, but it does not entitle anyone to produce social-science generalizations about car-driving on highways or about any kind of human behavior.

2.3 Empirical agreement - theoretical models

Whereas the validation of phenomenological models does not involve the confirmation of any theoretical hypothesis, it is the distinctive feature of theoretical models to include theoretical hypotheses so that their validation may involve the confirmation of such hypotheses -- however, it is not always the case. In this section, we present a simple, even simplistic example to shed light on the subtle relationship between empirical agreement and the confirmation of theoretical hypotheses in theoretical models.

Our example is the model of a free falling body. This model includes Newton's second law as a theoretical principle. It is applied to a body of mass m in free fall. One considers that the body is dropped at zero velocity at time t = 0 from height h. Besides, an idealization is often introduced: one assumes that the fall occurs without air friction, i.e., in void. One then solves the equation provided by Newton's second law for which the sum of the forces on the body is equal to its mass multiplied by its acceleration ($\Sigma \mathbf{f} = m\mathbf{a}$). The solution of the free fall time is $t = \sqrt{(2h/g)}$. A numerical application allows one to compare the possible results with corresponding measures (that could have been obtained during an experiment which, for example, would have consisted in dropping the body from the top of a building and timing the free fall).

What conclusion can be drawn from the comparison of the numerical results with measurement results? In this case, it is possible to bypass Duhem problem as we now show. Strictly speaking, the comparison leads on the confirmation of Newton's second law and the idealization that there is no air friction, all at once. However, it would be awkward to stick to this conclusion as Newton's law is already highly confirmed. As a result, the correct conclusion is that the "no air friction" idealization is correct. Conversely, empirical disagreement would make one blame the no friction idealization. Thus, in this example, empirical disagreement points to the source of a possible misrepresentation in the model either among the simplifying assumptions or the approximations (due to the numerical scheme). (In this example, we deliberately let aside the case where measurement procedures depend on the theory or hypothesis being tested).

2.4 Validating the computational process

As clear from the previous section, even though the enterprise of model validation faces Duhem problem, it is not blocked by its effects. It is even fair to say that a major part of the activity of modeling is devoted to fighting against Duhem problem, with some success. For instance, as we have seen, some idealizations are relatively easy to isolate from other components.

As another example of a successful fight against Duhem problem threatening the validation of a theoretical model, let us examine a case in which empirical agreement has been used to check the quality of the numerical scheme involved in a computational model. In order to do so, Shirayama and Kuwahara have compared von Kármán vortex patterns in a flow past a cylinder (Shirayama and Kuwahara 1990). Figure 1 presents the experimental and computed streak-lines in flow past a circular cylinder at Reynolds number equal to 140. On the left, one can see the pattern of particle traces that correspond to the experimental streak-lines obtained using the electrolytic precipitation method. On the right, the simulation output; the simulation was done by representing the same physical properties (e.g., the same fluid viscosity) and conditions (e.g., same Reynolds number).

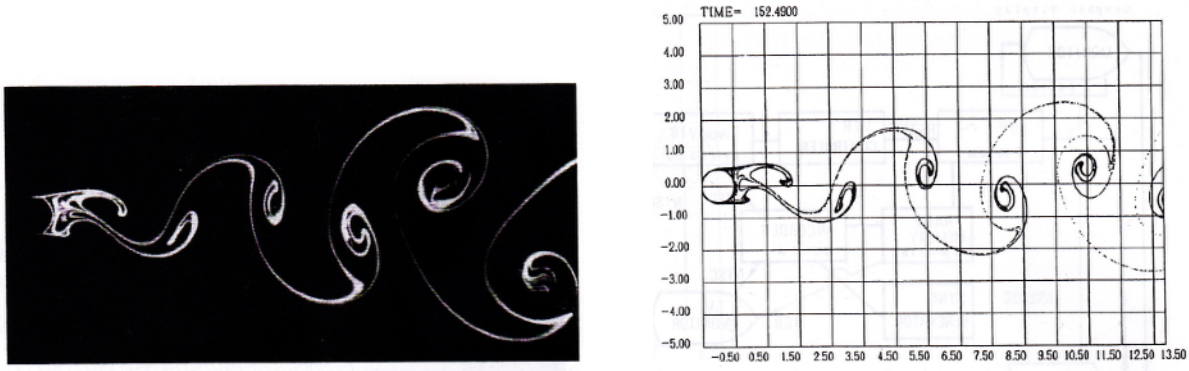

Figure 1. On the left, experimental streaklines in flow past a circular cylinder at Re = 140. On the right, pattern of particle traces at Re = 140 (computed streaklines) (from Shirayama and Kuwahara 1990, p. 73)

