Benchmarking hydrological models for low-flow simulation and forecasting on French catchments


To cite this version:


HAL Id: hal-01061927

https://hal-brgm.archives-ouvertes.fr/hal-01061927

Submitted on 8 Sep 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Benchmarking hydrological models for low-flow simulation and forecasting on French catchments

P. Nicolle\textsuperscript{1}, R. Pushpalatha\textsuperscript{1}, C. Perrin\textsuperscript{1}, D. François\textsuperscript{2}, D. Thiéry\textsuperscript{3}, T. Mathevét\textsuperscript{4}, M. Le Lay\textsuperscript{4}, F. Besson\textsuperscript{5}, J.-M. Soubeyroux\textsuperscript{5}, C. Viet\textsuperscript{5}, F. Regimbeau\textsuperscript{5}, V. Andréassian\textsuperscript{1}, P. Maugis\textsuperscript{6}, B. Augeard\textsuperscript{7}, and E. Morice\textsuperscript{8}

\textsuperscript{1}Irstea, HBAN, Antony, France
\textsuperscript{2}Université de Lorraine, LOTERR, Metz, France
\textsuperscript{3}BRGM, Orléans, France
\textsuperscript{4}EDF-DTG, Grenoble, France
\textsuperscript{5}Météo-France, Direction de la Climatologie, Toulouse, France
\textsuperscript{6}IPSL, LSCE, Gif-sur-Yvette, France
\textsuperscript{7}ONEMA, Vincennes, France
\textsuperscript{8}Direction de l’eau et de la Biodiversité, Ministère de l’écologie, du développement durable et de l’énergie, La Défense, France

Correspondence to: P. Nicolle (pierre.nicolle@irstea.fr)

Received: 28 October 2013 – Published in Hydrol. Earth Syst. Sci. Discuss.: 15 November 2013
Revised: 12 June 2014 – Accepted: 16 June 2014 – Published: 5 August 2014

Abstract. Low-flow simulation and forecasting remains a difficult issue for hydrological modellers, and intercomparisons can be extremely instructive for assessing existing low-flow prediction models and for developing more efficient operational tools. This research presents the results of a collaborative experiment conducted to compare low-flow simulation and forecasting models on 21 unregulated catchments in France. Five hydrological models (four lumped storage-type models – Gardenia, GR6J, Mordor and Presages – and one distributed physically oriented model – SIM) were applied within a common evaluation framework and assessed using a common set of criteria. Two simple benchmarks describing the average streamflow variability were used to set minimum levels of acceptability for model performance in simulation and forecasting modes. Results showed that, in simulation as well as in forecasting modes, all hydrological models performed almost systematically better than the benchmarks. Although no single model outperformed all the others for all catchments and criteria, a few models appeared to be more satisfactory than the others on average. In simulation mode, all attempts to relate model efficiency to catchment or streamflow characteristics remained inconclusive. In forecasting mode, we defined maximum useful forecasting lead times beyond which the model does not bring useful information compared to the benchmark. This maximum useful lead time logically varies between catchments, but also depends on the model used. Simple multi-model approaches that combine the outputs of the five hydrological models were tested to improve simulation and forecasting efficiency. We found that the multi-model approach was more robust and could provide better performance than individual models on average.

1 Introduction

1.1 Why anticipate low flows?

In many countries, rivers are the primary supply of water. In France, where this research was conducted, 81% of the 33 km$^3$ of total water withdrawals in 2009 came from rivers (Commissariat général au développement durable, CGDD, 2012). Municipal water supply, irrigation, navigation, hydropower and thermal power plant cooling are highly dependent on streamflow and can be strongly affected by water shortages in rivers (Bousquet et al., 2003). Increasing efforts to maintain minimum environmental flows in rivers make the...
Early anticipation of low-flow periods is needed to improve water management and take more timely measures to mitigate the socio-economic and ecological impacts of water shortages (Chiw and McMahon, 2002; Hamlet et al., 2002; Karamouz and Araghinejad, 2008). Extreme droughts, which occurred in Western Europe in 1921 (Duband et al., 2004), 1949 (Duband, 2010), 1976 (Brochet, 1978; Gazelle, 1979) and more recently in 2003 (Moreau, 2004; Vidal et al., 2010b), underline the need for anticipation systems. In addition, the current trend and/or perspective of more severe summer low flows in the context of climate change further highlights the need for appropriate management tools for low flows (Svensson et al., 2005; Manoha et al., 2008; Feyen and Dankers, 2009).

