Hydrological Sciences Journal – Journal des Sciences Hydrologiques, 60 (7–8) 2015
http://dx.doi.org/10.1080/02626667.2014.967248
Special issue: *Modelling Temporally-variable Catchments*

Hydrology under change: an evaluation protocol to investigate how hydrological models deal with changing catchments

G. Thirel[1], V. Andréassian[1], C. Perrin[1], J.-N. Audouy[2], L. Berthet[2], P. Edwards[3], N. Folton[4], C. Furusho[1], A. Kuentz[1,5,6], J. Lerat[7], G. Lindström[8], E. Martin[9], T. Mathevet[5], R. Merz[10], J. Parajka[11], D. Ruelland[12] and J. Vaze[13]

[1]*Irstea, Hydrosystems and Bioprocesses Research Unit (HBAN), Antony, France*
guillaume.thirel@irstea.fr

[2]*Allier Basin Flood Forecasting Centre (DREAL d'Auvergne), Clermont-Ferrand, France*

[3]*Northern Research Station, USDA Forest Service, Northern Research Station, Parsons, West Virginia, USA*

[4]*OHAX Research Unit, Irstea, Aix-en-Provence, France*

[5]*EDF DTG, Grenoble, France*

[6]*LTHE, Grenoble, France*

[7]*Bureau of Meteorology, Canberra, Australia*

[8]*SMHI, Norrköping, Sweden*

[9]*CNRM-GAME, Météo-France, CNRS, Toulouse, France*

[10]*Department Catchment Hydrology, Helmholz Centre for Environmental Research (UFZ), Halle, Germany*

[11]*Institute of Hydraulic and Water Resources Engineering, TU Vienna, Vienna, Austria*

[12]*CNRS, HydroSciences Montpellier, Montpellier, France*

[13]*Land and Water Flagship, CSIRO, Canberra, Australia*

Received 2 January 2014; accepted 18 August 2014

Editor D. Koutsoyiannis; **Associate editor** A. Efstratiadis

Abstract Testing hydrological models under changing conditions is essential to evaluate their ability to cope with changing catchments and their suitability for impact studies. With this perspective in mind, a workshop dedicated to this issue was held at the 2013 General Assembly of the International Association of Hydrological Sciences (IAHS) in Göteborg, Sweden, in July 2013, during which the results of a common testing experiment were presented. Prior to the workshop, the participants had been invited to test their own models on a common set of basins showing varying conditions specifically set up for the workshop. All these basins experienced changes, either in physical characteristics (e.g. changes in land cover) or climate conditions (e.g. gradual temperature increase). This article presents the motivations and organization of this experiment—that is—the testing (calibration and evaluation) protocol and the common framework of statistical procedures and graphical tools used to assess the model performances. The basins datasets are also briefly introduced (a detailed description is provided in the associated Supplementary material).

Key words non-stationarity; IAHS workshop; calibration; evaluation; protocol; split-sample test; changing catchments dataset

Hydrologie sous changement: un protocole d'évaluation pour examiner comment les modèles hydrologiques s'accommodent des bassins changeants

Résumé: Tester les modèles hydrologiques pour des conditions changeantes est essentiel pour évaluer leur capacité à faire face à des bassins changeants et leur pertinence pour les études d'impact. Un atelier, dédié à cette problématique et pendant lequel les résultats d'une expérimentation de tests conjoints ont été présentés, s'est tenu dans cette perspective en juillet 2013 à Göteborg lors de l'Assemblée Générale 2013 de l'AISH. Avant l'atelier, les participants avaient été invités à tester leurs propres modèles sur un jeu commun de bassins qui montraient des conditions changeantes et qui avait été spécifiquement préparé pour cet atelier. Tous ces bassins ont subi des changements, soit de leurs caractéristiques physiques (par exemple l'évolution de la couverture du sol), soit des

conditions climatiques (par exemple une augmentation graduelle de la température). Cet article présente les motivations et l'organisation de cette expérimentation, c'est-à-dire le protocole de tests (calage et évaluation) et la manière dont les résultats ont été analysés dans un cadre commun utilisant des mesures statistiques d'efficacité et des outils graphiques. Les jeux de données des bassins sont également brièvement présentés (une description détaillée est fournie dans le document complémentaire associé à cet article).

Mots clefs non-stationnarité ; atelier AISH ; calage ; évaluation ; protocole ; test de calage-contrôle ; jeu de données de bassins changeants

1 MOTIVATION AND SCOPE

1.1 On changing catchments

Better understanding how hydrological systems respond to changing conditions is a key question in the scientific community that may help improve the prediction of the impacts of various future environmental changes (Peel and Blöschl 2011). However, the notion of change through time is relative and depends on the time window that is considered (Koutsoyiannis 2013). 'Change' may have several meanings for catchments. Here we define a changing catchment as one that undergoes a significant change in land cover or climate conditions (a systematic deviation that is outside the range of the historic record). In addition, water management changes due to the construction of a storage dam are also considered, but other human-induced changes that can have a significant impact on flow dynamics (e.g. water withdrawals, dike construction, or streambed gravel mining) are not included in the analysis. Note that a change in land cover or climate does not necessarily imply a significant change in the hydrological behaviour of the catchment.

1.2 Should we fear changing catchments?

During the early days of hydrological modelling (in the 1960s and 1970s), there was a belief that computers eventually would be able to solve all arising problems, so there was little concern about changing catchments. Linsley (1982), who created one of the first modern hydrological models at Stanford University, considered that a good generic model should be able to adapt to the breadth of catchment conditions and events, provided that it would "*represent the various processes with sufficient fidelity so that irrelevant processes can be 'shut off' or will simply not function*". In those years, the idea was that calibration was a safe way to parameterize models so that they would adequately reproduce the catchment behaviour.

These happy days ended with increasing doubts on calibration. Already in the early 1970s, Johnston and Pilgrim (1973) had described the numerous disappointments caused by an extensive search for the optimum values of the parameters of Boughton's catchment model. They developed a comprehensive list of problems that have since been recognized as the major impediments to the calibration of hydrological models, including discontinuities of the response surface, multiplicity of equivalent optima, un-identifiability, and lack of robustness of calibrated parameter values. Eventually it became obvious that improvement in search algorithms could reduce the *numerical* mis-calibration problems but the *hydrological* over-calibration problems remained to be addressed (Andréassian *et al*. 2012). Obviously, model calibration could "*work to accommodate reality, often in a subtle way*" (Beven 1977), but came at the cost of the model's predictive capacity.

When, on a perfectly steady catchment, parameters can change for (apparently) no reason from one calibration period to another, it seems unlikely that we can ever address the issue of fully representing responses in changing catchments. To be able to extrapolate to other climates or other land uses, it is essential to understand the causes of the dynamic behaviour of the catchment; otherwise the model will remain a "*mathematical marionette*", only able to "*dance to a tune it has already heard*" (Kirchner 2006).

1.3 Towards solutions for dealing with change

The International Association of Hydrological Sciences (IAHS) decade on Prediction on Ungauged Basins (PUB; 2003–2012) (Hrachowitz *et al*. 2013), in combination with the initiation of the 'Panta Rhei' decade (2013–2022) (Montanari *et al*. 2013), gave a new impetus to attempts to assess the extrapolation capacity of hydrological models in space and time. As Boughton (2005) put it, "*the problem of estimating runoff from ungauged catchments* [...] *is closely related to the problem of estimating the change in runoff that will occur when the land use of a catchment changes*". Recent studies (Seibert 2003, Coron *et al*. 2011, 2012, Merz *et al*. 2011, Refsgaard *et al*. 2013) have attempted to distinguish *apparent*

changes, i.e. in observable properties, such as land cover and climate conditions (but that may not necessarily induce hydrological changes), from *intrinsic* changes, i.e. in the catchment hydrological behaviour (understood as an input–output relationship), that may be difficult to identify or attribute to specific causes on the sole basis of observations and that require the use of some kind of hydrological model.

However, we believe that hydrological modellers need to strive to respond collectively to the growing key questions in the challenging research field of hydrology under change (Peel and Blöschl 2011). Towards this end, the Workshop *Testing simulation and forecasting models in non-stationary conditions* was held during the IAHS General Assembly in Göteborg, Sweden (July 2013), to foster the development of relevant hydrological parameterization methods for application to changing catchments. Before this workshop, the participants were asked to carry out calibrations of their models over successive and contrasted test periods. An evaluation protocol was defined, and a set of metrics was proposed as a means for analysing the models' performance and parameter transferability. To promote comparability, we gathered datasets from changing catchments and asked the modellers to use them preferentially as test beds. With the proposed framework, the participants were asked to contribute answers to the following questions during the Göteborg workshop:

- Can changes in model parameters calibrated over different periods tell us whether a catchment is changing, or are there too many numerical artefacts to answer this question (due to poor parameter identifiability, model over-parameterization, etc.)?
- Are our models robust and/or flexible enough to be used under changing conditions?
- What approaches should be tried in the coming years to better handle hydrological modelling under change?