Shirayama and Kuwahara obtain good agreement, assessed by pattern resemblance, between numerical and experimental results despite the strong nonlinearity of the flow field. For them, “this indicates that the computation is of high quality. The conclusion that the coincidence of two streak-line-patterns verifies not only the spatial accuracy but also that the time accuracy can be derived from the nature of the streak-line patterns, which reflect past behavior.” (1990, p. 73-74) Pattern resemblance means that the velocities in the model agree with the data. What can we infer from pattern resemblance? In this case, it is not used as a basis to infer that the fluid equations correctly represent the phenomenon and predict the velocities, simply because these equations are already known to be well-confirmed. It is used as a basis to infer that the numerical scheme is well adapted to the phenomenon at hand.

In this example, if empirical disagreement occurred, it would have been the best guide toward the resolution of computation or modeling problems since it would have helped one to discriminate the source of a potential misrepresentation in the model within idealizations and approximations.

The first conclusion we can draw from sections 2.3 and 2.4 is that whereas Duhem problem impinges upon the validation process of all models, some components of theoretical models are nevertheless more "Duhem-problem sensitive" than others. The latter can be more easily isolated from the remaining of the model and tested out. As a result, it is possible to discover ways out of confirmation and refutation holism and a large part of the work of modelers is precisely devoted to search for these ways.

2.5 A well-grounded theoretical model

Contrary to the simplistic example we have presented in section 2.3, some theoretical models are used as vehicles for the confirmation of theoretical hypotheses. In this section, we discuss the conditions at which it is possible. Before doing so, we want to make clear that we put aside the question of the legitimacy of data as evidence and the question of their reliability because, although they do pertain to the confirmation task, we want to keep our discussion of empirical agreement isolated from these larger themes.

Let us first emphasize that model validation is a different operation from hypothesis confirmation. The difference is related with the holistic character of model validation, which is actually an obstacle to hypothesis confirmation. However, as a matter of fact, it is often necessary to use model validation as an intermediary step toward hypothesis validation. Duhem problem forbidding any simple inference from the former to the latter, it is now time to look into uses of empirical agreement when hypothesis confirmation is the task to perform.

Empirical agreement, when it is not only due to calibration or the introduction of *ad hoc* terms, indicates that the model, as a whole, is valid in the limited domain within which the comparison with available data has been performed. However, to say that a component hypothesis has been tested implies that its domain of validity is larger than the limited domain in which empirical agreement has been verified.

What does "larger" mean in this context? Two interpretations can be put forward. According to the first, a "larger" domain of validity means that the hypothesis allows for more numerical predictions for variables and parameters about which no empirical data are available. The larger domain is obtained by extending the range of variables for which the model can provide outputs. The second interpretation relies on the capacity of certain hypotheses to *explain* phenomena besides allowing for predictions. This capacity is usually considered to hold in a larger domain than the one in which empirical agreement obtains.

The distinction between the two interpretations of a "larger" domain of validity than the one determined by empirical agreement is easy to establish in retrospect or "from above", that is, from a towering point of view that is inaccessible to scientists engaged in a research work. (We borrow the expressions "from above" and "from within" to van Fraassen (2008)). At the frontier of research, however, it is by definition not known with certainty which hypotheses are explanatory and which are only predictive. When one adopts the perspective "from within", that is, the perspective of the scientists who build up a new model and try to assess the quality of the hypotheses they have used therein, without being able to rely on established knowledge, the distinction is not readily apparent. Is it however possible to find out elements allowing modelers to decide whether the first or the second interpretation is relevant in the case at hand.

A tentative answer is that hypotheses remaining valid under systematic variations are better candidates than hypotheses that do not yield empirically plausible results when twisted in order to explore domains neighbouring the original domain. To put it in other terms, robust hypotheses are better candidates than phenomenological hypotheses. Here, robust hypotheses are not only assessed within a set of models that all represent a same target system (as in climate science; see Parker 2010a, 2010b, 2013) but more generally within as broad as possible a set of models used to represent distinctive target systems for different purposes. This distinction would solve the problem if only it could be defined unequivocally. However, assessing the robustness of a hypothesis can no more be done by any algorithm than any kind of model validation. This assessment is a matter of the modeler's judgment and is always open to discussion and criticism.

