



INSTITUT DE FRANCE
Académie des sciences

Comptes Rendus

Géoscience

Sciences de la Planète

Vazken Andréassian

On the (im)possible validation of hydrogeological models

Volume 355, Special Issue S1 (2023), p. 337-345


Online since: 27 September 2022

Issue date: 23 February 2024

Part of Special Issue: Geo-hydrological Data & Models

Guest editors: Vazken Andréassian (INRAE, France),
Valérie Plagnes (Sorbonne Université, France), Craig Simmons (Flinders University,
Australia) and Pierre Ribstein (Sorbonne Université, France)

<https://doi.org/10.5802/crgeos.142>

 This article is licensed under the
CREATIVE COMMONS ATTRIBUTION 4.0 INTERNATIONAL LICENSE.
<http://creativecommons.org/licenses/by/4.0/>



*The Comptes Rendus. Géoscience — Sciences de la Planète are a member of the
Mersenne Center for open scientific publishing*

www.centre-mersenne.org — e-ISSN : 1778-7025



Research article

Geo-hydrological Data & Models

On the (im)possible validation of hydrogeological models

Vazken Andréassian ^a

^a Université Paris-Saclay, INRAE, HYCAR Research Unit, Antony, France

E-mail: vazken.andreassian@inrae.fr

Abstract. This paper revisits the controversy on the validation of hydrogeological models, 30 years after it broke out with the publications by Konikow and Bredehoeft [1992a] and de Marsily et al. [1992]. In that debate, Konikow and Bredehoeft [1992a] argued that the word “valid” was misleading to the public and should not be used with respect to models. de Marsily et al. [1992] answered that while the bases of hydrogeological models (conservation of mass and Darcy’s law) were uncontested and unconditionally valid, specific validation exercises were dearly needed to evaluate the parameters and the geometry of these models (confronting the models with data they had not seen during the calibration phase). By updating and extending the literature review, we reanalyze this debate and the arguments presented and conclude by proposing an extension of de Marsily’s position, which underlines the necessity to look at validation from two distinct viewpoints, i.e. the point of view of the model’s explanatory power (theoretical content) and the point of view of its predictive power. The explanatory and predictive dimensions of model validation are to be considered separately.

Keywords. Hydrogeological model, Model validation, Corroboration, Falsifiability, Ghislain de Marsily.

Manuscript received 24 March 2022, accepted 11 July 2022.

1. Introduction

While validation exercises cannot ensure perfection, they help hydrogeologists achieve their level best by increasing their confidence in the model used: This is how one could summarize the position of de Marsily et al. [1992] in the model validation debate that set them in opposition to Konikow and Bredehoeft [1992a]. The argumentation of de Marsily et al. [1992] was quite straightforward: They insisted that model validation was an essential exercise for hydrogeology, and that it was excessive to call upon the Popperian vision of falsifiability [Popper, 1959] to renounce testing exhaustively hydrogeological models: “Groundwater flow models rely essentially on two concepts: (i) mass balance, (ii) Darcy’s law. The former is a principle, not a theory. No one is going to seriously argue that the mass conservation principle may one day be invalidated. [...] Darcy’s law is not a theory; it is an empirical observation, which is applied in a huge number of cases (although it can be in error in a few very special cases, and even so, the departure from the linear Darcy law will be of little significance in most applications)” [de Marsily et al., 1992, p. 367].

Beyond the principles that they considered pointless to contest, de Marsily et al. [1992] insisted that both the parameters and the geometry (the structure) of a hydrogeological model remain uncertain and this is precisely why the so-called *validation* exercises were needed: to either refine the model progressively, or to confirm the robustness of past parametric and structural choices. They argued that the validation exercises were meaningful and that they necessitated using the model in a predictive mode and confronting it with data it had not seen during the calibration phase. This process “increases the confidence” in the model in question, and even if certainty and perfection remain out of reach, this is already a worthy result.

In their response, Konikow and Bredehoeft [1992b] wrote that “using the word ‘valid’ with respect to models misleads the public” and makes hydrogeologists “look foolish to our scientific colleagues”. However, they agreed that the exercise they called “postaudit” (and which consists in revisiting past predictions after a few years) was useful.

