Physics and Chemistry of the Earth 42–44 (2012) 70–76

Contents lists available at SciVerse ScienceDirect

Physics and Chemistry of the Earth

journal homepage: www.elsevier.com/locate/pce

Validation of hydrological models: Conceptual basis, methodological approaches and a proposal for a code of practice

Daniela Biondi [a,*], Gabriele Freni [b], Vito Iacobellis [c], Giuseppe Mascaro [d], Alberto Montanari [e]

[a] Dipartimento Difesa del Suolo "V. Marone", Università della Calabria, Italy
[b] Facoltá di Ingegneria ed Architettura, Università di Enna "Kore", Italy
[c] Dipartimento di Ingegneria delle Acque e di Chimica, Politecnico di Bari, Italy
[d] Dipartimento di Ingegneria del Territorio, Università di Cagliari, Italy
[e] Dipartimento DICAM, Università di Bologna, Italy

ARTICLE INFO

Article history:
Available online 5 August 2011

Keywords:
Hydrological model
Validation
Performance indexes
Model diagnostic
Calibration

ABSTRACT

In this paper, we discuss validation of hydrological models, namely the process of evaluating performance of a simulation and/or prediction model. We briefly review the validation procedures that are frequently used in hydrology making a distinction between scientific validation and performance validation. Finally, we propose guidelines for carrying out model validation with the aim of providing agreed methodologies to efficiently assess model peculiarities and limitations, and to quantify simulation performance.

1. Introduction

The term validation is well known in hydrology and environmental modelling and is commonly used to indicate a procedure aimed at analysing performance of simulation and/or forecasting models. In the scientific context, the term validation has a broader meaning including any process that has the goal of verifying the ability of a procedure to accomplish a given scope. As an example, it can indicate the verification of a preliminary hypothesis or the security assessment of a computer network.

The need for agreed and standardised validation protocols in hydrological modelling has become progressively more urgent. In fact, in the last 30 years, hydrologic modelling has been greatly improved thanks to the increasing availability of computational resources, the advancement in the process understanding as well as the availability of spatially distributed data, mainly provided by remote sensors (Smith et al., 2004). The scientific literature continuously proposes new, sophisticated modelling solutions aimed at reproducing the hydrological cycle at multiple scales (e.g., field, watershed and even global scale) and for several goals, including research-oriented objectives, such as advancing the knowledge of physics of water movement, and more practical scopes, like water resources evaluation, flood protection and design of civil infrastructures. These numerical models adopt approaches and computational schemes that may be widely different. For this reason, validation protocols are required to (i) facilitate model inter-comparison, (ii) improve development of superior models, as well as their coupling and integration with data assimilations schemes, and (iii) help forecast users optimise their decision making.

Another important reason for developing standard validation criteria is the progressive mismatch between the complexity of modelling tools and the capacity of modellers and practitioners to rigorously assess the reliability of modelling application (Hug et al., 2009). This difficulty is exacerbated by the lack of sufficiently informative data. As an example, measured variables are often point values while simulated variables are frequently averaged in time and/or in space. Moreover, measured variables are affected by uncertainty due to the monitoring technology (e.g., Di Baldassarre and Montanari, 2009). The problem with data availability and uncertainty has been highlighted since the first approaches to environmental modelling validation and has often limited the possibility to adopt techniques successfully used in other scientific disciplines (Santhi et al., 2001).

A number of notable efforts have been recently devoted towards the development of shared modelling methodologies and verification standards in hydrology and close disciplines. The US National Weather Service (NWS) created a team of researchers, named Hydrologic Verification System Requirements Team, which had, among his tasks, the goal of establishing requirements for a

* Corresponding author.
E-mail addresses: dbiondi@dds.unical.it (D. Biondi), gabriele.freni@unikore.it (G. Freni), v.iacobellis@poliba.it (V. Iacobellis), gmascaro@unica.it (G. Mascaro), alberto.montanari@unibo.it (A. Montanari).

1474-7065/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.pce.2011.07.037

comprehensive national system to verify hydrologic forecast (see the website http://www.weather.gov/oh/rfcdev/projects/hvsrt_charter_05.htm). Moreover, Theme 3 ''Advance the learning from the application of existing models, towards uncertainty analyses and model diagnostics'' of the Predictions in Ungaged Basins (PUB) initiative also promotes harmonisation of model evaluation techniques. Similar standardisation processes have been started in other research fields in which mathematical modelling has become a common practice such as environmental quality assessment and water resources management (Belia et al., 2009; Muschalla et al., 2009). Particularly, Belia et al. (2009) proposed a road map for the definition of standard modelling approaches and model evaluation protocols starting from the creation of a common knowledge base to which water quality modellers can refer for their applications. Muschalla et al. (2009) defined an open protocol to apply to wastewater process modelling, highlighting the importance of model validation and uncertainty analysis especially in those cases where model complexity is not supported by sufficient data availability. In addition, a relevant effort was provided by the EU research project HarmoniQuA (Harmonizing Quality Assurance in model based catchments and river basin management) aimed to the development of modelling support tools that investigate the reliability of modelling responses at catchment scale (Refsgaard et al., 2005; Scholten et al., 2007).