The notion of robustness can be further analyzed in the following way. As said above, hypothesis testing is often only possible via model testing. Now, model validation is local. In order to test the validity of a hypothesis beyond the scope of the original model, it is thus necessary to try and validate several models within different, possibly overlapping domains. If a hypothesis is included into several, distinct models which have been validated within different domains, it may be said robust. However, as already made clear, such a judgement cannot be definite.

2.6 Conclusion to section 2

In this section, we have seen that despite Duhem problem, empirical agreement can play meaningful and important roles in the process of model validation. Inferences can be made on this basis if sufficient control is guaranteed on other components of the model. This is usually not easy, but it seems clear that Duhem problem is no insurmountable obstacle for the task of model validation when validation is conceived as a matter of judgement rather than as susceptible to receive any definite answer.

3 The strength of empirical agreement

In this section, we go further into our analysis of the meaning of empirical agreement in modeling practice. After having pointed out that Duhem problem, although pervasive, is nevertheless the object of systematic, deliberate, and relatively successful work from modelers, we now turn to another problem that has more seldom been identified. This problem emerges as a potential conflict between empirical agreement and established knowledge, when empirical agreement is unexpected or left unexplained. We shall present this problem through an example.

The example we focus on in this section is a computational version of Turing’s model of morphogenesis. Morphogenesis is the creation of patterns, forms or structures in living organisms during their development, like, for example, the different stages in embryogenesis, the outbreak of pigmented spots on skins and furs. In 1952, Turing proposed a mathematical model of morphogenesis in which a system of chemical substances called the "morphogens" react together after an instability created by random perturbations and diffuse through a tissue in accordance with reaction-diffusion equations. Although the morphogens are initially distributed homogeneously in the tissue, a pattern or structure may later be developed further to random disturbances.

Computational versions of Turing model have been developed and agree with observations (Kondo and Arai 1995; Watanabe and Kondo, 2012). For example, Kondo and Arai (1995) have managed to successfully reproduce the visible skin pattern of the marine angelfish Pomacanthus, which has a stripe pattern. The stripes of the fish are parallel, but a small perturbation may modify their structure. Nevertheless, after a time, the stripes rearrange themselves and recover their initial configuration. The Turing model predicts this rearrangement.

The reason why we have chosen this example is that at first sight, it looks as a bona fide phenomenological model as the mechanism that is introduced therein to produce the skin pattern is *fictitious*. For all we know, there aren't any morphogens in animals. Thus, we have here an example in which empirical agreement is good without the underlying hypotheses being considered plausible or even admissible.

What are we to conclude from such an example in which empirical agreement is very good, but for bad reasons, relative to our background knowledge? The authors conclude that “the striking similarity between the actual and simulated pattern rearrangement strongly suggests that a reaction-diffusion wave is a viable mechanism for the stripe pattern of Pomacanthus” (p. 765). Therefore, to their view, empirical agreement has the power to transform what has commonly been viewed as a phenomenological model into a theoretical model including an explanatory hypothesis.

Is empirical agreement endowed with such a power to transform a phenomenological model into a theoretical model? In order to answer this question, let us first recall how we analyze the difference between the two types of models. A phenomenological model may include assumptions about underlying processes producing the phenomenon to be represented, in order to achieve empirical agreement with available data, but these processes are *not* conceived as being explanatory. The main difference between phenomenological and theoretical models, according to our view, is that theoretical models include hypotheses that are consistent with established knowledge and attempt at developing it, whereas the processes postulated in phenomenological models can contradict established knowledge. For sure, it is not always clear whether an assumption is consistent with established knowledge or diverges therefrom. It may be a matter of plausibility assessment to decide on which class a model belongs. However, the original Turing model clearly belonged to the class of phenomenological models because nothing in the relevant background implied the existence of morphogens or the prevailing importance of diffusion mechanisms.

Turing model turned into a theoretical model from the effect of surprisingly good empirical agreement is a striking example of the somewhat confused status of empirical agreement. However, phenomenological models are only seldom turned into theoretical ones. This is only the case when no other convincing explanation is available. This situation holds for Turing model, but not generally (e.g., it does not hold for the traffic flow model). Thus the possibility that a model turn theoretical is contingent upon the existence of competing models.

To conclude, empirical agreement, in some cases, in much more powerful than one can infer from the sole consideration of Duhem problem. It is a genuine driver of scientific change in spite of confirmation and refutation holism.

Conclusion

We have claimed that empirical agreement is not a sufficient criterion for a model's validity, because it faces Duhem problem of refutation and confirmation holism. However we argued that models do not all suffer in the same way from this problem.