As in all controversies, the vocabulary used is not always well defined. “Valid” comes from the Latin

“validus”, meaning “strong, healthy”. The concept of validity has a precise definition in logic, that of a univocal link between the premises and the conclusion of an argument (i.e., if the premises are true, then the conclusion has to be true). It is a well-established concept in law, where a norm is valid if conditions of form (the procedure is respected) and substance (the superior rules of law are respected) are ensured. It is not precisely defined in hydrogeology, where neither Konikow and Bredehoeft [1992a] nor de Marsily et al. [1992] provided a clear definition (we would not consider Konikow and Bredehoeft’s definition of validation as “a process that can guarantee that a model is a correct representation of the physical world” to be precise, because the term “correct” is as vague as the term “valid” was in the first place).

Thirty years have passed since the publication of the articles by de Marsily et al. [1992] and Konikow and Bredehoeft [1992a], and we posit that it is time to propose a critical appraisal of this debate, in the light of more recent contributions on the model validation issue.

2. Further contributions from hydrogeologists

Several hydrogeologists brought a further contribution to the debate: Carrera et al. [1993] started by stressing that to them an accurate characterization of geological media was “absurdly utopic”, adding that due to the numerous unknowns and uncertainties in both physical processes and underground media properties, validation was a “rather elusive concept, probably more controlled by the modeler’s background and views of reality than by actual facts”. They insisted on the fact that the qualitative nature of many observations necessarily results in a somewhat subjective conceptualization by the modeler, resulting in several equally likely alternative models. At that point, it was essential to agree on an objective model selection process, and the authors proposed a selection process involving (i) an analysis of model residuals, (ii) an analysis of model parameters (with the aim of having “reasonable” values), and (iii) the computation of theoretical measures of model validity. Acknowledging the difficulty of linking model parameter values with field measurements (because of the scale issue), they insisted on aiming at parameter stability and added that parsimony was a

good means to obtain robust parameters. In conclusion, they underlined that different people perceive the validation process differently, and suggested that models be seen as simple theories about the behavior of the natural systems, to reduce the “drama and controversy often associated with the concept of validation”.

Gorokhovski and Nute [1996] also contributed to the debate: considering the “Popperian” validation of hydrogeological models impossible, they proposed to focus on improving the evaluation of modelling uncertainties using full models and surrogate models, in what they name a “two-level modelling approach”.

The vision of Doherty [2011] is also worth of mention in the frame of this debate: in an editorial of *Groundwater*, he discussed the relative merits of complex (“picture-perfect”) and simple (“abstract”) models, which should both have a role to play for the sake of extracting as much information as possible from historical data. He added that the abstract models are too-often discarded “just because the model does not ‘look like’ what we imagine reality to look like”, while “a model deserves criticism only when it fails to achieve the only thing that it has a right to claim—quantification of uncertainty and maximum reduction of uncertainty through optimal processing of environmental data”.

3. The model validation debate on the other side of the hydrological fence (among the “surface” hydrologists)

We all agree that there is only one unique water cycle, and that the border between *hydrogeology* and *surface hydrology* is only cultural, mostly an inheritance of the too-narrow disciplinary teaching of the 20th century. There are however different traditions in surface hydrology (where models focus on reproducing the precipitation-streamflow relationship without mentioning groundwater levels most of the time), and hydrogeology (where the reproduction of piezometric levels is of primary importance and the surface processes are only considered under a “recharge” perspective). Let us now see how the issue of model validation has been dealt with by the “surface” hydrologists.

The loudest voice on this topic has unarguably been that of Vit Klemeš, former president of the International Association of Hydrological Sciences. He entered the debating arena with a paper published few years before the article by de Marsily et al. [1992]. Klemeš [1986] defended the generalization of what he called *split sample tests*¹ (SSTs), and proposed a progressive four-level calibration–validation testing scheme to assess hydrological models. Klemeš’s SST focuses on model transposability in time and space, with increasing difficulties presented to the model: (i) the elementary SST is based on calibrating and validating the model on two independent periods, (ii) the proxy-basin SST is based on transferring parameters between neighboring catchments, (iii) the differential SST is based on calibrating and validating the model on two independent and contrasting (dry/wet or cold/warm) periods, and (iv) the proxy-basin differential SST is based on transferring parameters between neighboring catchments on contrasting periods. Klemeš’s hope was that a wider adoption of SST practices could lead to reducing “the most glaring abuses of simulation models” and in promoting realistic assessments among modelers by avoiding “exaggerated claims regarding model capabilities.”² All this is quite similar to de Marsily’s objective: “increasing confidence”.

