Crash tests for a standardized evaluation of hydrological models

Vazken Andréassian, Charles Perrin, Lionel Berthet, Nicolas Le Moine, Julien Lerat, Cécile Loumagne, Ludovic Oudin, Thibault Mathevet, Maria-Helena Ramos, Audrey Valéry

Hydrol. Earth Syst. Sci., 13, 1757–1764, 2009
www.hydrol-earth-syst-sci.net/13/1757/2009/

Hydrology and Earth System Sciences

HESS Opinions
"Crash tests for a standardized evaluation of hydrological models"

V. Andréassian[1], C. Perrin[1], L. Berthet[1,6], N. Le Moine[2], J. Lerat[3,*], C. Loumagne[1], L. Oudin[4], T. Mathevet[5], M.-H. Ramos[1], and A. Valéry[1]

[1]Cemagref, Hydrosystems and Bioprocesses Research Unit, Antony, France
[2]EDF-LNHE, Chatou, France
[3]Cemagref, G-EAU Research Unit, Montpellier, France
[4]Université Pierre et Marie Curie, UMR Sisyphe, Paris, France
[5]EDF-DTG, Grenoble, France
[6]AgroParisTech, Paris, France
[*]now at: CSIRO, Canberra, Australia

Received: 7 April 2009 – Published in Hydrol. Earth Syst. Sci. Discuss.: 4 May 2009
Revised: 31 August 2009 – Accepted: 2 September 2009 – Published: 1 October 2009

Abstract. As all hydrological models are intrinsically limited hypotheses on the behaviour of catchments, models – which attempt to represent real-world behaviour – will always remain imperfect. To make progress on the long road towards improved models, we need demanding tests, i.e. true *crash tests*. Efficient testing requires large and varied data sets to develop and assess hydrological models, to ensure their generality, to diagnose their failures, and ultimately, help improving them.

1 Introduction

Since this opinion paper deals with hydrological models, let us first define what we call a *model*: we restrict the discussion to *model structures*, which are also named *model codes* in the terminology of Refsgaard and Henriksen (2004). In that case, a model is defined by a set of equations allowing streamflow simulation based on input data and a parameter set (which varies from one catchment to another). It differs from a site-specific model, built for a particular area.

Correspondence to: V. Andréassian
(vazken.andreassian@cemagref.fr)

1.1 Hydrological models and the quest for an impossible validation

When developing a model, hydrologists seek a better understanding of physical processes and/or a gain in their ability to predict flow or other hydrological variables. But whatever its purpose, a model needs to be validated at some point. How the term *validation* should be defined and how *validity* should be measured, remains a matter of debate, which has been well summarized by Refsgaard and Henriksen (2004). The philosopher Popper (1959) considered that a model could only be *corroborated* or *refuted* (falsified). Klemeš (1986) proposed to speak about the *operational adequacy* of a model, rather than about its validity. Konikow and Bredehoeft (1992) argued that the word validation should not be used because it gives a false impression of model capability, while de Marsily et al. (1992) rejected the semantic debate, considering that hydrologists are never "striving for certainty and perfection", but only to do their "level best". Oreskes et al. (1994) underlined that models can only be evaluated in relative terms.

Whatever terminology we adopt, we need a method to evaluate models. As a starting point for this discussion, we propose to follow Klemeš (1986) in considering that a few necessary conditions to warrant model adequacy are:

– *model transposability in time* (i.e. whether the model can yield similar levels of errors, under both similar or very different climate conditions);

Published by Copernicus Publications on behalf of the European Geosciences Union.