These questions guided the searches for improving model results on changing catchments, as reported in the special issue of *Hydrological Sciences Journal* to which this paper contributes (Thirel *et al*. 2015).

This paper presents the protocol (Section 2) and the numerical graphical tools elaborated for analysing the model results (Section 3). The datasets of the changing catchments are briefly introduced in Section 4 and described in more detail in the Supplementary material.

2 CALIBRATION AND EVALUATION PROTOCOL

2.1 On differential split-sample testing (DSST)

Due to the difficulty of exhaustively describing all the physical processes involved in the transformation of precipitation into streamflow, many hydrological models of a wide range of complexity have been proposed in the literature. These models require the calibration of several free parameters against discharge observations, because these parameters often cannot be linked to physical characteristics directly (Abebe *et al*. 2010). However, calibration may be unable to reach a good representation of streamflow in all conditions, whatever the quantity of data used for calibration (Andréassian *et al*. 2012), and may show unstable behaviour between different calibration periods. The modeller should try to solve this problem by seeking ways to obtain stable parameters.

The diagnosis of model stability is possible through testing procedures that have been proposed for assessing the dependence of model parameters on climate or other potentially changing factors. The most famous example, which is also the most frequently used method, is the Split-Sample Test (SST) proposed by Klemeš (1986) within a hierarchical scheme for systematic testing of hydrological models. The temporal transposability of models is assessed through cross-calibration and validation tests of the models over two different periods. The even more demanding Differential Split-Sample Test (DSST) described by Klemeš (1986) seems particularly relevant for evaluating models under changing conditions. Within the DSST, the model should be calibrated and validated over contrasted periods: for example, calibrated over a dry period and validated during a wet period. The Klemeš testing scheme also includes the evaluation of model spatial transposability by testing neighbouring catchments (proxy-basin test), possibly with contrasted conditions (differential split-sample proxy-basin test) (Klemeš 1986, 2009).

In many cases, models fail when the most demanding test described above is applied and, for this reason, the full Klemeš testing scheme is seldom applied (Andreassian *et al*. 2009, Klemeš 2009). Actually, only a few studies have applied it (Refsgaard and Knudsen 1996, Donnelly-Makowecki and Moore 1999, Stisen *et al*. 2012). However, the less-demanding DSST still seems a suitable tool to evaluate models under changing conditions. For this reason, the protocol described in Section 2.2 is based on this test. The DSST is now more commonly used by hydrologists (Seibert

2003, Vaze *et al*. 2010, Brigode *et al*. 2013, Refsgaard *et al*. 2013) and has even been generalized by Coron *et al*. (2012), who proposed the use of multiple test periods defined by a time-sliding window over the full available period. This testing approach also relates to other recent investigations on model parameter stability in time (see e.g. Gharari *et al*. 2013, Razavi and Tolson 2013).

2.2 Calibration and evaluation periods

For each dataset six calibration periods were defined. The first period, called the "complete period", was set in such a way that it made use of as much data as possible. At least two years were kept at the beginning of the complete period for model warm-up. Calibration on the complete period represents Level 1 of the protocol. Since this test level does not include proper model validation, it should be considered as a preliminary step for modellers to check the general model suitability for the studied catchment. Then five sub-periods, nested within the complete period and called P1, P2, P3, P4 and P5 were defined. These five periods have the same number of years and were chosen to cover the span of the complete period as fully as possible. Calibration on the five sub-periods represents Level 2 of the protocol. The details of the six test periods (Complete Period plus P1–P5) are defined in the Supplementary material for each test catchment.

The modellers were asked to calibrate their hydrological models on each test period and to make simulations over the complete period using each of the six parameter sets successively. Simulation performance was then analysed over the six periods. The principle of this protocol is summarized in Fig. 1. Modellers were asked to calculate performance criteria on the complete period, as well as on each of the five sub-periods to obtain common bases for the evaluation of models simulations.

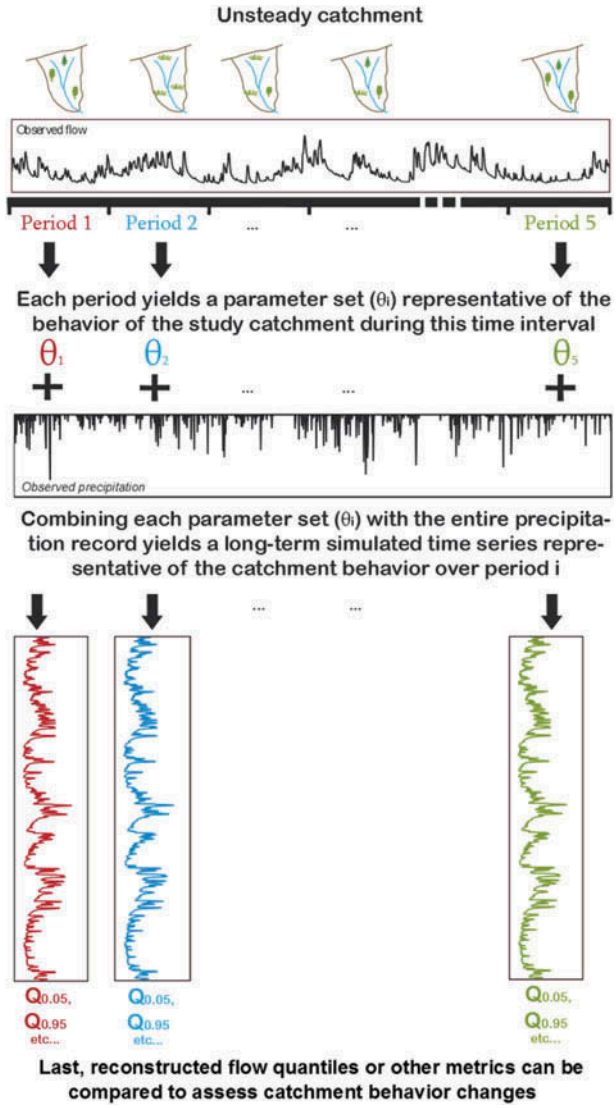

Fig. 1 Principle of the calibration and evaluation protocol over the five sub-periods.

2.3 Evaluation criteria

Prior to the workshop, a list of statistical scores was established, which are described in Table 1. We classified the scores into two groups:

- primary scores, which we considered to be the basis for a careful hydrological evaluation, since they evaluate simple expected properties in terms of water balance, representation of high or low flow, etc.; and
- secondary scores, the examination of which was not mandatory, but which can give valuable information on how well the models could handle the proposed changing catchments.

2.3.1 Primary scores The Nash-Sutcliffe efficiency criterion (NSE; Nash and Sutcliffe 1970) is a quadratic score commonly used to assess the quality of hydrological models. The NSE value ranges from $-\infty$ to 1, with negative values when the model performance falls below that of a simple model that would simulate a constant value equal to the average of the observed discharges. The maximum score of 1 means that the simulated discharges equal the observed discharges. Due to its quadratic nature, the NSE principally gives information about the capacity

Table 1 Definition of the evaluation criteria. n: number of days of the evaluation period; Q_{sim} and Q_{obs}: simulated and observed discharges; μ: arithmetic mean over the evaluation period; and σ: standard deviation. The parameters r, α_{NSE}, β_{NSE}, KGE′ι, $\beta_{KGE'ι}$ and $\gamma_{KGE'}$ can be adapted to low flows in the same way as NSE_{LF} is an adaptation of NSE to low flows.