Even if the need for common validation criteria is widely accepted in the hydrology and in the environmental science context, piecemeal contributions to this aim were presented in the specialised literature (Klemeš, 1986; Andréassian et al., 2009; Krause et al., 2005; Schaefli and Gupta, 2007; Gupta et al., 2009). In addition, validation protocols have been so far rarely applied in practical cases. For example, the NWS recently conducted two experiments, named DMIP (Distributed Model Intercomparison Project) and DMIP2, where several distributed hydrological models have been applied to common benchmark cases, consisting of well-instrumented basins located in diverse regions of US with contrasting climatic and landscape conditions (Smith et al., 2004). These initiatives were very successful, with the contribution of a notable number of different models. However, no standardised validation protocols have been utilised to analyse results. Verification tools would have been instead extremely important to intercompare the models, by quantifying their capability to reproduce specific hydrological processes, or to assess their robustness, i.e., whether if they may be applied in a broad range of conditions and climates, or only in specific regimes.

The considerations outlined above indicate that, despite recent efforts, an agreed and rigorous validation approach has not yet been proposed, due to the strong difficulty in identifying a unique and general protocol applicable to the large number of existing models and kinds of applications proposed in hydrology. As a matter of fact, in the majority of cases, validation is limited to analyse one or two events (e.g., intense floods), by simply comparing times series of simulated versus observed variables, and computing few lumped metrics that are able to capture only some attributes characterising model performance. The present paper aims at overcoming these limitations, by delineating a first proposal for a validation protocol in hydrology that explicitly distinguishes two phases: (i) the quantitative evaluation of model performances, and (ii) the qualitative evaluation of model structure and science foundation. The guidelines here proposed are intended to aid the work of researchers and practitioners/engineers while developing and applying numerical modelling tools in hydrology. After clarifying some definitions that are used in the paper (Section 2), we provide a summary of the state of the art of validation techniques for surface hydrological modelling including metrics and graphical tools, whose combined use is suggested in the proposed protocol (Section 3). The validation protocol with relative guidelines are presented in Section 4, while conclusions are drawn in Section 5.

2. Definitions and principles of evaluation theory

Prior to defining some basic concepts that have been used throughout this paper, we underline that high uncertainty exists in the terminology adopted in the present literature focused on the general process of evaluating the usefulness of a model for a given purpose. This implies that, in diverse contexts (e.g., environmental sciences, economics, meteorology, and medicine), the same word or expression is referred to indicate different activities. For example, the word *verification* is currently utilised in atmospheric science in the expression *forecast verification* to indicate the procedures aimed at measuring the ability of a meteorological model to predict the future weather (e.g., Jolliffe and Stephenson, 2003). Alternative expressions in this field are *forecast evaluation*, *validation* or *accuracy*. In the broader field of environmental modelling, some authors used the expression *model verification* to define a procedure for establishing that the model code correctly solves the set of mathematical equations adopted to simulate the real world (Matott et al., 2009). Since a discussion on the uncertainty in terminology and taxonomy would be too long and out of the scope of this paper, here we underline the existence of this problem and refer the reader to, e.g., Anderson and Bates (2001) and Matott et al. (2009) for more details.

The definitions adopted in this work are based on the consideration that the most frequent validation procedures, used in hydrologic and environmental modelling in general, proposed to split model evaluation in three complementary phases (Gupta et al., 2008): (a) quantitative evaluation of model performance; (b) qualitative evaluation of model performance; (c) qualitative evaluation of model structure and scientific basis. In the following, we alternatively use the expressions *model validation* or *performance validation* to indicate the concepts (a and b). This would be equivalent to the definition of model validation proposed by Matott et al. (2009). We instead adopt the expression *scientific validation* to refer to the activities described in point (c).

In what follows, the term *model* will be used to indicate a numerical tool for simulating input, state and output variables of a specific process. The user confidence into model results is strictly connected with the simulation reliability, which is assessed by comparing modelling output with data observed in the real world. The comparison is in turn made by means of user-defined criteria depending on the aim of the specific application. In the subsequent sections, the main performance and scientific validation procedures presented in literature are briefly reviewed.

3. Overview of techniques and methodological practice

3.1. Performance validation: graphical techniques and performance metrics

The typical approach adopted to evaluate model performance requires the comparison between simulated outputs on a set of observations that were not used for model calibration. This procedure coincides with the so-called split sample test in the classic hierarchical validation scheme proposed by Klemeš (1986), as well as with the first level of the theoretical scheme of Gupta et al. (2008).

Model performance can be addressed by means of qualitative and quantitative criteria. The former essentially rely on the graphical comparison between observed and simulated data, whereas the latter are based on numerical performance metrics. Both approaches are fundamental tools to be used in complementary fashion, since they are able to capture distinct aspects of model performance.

The choice of the validation criteria is guided by several factors. It depends on the nature of the simulated variables and the main model purpose. It is also affected by the fact that model simulations can be either deterministic or probabilistic. In the traditional deterministic approach, a unique best output is produced. More recently, a number of techniques, such as ensemble forecasting (Schaake et al., 2007), have been proposed to account for the different sources of uncertainty associated with input data, model structure and parameterization. Through these approaches, probabilistic hydrological forecasts are produced that attempt to explicitly quantify uncertainty.

Many criticisms have been addressed to traditional lumped metrics for their lack of diagnostic power or inability to capture differences between different model or parameter sets leading to ambiguous situations characterised by equifinality. As a result, more powerful evaluation tool like multi-objective methods that combines different (weighted) performance metrics into one overall objective function (e.g., Gupta et al., 1998) have been proposed. Another notable issue is that metric interpretation in not always straightforward. A common approach to address this problem consists of evaluating the metrics computed from model outputs with a benchmark value or a reference forecast that is generally an unskilled forecast (such as random chance and persistence).