A few years after the paper by de Marsily et al. [1992], Refsgaard and Knudsen [1996] published a

¹Note that Klemeš never claimed to have invented the concept [see, e.g., Larson, 1931, Mosteller and Tukey, 1988]: he wrote that the SST “contains no new and original ideas; it is merely an attempt to present an organized methodology based on standard techniques, a methodology that can be viewed as a generalization of the routine split sample test”. But hydrologists still refer very often to his article, which is by far the most cited of his papers (over 750 citations as of December 2021), and SST during the last decade has seen a resurgence of interest [see, e.g., Coron et al., 2012, Seifert et al., 2012, Teutschbein and Seibert, 2013, Thirel et al., 2015, Dakhlaoui et al., 2019, Nicolle et al., 2021].

²Many years after publishing his famous paper, Klemeš (personal communication) wrote to us that he had in fact always been skeptical about the capacity of hydrologists to validate rigorously their model. He wrote that he knew in advance that the tests he had suggested would be “avoided under whatever excuses available because modelers, especially those who want to ‘market’ their products, know only too well that they would not pass it.” He concluded: “I had no illusions in this regard when I wrote my paper, but the logic of modelling led me to develop the ‘testing principle’ to its, let’s say, ‘theoretical limit.’”

paper entitled “Operational validation and intercomparison of different types of hydrological models.” They applied Klemeš’s four-level SST scheme to three models of increasing complexity, as their aim was to study the comparative robustness of different models. In this paper, they provide their own definition of “model validation”, which appears to be a nice synthesis of the opinions of de Marsily et al. [1992] and Konikow and Bredehoeft [1992a]: “Model validation is here defined as the process of demonstrating that a given site-specific model is capable of making accurate predictions for periods outside a calibration period. A model is said to be validated if its accuracy and predictive capability in the validation period have been proven to lie within acceptable limits or errors. It is important to notice that the term model validation refers to a site-specific validation of a model. This must not be confused with a more general validation of a generalized modelling system which, in principle, will never be possible” (p. 2190). The same group of authors developed their vision on the subject in subsequent papers [Refsgaard and Henriksen, 2004, Henriksen et al., 2003].

Over the past three decades, Professor Keith Beven has actively discussed the model validation issue. He, however, advocates and promotes a *rejectionist* approach, where “the question is not really validation but rather on what basis should a model run survive invalidation” [Beven, personal communication]. In a recent synthesis [Beven, 2019a], he defends the idea that “a simulation model should be shown to be fit-for-purpose, corroborated against some kind of observation or judgment, even if there are few rules about precisely what constitutes ‘fit’ and ‘purpose’, such that its use can be justified.” He proposes for model evaluation an approach called “limits of acceptability”, considering that there will be “a gradation of acceptability from the ‘best’ models that can be found, to those that are clearly not acceptable as simulators of the system of interest: in this context, the equifinality concept is intrinsically linked to model calibration and validation. The equifinality thesis suggests that there will be no single model representation of an environmental system, but rather an evolving ensemble of models that are considered acceptable in the sense of being useful in prediction as new information becomes available” [Beven, 2019b].

Let us mention here also our own past contribu-

tion to this debate [Andréassian et al., 2009]: While avoiding the terminological debate on the possible or impossible validation of hydrological models, we did argue that it was important to test models as exhaustively and vigorously as possible, with truly demanding tests that we proposed to call *crash tests*: Since the car industry can learn by destroying on purpose an exemplary of their production, we hydrologists should not be ashamed of taking our models to their limits and even a little beyond. We also underlined that the validation of a given model structure would require that tests be conducted on large sets of catchments, as large and varied as possible [see also on this topic Andréassian et al., 2006, Gupta et al., 2014]. A few years later, Biondi et al. [2012] proposed two “code of practices”, one for the validation of the performances of hydrological models, and another for what they call the “scientific validation” of the model. They insist on discussing model limitations “with the same detail that is dedicated to model strengths”, taking the example (Table 1) of the well-known SWOT analysis [on this issue of valuing the evaluation of model failures, see also our discussion in Andréassian et al., 2010].