Criterion	Mathematical formula	Description	Best value
NSE	$1 - \frac{\sum_{t=1}^{n}(Q_{obs}(t) - Q_{sim}(t))^2}{\sum_{t=1}^{n}(Q_{obs}(t) - \mu(Q_{obs}))^2}$ $= 2\alpha_{NSE} \times r - (\alpha_{NSE})^2 - (\beta_{NSE})^2$	Nash-Sutcliffe efficiency and decomposition	1
NSE_{LF}	$1 - \frac{\sum_{t=1}^{n}\left(\frac{1}{Q_{obs}(t)+\varepsilon} - \frac{1}{Q_{sim}(t)+\varepsilon}\right)^2}{\sum_{t=1}^{n}\left(\frac{1}{Q_{obs}(t)+\varepsilon} - \mu\left(\frac{1}{Q_{obs}+\varepsilon}\right)\right)^2}$	Nash-Sutcliffe efficiency on low flows	1
Bias (or $\beta_{KGE'}$)	$\frac{\mu(Q_{sim})}{\mu(Q_{obs})}$	Bias	1
$Freq_{LF}$	$freq(Q_{sim} < 0.05\mu(Q_{obs}))$	Frequency of low flows	Same as for Q_{obs}
r	$\frac{\sum_{t=1}^{n}(Q_{obs}(t) - \mu(Q_{obs}))(Q_{sim}(t) - \mu(Q_{sim}))}{\sigma(Q_{sim})\sigma(Q_{obs})}$	Linear correlation coefficient	1
α_{NSE}	$\frac{\sigma(Q_{sim})}{\sigma(Q_{obs})}$	Relative variability in the simulated and observed discharges	1
β_{NSE}	$\frac{\mu(Q_{sim}) - \mu(Q_{obs})}{\sigma(Q_{obs})}$	Bias normalized by the standard deviation of the observed discharges	0
KGE′	$1 - \sqrt{(r-1)^2 + (\beta_{KGE'} - 1)^2 + (\gamma_{KGE'} - 1)^2}$	Modified Kling-Gupta efficiency	1
$\gamma_{KGE'}$	$\frac{\sigma(Q_{sim})/\mu(Q_{sim})}{\sigma(Q_{obs})/\mu(Q_{obs})}$	Variation coefficient ratio	1

of the model to simulate high flows. For this reason, we also calculated a modified version of the NSE (NSE_{LF}), calculated with inverse transformed flows $1/(Q + \varepsilon)$, where Q is the flow and ε is a small constant introduced to avoid divisions by 0 in the case of zero flows. Here ε was set to one-hundredth of the mean observed flows ($\varepsilon = \mu(Q_{obs})/100$). This score has proven to be adequate for low flow estimation (Pushpalatha *et al*. 2012).

Bias compares the mean simulated discharges and the mean observed discharges over the test period. Here the bias was computed as the ratio between mean simulated discharges and mean observed discharges. When model simulations are not biased, the bias value is 1. Values lower than 1 indicate underestimation of the mean discharge, while values larger than 1 indicate overestimation. The evolution of this criterion over contrasted periods often provides useful information on a model's capacity to reproduce the water balance over time.

2.3.2 Secondary scores One model can perform well for a specific range of discharges (e.g. low flows), but not for the other magnitudes of discharges (e.g. high flows). The NSE and NSE_{LF} criteria already give an indication of this kind of behaviour, but the distribution of the simulated discharges can give additional information. This is especially the case when an infrequent event occurs, because these strongly influence the NSE or NSE_{LF} values, due to their quadratic formulation. Visually comparing the simulated flow duration curves (FDCs) with the observed FDC is one way to compare the distributions (see Section 3). Four specific quantiles, focusing on high and low flows, i.e. 95% ($Q_{0.95}$), 85% ($Q_{0.85}$), 15% ($Q_{0.15}$) and 5% ($Q_{0.05}$) quantiles, are also calculated, where $Q_{0.95}$ represents the discharge value *not* exceeded 95% of the time (i.e. it is a high-flow indicator), while $Q_{0.05}$ represents the discharge value exceeded 95% of time (i.e. it is a low-flow indicator). Comparing simulated and observed quantiles over different periods gives an indication of how

close the modelled flow distribution is to the observed distribution. However, some intermittent-flow catchments are present in our dataset. For them, null discharges are quite common and the observed $Q_{0.05}$ and $Q_{0.15}$ are equal to 0; this also may be the case for the simulated quantiles. For this reason, the frequency of days with simulated discharges lower than 5% of the average observed discharge ($Freq_{LF}$) provides an alternative. The same formula can be applied with Q_{obs} instead of Q_{sim} so the frequency of low flows can be compared for observed and simulated discharges.

Although the NSE is a widely used criterion among hydrologists, due to representation of periodicity, NSE may take high values, thus providing misleading conclusions about the actual model capacity, for example for snowmelt-driven catchments or for catchments facing strong seasonal meteorological gradients. As a result, some authors have suggested modified versions of NSE (Mathevet *et al*. 2006, Criss and Winston 2008). The NSE can be decomposed as a sum of different criteria (Murphy 1988, Węglarczyk 1998), which makes it easier to understand the origin of poor model performance. In this context, the three components (correlation coefficient r, ratio of standard deviations α_{NSE} and relative bias β_{NSE}) of the NSE decomposition proposed by Gupta *et al*. (2009) in their Equation 4 (see Table 1) were used. The ideal values for the three elements of the NSE decomposition are $r = 1$, $\alpha_{NSE} = 1$ and $\beta_{NSE} = 0$. Due to the double appearance of α_{NSE} in the NSE decomposition, Gupta *et al*. (2009) showed that the maximum NSE is obtained when $\alpha_{NSE} = r$. As a result, the α_{NSE} value selected for optimizing the NSE leads to an underestimation of the variability of simulated flows. To overcome this problem, Gupta *et al*. (2009) proposed a new criterion, the Kling-Gupta efficiency (KGE), later slightly modified by Kling *et al*. (2012) (and called KGE′) to avoid interactions between the bias and variation coefficient ratios. In the proposed protocol, we decided to use this modified criterion (KGE′) and its decomposition. The NSE decomposition, the KGE′ and its decomposition (which is different from the NSE one) can be computed for low flows in the same way as NSE_{LF} is calculated from NSE.

3 GRAPHICAL ANALYSIS

Several types of plot were suggested by the workshop organizers to help modellers analyse their results through a common framework. Here we provide a detailed description of these graphical tools. Having a common graphical representation will indeed help readers to compare the results of different studies. In the following, the graphs describe the scores calculated on two time spans: a sub-period-based time span (Section 3.1) and an annual time span (Section 3.2). Presenting the scores described in Section 2.3 over these two time spans provides a complementary analysis of the results. While the plots given on sub-periods provide a means to analyse general trends over periods of several years, the annual scores highlight the effect of specific years on the scores. Most of the plots presented in this article comprise not only the scores over each evaluation period (i.e. the sub-periods or every year), but also the scores for each calibration made by the modellers. This two-dimensional analysis is necessary to analyse both the effect of the calibration period and the evaluation period on scores.

For both analyses, two ways of representing the results were proposed in most cases: classical curve-based plots and two-dimensional (2-D) tables. Both representations can identify good or poor model performance. One difference is, however, the possibility of easily visualizing the scores' stability across periods when using the curve-based plots (note that the stability of scores is an indicator of temporal transferability of hydrological parameters). The 2-D tables ease the identification of the length of time between the calibration and the evaluation periods.

3.1 Sub-period analysis

Recall that the Complete Period, which is the longest period we could use for each catchment, was split into five sub-periods for both the calibration and the evaluation of the models (see Section 2.2). In other words, six simulations were made with each model (one simulation for each calibration) and each of them can be analysed over six periods. In this paper we present the plots obtained for an evaluation of the six simulations on the five sub-periods and on the complete period.

The NSE, KGE′, their decompositions and the bias could be represented using the type of graph shown in Fig. 2, where six coloured curves are drawn: each curve corresponds to the criterion values over the evaluation periods for a single model calibration (NSE is considered in Fig. 2). The x-axis presents six ticks: each of them is a NSE value for an evaluation on a given sub-period or on the

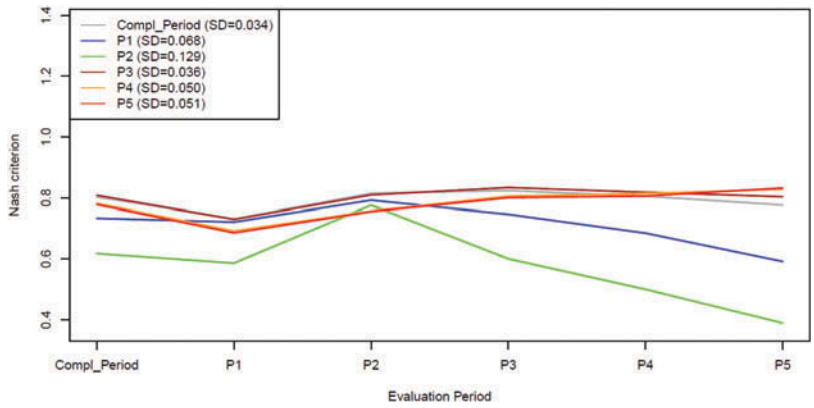

Fig. 2 Example of a plot for the five sub-periods plus the Complete Period. Here the NSE is represented. Each curve corresponds to the evolution of the score over the periods for a given calibration. SD: standard deviation of each curve from its mean.

complete period. The *y*-axis actually represents the NSE values. The standard deviation (SD) of each curve is also given as an additional indicator of the stability of the score value of the model simulations. However, providing a good simulation over a period is not sufficient to define a good model for simulating changing catchments: it is also necessary to reproduce this good performance for all evaluation periods. Please note that for the bias plot (not shown here), the $y = 1$ line (i.e. the perfect value) is also drawn, represented by a dashed line.