In the review presented in this section, we mainly consider a hydrological model simulating the streamflow in a river basin, thus focusing on metrics and graphical techniques primarily applied to time series. In particular, we present a review of robust performance validation methods that can be utilised for both long-term multi-seasonal simulations and periods characterised by specific dominant processes (e.g., extremes or snow melting periods). In each of the following subsection, we first introduce the techniques useful for validating performance of deterministic simulations and, then, we discuss the methods utilised for assessing accuracy of probabilistic hydrological predictions. We highlight that, in this last case, most of the verification techniques are based on tools developed in applied meteorology, a discipline that has historically devoted consistent efforts on model validation and forecast verification.

3.1.1. Graphical techniques

Graphical techniques allow a subjective and qualitative validation. Despite the pletora of exiting goodness-of-fit metrics, visual inspection still represents a fundamental step in model validation as it allows the study of temporal dynamics of model performance and facilitate the identification of patterns in error occurrence. In most cases, they are based on a graphic comparison of simulated and measured time series (Fig. 1a). This kind of plots can be difficult to read, especially when the observation period is long. Scatterplots of simulated versus observed discharge are more easily interpretable and provide an objective reference given by the 1:1 line of perfect fit (Fig. 1b). Other common graphical representations are residual plots (Fig. 1c) and the comparison of streamflow duration curves as well as flood frequency distributions. Recently, the use of ensemble forecast techniques in hydrological models has led to the adoption of graphical methods developed and typically used in applied meteorology to evaluate probabilistic forecasts, like the reliability diagram (Fig. 1d) and the verification rank histogram (Fig. 1e) (Wilks, 2006; Mascaro et al., 2010).

3.1.2. Performance metrics

The performance metrics (or indexes) provide a quantitative and aggregate estimate of model reliability and are generally expressed as a function of the simulation errors. Some metrics have a statistical foundation, as the likelihood functions (Beven et al., 2001; Romanowicz and Beven, 2006), the AIC (Akaike Information Criterion), the BIC (Bayesian Information Criterion) and the KIC (Kashyap Information Criterion). The last three statistic criteria account for the mathematical complexity of the model by including the number of model parameters in the metric computation.

A wide number of metrics is derived from the general expression (Van der Molen and Pintér, 1993):

$$F = \left[\frac{1}{N}\sum_{t=1}^{N}|y_{s,t} - y_{o,t}|^{\tau}\right]^{1/b}, \quad \tau \geqslant 1,\ b \geqslant 1 \tag{1}$$

or from the analogous relation based on the relative deviations,

$$F = \left[\frac{1}{N}\sum_{t=1}^{N}\left|\frac{y_{s,t} - y_{o,t}}{y_{o,t}}\right|^{\tau}\right]^{1/b}, \quad y_{o,t} \neq 0,\ \tau \geqslant 1,\ b \geqslant 1 \tag{2}$$

where F is the performance metric, N is the number of observations, while $y_{s,t}$ and $y_{o,t}$ are the simulated and observed values at time t, respectively.

In particular, the metrics related to (2) are dimensionless and, thus, provide a more balanced evaluation of model performance over the entire study period. Metrics derived from both expressions (1) and (2) do not have an upper boundary while a null value indicates a perfect fit.

According to the values assumed by τ and b parameters, the two expressions provide different metrics, some of which are listed in Table 1. For higher τ, the metric is more sensitive to large differences between simulated and observed values. Several performance metrics adopt $\tau = 2$ and are therefore based on squared deviations.

To assess the quality of the model fit, other indexes, for example the Janus coefficient (Power, 1993), compare the model errors in the validation and the calibration period. Other goodness-of-fit metrics are based on regression operations between simulated and observed data (Legates and McCabe, 1999). In this category, we include (Table 1): the coefficient of determination R^2, the index of agreement D (Wilmott et al., 1985), and the coefficient of efficiency NSE introduced by Nash and Sutcliffe (1970), which is by far the most utilised index in hydrological applications. Differently from the metrics previously described, the perfect agreement is achieved when R^2, D and NSE are equal to unity.

Due to its large popularity, it is worthy to focus on Nash–Sutcliffe coefficient, whose main characteristics are as follows: (i) it measures the departure from unity of the ratio between the mean squared error and the variance of the observations; (ii) it varies between $-\infty$ and 1; (iii) a null value is obtained when the simulation is identically equal to the mean value of the observed series.

The diagnostic properties of the Nash–Sutcliffe efficiency have been recently investigated in detail by Gupta et al. (2009) through the decomposition into more meaningful components. These authors show that using NSE is equivalent to check model capability to reproduce the following statistics: (i) mean value and (ii) variance of the discharge time series, and (iii) coefficient of correlation between simulated and observed time series. The weight attributed to each of the above components depends on the magnitude of the observed data, but is mainly concentrated on correlation. Basing on this evidence, Gupta et al. (2009) proposed an innovative index, called KGE (Kling-Gupta Efficiency), expressed as an explicit function of the three statistics mentioned above.