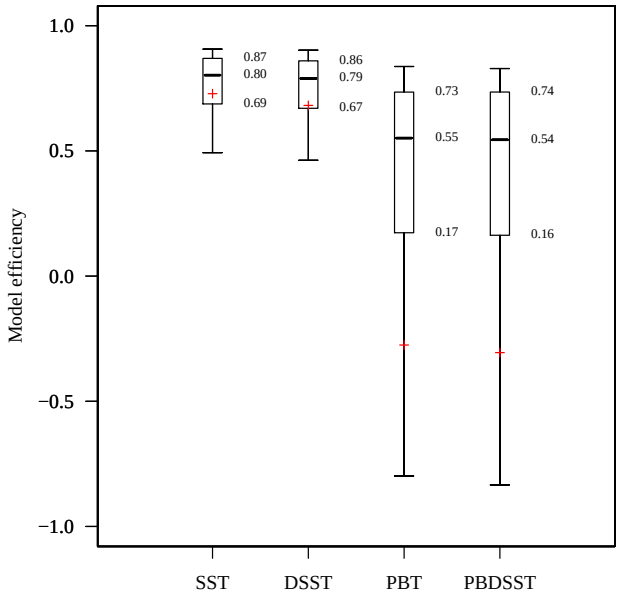

Fig. 1. Boxplot of the GR4J model performance over a set of 600 catchments in France (model efficiency: Nash-Sutcliffe criterion calculated on root square flows during validation period). SST: split sample test; DSST: differential split sample test; PBSST: proxy-basin split sample test; PBDSST: proxy-basin differential split sample test. Percentiles shown on the boxplot are: 0.1, 0.25, 0.5, 0.75, 0.9; crosses indicate the mean).

– *model transposability in space* (i.e. whether the model can yield similar levels of errors on different catchments, again under similar or very different climate conditions).

Klemeš (1986) proposed a four-level testing scheme aimed at assessing the general transposability of a model, thus extending the split-sample test (calibration/validation) that was in common use at that time. Although Klemeš wrote that the power of this four-level testing scheme was "rather modest, and [that] even a fully successful result [could] be seen only as a necessary, rather than a sufficient, condition for model adequacy vis-à-vis the specific modelling objective", it is patently obvious that this *modest* testing scheme has been left on the shelf (Refsgaard and Henriksen, 2004). More than 20 years later, only the first level (i.e. the split-sample test) is in standard use in evaluating hydrological models. The three remaining tests (the proxy-basin test, the differential split sample test, and the proxy-basin differential split sample test) are rarely applied.

A few exceptions should be mentioned: Refsgaard and Knudsen (1996), Donelly-Makowecki and Moore (1999) and Xu (1999) used the full four-level test; Seibert (2003) used the differential split-sample test. Recently, Le Moine (2008) applied the full test to the GR4J model on 600 French catchments. The results, summarized in Fig. 1, may partly explain the disregard of the modelling community: the drop in performance – when going from the split sample test to the other tests – is drastic.

The full Klemeš test can indeed be so demanding (i.e. so disappointing for enthusiastic model developers) that it has had a repulsive effect. Does it mean that it is useless, or that modellers did not fully consider what could be learnt in applying this full test? We favour the second option.

1.2 Model testing on large catchment sets: necessity or bulimia?

Oreskes (1998) noted with surprise that "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". Indeed, it sometimes seems as difficult for a hydrologist to publically admit the limitations of his creation as it is for an alcoholic to acknowledge his addiction. We consider that one way to overcome this is to develop and evaluate hydrological models on large and diversified catchment sets, and to always present the results of model-related discussions with distributions of model performance, obtained on a significant number (a few hundred or more) of catchments (see examples in Perrin et al., 2001; Le Moine et al., 2007; Oudin et al., 2008). By doing so, it will be possible to check that the proposed models have a general capacity to represent hydrological behaviour, and thus that the application spectrum is not limited to a few catchments and to stationary space-time conditions (hopefully there will always be a "willing" catchment to give acceptable results, so that nobody will lose face). There is nothing very original about this proposal: Roche (1971) and Linsley (1982) had already raised this point of view long ago and spread the idea that large sets of catchments provide a useful and informative way to test hydrological models. More recently, we have defended a similar point of view in Andréassian et al. (2007).

However, there may be misunderstandings on the objectives followed by using large data sets to develop and evaluate models. Some modellers may consider this approach *bulimic* modelling. Others consider that this would mean searching for a *universal* model. This is obviously not the case, as it would be naïve to think that at the present stage of hydrological modelling, a single model could work well in all places and conditions. But we are convinced that large catchment sets are the only possible way to learn from the variety of catchments, simply because they make it easier to falsify (refute) the models we wish to test (Popper, 1959). A few modellers seem to share this point of view: they also have published model tests based on large catchment sets, particularly in the perspective of modelling ungauged catchments (see among others Nathan and McMahon, 1990; Vandewiele et al., 1992; Merz and Blöschl, 2004; McIntyre et al., 2005; Kay et al., 2006; Young, 2006; Boughton and Chiew, 2007).