Another representation of the same scores consisted in using 2-D table plots (see Fig. 3 for an example on NSE_{LF}), in which the score values are given, not by curves, but by coloured squares. In these tables, a row corresponds to an evaluation period and a column to a calibration period. In other words, if one wishes to assess the impact on a score of the period used for the model calibration over one of the evaluation periods (see Fig. 3), the row corresponding to this evaluation period will give that information. By contrast, for a single calibration, to check the change in the score *vs* the evaluation period, it is necessary to read the table in columns. To simplify the interpretation of these tables, the squares for which the evaluation period is included in the

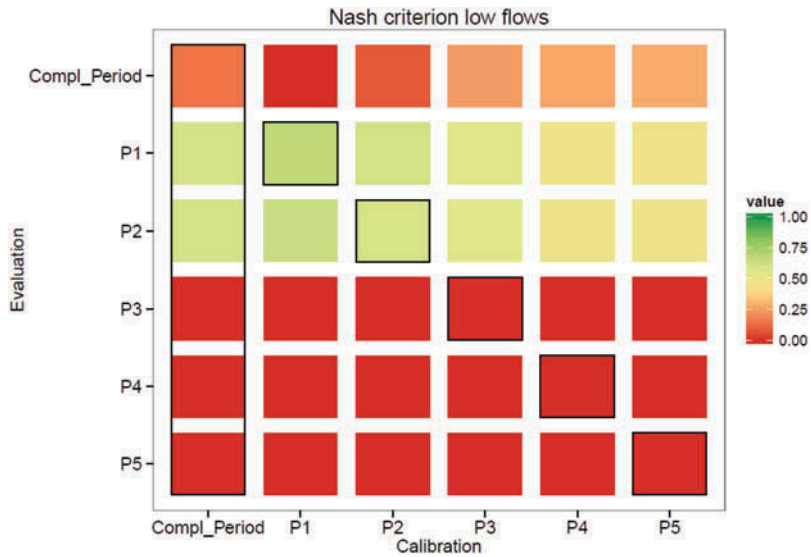

Fig. 3 Example of a table plot for the five sub-periods plus the Complete Period. Here the NSE_{LF} is represented. Rows give the scores of the six calibrations on a given evaluation period; columns give the scores of one calibration over each of the six evaluation periods. Boxes outlined in black represent ones for which the evaluation period is included in the calibration period.

calibration period are outlined in black (elements of the diagonal plus the first column). Thus Fig. 2 makes it easier to identify parameter sets that are stable over time, and Fig. 3 makes it easier to identify the impact of the distance between the calibration and evaluation periods considered.

The different quantiles ($Q_{0.95}$, $Q_{0.85}$, $Q_{0.15}$ and $Q_{0.05}$) and the Freq_{LF} can also be represented through the graph of Fig. 2. The information added to these plots is a curve for the observed discharge score, represented by a dashed black line (see Fig. 4). Since the objective of these criteria is not to be as stable and as good as possible, but to be as close as possible to the observations, the SD metric is replaced here with the correlation between each simulated curve and the observed curve. This type of representation includes one graph for each quantile considered or for Freq_{LF}, and each graph shows all the evaluation periods. The plots of monthly discharge regimes include one graph for each evaluation period.

The quantiles and the Freq_{LF} can also be drawn with 2-D tables. The only difference compared to the NSE 2-D tables is the addition of a column for observations (e.g. see the first column "Obs." in Fig. 5). We did not present the discharge regime curves (mean monthly values) in this way.

Simulated and observed FDCs are plotted as shown in Fig. 6. In this graph, each coloured line is obtained from a different model calibration, and the observed FDC is given with a dashed curve. The discharge values (m^3/s) are plotted against their

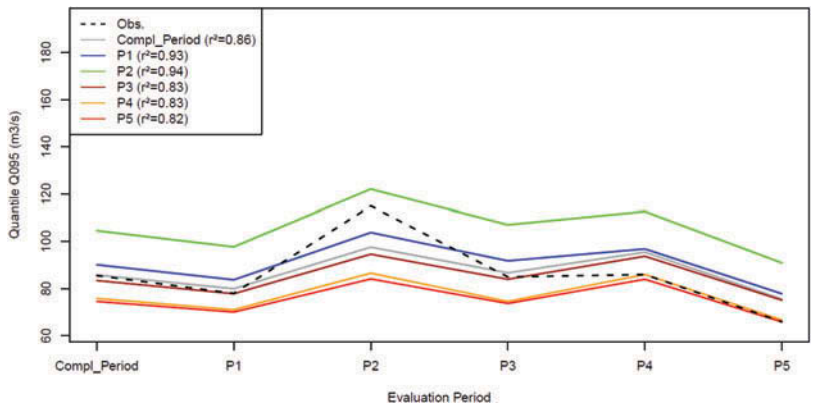

Fig. 4 Example of a plot for the five sub-periods plus the Complete Period, including the observed discharge score (dashed line). Here the $Q_{0.95}$ is represented.

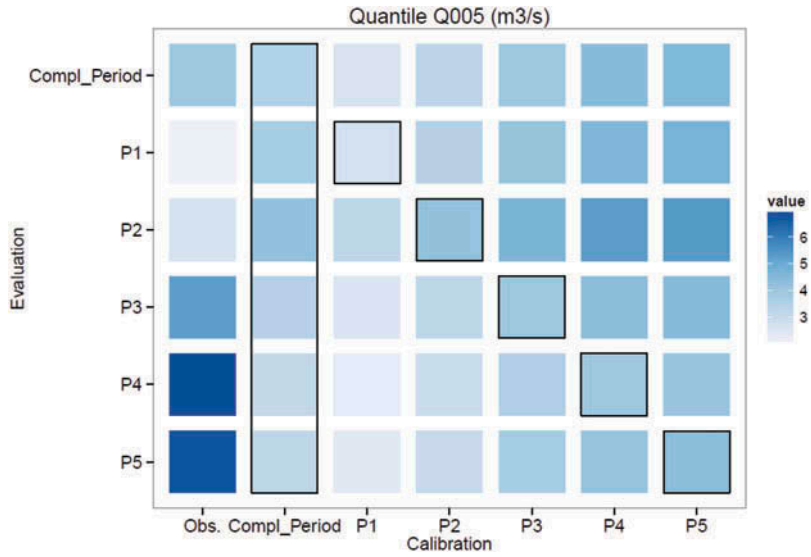

Fig. 5 Example table plot for the five sub-periods plus the Complete Period and the observation. Here the $Q_{0.05}$ is represented.

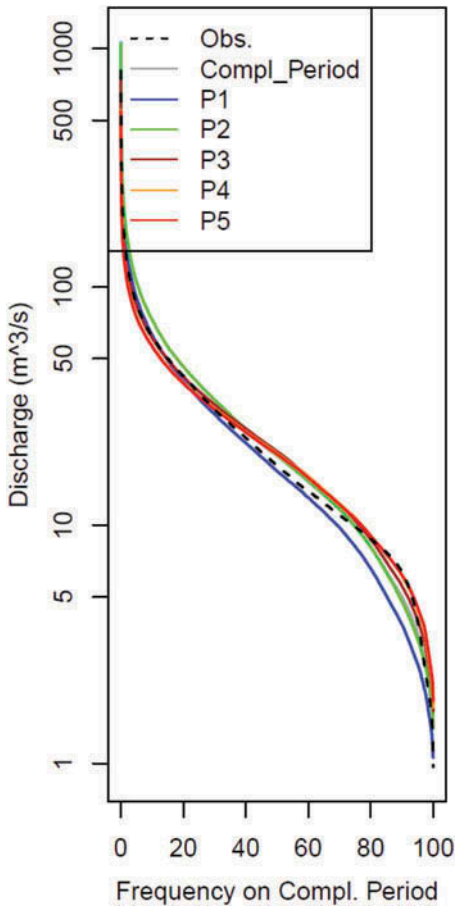

Fig. 6 Example of a plot for the five sub-period calibrations plus the Complete Period representing the flow duration curve (FDC). The observed FDC is given by a dashed line. For this plot the evaluation is done on the complete period only.

frequency of exceedence. In total, six similar graphs are plotted, one for each evaluation period. There is no two-dimensional table version of these graphs.