Additional modifications of the NSE have been proposed in the literature, including those based on transformed variables, others using relative instead of absolute errors, and those adopting reference value different from the mean (Krause et al., 2005; Chiew and McMahon, 1994; Romanowicz et al., 1994; Freer et al., 1996). Other frequently used indexes include those based on the rank correlation criteria, such as the Spearman and Kendall coefficients. Re-

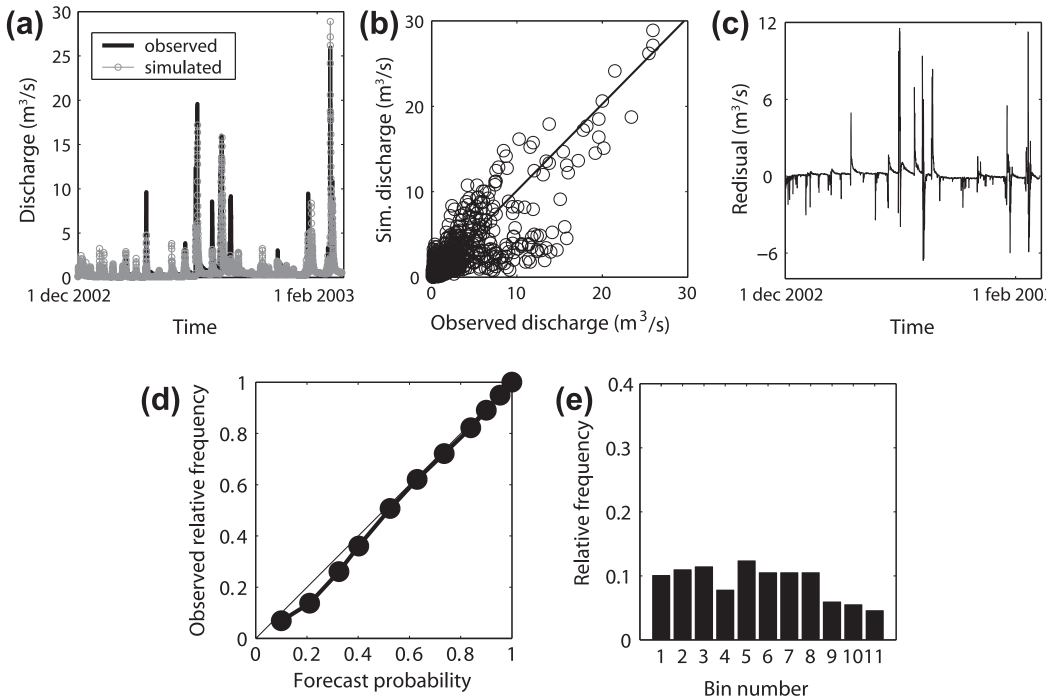

Fig. 1. Graphical methods used to evaluate model performance. Deterministic forecast: (a) observed and simulated time series; (b) scatter plot; (c) residual plot. Probabilistic forecast: (d) reliability diagram; (e) verification rank histogram.

Table 1
Numerical metrics used to evaluate model performance.

Performance metric	Expression
Mean Absolute Error (MAE)	$F_1 = \frac{1}{N}\sum_{t=1}^{N}\lvert y_{s,t} - y_{o,t}\rvert$
Mean Square Error (MSE)	$F_2 = \frac{1}{N}\sum_{t=1}^{N}\lvert y_{s,t} - y_{o,t}\rvert^2$
Root Mean Square Error (RMSE)	$F_3 = \left[\frac{1}{N}\sum_{t=1}^{N}\lvert y_{s,t} - y_{o,t}\rvert^2\right]^{1/2}$
Minimax objective function	$F_4 = \frac{1}{N}\max\lvert y_{s,t} - y_{o,t}\rvert$
Average Absolute Percentage Error (AAPE)	$F_5 = 100\frac{1}{N}\sum_{t=1}^{N}\left\lvert\frac{y_{s,t}-y_{o,t}}{y_{o,t}}\right\rvert$
Mean Square Relative Error (MSRE)	$F_6 = 100\frac{1}{N}\sum_{t=1}^{N}\left\lvert\frac{y_{s,t}-y_{o,t}}{y_{o,t}}\right\rvert^2$
Coefficient of determination (R^2)	$F_7 = \left\{\frac{\sum_{t=1}^{N}(y_{o,t}-\bar{y}_o)(y_{s,t}-\bar{y}_s)}{\left[\sum_{t=1}^{N}(y_{o,t}-\bar{y}_o)^2\right]^{0.5}\left[\sum_{t=1}^{N}(y_{s,t}-\bar{y}_s)^2\right]^{0.5}}\right\}^2$
Index of agreement (D)	$F_8 = 1 - \frac{\sum_{t=1}^{N}(y_{s,t}-y_{o,t})^2}{\sum_{t=1}^{N}\left(\lvert y_{s,t}-\bar{y}_o\rvert + \lvert y_{o,t}-\bar{y}_o\rvert\right)^2}$
Nash–Sutcliffe Efficiency coefficient (NSE)	$F_9 = 1 - \frac{\sum_{t=1}^{N}(y_{s,t}-y_{o,t})^2}{\sum_{t=1}^{N}(y_{o,t}-\bar{y}_o)^2}$

cently, new validation criteria based on fuzzy measures or on mechanism to account for expert knowledge and "soft data" have become popular, even if they introduce a larger degree of subjectivity in performance evaluation (Seibert and McDonnell, 2002; Beven, 2006). The review presented so far has been focused on metrics mainly applicable to deterministic simulations and is not intended to be exhaustive. The reader is referred to Jachner and van den Boogaart (2007) and Dawson et al. (2007) for a detailed survey.