4. Other relevant contributions from the fields of science history, ecology, and statistics

Science historian Naomi Oreskes made several relevant contributions to the debate, with some explicit references to the dialogue between de Marsily et al. [1992] and Konikow and Bredehoeft [1992a]. In an initial paper, Oreskes et al. [1994] argued that models can only be evaluated in relative terms (i.e., a model should not be declared “good” but only “better” than an alternative one). They underlined that “the term validation does not necessarily denote an establishment of truth. Rather, it denotes the establishment of legitimacy typically given in terms of contracts, arguments and methods. A valid contract is one that has not been nullified by action or inaction. A valid argument is one that does not contain obvious errors of logic. By analogy, a model that does not contain known or detectable flaws and is internally consistent can be said to be valid.” Oreskes et al. [1994] explicitly referred to the position of de Marsily et al., which they commended as *honest* (but not easily marketable...), considering that it fell under the

Table 1. Schematic representation of a SWOT analysis for models [modified from Biondi et al., 2012]

		Factors related to the model's predictive power	
		Strengths	Weaknesses
Factors related to the model's explanatory power	Opportunities	Highlight model strengths and related opportunities	Highlight model weaknesses and how they can be mitigated
	Risks	Highlight how model strengths allow avoiding risks	Highlight which risks are caused by model weaknesses

van Fraassen school of thought, i.e., *constructive empiricism*, where the goal of a scientific theory cannot be truth (unobtainable) but rather what van Fraassen names *empirical adequacy*.

In a second paper, Oreskes [1998] returned to the topic of validation in order to address issues related to models used to evaluate/support public policies: There, the semantic debate becomes overwhelming and Oreskes argued that “rather than talking about strategies for validation, we should be talking about means of evaluation”. A very interesting point in Oreske's [1998] paper is a remark on the surprising reluctance of most scientists toward evaluation tests: “Most scientists are aware of the limitations of their models, yet this private understanding contrasts the public use of affirmative language to describe model results.”

In a third paper, Oreskes and Belitz [2001] first expressed semantic regrets—“the term ‘validation’ is an unfortunate one”—then underlined that the main problem lies with the extrapolation capacity of models: “Models may match available observations, yet still be conceptually flawed. Such models may work in the short run, but later fail. [...] Rather than think of models as something to accept or reject [...] it may be more useful to think of models as tools to be modified in response to knowledge gained through continued observation of the natural systems being represented.”

For the ecological sciences, Caswell [1976] discussed the model validation issue and argued that validation should be looked at differently depending on the purpose of the model: He considered it essential to distinguish between predictive models and theoretical models (i.e., models aimed at providing insight into how the system operates). Caswell deemed that theoretical models should be examined according to the Popperian sequence of “conjectures and refutations”, and proposed reserving the term

“validation” for predictive models *only* (and to use the *Popperian* term of corroboration for theoretical models). He explained that the same model can be judged on both grounds, and can eventually be simultaneously declared predictively validated and theoretically refuted.

Two decades later, Power [1993] suggested a two-step approach to validation that would first check that candidate models are able to reproduce the statistical properties of the observations, in order to eliminate models with poor statistical properties. Only in a second phase would the models predictive properties be evaluated. Rykiel [1996] published an exhaustive review of model testing and validation practices in the field of ecological modeling, and his review shows that ecologists do not agree on the semantics or on the practices: In this way, they do not differ from the hydrogeologists! From an ecological research perspective, Rykiel [1996] considered that “the validation problem reflects ambiguity about how to certify the operational capability of a model versus how to test its theoretical content. The crux of the matter is deciding (1) if the model is acceptable for its intended use, i.e., whether the model mimics the real world well enough for its stated purpose, and, (2) how much confidence to place in inferences about the real system that are based on model results. The former is validation, the latter is scientific hypothesis testing. [...] Models can indeed be validated as acceptable for pragmatic purposes, whereas theoretical validity is always provisional.” In conclusion, the author insisted that “validation is not a procedure for testing scientific theory or for certifying the ‘truth’ of current scientific understanding, nor is it a required activity of every modelling project. Validation means that a model is acceptable for its intended use because it meets specified performance requirements.”