3.2 Annual analysis

In addition to examining five contrasted sub-periods plus the Complete Period, we decided to provide a more detailed graphical analysis computed on an annual basis to show how changes in scores can highlight sudden changes or outlier years. It should be recalled that model calibration was no different to that for the plots presented in Section 3.1 (also based on the Complete Period plus periods P1–P5). No further calibration (e.g. on individual years) was performed by the modellers.

Each of the graphs presented for the sub-periods could easily be extended to the annual basis. An example of the extension of Fig. 2 to an annual basis is presented in Fig. 7: the six curves (i.e. the six different model calibrations) are plotted, but the *x*-axis, which gives the evaluation periods, is extended to each year of the Complete Period. Note that the values of the score for the Complete Period are still indicated as the first element of the *x*-axis. Combining Fig. 7 and Fig. 2 shows that the poor performance of calibrations on P1 and P2 (blue and green curves) on the P3–P5 periods is not due to unusually poor event simulations influencing the NSE metrics, but to a systematic deficiency of the simulations (as shown by the low NSE values for P1 and P2 calibrations from 1979 to 2008, which include the P3–P5 periods).

Figure 8 is the extension of Fig. 3 on an annual basis. Here there is one row for each evaluation year (plus one for the Complete Period). Since the extension on an annual basis of Fig. 4 is similar to that of Fig. 2, and the extension of Fig. 5 is similar to that of Fig. 3, they are not shown here. For Fig. 6, instead of having six different FDC plots, we provided as many

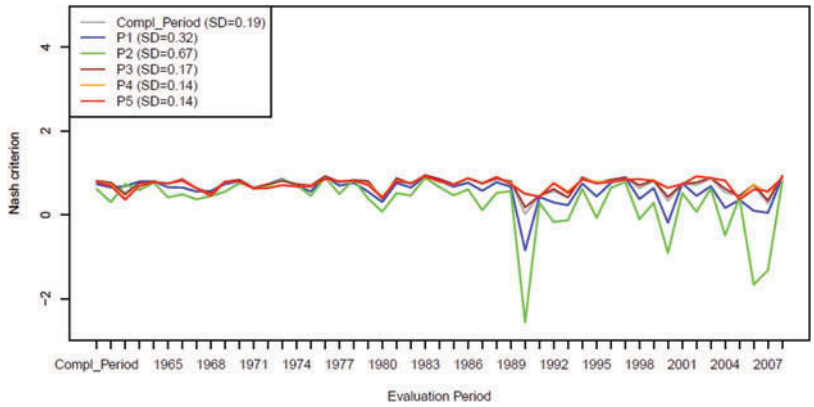

Fig. 7 Example of a plot showing the year-to-year variations of an efficiency criterion (NSE).

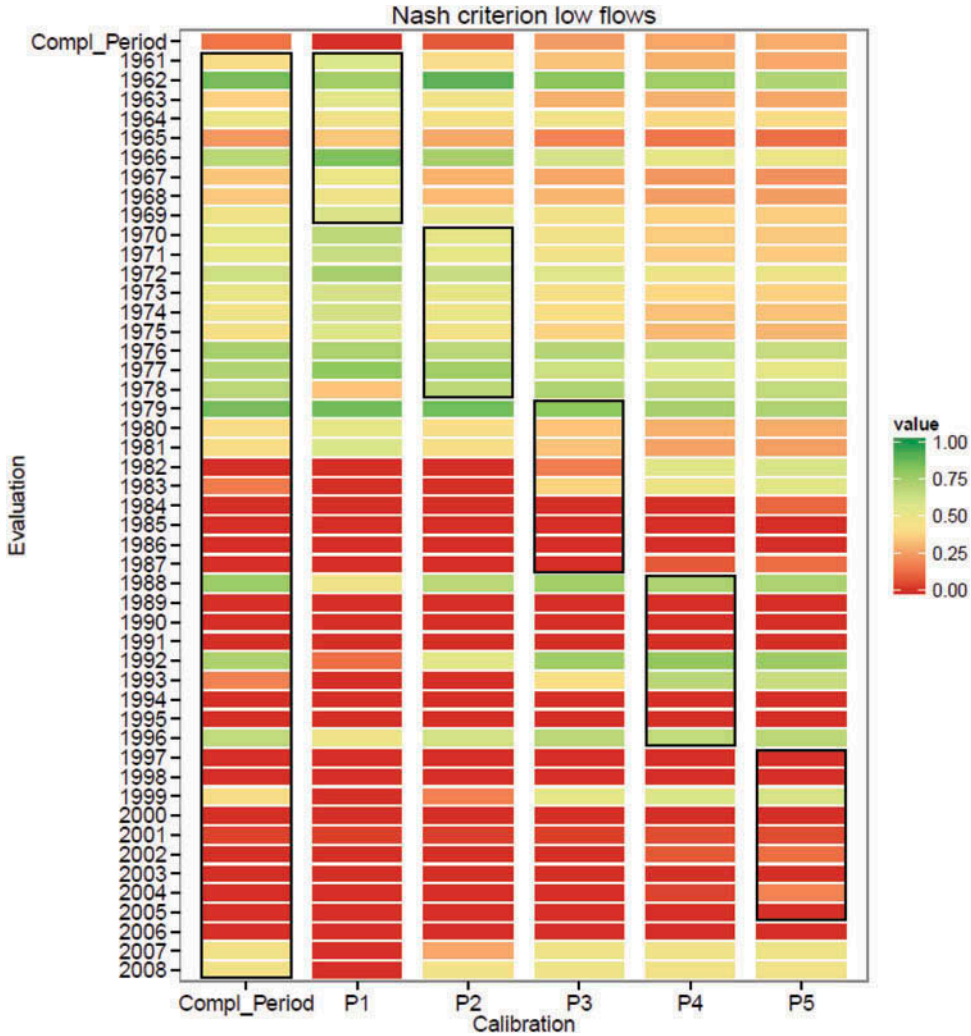

Fig. 8 Example of a table plot on an annual basis (NSE_{LF}).

plots as years in the Complete Period (plus one plot for the Complete Period).

3.3 Comparing the models

From the plots presented in the previous sections, the modellers received a substantial amount of information to assist in analysing their simulations. However, we did not wish to limit this workshop to a diagnosis of hydrological-model illnesses. We also wished to encourage modellers to rid their models of the deficiencies identified and propose solutions (e.g. by explicitly accounting for the causes of change in their models, see e.g. Nalbantis *et al*. (2011). Testing the possible solutions to remedy the models required having a simple visual check as to whether a solution improved or deteriorated model efficiency. Therefore, we also developed a way to represent the differences between the scores of several models or versions of models. The comparison graphs proposed to the modellers were of the two-dimensional table type. An example is given in Fig. 9 for a comparison of the NSE_{LF} values for two models, called Model 1 and Model 2. Since the structure of this representation is quite complex, a detailed description is provided here:

- First, note that on this graph, no NSE_{LF} values are actually given. The colour of the small squares corresponds to the difference between the NSE_{LF} values of two models. The upper right-hand block of the graph indicates NSE_{LF} for Model 2 minus NSE_{LF} for Model 1, and the lower left-hand block of the graph is NSE_{LF} for Model 1 minus NSE_{LF} for Model 2. In the colour scale, blue represents positive differences, white

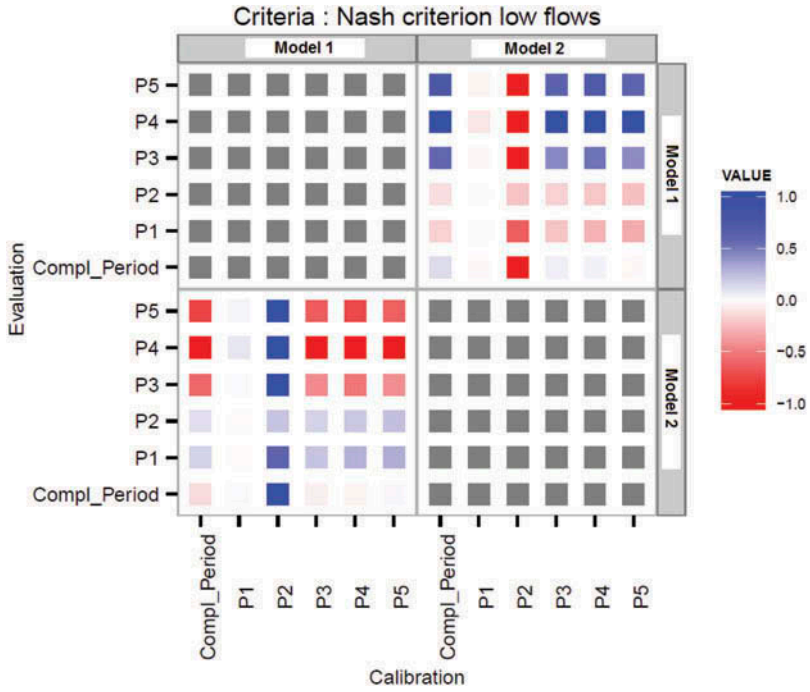

Fig. 9 Comparison of NSE_{LF} between Model 1 and Model 2. The colours represent the difference between the NSE_{LF} of the two models.