When dealing with probabilistic forecasts, traditional goodness-of-fit metrics (like those mentioned for deterministic simulations) do not allow a complete and fair evaluation of the forecast performance. In his essay on the nature of goodness in weather forecasting, Murphy (1993), considering a distribution-based approach, distinguishes nine attributes that contribute to the fullest description of the multi-faceted nature of the probabilistic forecast quality: Bias, Association, Accuracy, Skill, Reliability, Resolution, Sharpness, Discrimination and Uncertainty. Each of these attributes carries a fundamental information about the forecast performance and, only recently, some techniques have been specifically designed to quantify their weight in the process of verifying hydrologic probabilistic forecasts. Contributions in this field include the work of Welles (2005), Welles et al. (2007), Laio and Tamea (2007), Engeland et al. (2010), Mascaro et al. (2010) and the technical report released by NWS downloadable at http://www.nws.noaa.gov/oh/rfcdev/docs/Final_Verification_Report.pdf.

Moreover, taking again inspiration to applied meteorology, both deterministic and probabilistic hydrological forecasts can be transformed into a categorical yes/no forecasts according to some critical value or probability threshold (e.g., probability that the streamflow accumulated in a given duration will exceed a certain threshold). The contingency table, which shows the frequency of yes and no forecasts and occurrences and the frequency of their combinations (hit, miss, false alarm, correct negative), is a common way to analyse what types of errors are being made. A large variety of categorical statistics can be computed from the elements of the contingency table to describe particular aspects of forecasts performance, including probability of detection, false alarm rate, critical success index, Gilbert skill score, Peirces skill score, Heidke skill score, among the others. A commonly adopted verification tool in case of probabilistic forecast for binary (yes/no) events is the Brier score, which takes the form of the more general Ranked Probability Score (RPS) when it is intended to be applicable to multi-category forecasts. A further generalisation of the RPS to an infinite number of classes led to the definition of the Continuous Ranked Probability Score, a metric particularly useful to verify reliability, resolution and uncertainty attributes of ensemble streamflow forecasts.

3.2. Scientific validation

The concept of scientific validation has been originated from the idea that verifying the model performance by simply comparing outputs and observations does not assure that the model is correct from a scientific point of view. In other words, this limitation does not allow us to assess if the model structure and parameterization are consistent with the physics of the simulated processes (Oreskes et al., 2003). It is well known that every model provides a simplified representation of reality, which depends on availability of observations, knowledge of phenomena, computational capability, and final purposes of the application. Given these limitations, the scientific validation aims at evaluating the consistency, and the coherence with real world, of the model thought as an ISO (input-state-output) system. In this framework, the quantification of the different sources of uncertainty (e.g., observations, process parameterization, model structure) is crucial (Todini, 2007). Synthesizing, the scientific validation has the goal of verifying that right outputs are produced for the right reason (Kirchner et al., 2006; Aumann, 2007).

The scientific validation may include and extend the performance validation and is specifically required in particular cases, including: (a) when the quality and quantity of the observations used for comparison with model outputs are not sufficient to allow an adequate performance validation; (b) when the model is utilised with the goal of advancing the knowledge of physical processes, rather than to make predictions; (c) when the hydrological model is not the "focus" of a given application, but just a "means" to characterise the initial conditions or quantify variables needed to study other physical, chemical, and/or biological processes.

So far, no agreed methodological approach has been proposed by the international scientific community to deal with scientific validation. However, beyond the general principles above stated, a large number of techniques are already developed in hydrology as well as in other disciplines, which can be considered as applications in the field of scientific validation.

One goal of scientific validation is the assessment of model hypotheses. This task often relies on the identification of the main processes that affect the real world and that model should carefully account for. With this aim Aumann (2007), in the field of ecology, suggests to conduct a system analysis aimed at detecting the processes, occurring at different scales, that can be considered as the dominant processes or emergent properties across different hierarchical levels of the model. In general, the observation of the same natural processes at different scales can lead to useful insights in process knowledge. In this framework techniques based on the upscaling/downscaling of state variables, model parameters, input variables and model conceptualizations (Bierkens et al., 2000) provide a recognised, model-oriented approach for coping with the scale transfer problem (e.g., Bloschl and Sivapalan, 1995). An upscaling/downscaling technique is also used for model diagnostic in meteorology (Hantel and Acs, 1998). Coming back to hydrology, strong enhancements in recognising the main process controlling water balance and rainfall–runoff processes may be achieved by jointly exploiting (i) lumped, semi-distributed and distributed models; (ii) parcel, hillslope and catchment models and observations; (iii) regional and at-site estimates of hydrological random variables.

Besides the assessment of model hypotheses, scientific validation aims at providing the proof of model adequacy to the representation of real world beyond (or together with) the result of validation tests. In fact, one model could be right for the wrong reasons, for example, by compensating error in model structure with errors in parameter values (Refsgaard and Henriksen, 2004).

This argument may lead to recognising the equifinality problem posed by Beven et al. (2001) and Beven (2006) but, rather, scientific validation points to the identification of model selection and parameter estimation pursuing the inequifinality concept proposed by Todini (2007). Casted in the Bayesian framework, the inequifinality concept is expressed by stating that model structure, parameter vector and predictions can be chosen as those more likely than others, i.e. with posterior densities characterised by more pronounced peaks and smaller predictive uncertainties.

In many research fields, a multi-criteria approach has been proposed where the behaviour of different state variables, internal to the model, is analysed and exploited in order to verify and diagnose the model. This kind of approach can take advantage of information and data obtained from remote sensing and/or field based observations of physical quantities related to vegetation states, air temperature, soil moisture, etc. (Castelli, 2008). On the other hand, this philosophy leads to increasing model complexity. In fact, multi-site validation is possible if simulations of spatial patterns are accounted for. Also, multi-variable checks are source of precious information if predictions of the behaviour of individual subsystems within a catchment are performed (Refsgaard and Henriksen, 2004). Thus, another important checkpoint of scientific validation lies in the assessment of the equilibrium between model purpose, model complexity and availability of data sources and information. In this perspective, scientific and performance validation may be merged. This happens, for example, in cases when the performance metrics combine various measures aimed at making diagnosis or providing information to correct the model at the appropriate level.