In spite of early attempts to develop models for applications on low flows (Riggs, 1953; Bernier, 1964; Popov, 1964; Singh and Stall, 1971; Laras, 1972; Oberlin and Michel, 1978), low-flow forecasting has received only limited attention in the literature compared to flood forecasting (see e.g. reviews by Cloke and Pappenberger, 2009; Hapuarachchi et al., 2011). Although quite similar in essence, the two exercises have marked differences, essentially due to the different dynamics of floods and low flows. Indeed, low flows are long-lasting phenomena with slow dynamics, contrary to floods. Besides, expectations are different in terms of forecast lead times, which are longer in the case of low flows, typically ranging from a few days to a few weeks. Therefore there is a need to assess the ability of existing forecasting tools to anticipate low-flow situations both in terms of magnitude and lead time.

### 1.2 Hydrological models for low-flow forecasting

Most models proposed for low-flow forecasting can be considered as hydrological models, in the sense that they try to simulate the catchment response to given meteorological conditions. A few of them also try to incorporate upstream information, e.g. dam operations. Early modelling attempts include linear Autoregressive Moving Average (ARMA) type models, propagation models and recession curves (Lefèvre, 1974; Yates and Snyder, 1975; Avalos Lingan, 1976; Guilbot et al., 1976; Girard, 1977; Oberlin and Michel, 1978; Miquel and Roche, 1985; Rivera-Ramirez et al., 2002). Data-driven approaches like neural networks and conceptual rainfall-runoff models are also more and more widely used (Campolo et al., 1999; Garçon et al., 1999; Stravs and Brilly, 2007). Some of these models make simplifying assumptions, e.g. hypothesizing no-rainfall future conditions in the case of recession models. This is the most pessimistic case in a low-flow forecasting context, but often a not entirely realistic one when lead times of a few weeks are considered.

To make more reliable forecasts and extend to longer lead times, it is necessary to account for future meteorological conditions (e.g. Coulibaly, 2003). To account for the uncertainty in the future conditions (mainly in terms of temperature and precipitation), the typical methodology consists in simulating an ensemble of low-flow forecasts (similar to ensemble flood forecasts), using a hydrological model fed by an ensemble of meteorological scenarios. These forecasts are then statistically analysed for the target time period (see e.g. Garçon et al., 1999; Perrin et al., 2001; Demirel et al., 2013a).

### 1.3 Limits of existing tools

Low-flow forecasting with hydrological models is actually a difficult task since processes conditioning low flows may depend on the region, season or lead time. For example, Demirel et al. (2013b) investigated the role of five indicators (precipitation, potential evapotranspiration, groundwater storage, snow storage and lake storage) on the Rhine basin low flows and found that their relative magnitude varies with the forecast lead time. Singla et al. (2012) also showed that the predictability of flows in the spring season strongly depends on snow cover in the mountainous regions. The relation between surface water and groundwater in low-flow conditions was also investigated by many authors, showing the need to account for this in low-flow forecasting models (Tajjar, 1993; Pointet et al., 2003; Rassam, 2011). Clearly, the applicability of hydrological models for low-flow forecasting depends on the way these various processes are accounted for in the model. For example, the work of Staudinger et al. (2011) illustrates the sensitivity of summer low-flow simulation to the formulation of the model structure. A number of techniques can be used in conjunction with a hydrological model to improve its forecasting efficiency and decrease modelling uncertainty. Assimilation of observed data (e.g. observed streamflow or soil moisture) available at the time the forecast is issued may be one option. Using post-processing techniques to correct the bias or the spread of model outputs may also prove useful (see e.g. the discussion by Demirel et al., 2013a), as well as multi-model approaches (Georgakakos et al., 2004; Velázquez et al., 2011).

Our literature review showed that there are very few studies comparing the performance of existing hydrological models so that it is difficult to know their respective strengths and weaknesses in a low-flow forecasting perspective. A noteworthy exception is the study by Demirel et al. (2013a), who compared the HBV and GR4J models and found that the former provides better forecasts than the latter. These authors also indicate that parameter estimation is a major source of uncertainty for medium-range (10 days ahead) low-flow forecasts.
1.4 Scope of the paper