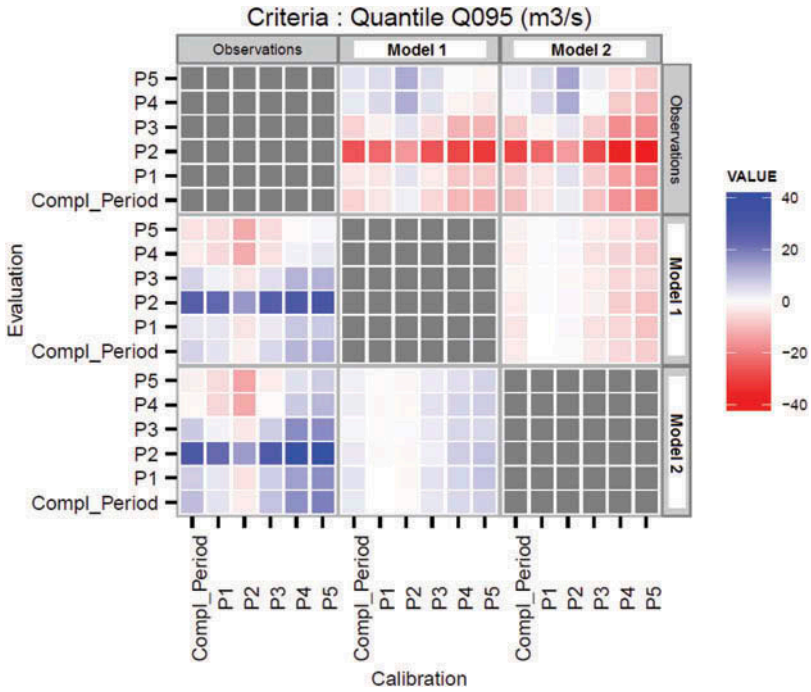

Fig. 10 Comparison of modelled and observed $Q_{0.95}$. The colours represent the difference between the $Q_{0.95}$ of the two models.

indicates that the two scores are very similar, and red denotes negative differences. Thus, the small blue squares drawn in the upper right-hand block of the graph indicate that Model 2 performs better than Model 1, and the small red squares indicate that Model 1 performs better than Model 2. For the lower left-hand block, the colours have the opposite meaning.

- The two blocks on the diagonal are filled with small grey squares, because they each compare one model with itself. Note that there is redundant information and that only one-fourth of these graphs would be sufficient. Nevertheless, as explained below, this graph can be extended, if needed, to compare more than two models or to include observations.

Now that the definition of the information content of this type of graph has been given, it is essential to explain how to read it correctly. The first way is of course to check the colour trends of the blocks: if one of the blocks is dominated by blue or red squares, then one of the models outperforms the other. Due to its two-dimensional structure, this table can also answer two questions:

1. For a given calibration period, how do the two models compare when evaluated over different evaluation periods?
2. For a given evaluation period, how do the two models compare when calibrated over different calibration periods?

Point (1) can be addressed by analysing the columns in Fig. 9. For example, let us imagine that we wish to compare how both models perform when they are calibrated over P2. To do so, we have to look at the P2 column in Fig. 9 (either on the lower left-hand block or on the upper right-hand block). From the study of this column, it is clear that Model 1 outperforms Model 2 for all evaluation periods, because this column is totally blue (lower left-hand block) or red (upper right-hand block). Only when the models are evaluated over P2 does their performance seem to be closer. This is an indication that both models have a similar performance over the period that is used for their calibration, but that the performance of Model 2 collapses when it is evaluated over another period (i.e. Model 2 is not robust when it is calibrated on P2).

Let us now focus on the second (2) type of comparison: for example, how the models behave on P3. For this, it is sufficient to study one of the P3 rows (let us take the one belonging to the upper right-hand block). From this row, it is clear that both models have a similar performance when calibrated on P1 (white square), that Model 1 performs better when the models are calibrated on P2 (red square), and that Model 2 performs better when they are calibrated over the other periods (blue squares). This may indicate that the P2 calibration of Model 2 fails on P3 for some reason.

For some metrics, such as the quantiles and $Freq_{LF}$, a comparison with the observations is necessary. For these metrics, the 2 × 2 block table (e.g. Fig. 9) becomes a 3 × 3 block table (see Fig. 10 for $Q_{0.95}$). Again in Fig. 10, only the differences between $Q_{0.95}$ values are given (and not $Q_{0.95}$ values themselves). When using this type of table, direct comparisons between each model and the observations are allowed and are not limited to comparisons between models. The approach to interpreting Fig. 10 is similar to that for Fig. 9 (i.e. for evaluating a single calibration, or several calibrations over all the periods), so it is not detailed further here. Again, it could be argued that instead of using a 3 × 3 block table, we could have limited it to a 2 × 2 block matrix (the four blocks of the upper right-hand part or the four blocks of the lower left-hand part) because of the reverse symmetry of the complete table. However, this version proved to be much easier to explain to users.

4 THE BASINS DATASET

Assembling a representative dataset of changing catchments was necessary for the success of this experiment. The dataset provided to the workshop participants consisted of 14 catchments that have experienced apparent changes. The locations of the catchments are shown in Fig. 11, and their main characteristics are given in Table 2. Each catchment and its corresponding dataset are described in detail in the Supplementary material. Summaries of mean monthly precipitation and streamflow data are given in Fig. 12. The hydrological regimes range from rain-fed (Axe Creek, rivers Bani, Flinders, Gilbert, Rimbaud and Wimmera) to snow-fed or pluvio-nival (rivers Allier, Durance, Garonne, Watershed 6 of the Fernow Experimental Forest, and Blackberry, Ferson and Obyån creeks).

The variety of changes affecting the 14 catchments offers a wide test bed to evaluate the capacity of hydrological models to deal with these changes. Several types of change are apparent. Temperature increase is the main factor of change in hilly or mountainous parts of Europe. This was the case for the rivers Garonne, Allier, Durance and Kamp. In the River Allier, a second type of change is the construction of a dam for sustaining low flows. Afforestation, deforestation and reforestation also modify the response of catchments in a substantial way. This type of land cover modification affected the River Rimbaud (forest fire), Obyån Creek (Gudrun storm destroying the forest) and Watershed 6 of the Fernow

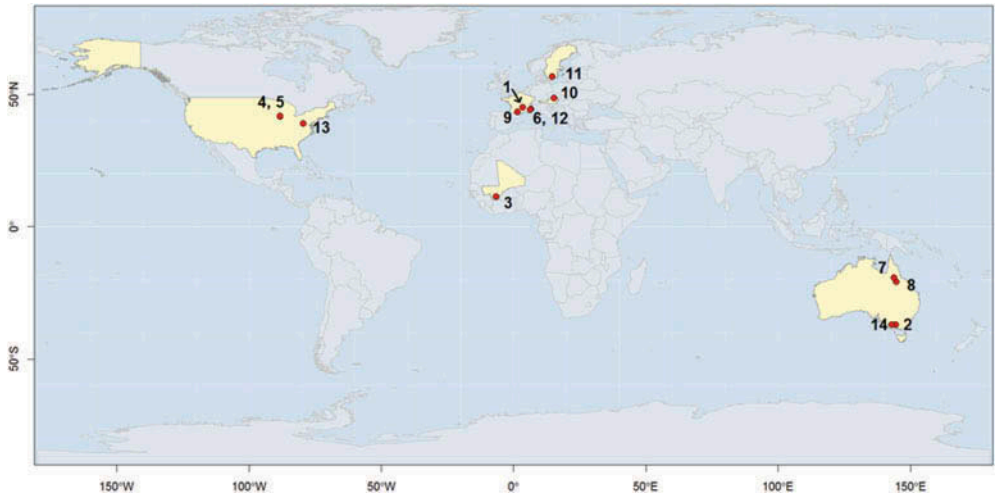

Fig. 11 Location of the 14 catchments. The numbers correspond to the first column of Table 2.