More recently, the statistician Shmueli [2010] pub-

lished a synthesis paper entitled “To explain or to predict?” where he discussed in much detail the distinction between explanatory and predictive models. This distinction seems to be central in the model validation debate; indeed, an explanatory model is to be validated qualitatively (and not necessarily quantitatively), while a predictive model is to be validated quantitatively (and could possibly be a “black-box” model, without any explicit explanatory capacity): “Predictive models are advantageous in terms of negative empiricism: a model either predicts accurately or it does not, and this can be observed. In contrast, explanatory models can never be confirmed and are harder to contradict.” Shmueli [2010] argued that misunderstandings arise from the frequent conflation between explanatory power and predictive power in science: “While explanatory power provides information about the strength of an underlying causal relationship, it does not imply its predictive power.” To conclude, the author suggested considering explanatory and predictive abilities as two dimensions: “explanatory power and predictive accuracy are different qualities and a model will possess some level of each.”

5. Discussion

5.1. *Validation from a model uncertainty perspective*

Over the past 30 years, uncertainty assessments have progressively become an inseparable part of modeling practice. The estimation of predictive uncertainty is seen as a kind of “quality insurance” [Refsgaard et al., 2005] and is as such considered good practice for any environmental modeling activity [Refsgaard et al., 2007]. In groundwater modeling, the uncertainty topic has obviously been discussed for years [de Marsily, 1978, Delhomme, 1979] but no general agreement has yet been reached on how to adequately quantify it; see, for example Barnett et al. [2012] and Guillaume et al. [2016] for a review. Notwithstanding the present popularity of uncertainty assessment exercises, which are now becoming part of the common modeling evaluation practice, it is important to stress here that they can only be seen as a necessary but not sufficient means for model validation, because they only refer to the predictive dimension of models (cf. the aforementioned discussion of the 2010 Shmueli paper). And

one can find in the history of science models that were “right but for the wrong reason” [e.g., the Ptolemaic planetary model and its famous epicycles, Klemesš, 1986].

5.2. *Validation from a sensitivity analysis perspective*

Sensitivity analysis (SA) is as old as model construction, but the last three decades have seen a renewed interest in the use of SA techniques. Keeping a model slim is not enough to make it a good model, but it can definitely contribute to turn the model validation process more efficient. According to Saltelli et al. [2000], SA can help investigate “whether a model resembles the system or processes under study; the factors that most contribute to the output variability and that require additional research to strengthen the knowledge base; the model parameters (or parts of the model itself) that are insignificant, and that can be eliminated from the final model; if there is some region in the space of input factors for which the model variation is maximum; the optimal regions within the space of the factors for use in a subsequent calibration study; if and which (group of) factors interact with each other.”

5.3. *Validation from a data availability perspective*

Over the past 30 years, the type and amount of data available for model validation has evolved, and this has had an impact on the “feasibility” of validation exercises. On the positive side, distributed data from satellites are now available, sometimes at high frequency. New measurements have appeared, allowing evaluating models at a regional scale rather than at a point scale: one can mention here NASA’s Gravity Recovery and Climate Experiment (GRACE), which provides since 2002 a quantitative measurement of terrestrial water storage changes, allowing the estimation of groundwater storage changes [Tapley et al., 2004]. Other satellite products offer information on actual evaporation and snow extent, and while the quality of satellite precipitation estimates remains rather modest, it has improved too. Water quality and water temperature sensors are also increasingly available, so that in many regions of the world

the possibilities for quantitative validation of hydrogeological predictions have increased. Of course, there is another side to every coin... and one should also mention that in many areas of the world, the density of ground stations (measuring either streamflow, piezometric level or precipitation) has actually decreased...