4. Proposal of a validation protocol: general guidelines

The scientific literature continually suggests new advances about model validation. However, most of the new proposals are dedicated to specific technical details, like, for example, the optimal combination of performance metrics (Reusser et al., 2009) and model diagnostic issues. Less consideration has been devoted to the ethic and philosophical principles that should guide the development of innovative validation techniques. The principal reason for this limited attention is probably the potential subjectivity of these guiding principles. We instead believe that the latter should become the main subject of the discussion. Model validation should be intended as a modeller self-training tool more then a way to objectively show the performance of the model.

The leading principle that we would like to emphasise is that the value of a simulation study should be associated not only with the quality of results, but also (and perhaps more importantly) with their scientific interest. We would like to re-elaborate the idea that a model performs well only if it returns satisfactory results. In fact, it is widely recognised that the good score returned by a performance metric only provides a limited view of the practical utility and scientific value of a model. It is also recognised that providing examples of poor model performance is very useful to highlight model weaknesses. These principles are valid both for scientific model development and for practical applications. Especially in the latter case, the knowledge of model limitations is even more important than the investigation of its best performance because they should affect design strategies and safety factors. Model validation in engineering practice is often perceived only as a quality assurance issue while it may have a broader pro-active impact on modelling choices guiding monitoring campaigns, specific investigations or simply a wiser choice of the modelling tools. This misunderstanding, along with the obvious costs of the procedure, took model validation to be neglected in the most part of practical applications.

We believe an important guideline for the validation process could be given by the so-called SWOT analysis (Hill and Westbrook, 1997), namely, a tool for strategic planning that can be used to evaluate Strengths, Weaknesses, Opportunities and Threats associated with a model and its application. A possible schematic of the SWOT analysis applied to hydrological validation is presented in Table 2. This approach allows assessing which opportunities can be gained and which risks can be avoided through the strengths of the model, as well as which risks can be caused by the model weaknesses and how they can be mitigated. The under-

Table 2
Proposal of a SWOT analysis applied to a hydrological model.

SWOT analysis		Internal factors	
		Strenghts	Weaknesses
External factors	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

lying philosophy is that model limitations should be discussed with the same detail that is dedicated to model strengths. When a particular model is chosen strengths are usually discussed in view of the scope of the analysis (therefore highlighting the opportunities). However, any model only provides an approximation of reality and therefore weaknesses are unavoidably present. We believe that these limitations should be discussed as well, along with the related risks. Actually, model weaknesses are rarely mentioned in scientific studies. The systematic use of the SWOT approach in hydrology would stimulate a more insightful scientific evaluation.

4.1. Guidelines for performance validation

The basic idea in performance validation is to provide several elements that can be used by researchers and practitioners/engineers to clarify different and complementary issues related to model performance. The guidelines are summarised by the following points.

1. Provide clear and unequivocal indications about model performance in real world applications.
2. Apply the validation procedure by using independent information with respect to what was used for model calibration.
3. Perform validation and discussion of data reliability, and possibly implement a combined validation of models and data.
4. Use graphical techniques and several numerical performance metrics to evaluate different aspects of model performance. Among the available graphical techniques, we suggest the use of scatter plots of observed versus simulated values for their immediate readability. The use of the logarithmic scale should be properly justified. The selected metrics should be justified.
5. When dealing with probabilistic simulations, use rigorous techniques that test several attributes of forecast quality.
6. When presenting results, do not focus only on a few cases (e.g., a single intense flood event), but consider a statistically significant number of cases including those where the model did not return satisfactory results. Indications about worst performance should be provided, discussing the possible reasons that are responsible for the obtained performance level.
7. If possible, extend the validation to model input and state variables.
8. If possible, validate the model over different temporal and spatial scales.
9. Evaluate the opportunity to apply jack-knife techniques to create confidence intervals (Shao and Tu, 1995; Castellarin et al., 2004; Brath et al., 2003).

The above list is meant to be the basis for a code of practice which is based on the principle of integrating different validation methods for comprehensively evaluating model strengths and limitations.

4.2. Guidelines for scientific validation

Guidelines for the scientific validation protocol, which are mainly useful for research and model development, can be summarised as follows:

1. Clearly identify the model purpose(s) and check if the adopted model addresses it (them).
2. List and discuss all the assumptions; describe the validation procedure and the relative hypotheses.
3. Analyse the reliability of theoretical fundaments; justify the degree of complexity and the computational burden.
4. Evaluate and discuss possible alternative modelling hypotheses.
5. Use all the possible knowledge (physical processes, observations) to support model development and application. Underline the coherence of the solution with the physical basis of the simulated processes.
6. Analyse the entire ISO system pointing out the uncertainty associated with input and output components.
7. Make data and numerical codes publicly available, make the scientific community able to reproduce results. If this is not be possible (data ownership and so on), provide a detailed description of the study to support repeatability.
8. Identify strengths and weaknesses of the model and highlight their interactions with risks and opportunities, as proposed by the SWOT analysis. Provide support to scientific review with a detailed discussion of the critical points.