Table 2 Main characteristics of the 14 basins included in the database. See Fig. 11 for locations. The Figure numbers refer to those of the Supplementary material.

No.	River and gauging station	Figure number	Country	Catchment area (km^2)	Altitude range (m)	Annual precipitation (mm)	Available period	Type of change
1	River Allier at Vieille-Brioude	1	France	2 267	436–1551	900	1958–2008	Temperature increase, storage dam after 1983
2	Axe Creek at Longlea	2	Australia	237	175–710	625	1970–2011	Millennium Drought during the 2000s
3	River Bani at Douna	3	Mali, Ivory Coast and Burkina Faso	103 390	270–700	1075	1959–1990	West-Africa drought
4	Blackberry Creek at Yorkville	4	USA	182	185–310	975	1980–2011	Growing urbanization
5	Ferson Creek at St. Charles	4	USA	134	215–325	1000	1980–2011	Growing urbanization
6	River Durance at La Clapière	5	France	2 170	787–4102	1275	1901–2010	Temperature increase
7	River Flinders at Glendower	6	Australia	1 912	390–950	625	1967–2011	Arid catchment under cyclonic heavy rainfall influence Major floods in 1974 and 2009
8	River Gilbert at Gilberton	6	Australia	1 907	480–1070	750	1963–1988	Arid catchment under cyclonic heavy rainfall influence Major flood in 1974
9	River Garonne at Portet-sur-Garonne	7	France	9 980	140–3200	1125	1958–2008	Temperature increase
10	River Kamp at Zwettl	8	Austria	622	500–1000	750	1976–2008	Increase in temperature, flood in August 2002
11	Obyån Creek at Lissbro	9	Sweden	97	140–225	800	1981–2010	Loss of forest due to the Gudrun storm (2005)
12	River Rimbaud at Collobrières	10	France	1.4	470–622	1075	1966–2006	Forest fire in 1990
13	Watershed 6 of the Fernow Experimental Forest	11	USA	0.2	730–860	1425	1956–2009	Deciduous forest cut, and converted to conifers
14	River Wimmera at Glenorchy Concrete Weir Tail	12	Australia	2 000	170–950	550	1960–2009	Millennium Drought during the 2000s

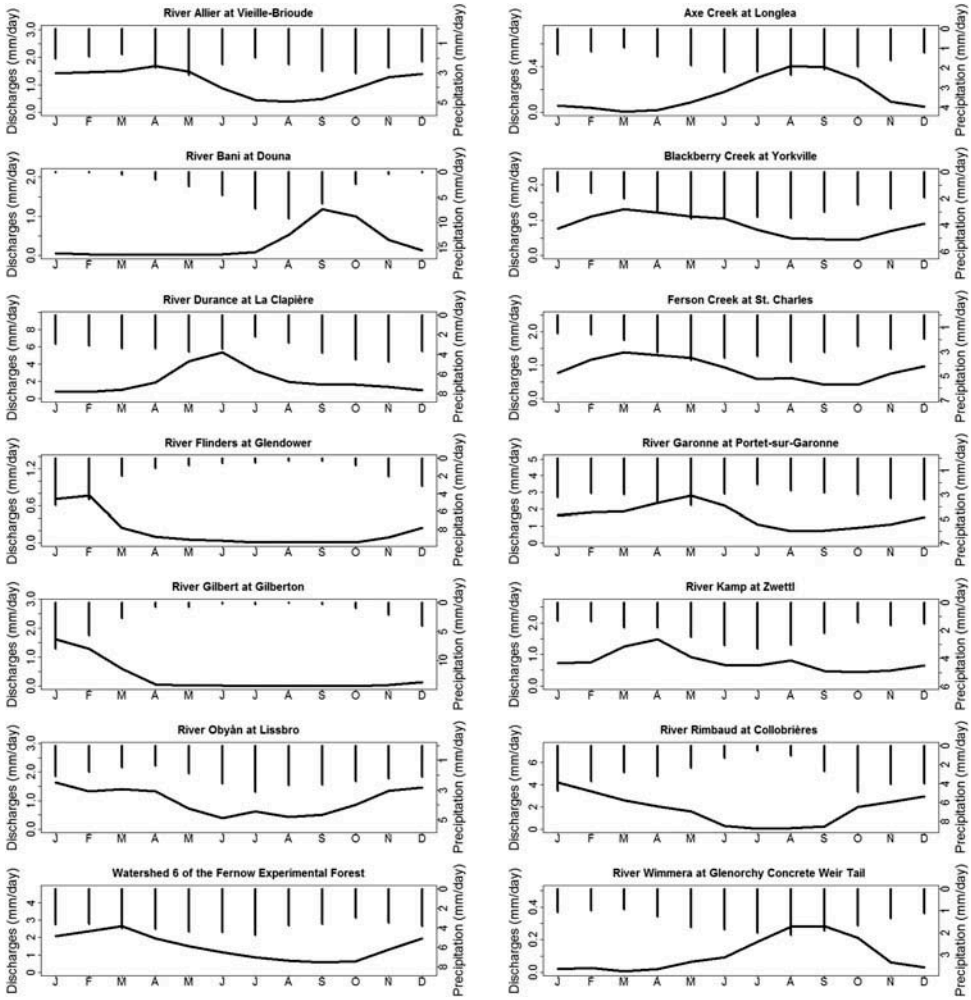

Fig. 12 Mean monthly streamflow and precipitation for the 14 basins over the Complete Period. Discharge (mm d^{-1}) is represented by lines and precipitation (mm d^{-1}) by histograms.

Experimental Forest (deforestation of the native deciduous forest followed by planting and establishment of conifers). Another type of land cover change is urbanization. Two American catchments, Blackberry Creek and Ferson Creek, were subject to such changes. The case of prolonged drought periods (extending over several years), which are known to affect the calibration of hydrological models, is illustrated by the "Millennium Drought" (River Wimmera and Axe Creek) and the West-Africa drought (River Bani). Finally, two semi-tropical basins (rivers Gilbert and Flinders) were included in the dataset. The highly variable regimes of these catchments is challenging for hydrological modelling.

Meteorological data (precipitation, temperature, potential evapotranspiration) were provided to the participants as lumped values at the basin scale (no spatially-distributed data were available). These meteorological data were obtained from different sources, including point data from a single meteorological station, spatially-interpolated data from several meteorological stations, a mix of meteorological stations and meteorological model re-analysis, and even climate reconstruction. Discharge data were provided for the outlet of each basin. All data were at the daily time step, except for the River Rimbaud, for which hourly data were available (the hourly time step is more useful to simulate this basin).

5 CONCLUSIONS

Improving our capacity to work with changing catchments is a necessity for a variety of hydrological applications. Despite the efforts of modellers over the past few years, no joint action had been initiated previously to specifically address this issue on common basin datasets with different hydrological models applied to a common calibration and evaluation

framework. The exercise, conducted during the 2013 IAHS workshop, was an attempt to address this deficiency. With the methodological framework described herein, we aimed to provide the basis for a stimulating and participative workshop, with a full set of criteria, including graphical metrics, specifically designed to easily analyse the stability properties of model simulation over multiple-period tests. This testing and analysis framework was applied to the set of catchments selected for the workshop. We hope that this attempt to rally the community of hydrological modellers around common questions of prediction under change will stimulate new initiatives during the start of the Panta Rhei decade.

Acknowledgements The authors would like to thank Francis Chiew of CSIRO for discussions to define the calibration and evaluation protocol and Laurent Coron (Irstea) for fruitful discussions on this article. We also thank IAHS and its StaHy (ICSH) and surface water (ICSW) commissions for their support in organizing this workshop. Peter Krause and an anonymous reviewer, as well as the Associate Editor Andreas Efstratiadis, are thanked for comments that improved this manuscript and the supplementary material containing the catchment and dataset descriptions.

Disclosure statement No potential conflict of interest was reported by the author(s).