5.4. *Validation or evaluation?*

Among the criticisms made to the 1992 model validation debate, one is full of good sense: since there is so much controversy around the word “validation”, let us choose another softer one and give it a precise definition. This is the point of view developed by Oreskes [1998]: “rather than talking about strategies for validation, we should be talking about means of evaluation. That is not to say that language alone will solve our problems or that the problems of model evaluation are primarily linguistic. The uncertainties inherent in large, complex models will not go away simply because we change the way we talk about them. But this is precisely the point: calling a model validated does not make it valid.” This is certainly right, but on the other side we must acknowledge that it is extremely complicated to fight language habits! For example, the French language is full of undesirable anglicisms, which the Académie Française is fighting against... with limited success. We have been able to introduce “ordinateur” to replace “computer”, but we keep using the English language “sport” instead of its old French equivalent “desport”. If we decide to wait for our colleagues to accept and adopt our naming conventions... we may need a lot of patience.

6. Conclusion

Thirty years after the publication of the article by de Marsily et al. [1992], our literature review has allowed us to shed some new light on the model validation debate. For de Marsily et al. [1992], model validation exercises were meant to *increase the confidence* that a hydrogeologist would have in his/her model. This notion of *confidence* was multifactorial, as a model was to hold both explanatory power (cf. the reference to Darcy’s law and the principle of mass conservation) and predictive power (cf. the reference to success obtained in tests on an independent period).

This distinction between the predictive and explanatory dimensions of validation [underlined among others by Caswell, 1976, Beven, 2001 and Shmueli, 2010] is essential: With regard to our model validation debate, it implies that model validation can have two dimensions (hence the possibility of misunderstandings for those who did not realize it in the first place). It also implies the possibility of searching for compromises between these two dimensions: A “strong” predictive model could be preferred to a “weak” explanatory model, and vice versa. Obviously, validation becomes a multi-objective endeavor, and as such, it will require hydrogeologists to look for compromises (which may remain a matter of debate among them).

To conclude this conclusion, we would like to propose our own definition of model validation by extending that of de Marsily; and while this new definition is unavoidably shifted towards the way surface hydrologists look at models, we do believe that it retains enough generality to be of common interest to the hydrological and hydrogeological sciences:

- (1) The validation of models is possible and necessary;
- (2) When judging the validity of a model, one needs to keep in mind that a model remains an abstraction and a simplification;
- (3) Judging the validity of a hydrological model requires one to consider the model’s objectives as well as its space and time scale;
- (4) Validity can be considered from the point of view of the model’s explanatory power (theoretical content) and/or from the point of view of its predictive power. The explanatory and predictive dimensions of model validation must be considered separately: A model can eventually be simultaneously declared predictively validated and theoretically refuted;
- (5) When validity cannot be assessed in an absolute way, the value of a model can be examined from a comparative perspective;
- (6) When judging a model’s predictive power, the quantitative predictions are at least to be judged based on measurements that have not been used for model calibration, and possibly on measurements requiring a higher extrapolation capacity;

- (7) An assessment of the model's predictive uncertainty can be helpful with the validation process.

Conflicts of interest

The author has no conflict of interest to declare.

Acknowledgements

The author acknowledges the review of two anonymous referees, which helped him improve significantly his manuscript.