1.3 Scope

In this paper, we suggest that large and varied data sets are needed to develop and test hydrological models, to ensure their generality, to diagnose their failures, and to improve them. After reviewing the main arguments of those supporting and opposing the use of large catchment sets, we discuss our reasons for advocating why a model should be tested in a way comparable to the crash test used in the automobile industry.

2 Catchment monographs or studies on large catchment sets?

2.1 Arguments in favour of catchment monographs

At the present time, work on a single basin – or a very limited number of basins – remains the rule for most of the hydrological modelling studies reported in the literature. There are several reasons for this:

- First of all, many hydrologists look at hydrological modelling in a bottom-up, mechanistic manner. It is therefore natural to think that a single case-study could be enough to discover and dissect the main small-scale physical processes controlling the movement of water in a catchment.
- Second, in practice, it remains difficult to apply models whose parameterization is data-demanding or time-consuming to large catchment sets.
- Third, some hydrologists who defend the downward modelling philosophy do favour model structures customized on a catchment-by-catchment basis: see for example the "flexible" model philosophy of Fenicia et al. (2008a).
- Fourth, it is a widespread belief among hydrologists that the structure of a catchment model is necessarily climate- or region-specific, as a consequence of the prescriptions of the conceptual approach, which advocates keeping only in a model those driving processes that the modeller deems active in a given catchment.
- Last, measurements may be viewed with suspicion. In these conditions, confidence in a model cannot come from a confrontation with measurements (which may be considered too uncertain), but should instead come from the physical realism of the equations embedded in the structure of a model. However, this has limitations, as discussed by Beven (2001) and Silberstein (2005).

2.2 Arguments in favour of the use of large data sets

Why should a model developed on a given catchment be directly applicable to another one? After all, the components of the model structure are likely to be over-specialized, i.e. to reflect the peculiarities of the catchment used during the model-development phase.

On this topic, one of the pioneers of hydrologic modelling, Ray Linsley (1982), argued that "because almost any model with sufficient free parameters can yield good results when applied to a short sample from a single catchment, effective testing requires that models be tried on many catchments of widely differing characteristics, and that each trial cover a period of many years" (p. 14–15).

Other modelling pioneers have been defending the same point of view. Moore and Mein (1975) stressed that different climatic zones should be covered in a model test set, and they insisted that "the catchments on which the original versions of the models were developed should not be included to ensure independence of the test" (p. 123). Klemeš (1986) stressed that the use of "more test basins, more extensive split-sample schemes, etc., would increase the credibility standing of a model, and [...] lead to meaningful generalizations" (p. 22). Bergström (1991) insisted that to improve our confidence in hydrological modelling, we need to apply models "under a span of different geographical, climatological and geological conditions" (p. 127).

As for model improvement, one can also cite Andersson (1992), who reminded us that "a certain change of model structure can improve the model performance for some basins whereas it is unchanged or deteriorated for other basins. Improvements can also occur only for certain periods. It is therefore important to test the new model for a large set of basins and for long time series before drawing conclusions of a general model improvement" (p. 330).

More recently, Mouelhi et al. (2006) discussed the use of large data sets to develop a downward hydrological model. In particular, they stressed the need to have test catchments that are as climatically diverse as possible as the only way to test the ability of a model to represent the non-linearities of catchment behaviour. They also underlined that a large test set gives an opportunity to look for the features shared by catchments where the model fails, to better understand the causes of these failures and propose general remedies rather than only ad hoc solutions that could well be valid on only a single catchment. In that sense, large datasets may be seen as good safeguards against the development of overly complex models or as a way to identify catchment emerging properties as defined by McDonnell et al. (2007).

Oreskes et al. (1994) argued that "models are most useful when they are used to challenge existing formulations, rather than validate or verify them." Beven (2007) added that "more may be learned from model rejection than acceptance; rejection of a hypothesis, when properly justified, is an important stage in model development and improvement." We believe that using large catchment test sets provides a perfect opportunity to analyse model failure in a general way.