As previously highlighted, the scientific validation is contiguous to performance validation and can be enhanced by points listed in the previous subsection, whenever they can be applicable in relation to data availability and model goals.

5. Conclusions

This paper intends to provide a contribution towards the identification of agreed principles for model validation. The basic idea is that validation should provide an exhaustive evaluation of both model scientific basis and performance. For this purpose, it is necessary to highlight not only the model strengths but also the weaknesses, according, for example, to the principles suggested by the SWOT analysis. A first suggestion for a model validation protocol is presented, by providing recommendations to structure the validation process and to produce a comprehensive and comprehensible validation. We believe the on-going development of new modelling tools and applications requires focusing on the criteria for a transparent presentations of models and results.

Acknowledgements

The authors thank three anonymous reviewers whose comments helped to improve the quality of the manuscript. The authors thank Prof. Pasquale Versace and Prof. Riccardo Rigon whose suggestions and encouragements have been very helpful.

References

Anderson, M.G., Bates, P.D., 2001. Model Validation: Perspectives in Hydrological Science. John Wiley & Sons, Inc..
Andréassian, V., Perrin, C., Berthet, L., Le Moine, N., Lerat, J., Loumagne, C., Oudin, L., Mathevet, T., Ramos, M.-H., Valéry, A., 2009. HESS opinions crash tests for a standardized evaluation of hydrological models. Hydrol. Earth Syst. Sci. 13, 1757–1764.
Aumann, C.A., 2007. A methodology for developing simulation models of complex systems. Ecol. Model. 202, 385–396. doi:10.1016/j.ecolmodel.2006.11.005.
Belia, E., Amerlinck, Y., Benedetti, L., Johnson, B., Sin, G., Vanrolleghem, P.A., Gernaey, K.V., Gillot, S., Neumann, M.B., Rieger, L., Shaw, A., Villez, K., 2009.