References

- Andréassian, V., Hall, A., Chahinian, N., and Schaake, J. (2006). Introduction and synthesis: Why should hydrologists work on a large number of basin data sets? In *Large Sample Basin Experiments for Hydrological Model Parameterization: Results of the Model Parameter Experiment—MOPEX*, volume 307, pages 1–5. IAHS Publ.
- Andréassian, V., Perrin, C., Berthet, L., Le Moine, N., Lerat, J., Loumagne, C., Oudin, L., Mathevet, T., Ramos, M. H., and Valéry, A. (2009). Crash tests for a standardized evaluation of hydrological models. *Hydrol. Earth Syst. Sci.*, 13, 1757–1764.
- Andréassian, V., Perrin, C., Parent, E., and Bardossy, A. (2010). Editorial—the court of miracles of hydrology: can failure stories contribute to hydrological science? *Hydrol. Sci. J.*, 55(6), 849–856.
- Barnett, B., Townley, R., Post, V., Evans, R., Hunt, R. J., Peeters, L., Richardson, S., Werner, A. D., Knapp, A., and Boronkay, A. (2012). Australian groundwater modelling guidelines. Report no 82. National Water Commission, Canberra.
- Beven, K. (2001). On explanatory depth and predictive power. *Hydrol. Process.*, 15, 3069–3072.
- Beven, K. (2019a). Invalidation of models and fitness-for-purpose: A rejectionist approach. In Beisbart, C. and Saam, N. J., editors, *Computer Simulation Validation*, pages 145–171. Springer, Cham.
- Beven, K. (2019b). Validation and equifinality. In Beisbart, C. and Saam, N. J., editors, *Computer Simulation Validation*, pages 791–809. Springer, Cham.
- Biondi, D., Freni, G., Iacobellis, V., Mascaro, G., and Montanari, A. (2012). Validation of hydrological models: Conceptual basis, methodological approaches and a proposal for a code of practice. *Phys. Chem. Earth*, 42–44, 70–76.
- Carrera, J., Mousavi, S. F., Usunoff, E. J., Sánchez-Vila, X., and Galarza, G. (1993). A discussion on validation of hydrogeological models. *Reliab. Eng. Syst. Saf.*, 42, 201–216.
- Caswell, H. (1976). The validation problem. In Patten, B., editor, *Systems Analysis and Simulation in Ecology*, volume IV, pages 313–325. Academic Press, New York, NY.
- Coron, L., Andréassian, V., Perrin, C., Lerat, J., Vaze, J., Bourqui, M., and Hendrickx, F. (2012). Crash testing hydrological models in contrasted climate conditions: An experiment on 216 Australian catchments. *Water Resour. Res.*, 48, article no. W05552.
- Dakhlaoui, H., Ruelland, D., and Trambly, Y. (2019). A bootstrap-based differential split-sample test to assess the transferability of conceptual rainfall-runoff models under past and future climate variability. *J. Hydrol.*, 575, 470–486.
- de Marsily, G. (1978). *De l'identification des systèmes hydro-géologiques*. Doctorat d'état thesis, Université Pierre et Marie Curie, Paris.
- de Marsily, G., Combes, P., and Goblet, P. (1992). Comment on 'Ground-water models cannot be validated', by L.F. Konikow and J.D. Bredehoeft. *Adv. Water Resour.*, 15, 367–369.
- Delhomme, J. P. (1979). Spatial variability and uncertainty in groundwater flow parameters: a geostatistical approach. *Water Resour. Res.*, 15, 269–280.
- Doherty, J. (2011). Modeling: Picture perfect or abstract art? *Groundwater*, 49(4), 455.
- Gorokhovski, V. and Nute, D. (1996). Validation of hydrogeological models is impossible: what's next? In *Calibration and Reliability in Groundwater Modelling*, volume 237, pages 417–424. IAHS Red Book.
- Guillaume, J. H. A., Hunt, R. J., Comunian, A., Blakers, R. S., and Fu, B. (2016). Methods for exploring uncertainty in groundwater management predictions. In Jakeman, A. J., Barreteau, O., Hunt, R. J., Rinaudo, J.-D., and Ross, A., editors, *Integrated Groundwater Management: Concepts, Approaches and Challenges*, pages 711–737. Springer International Publishing, Cham.
- Gupta, H. V., Perrin, C., Kumar, R., Blöschl, G., Clark, M., Montanari, A., and Andréassian, V. (2014).