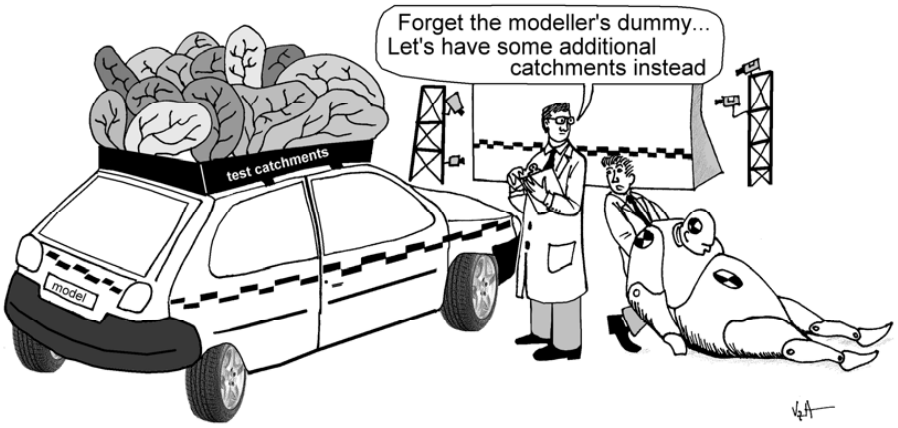

Fig. 2. Crash testing a rainfall-runoff model.

2.3 Any arguments in favour of a hybrid approach?

Although they may appear contradictory at first sight, there are definitely solutions to make single catchment analyses and large data set studies work together. Indeed, individual catchment analyses are an irreplaceable source of inspiration for hydrologists (both experimentalists and modellers) to develop new ideas and theories. But since monographs lack generality, it is necessary to systematically evaluate any such idea or theory on a larger catchment set. It may be worth remembering here the words of the famous French hydrologist Marcel Roche (1971), who insisted that a hydrologist "must above all be wary of one's own experience [...], how many hydrologists have actually believed they had a universal tool when they had only obtained a regional arrangement of elsewhere useless parameters."

When developed on a single catchment, models could be submitted to a process of sensitivity analysis and testing to identify those components or parameters that are not sensitive and can eventually be removed or fixed. This often leads to model structure simplification, a painful process which was meaningfully described by two authors: Bergström (1991) reported that "going from complex to simpler model structures requires an open mind, because it is frustrating to have to abandon seemingly elegant concepts and theories. It is normally much more stimulating, from an academic point of view, to show significant improvement of the model performance by increasing complexity" (p. 125). Martin (1996) stressed that "the prediction obtained with a complex model often points to a simpler model which could have been used in the first place. The challenge here is for the designer who has failed to keep his model simple to recognize the fact when confronted with it." Working on large data sets may be a way to avoid this process and to directly build more simple and general models.

3 Model developers, model users, and crash tests

Refsgaard and Henriksen (2004) detailed the different roles and expectations of model developers and model users. Most users are interested in a single or a limited number of catchments for which they wish to establish the best possible model. Therefore, a model developed on a large data set may appear irrelevant to their needs. Is this necessarily so? Let us try to clarify this situation by drawing an analogy with car testing.

3.1 Model developers should implement crash tests

Before launching a new car, automobile manufacturers systematically submit it to a crash test (see for example the NCAP website at www.euroncap.com). Crash tests have contributed to a true progress in transportation safety over the last few decades, and none of us would dream of driving a car – much less transporting family members – that had failed the minimum requirements of a crash test.

We consider that hydrological model developers have the same responsibility: they must perform a comprehensive crash test of their model to ensure that it is safe to use (Fig. 2). Of course, by submitting it to an extreme range of natural catchments and situations, they may lead it outside what they consider to be its range of application. But this is precisely what is needed: analysing model failures will make it possible to define its real limits of application (its "pedigree" as discussed by Beven, 2007), a necessary information for all potential users. It will also help propose new ways to improve model structure.

3.2 Model users should heed the results of crash tests when choosing a model

Let us continue with the car analogy: when choosing a new car, a responsible driver needs to identify "the right car".

This choice will depend on his objectives, taste, budget, sensitivity to advertisement, etc. (and perhaps also which models are on display at the dealership). Ideally, a responsible driver should never buy a car only for the emotions incited by the advertisement: one should compare performance, test the car, gather additional information from other users, etc. Being sure that one's car has been exhaustively tested is even more important because it is rare to have one car to commute to work in the city centre, one to travel to the countryside on the weekends and another one for winter vacation trips. Automobiles must serve a variety of purposes.