Wastewater treatment modelling: dealing with uncertainties. Water Sci. Technol. 60, 1929–1941.
Beven, K.J., 2006. A manifesto for the equifinality thesis. J. Hydrol. 320, 18–36.
Beven, K.J., Freer, J., 2001. Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology. J. Hydrol. 249, 11–29.
Bierkens, M.F.P., Finke, P.A., de Willigen, P., 2000. Upscaling and Downscaling Methods for Environmental Research. Kluwer Academic Publishers, Dordrecht, 190 pp.
Bloschl, G., Sivapalan, M., 1995. Scale issues in hydrological modelling: a review. Hydrol. Proc. 9, 251–290.
Brath, A., Castellarin, A., Montanari, A., 2003. Assessing the reliability of regional depth-duration-frequency equations for gaged and ungaged sites. Water Resour. Res. 39. doi:10.1029/2003WR002399.
Castellarin, A., Galeati, G., Brandimarte, L., Montanari, A., Brath, A., 2004. Regional flow-duration curves: reliability for ungauged basins. Adv. Water Resour. 27, 953–965.
Castelli, F., 2008. Sinergia fra modelli idrologici e osservazioni satellitari: la dinamica e il paradigma bayesiano. In: Proceedings of the 31th National Conference on Hydraulics and Hydraulic Works, Perugia, Italy, 2008.
Chiew, F.H., McMahon, T.A., 1994. Application of the daily rainfall–runoff model MODHYDROLOG to 28 Australian catchments. J. Hydrol. 153, 383–416.
Dawson, C.W., Abrahart, R.J., See, L.M., 2007. HydroTest: a web-based toolbox of evaluation metrics for the standardized assessment of hydrological forecasts. Environ. Model. Softw. 22, 1034–1052.
Di Baldassarre, G., Montanari, A., 2009. Uncertainty in river discharge observations: a quantitative analysis. Hydrol. Earth Syst. Sci. 13, 913–921.
Engeland, K., Renard, B., Steinsland, I., Kolberg, S., 2010. Evaluation of statistical models for forecast errors from the HBV model. J. Hydrol. 384, 142–155.
Freer, J., Beven, K.J., Ambroise, B., 1996. Bayesian estimation of uncertainty in runoff prediction and the value of data: an application of the GLUE approach. Water Resour. Res. 32, 2161–2173.
Gupta, H.V., Sorooshian, S., Yapo, P.O., 1998. Toward improved calibration of hydrological models: multiple and noncommensurable measures of information. Water Resour. Res. 34, 751–763.
Gupta, H.V., Wagener, T., Liu, Y., 2008. Reconciling theory with observations: elements of a diagnostic approach to model evaluation. Hydrol. Proc. 22, 3802–3813.
Gupta, H.V., Kling, H., Yilmaz, K.K., Martinez, F.G., 2009. Decomposition of the mean square error and NSE performance criteria: implications for improving hydrological modelling. J. Hydrol. 377, 80–91.
Hill, T., Westbrook, R., 1997. SWOT analysis: its time for a product recall. Long Range Plan. 30, 46–52. doi:10.1016/S0024-6301(96)00095-7.
Hantel, M., Acs, F., 1998. Physical aspects of the weather generator. J. Hydrol. 212–213, 393–411.
Hug, T., Benedetti, L., Hall, E.R., Johnson, B.R., Morgenroth, E.F., Nopens, I., Rieger, L., Shaw, A.R., Vanrolleghem, P.A., 2009. Mathematical models in teaching and training: mismatch between education and requirements for jobs. Water Sci. Technol. 59, 745–753.
Jachner, S., van den Boogaart, K.G., 2007. Statistical methods for the qualitative assessment of dynamic models with time delay (R Package qualV). J. Stat. Softw. 22.
Jolliffe, I.T., Stephenson, D.B., 2003. Forecast Verification: A Practitioner's Guide in Atmospheric Science. John Wiley and Sons, Chichester, ISBN 0-471-49759-2.
Kirchner, J.W., 2006. Getting the right answers for the right reasons: linking measurements, analyses, and models to advance the science of hydrology. Water Resour. Res. 42, W03S04.
Klemeš, V., 1986. Operational testing of hydrological simulation models. Hydrol. Sci. J. 31, 13–24.
Krause, P., Boyle, D., Bse, F., 2005. Comparison of different efficiency criteria for hydrological model assessment. Adv. Geosci. 5, 89–97.
Laio, F., Tamea, S., 2007. Verification tools for probabilistic forecasts of continuous hydrological variables. Hydrol. Earth Syst. Sci. 11, 1267–1277.
Legates, D.R., McCabe, G.J., 1999. Evaluating the use of goodness-of-fit measures in hydrologic and hydroclimatic model validation. Water Resour. Res. 35, 233–241.
Mascaro, G., Vivoni, E.R., Deidda, R., 2010. Implications of ensemble quantitative precipitation forecast errors on distributed streamflow forecasting. J. Hydrometeorol. 11, 69–86.
Matott, L.S., Babendreier, J.E., Purucker, S.T., 2009. Evaluating uncertainty in integrated environmental models: a review of concepts and tools. Water Resour. Res. 45. doi:10.1029/2008WR007301.
Murphy, A.H., 1993. What is a good forecast? An essay on nature of goodness in weather forecasting. Weather Forecast. 8, 281–293.
Muschalla, D., Schuetze, M., Schroeder, K., Bach, M., Blumensaat, F., Gruber, G., Klepiszewski, K., Pabst, M., Pressl, A., Schindler, N., Solvi, A.M., Wiese, J., 2009. The HSG procedure for modelling integrated urban wastewater systems. Water Sci. Technol. 60, 2065–2075.
Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models. J. Hydrol. 10, 282–290.
Oreskes, N., 2003. The role of quantitative models in science. In: Canham, C.D. et al. (Eds.), The Role of Models in Ecosystem. Science Princeton Univ. Press, pp. 13–31.
Power, M., 1993. The predictive validation of ecological and environmental models. Ecol. Model. 68, 33–50.
Refsgaard, J.C., 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., 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.
Reusser, D.E., Blume, T., Schaefli, B., Zehe, E., 2009. Analysing the temporal dynamics of model performance for hydrological models. Hydrol. Earth Syst. Sci. 13, 999–1018.
Romanowicz, R., Beven, K.J., Tawn, J.A., 1994. Evaluation of predictive uncertainty in nonlinear hydrological models using a Bayesian approach. In: Barnett, V., Turkman, F. (Eds.), Statistics for the Environment 2: Water Related Issues. Wiley.
Romanowicz, R.J., Beven, K.J., 2006. Comments on generalised likelihood uncertainty estimation. Reliab. Eng. Syst. Safe. 91, 1315–1321.
Santhi, C., Arnold, J.G., Williams, J.R., Dugas, W.A., Srinivasan, R., Hauck, L.M., 2001. Validation of the SWAT model on a large river basin with point and nonpoint sources. J. Am. Water Resour. Ass. 37, 1169–1188.
Schaake, J.C., Hamill, T.M., Buizza, R., Clark, M., 2007. The hydrological ensemble prediction experiment. Bull. Am. Meteorl. Soc. 88, 1541–1547.
Schaefli, B., Gupta, H.V., 2007. Do Nash values have value? Hydrol. Proc. 21, 2075–2080.
Scholten, H., Kassahun, A., Refsgaard, J.C., Kargas, T., Gavardinas, C., Beulens, A.J.M., 2007. A methodology to support multidisciplinary model-based water management. Environ. Model. Softw. 22, 743–759.
Seibert, J., McDonnell, J.J., 2002. On the dialog between experimentalist and modeler in catchment hydrology: use of soft data for multicriteria model calibration. Water Resour. Res. 38. doi:10.1029/2001WR000978.
Shao, J., Tu, D., 1995. The Jackknife and Bootstrap. Springer-Verlag, New York.
Smith, M.B., Georgakakos, K.P., Liang, X., 2004. The distributed model intercomparison project (DMIP). J. Hydrol. 298, 1–32. doi:10.1016/j.jhydrol.2004.05.001.
Todini, E., 2007. Hydrological catchment modelling: past, present and future. Hydrol. Earth Syst. Sci. 11, 468–482.
Van der Molen, D.T., Pintér, J., 1993. Environmental model calibration under different specifications: an application to the model SED. Ecol. Model. 68, 1–19.
Welles, E., 2005. Verification of River Stage Forecasts. PhD Dissertation, University of Arizona.
Welles, E., Sorooshian, S., Carter, G., Olsen, B., 2007. Hydrologic verification, a call for action and collaboration. Bull. Am. Meteorl. Soc. 88, 503–511. doi:10.1175/BAMS-88-4-503.
Wilks, D.S., 2006. Statistical Methods in the Atmospheric Sciences, 2nd ed. Academic Press, 627 pp.
Wilmott, C.J., Ackleson, S.G., Davis, R.E., Feddema, J.J., Klink, K.M., Legates, D.R., O'Donnell, J., Rowe, M.C., 1985. Statistics for the evaluation and comparison of models. J. Geophys. Res. 90, 8995–9005.