Supplementary material Supplementary material for this article can be accessed at:http://dx.doi.org/10.1080/02626667.2014.967248.

REFERENCES

Abebe, N.A., Ogden, F.L., and Pradhan, N.R., 2010. Sensitivity and uncertainty analysis of the conceptual HBV rainfall–runoff model: implications for parameter estimation. *Journal of Hydrology*, 389 (3–4), 301–310. doi:10.1016/j.jhydrol.2010.06.007.

Andréassian, V., *et al*., 2012. All that glitters is not gold: the case of calibrating hydrological models. *Hydrological Processes*, 26 (14), 2206–2210. doi:10.1002/hyp.9264.

Andreassian, V., *et al*., 2009. HESS opinions 'Crash tests for a standardized evaluation of hydrological models'. *Hydrology and Earth System Sciences*, 13 (10), 1757–1764. doi:10.5194/hess-13-1757-2009.

Beven, K., 1977. Response to General Reporter's comments on papers by Beven, and Beven and Kirkby. *In*: H. J. Morel-Seytoux *et al*. eds. *Surface and subsurface hydrology*. Littleton, CO: Water Resources Publications, 221–222.

Boughton, W., 2005. Catchment water balance modelling in Australia 1960–2004. *Agricultural Water Management*, 71, 91–116. Doi:10.1016/j.agwat.2004.10.012.

Brigode, P., Oudin, L., and Perrin, C., 2013. Hydrological model parameter instability: a source of additional uncertainty in estimating the hydrological impacts of climate change? *Journal of Hydrology*, 476, 410–425. Doi:10.1016/j.jhydrol.2012.11.012.

Coron, L., *et al*., 2011. Pathologies of hydrological models used in changing climatic conditions: a review. *In*: S.W. Franks *et al*., eds. *Hydro-climatology: Variability and change* [online]. Wallingford: International Association of Hydrological Sciences, IAHS Publ. 344, 39–44. Available from: http://iahs.info/uploads/dms/16759.10-39-44-344-19-CoronACCEPTED_JH02-FINAL-WITH-MODIF.pdf [Accessed 28 April 2015].

Coron, L., *et al*., 2012. Crash testing hydrological models in contrasted climate conditions: an experiment on 216 Australian catchments. *Water Resources Research*, 48 (5), doi:10.1029/2011wr011721.

Criss, R.E. and Winston, W.E., 2008. Do Nash values have value? Discussion and alternate proposals. *Hydrological Processes*, 22 (14), 2723–2725. doi:10.1002/hyp.7072.

Donnelly-Makowecki, L.M. and Moore, R.D., 1999. Hierarchical testing of three rainfall-runoff models in small forested catchments. *Journal of Hydrology*, 219 (3–4), 136–152. doi:10.1016/S0022-1694(99)00056-6

Gharari, S., *et al*., 2013. An approach to identify time consistent model parameters: sub-period calibration. *Hydrology and Earth System Sciences*, 17 (1), 149–161. doi:10.5194/hess-17-149-2013

Gupta, H.V., *et al*., 2009. Decomposition of the mean squared error and NSE performance criteria: implications for improving hydrological modelling. *Journal of Hydrology*, 377 (1–2), 80–91. doi:10.1016/j.jhydrol.2009.08.003.

Hrachowitz, M., *et al*., 2013. A decade of Predictions in Ungauged Basins (PUB)—a review. *Hydrological Sciences Journal*, 58 (6), 1198–1255. doi:10.1080/02626667.2013.803183.

Johnston, P.R. and Pilgrim, D.H., 1973. *A study of parameter optimisation for a rainfall–runoff model*. School of Civil Engineering, University of New South Wales, Kensington, 197.

Kirchner, J.W., 2006. Getting the right answers for the right reasons: linking measurements, analyses, and models to advance the science of hydrology. *Water Resources Research*, 42 (3), doi:10.1029/2005WR004362.

Klemeš, V., 1986. Operational testing of hydrological simulation-models. *Hydrological Sciences Journal*, 31 (1), 13–24. doi:10.1080/02626668609491024.

Klemeš, V., 2009 Interactive comment. *Hydrology and Earth System Sciences Discussions* [online]. Available from: http://www.hydrol-earth-syst-sci-discuss.net/6/C1069/2009 [Accessed 9 June 2015].

Kling, H., Fuchs, M., and Paulin, M., 2012. Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios. *Journal of Hydrology*, 424–425, 264–277. doi:10.1016/j.jhydrol.2012.01.011.

Koutsoyiannis, D., 2013. Hydrology and change. *Hydrological Sciences Journal*, 58 (6), 1177–1197. doi:10.1080/02626667.2013.804626

Linsley, R.K., 1982. *Hydrology for engineers (Water resources & environmental engineering)*. ISE Editions. New York, NY: McGraw-Hill Education, 512.

Mathevet, T., *et al*., 2006. A bounded version of the Nash-Sutcliffe criterion for better model assessment on large sets of basins. *In*: Andréassian *et al*., eds. *Large sample basin experiments for hydrological model parameterization: results of the MOdel Parameter EXperiment* [online]. Wallingford: International Association of Hydrological Sciences, IAHS Publ. 307, 211–

219. Available from: http://iahs.info/uploads/dms/13614.21–211-219-41-MATHEVET.pdf [Accessed 28 April 2015].

Merz, R., Parajka, J., and Blöschl, G., 2011. Time stability of catchment model parameters: implications for climate impact analyses. *Water Resources Research*, 47 (2), doi:10.1029/2010WR009505.

Montanari, A., *et al*., 2013. "Panta Rhei—everything flows": change in hydrology and society—the IAHS scientific decade 2013–2022. *Hydrological Sciences Journal*, 58 (6), 1256–1275. doi:10.1080/02626667.2013.809088

Murphy, A.H., 1988. Skill scores based on the mean square error and their relationships to the correlation coefficient. *Monthly Weather Review*, 116 (12), 2417–2424. doi:10.1175/1520-0493(1988)116<2417:SSBOTM>2.0.CO;2

Nalbantis, I., *et al*., 2011. Holistic versus monomeric strategies for hydrological modelling of human-modified hydrosystems. *Hydrology and Earth System Sciences*, 15 (3), 743–758. doi:10.5194/hess-15-743-2011

Nash, J.E. and Sutcliffe, J.V., 1970. River flow forecasting through conceptual models part I—a discussion of principles. *Journal of Hydrology*, 10 (3), 282–290. doi:10.1016/0022-1694(70)90255-6

Peel, M.C. and Blöschl, G., 2011. Hydrological modelling in a changing world. *Progress in Physical Geography*, 35 (2), 249–261. doi:10.1177/0309133311402550

Pushpalatha, R., *et al*., 2012. A review of efficiency criteria suitable for evaluating low-flow simulations. *Journal of Hydrology*, 420–421, 171–182. doi:10.1016/j.jhydrol.2011.11.055.

Razavi, S. and Tolson, B.A., 2013. An efficient framework for hydrologic model calibration on long data periods. *Water Resources Research*, 49 (12), 8418–8431. doi:10.1002/2012WR013442

Refsgaard, J.C. and Knudsen, J., 1996. Operational validation and intercomparison of different types of hydrological models. *Water Resources Research*, 32 (7), 2189–2202. doi:10.1029/96WR00896

Refsgaard, J.C., *et al*., 2013. A framework for testing the ability of models to project climate change and its impacts. *Climatic Change*. 1–12. doi:10.1007/s10584-013-0990-2.

Seibert, J., 2003. Reliability of model predictions outside calibration conditions. *Nordic Hydrology*, 34 (5), 477–492.

Stisen, S., *et al*., 2012. On the importance of appropriate precipitation gauge catch correction for hydrological modelling at mid to high latitudes. *Hydrology and Earth System Sciences*, 16 (11), 4157–4176. doi:10.5194/hess-16-4157-2012

Thirel, G., Andréassian, V., and Perrin, C., 2015. On the need to test hydrological models under changing conditions. Editorial to the Special issue of *Hydrological Sciences Journal*, 60 (7–8). doi:10.1080/02626667.2015.1050027

Vaze, J., *et al*., 2010. Climate non-stationarity—validity of calibrated rainfall–runoff models for use in climate change studies. *Journal of Hydrology*, 394 (3–4), 447–457. doi:10.1016/j.jhydrol.2010.09.018.

Węglarczyk, S., 1998. The interdependence and applicability of some statistical quality measures for hydrological models. *Journal of Hydrology*, 206 (1–2), 98–103. doi:10.1016/S0022-1694(98)00094-8.