In hydrology, choosing the right model should require the same precautions, particularly in the case of operational hydrologists, who often have to use the same model structure for many catchments and different applications. Users should require tests that have gone beyond the usual range of application: Renault does not perform crash tests adapted to French roads only; Volvo does not limit its tests to snowy Swedish roads, nor Fiat to narrow winding Italian roads! A hydrologist using a model should know the limits of the model structure, based on the implementation of a complete crash test. A site-specific model, developed on a single site, may be very successful, but the question is: will it remain so in the long run? As Koutsoyiannis et al. (2009) put it, "there is no reason that the [natural] system properties remain unchanged over time." An end-user may prefer to trust a model that obtains slightly lower performance on the time series at hand, but that has been more exhaustively validated.

3.3 Large data sets and data quality

In the sections above, we have been advocating the use of large data sets. However, the issue of data quality often runs into objections to this approach: with a large data set, it is difficult (or even impossible) to manually control the quality of data within reasonable time. Only rather simple automatic data screening algorithms are usable, and they can identify only the most obvious erroneous values. Unavoidably, a few inconsistencies will remain in the time series, and some unknown upstream influences may also exist. This is viewed by some modellers as a good reason to avoid working on large data sets.

However, this argument does not hold for three reasons:

– First, when working on a large set of catchments, data originate from regional or national hydrological and meteorological databases that have their own data quality check procedures, which should guarantee an acceptable (although obviously not perfect) level of data quality. In that sense, it is likely that a majority of data can be considered as good quality, and a minority as suspect, giving to the latter a limited impact on the results. The problem of data quality is probably more crucial when working on a single case study, as model development may be impacted by erroneous data to a larger extent in this case.

– Second, because model evaluation is only meaningful in a comparative framework (a model can only be ranked good in comparison with alternative models). Therefore, Linsley (1982, p. 13) is right when he objects that "if the data are too poor for the use of a good simulation model they are also inadequate for any other model." Therefore, in intercomparison studies, data errors should not spoil the conclusions on the relative efficiency of several models (or model versions).

– Last, if we now look at real-time operational conditions, let us recognize that data quality checks are then necessarily limited. Though, models will have to be applied in these conditions. Therefore, part of a model crash test could consist in testing how a model responds to a deterioration of input data quality. We naturally do not advocate the intentional use of poor input data, but we consider that we need to document the impact of the progressive failure of a model when it encounters more and more input errors or missing values during its application or its calibration (see e.g. Oudin et al., 2006; Perrin et al., 2007).

4 Conclusions

4.1 Is there any truly objective model assessment?

All hydrological models are hypotheses on the behaviour of catchments. All are intrinsically limited in their capacity to represent real-world behaviours. We do need to improve them, while acknowledging that they will always remain far from perfection. And in this improvement process, the only option is falsification: for this, we need to be merciless towards our own models and to apply the most demanding crash test, using a large and varied catchment set. This will allow us to assess model robustness and generality, will help us to define its limits of applicability and to quantify the magnitude and the distribution of its errors. Using large catchment sets will also provide opportunities to think in a more general way about model failures: the identification of common features between catchments where the model fails can be an opportunity to understand the reasons for these failures and therefore suggest general solutions for model improvement, rather than merely ad hoc solutions that are only valid on a single catchment.

It does not mean that developing models on large data sets is a panacea that will allow models to be applied blindly. Crash tests do not guarantee that these models will never fail: good cars that have successfully passed all crash test can still be involved in accidents. This approach just aims at minimizing the risks of failure. Obviously, when using the crash-tested model for a specific application, site-specific validation remains necessary to accept or reject the model.

In this paper, we have shown an example obtained by applying the four-level Klemeš Crash Test (KCT) to GR4J,

a daily lumped rainfall-runoff model, over a few hundred catchments (Fig. 1). The drop of performance from the first two KCT test levels (simple and differential split sample test) to the two subsequent levels (simple and differential proxy-basin tests) illustrates what we could modestly describe as "an application where there is still considerable room for progress."

4.2 Towards other crash tests

We are convinced that a widespread use of the KCT (and eventually new crash tests) is required for the progress of hydrology as a science. As Kirchner (2006) puts it, hydrology can only move forward if we develop ways to test models more comprehensively and incisively, "which is different from testing how nicely a mathematical marionette can dance to a tune it has already heard" (p. 3–4). But we would like to stress that even the best of the tests will not identify a good model in absolute terms: it will simply define which model (or model category) is safer to use (Michel et al., 2006). As Savenije (2009) underlined it, searching for the *best* model is meaningless: we should be satisfied with developing *better* models.