- Large-sample hydrology: a need to balance depth with breadth. *Hydrol. Earth Syst. Sci.*, 18, 463–477.
- Henriksen, H. J., Troldborg, L., Nyegaard, P., Sonnenborg, T., Refsgaard, J. C., and Madsen, B. (2003). Methodology for construction, calibration and validation of a national hydrological model for Denmark. *J. Hydrol.*, 280, 52–71.
- Klemeš, V. (1986). Operational testing of hydrological simulation models. *Hydrol. Sci. J.*, 31, 13–24.
- Konikow, L. F. and Bredehoeft, J. D. (1992a). Groundwater models cannot be validated. *Adv. Water Resour.*, 15, 75–83.
- Konikow, L. F. and Bredehoeft, J. D. (1992b). Reply to comment. *Adv. Water Resour.*, 15, 371–372.
- Larson, S. C. (1931). The shrinkage of the coefficient of multiple correlation. *J. Educ. Psychol.*, 22, 45–55.
- Mosteller, F. and Tukey, J. W. (1988). Data analysis, including statistics. In *The Collected Works of John W. Tukey: Graphics 1965-1985*, volume 5. CRC Press, Boca Raton.
- Nicolle, P., Andréassian, V., Royer-Gaspard, P., Perrin, C., Thirel, G., Coron, L., and Santos, L. (2021). Technical note: RAT – a robustness assessment test for calibrated and uncalibrated hydrological models. *Hydrol. Earth Syst. Sci.*, 25, 5013–5027.
- Oreskes, N. (1998). Evaluation (not validation) of quantitative models. *Environ. Health Perspect.*, 106, 1453–1460.
- Oreskes, N. and Belitz, K. (2001). Philosophical issues in model assessment. In Anderson, M. G. and Bates, P. D., editors, *Model Validation: Perspectives in Hydrological Science*, pages 23–41. John Wiley and Sons, Ltd, London.
- Oreskes, N., Shrader-Frechette, K., and Belitz, K. (1994). Verification, validation, and confirmation of numerical models in the earth sciences. *Science*, 263, 641–646.
- Popper, K. (1959). *The Logic of Scientific Discovery*. Routledge, London.
- Power, M. (1993). The predictive validation of ecological and environmental models. *Ecol. Model.*, 68, 33–50.
- Refsgaard, J. C. and Henriksen, H. J. (2004). Modelling guidelines—terminology and guiding principles. *Adv. Water Resour.*, 27, 71–82.
- Refsgaard, J. C., Henriksen, H. J., Harrar, W. G., Scholten, H., and Kassahun, A. (2005). Quality assurance in model based water management—review of existing practice and outline of new approaches. *Environ. Model. Softw.*, 20, 1201–1215.
- Refsgaard, J. C. and Knudsen, J. (1996). Operational validation and intercomparison of different types of hydrological models. *Water Resour. Res.*, 32, 2189–2202.
- Refsgaard, J. C., van der Sluijs, J. P., Hojberg, A. L., and Vanrolleghem, P. A. (2007). Uncertainty in the environmental modelling process—A framework and guidance. *Environ. Model. Softw.*, 22, 1543–1556.
- Rykiel, E. J. (1996). Testing ecological models: the meaning of validation. *Ecol. Model.*, 90, 229–244.
- Saltelli, A., Chan, K., and Scott, E. M. (2000). *Sensitivity Analysis*. John Wiley, Hoboken, NJ.
- Seifert, D., Sonnenborg, T. O., Refsgaard, J. C., Højberg, A. L., and Troldborg, L. (2012). Assessment of hydrological model predictive ability given multiple conceptual geological models. *Water Resour. Res.*, 48, article no. W06503.
- Shmueli, G. (2010). To explain or to predict? *Stat. Sci.*, 25, 289–310.
- Tapley, B. D., Bettadpur, S., Watkins, M. M., and Reigber, C. (2004). The gravity recovery and climate experiment; mission overview and early results. *Geophys. Res. Lett.*, 31(9), article no. L09607.
- Teutschbein, C. and Seibert, J. (2013). Is bias correction of regional climate model (RCM) simulations possible for nonstationary conditions? *Hydrol. Earth Syst. Sci.*, 17, 5061–5077.
- Thirel, G., Andréassian, V., Perrin, C., Audouy, J.-N., Berthet, L., Edwards, P., Folton, N., Furusho, C., Kuentz, A., Lerat, J., Lindström, G., Martin, E., Mathevet, T., Merz, R., Parajka, J., Ruelland, D., and Vaze, J. (2015). Hydrology under change: an evaluation protocol to investigate how hydrological models deal with changing catchments. *Hydrol. Sci. J.*, 60, 1184–1199.