Furthermore we have shown that the most important issue is not Duhem problem itself, but arises when empirical agreement conflicts with established knowledge. When established knowledge is missing, empirical agreement hardly helps indicate whether the model hypotheses can be said theoretical or whether they are merely phenomenological.

References

Epstein, B. and Forber, P. (2013). The perils of tweaking: how to use macrodata to set parameters in complex simulation models. *Synthese* 190:203–218

Godfrey-Smith (2009) “Models and Fictions in Science”, *Philosophical Studies* 143, 2009, 101-116.

Hitchcock, C., & Sober, E. (2004). Prediction versus accommodation and the risk of overfitting. *British Journal for the Philosophy of Science*, 55, 1–34.

Jones, M. R. and Cartwright N. (eds.) (2005) *Idealization XII*: *Correcting* the *Model*. *Idealization* and *Abstraction* in the *Sciences*. Amsterdam/New York, NY: Rodopi.

Jones, M. R. (2005) “Idealization and Abstraction: A Framework,” in Jones, M.R. and Cartwright, N. (2005), pp.173-217

Kondo, S., & Arai, R. (1995). A reaction-diffusion wave on the skin of the marine angelfish Pomacanthus. *Nature*, 376.

Laymon, R. (1989a). Applying Idealized Scientific Theories to Engineering. Synthese, 81(3):353–371.

Laymon, R. (1989b). Cartwright and the Lying Laws of Physics. Journal of Philosophy, 86(7):353–372.

Lenhard, J., & Winsberg, E. (2010). Holism, Entrenchment, and the Future of Climate Model Pluralism. Studies in History and Philosophy of Science Part B, 41(3):253–262.

Nagel, K. and Schreckenberg, M. (1992) “A cellular automaton model for freeway traffic”, *Journal of Physics I* France , 2:2221-2229.

Oberkampf, W. L., Trucano, T. G. (2002) Verification and Validation in Computational Fluid Dynamics. Rapport Sandia. SAND2002-0529

Oberkampf, W. L., Trucano, T. G., Hirsch, C. (2002) Verification, validation and predictive capacity in computational engineering and physics. Applied Mechanics Review, Volume 57, Issue 5, 345.

Parenthoën, M. (2006) *Animation phénoménologique de la mer. Une approche énactive*. PhD dissertation in computer science, Université de Bretagne Occidentale.

Parker, W. (2009) Confirmation and Adequacy-for-Purpose in Climate Modeling. Aristotelian Society Proceedings, Supp. Volume.

Parker, W.S. (2010a) Predicting Weather and Climate: Uncertainty, Ensembles and Probability. *Studies in History and Philosophy of Modern Physics* 41: 263-272.

Parker, W.S. (2010b) Whose Probabilities? Predicting Climate Change with Ensembles of Models. Proceedings of PSA08. Philosophy of Science 77(5): 985-997.

Parker, W. (2013) "Ensemble modeling, uncertainty and robust predictions", *Climate Change*, 4:213–223.

Ramsey, J. L. (1990). Beyond Numerical and Causal Accuracy: Expanding the Set of Justificational Criteria. Proceedings of the Biennial Meeting of the Philosophy of Science Association, 1:485–499.

Ramsey, J. L. (1992). Towards an Expanded Epistemology for Approximations. Proceedings of the Biennial Meeting of the Philosophy of Science Association, 1:154–164.

Redhead, M. (1980). Models in Physics. *British Journal for the Philosophy of Science*, 31(2):145–163.

Trucano, T. G. , Swiler, L. P., Igusa, T., Oberkampf, W. L. and Pilch, M. V (2005) Calibration, Validation, and Sensitivity Analysis: What's What. Submitted to the *Journal of Reliability Engineering and System Safety*

Turing, A. M. (1952). The chemical basis of morphogenesis. *Philosophical* Transactions of the Royal Society of London. Series B, Biological Sciences. 237(641):37–72.

van Fraassen, B. C. (2008). *Scientific Representation: Paradoxes of Perspective.* Oxford University Press.

Watanabe M. and Kondo S. (2012) Changing clothes easily: Connexin41.8 regulates skin pattern variation. *Pigment Cell Melanoma Res*. 25(3):326–330.

Weisberg, M. (2007) "Three Kinds of Idealization," *The Journal of Philosophy*, 104 (12) 639-59.

Winsberg, E. (1999). Sanctioning models: The epistemology of simulation. *Science in Context*, 12(2):275–92.

Winsberg, E. (2010). *Science in the Age of Computer Simulation*. The University of Chicago Press.