The full Klemeš Crash Test is a step forward toward more powerful tests (and probably more discriminative tests, but this remains to be verified by applying it simultaneously to several models). Different crash tests can also be proposed, depending on model applications. For example, Ewen and Parkin (1996) proposed approaches to test model ability to predict climate or land use change impacts. Crash tests might also consist in calibrating models using one objective function and evaluating them in validation using different criteria. The discussion of Clarke (2008b) on model intercomparison suggests that model testing by several operators and following other good experimental practices could also make valuable crash tests. We certainly need to imagine new, more demanding testing schemes to respond to the current challenges of hydrological science. One of these challenges is surely to predict the possible consequences of climate change more reliably, and this raises the difficult question of which models have the best extrapolation capacity.

It would also be valuable to set up an international data set (including countries under various latitudes and having various measurement network configurations and levels of data availability) against which models could be tested by the scientific community. This could extend initiatives like the WMO workshops (1975, 1986, 1992), the Distributed Model Intercomparison Project (DMIP, where a large number of models were applied simultaneously on the same catchments with a strict verification protocol) (Smith et al., 2004) or the MOPEX project (Schaake et al., 2006). Such a data set should include many catchments without pre-screening, i.e. retaining outlier catchments (for example karstic catchments, or groundwater-dominated catchments), in order to study how models cope with these obstacles. This could perhaps open ways to link the efficiency of model structures to the level of data availability (Fenicia et al., 2008b) or to catchment classifications.

In our view, further progress in hydrological modelling will come in part from intercomparisons based on large data sets: after all, similar efforts, in other scientific domains, have been shown to be fruitful. This will obviously require that more rigorous and demanding testing schemes be routinely implemented, as suggested by Clarke (2008a).

Last, model tests on large data sets probably require designing new evaluation criteria that can extract consistent and interpretable information from the large amount of results they produce. Sets of assessment criteria were proposed in the past (see e.g. Dawson et al., 2007), but some of them may be difficult to interpret when applied on large datasets and new formulations may be necessary (see e.g. Mathevet et al., 2006). However, model assessments on large data sets can provide added value by not restricting them to the statistics classically used when testing models on individual catchments. Other types of information could also be analysed: spatial efficiency patterns on nested or neighbouring catchments; spatial coherence on extreme values computation; existing links between efficiency and catchment characteristics, etc. Methodological developments on these issues are needed.

Acknowledgements. We wish to thank John Schaake and an anonymous reviewer for their review comments on the text, as well as Robin Clarke, Andreas Efstratiadis, Neil McIntyre, Vit Klemeš, Nikos Mamassis, Kieran O'Connor and Jens Christian Refsgaard for their fruitful comments and suggestions during the open discussion on the manuscript, which were food for thoughts while finalizing this article.

Edited by: H. H. G. Savenije

References

Andersson, L.: Improvement of runoff models. What way to go?, Nord. Hydrol., 23, 315–332, 1992.

Andréassian, V., Loumagne, C., Mathevet, T., Michel, C., Oudin, L., and Perrin, C.: What is really undermining hydrologic science today? Hydrol. Process., 21, 2819–2822, 2007.

Bergström, S.: Principles and confidence in hydrological modelling, Nord. Hydrol., 22, 123–136, 1991.

Beven, K.: On explanatory depth and predictive power, Hydrol. Process., 15, 3069–3072, 2001.

Beven, K.: Towards integrated environmental models of everywhere: uncertainty, data and modelling as a learning process, Hydrol. Earth Syst. Sci., 11, 460–467, 2007, http://www.hydrol-earth-syst-sci.net/11/460/2007/.

Boughton, W. and Chiew, F.: Estimating runoff in ungauged catchments from rainfall, PET and the AWBM model, Environ. Modell. Softw., 22, 476–487, 2007.

Clarke, R. T.: A critique of present procedures used to compare performance of rainfall-runoff models, J. Hydrol., 352, 379–387, 2008a.

Clarke, R. T.: Issues of experimental design for comparing the performance of hydrologic models, Water Resour. Res., 44(1), W01409, doi:10.1029/2007WR005927, 2008b.

de Marsily, G., Combes, P., and Goblet, P.: Comment on "Groundwater models cannot be validated", by Konikow, L. F. and Bredehoeft, J. D., Adv. Water Resour., 15, 367–369, 1992.

Donelly-Makowecki, L. M. and Moore, R. D.: Hierarchical testing of three rainfall-runoff models in small forested catchments, J. Hydrol., 219, 136–152, 1999.

Ewen, J. and Parkin, G.: Validation of catchment models for predicting land-use and climate change impacts, 1, Method. J. Hydrol., 175, 583–594, 1996.

Fenicia, F., Savenije, H. H. G., Matgen, P., and Pfister, L.: Understanding catchment behavior through stepwise model concept improvement, Water Resour. Res., 44, W01402, doi:10.1029/2006WR005563, 2008a.

Fenicia, F., McDonnell, J. J., and Savenije, H. H. G.: Learning from model improvement: On the contribution of complementary data to process understanding, Water Resour. Res., 44(6), W06419, doi:10.1029/2007WR006386, 2008b.

Kay, A. L., Jones, D. A., Crooks, S. M., Calver, A., and Reynard, N. S.: A comparison of three approaches to spatial generalization of rainfall-runoff models, Hydrol. Process., 20, 3953–3973, 2006.

Kirchner, J.: Getting the right answers for the right reasons: linking measurements, analyses, and models to advance the science of hydrology, Water Resour. Res., 42, W03S04, doi:10.1029/2005WR004362, 2006.

Klemeš, V.: Operational testing of hydrologic simulation models, Hydrolog. Sci. J., 31, 13–24, 1986.

Konikow, L. F., and Bredehoeft, J. D.: Ground-water models cannot be validated, Adv. Water Resour. 15, 75–83, 1992.

Koutsoyiannis, D., Makropoulos, C., Langousis, A., Baki, S., Efstratiadis, A., Christofides, A., Karavokiros, G., and Mamassis, N.: *HESS Opinions*: "Climate, hydrology, energy, water: recognizing uncertainty and seeking sustainability", Hydrol. Earth Syst. Sci., 13, 247–257, 2009, http://www.hydrol-earth-syst-sci.net/13/247/2009/.

Le Moine, N., Andréassian, V., Perrin, C., and Michel, C.: How can rainfall-runoff models handle intercatchment groundwater flows? Theoretical study over 1040 French catchments, Water Resour. Res., 43, W06428, doi:10.1029/2006WR005608, 2007.

Le Moine, N.: Le bassin versant de surface vu par le souterrain: une voie d'amélioration des performances et du réalisme des modèles pluie-débit? Université Pierre et Marie Curie, Paris, PhD thesis, 322 pp., 2008.

Linsley, R. K.: Rainfall-runoff models-an overview, in: Proceedings of the international symposium on rainfall-runoff modelling, edited by: Singh, V. P., Water Resources Publications, Littleton, CO, 3–22, 1982.

Martin, P. H.: Physics of stamp-collecting? Thoughts on ecosystem model design, Sci. Total Environ., 183, 7–15, 1996.

Mathevet, T., Michel, C., Andréassian, V., and Perrin, C.: A bounded version of the Nash-Sutcliffe criterion for better model assessment on large sets of basins, IAHS Red Books Series, 307, 211–219, 2006.

McDonnell, J. J., Sivapalan, M., Vache, K., Dunn, S., Grant, G., Haggerty, R., Hinz, C., Hooper, R., Kirchner, J., Roderick, M. L., Selker, J., and Weiler, M.: Moving beyond heterogeneity and process complexity: a new vision for watershed hydrology, Water Resour. Res., 43(7), W07301, doi:10.1029/2006WR005467, 2007.

McIntyre, N., Lee, H., Wheater, H., Young, A., and Wagener, T.: Ensemble predictions of runoff in ungauged catchments, Water Resour. Res., 41, W12434, doi:10.1029/2005WR004289, 2005.

Merz, R. and Blöschl, G.: Regionalisation of catchment model parameters, J. Hydrol., 287, 95–123, 2004.

Michel, C., Perrin, C., Andréassian, V., Oudin, L., and Mathevet, T.: Has basin-scale modelling advanced beyond empiricism?, in: Large sample basin experiments for hydrological model parameterization, results of the Model Parameter Experiment – MOPEX, IAHS Publication 307, edited by: Andréassian, V., Hall, A., Chahinian, N., and Schaake, J., IAHS, Wallingford, 2006.

Moore, I. D. and Mein, R. G.: An evaluation of three rainfall-runoff models, Hydrological Symposium, Sydney, 122–126, 1975.

Mouelhi, S., Michel, C., Perrin, C., and Andréassian, V.: Linking stream flow to rainfall at the annual time step: the Manabe bucket model revisited, J. Hydrol., 328, 283–296, 2006.

Nathan, R. J. and McMahon, T. A.: The SFB model Part I – validation of fixed model parameters, Civ. Eng. Trans., Inst. Engs. Australia, CE32, 157–161, 1990.

Oreskes, N., Shrader-Frechette, K., and Belitz, K.: Verification, validation, and confirmation of numerical models in the earth sciences, Science, 263, 641–646, 1994.

Oreskes, N.: Evaluation (not validation) of quantitative models, Environ. Health Persp., 106, 1453–1460, 1998.

Oudin, L., Perrin, C., Mathevet, T., Andréassian, V., and Michel, C.: Impact of biased and randomly corrupted inputs on the efficiency and the parameters of watershed models, J. Hydrol., 320, 62–83, 2006.

Oudin, L., Andréassian, V., Lerat, J., and Michel, C.: Has land cover a significant impact on mean annual streamflow? An international assessment using 1508 catchments, J. Hydrol., 357, 303–316, 2008.

Perrin, C., Michel, C., and Andréassian, V.: Does a large number of parameters enhance model performance? Comparative assessment of common catchment model structures on 429 catchments, J. Hydrol., 242, 275–301, 2001.

Perrin, C., Oudin, L., Andréassian, V., Rojas-Serna, C., Michel, C., and Mathevet, T.: Impact of limited streamflow knowledge on the efficiency and the parameters of rainfall-runoff models, Hydrolog. Sci. J., 52, 131–151, 2007.

Popper, K.: The logic of scientific discovery, Routldege, London, 513 pp., 1959.

Refsgaard, J. C. and Knudsen, J.: Operational validation and intercomparison of different types of hydrological models, Water Resour. Res., 32, 2189–2202, 1996.

Refsgaard, J. C. and Henriksen, H. J.: Modelling guidelines – terminology and guiding principles, Adv. Water Res., 27, 71–82, 2004.

Roche, M.: Les divers types de modèles déterministes, La Houille Blanche, 111–129, 1971.

Savenije, H. H. G.: *HESS Opinions* "The art of hydrology", Hydrol. Earth Syst. Sci., 13, 157–161, 2009, http://www.hydrol-earth-syst-sci.net/13/157/2009/.

Schaake, J., Duan, Q., Andréassian, V., Franks, S., Hall, A., and Leavesley, G.: The model parameter estimation experiment – MOPEX, J. Hydrol., 320, 1–2, 2006.

Seibert, J.: Reliability of model predictions outside calibration conditions, Nord. Hydrol., 34, 477–492, 2003.

Silberstein, R. P.: Hydrological models are so good, do we still need data? Environ. Modell. Softw., 21, 1340–1352, 2005.

Smith, M. B., Seo, D. J., Koren, V. I., Reed, S. M., Zhang, Z., Duan, Q., Moreda, F., and Cong, S.: The distributed model intercomparison project (DMIP): motivation and experiment design, J. Hydrol., 298, 4–26, 2004.

Vandewiele, G. L., Xu, C. Y., and Win, N. L.: Methodology and comparative study of monthly models in Belgium, China and Burma, J. Hydrol., 134, 315–347, 1992.

WMO: Intercomparison of conceptual models used in operational hydrological forecasting, Operational Hydrology Report no. 7 and WMO no. 429, World Meteorological Organization, Geneva, 1975.

WMO: Intercomparison of models of snowmelt runoff, Operational Hydrology Report no. 23 and WMO no. 646, World Meteorological Organization, Geneva, 1986.

WMO: Simulated real-time intercomparison of hydrological models, Operational Hydrology Report no. 38 and WMO no. 779, World Meteorological Organization, Geneva, 1992.

Xu, C. Y.: Operational testing of a water balance model for predicting climate change impacts, Agr. Forest Meteorol., 98(9), 295–304, 1999.

Young, A. R.: Stream flow simulation within UK ungauged catchments using a daily rainfall-runoff model, J. Hydrol., 320, 155–172, 2006.