Given this lack of common evaluation of low-flow forecasting models and the need to provide end-users with advanced forecasting tools, the French national agency for water and aquatic environments (ONEMA), and the Ministry for Ecology (MEDDE) jointly launched in 2010 a comparative study for evaluating existing operational (or pre-operational) low-flow forecasting models on basins covering a variety of French hydroclimatic contexts. The project, called PREMHYCE, was designed as an open experiment. Each participant was invited to follow a single testing protocol to run his own model on a common database set up for the project. Since the experience of the modeller may play a role in the quality of the model’s implementation, this placed the models in the best conditions for obtaining optimal results. The test set intentionally included a wide variety of conditions to draw more general conclusions (Andréassian et al., 2009; Gupta et al., 2014). Although the project was restricted to the French context and limited to French participants for practical reasons, the results are likely to be of wider interest for the community of researchers and managers working on these issues. The project mainly intended to identify the respective advantages of the models on the selected catchments for low-flow simulation and forecasting objectives. Here, following the definitions given by Beven and Young (2013), the simulation is understood as “the quantitative reproduction of the catchment behaviour, given defined inputs but without reference to any observed outputs,” whereas forecasting is “the quantitative reproduction of the catchment behaviour ahead of time, but given observations of the inputs, state variables (where applicable), and outputs up to the present time (the forecasting starting point).” As forecast inputs are likely the most important source of uncertainties in streamflow forecasting, it seems important to first analyse hydrological models in simulation mode to better understand their performance differences.

The aim of this paper is to present the main outcomes of the PREMHYCE project. In the next section, we present the catchments and data used for this research, the tested models and an overview of the testing protocol, including evaluation criteria. Section 3 details the main results obtained on the catchment set in simulation and forecasting modes and analyses the differences between models. Section 4 opens the discussion on three questions, namely (1) within a set of models, is a better low-flow simulation model also a better forecasting model? (2) Which maximum lead time can be expected in low-flow forecasting? (3) Can models be efficiently combined in a multi-model approach? The last section presents the main lessons and perspectives of this work.

Figure 1. Location of the 21 selected catchments in France. Each outlet is shown by a dot and referred to with the HYDRO code (see details in Table 1).

2 Material and methods

The approach followed in the PREMHYCE project was largely inspired by modelling experiments carried out in the past few years, in which participants had been invited to run their models on a common data set. WMO (1975, 1986, 1992) was among the first to organise such experiments to evaluate model running for simulation, snowmelt or flood forecasting purposes. More recently, the Distributed Model Intercomparison Project (DMIP) experiments (Smith et al., 2004, 2012) carried out by the NOAA in the USA to evaluate distributed simulation models provide excellent examples of testing protocols. However, to our knowledge, none of these experiments were designed to evaluate models for a low-flow forecasting objective. Therefore, we built our own common testing protocol to evaluate the relative efficiency of several models currently used in France in operational or pre-operational conditions.

2.1 Catchment set and data

2.1.1 Selection of catchments

A set of 21 catchments distributed over continental France was built to serve as the test bed. The catchments were selected based on several criteria. We intended to have (1) a wide diversity of physical and climate conditions representative of the diversity of conditions found in France; (2) sufficiently long time series from gauging stations that include a variety of low-flow events, with data deemed to be good quality by the operational hydrometric services and with
Table 1. Summary of the main characteristics of the 21 selected catchments.

<table>
<thead>
<tr>
<th>No.</th>
<th>HYDRO code</th>
<th>River at station</th>
<th>Area (km²)</th>
<th>Median elevation (m)</th>
<th>Starting date for flow series</th>
<th>Ending date for flow series</th>
<th>Flow availability (years)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A1080330</td>
<td>Ill at Didenheim</td>
<td>657</td>
<td>390</td>
<td>1 Nov 1973</td>
<td>2 Mar 2010</td>
<td>36</td>
</tr>
<tr>
<td>2</td>
<td>B2220010</td>
<td>Meuse at Saint-Mihiel</td>
<td>2542</td>
<td>350</td>
<td>1 Jul 1968</td>
<td>3 Jan 2010</td>
<td>42</td>
</tr>
<tr>
<td>3</td>
<td>H2342020</td>
<td>Serein at Chablis</td>
<td>1121</td>
<td>309</td>
<td>1 Aug 1954</td>
<td>3 Mar 2010</td>
<td>56</td>
</tr>
<tr>
<td>4</td>
<td>H4252010</td>
<td>Orge at Morsang-sur-Orge</td>
<td>927</td>
<td>133</td>
<td>1 Oct 1967</td>
<td>7 Mar 2010</td>
<td>43</td>
</tr>
<tr>
<td>5</td>
<td>H7401010</td>
<td>Oise at Semigny</td>
<td>4316</td>
<td>137</td>
<td>1 Jan 1955</td>
<td>2 Mar 2010</td>
<td>55</td>
</tr>
<tr>
<td>6</td>
<td>H8212010</td>
<td>Andelle at Vascœuil</td>
<td>379</td>
<td>159</td>
<td>1 Jan 1973</td>
<td>27 Feb 2010</td>
<td>36</td>
</tr>
<tr>
<td>7</td>
<td>I5221010</td>
<td>Vire at Saint-Lô</td>
<td>868</td>
<td>159</td>
<td>1 Jan 1971</td>
<td>3 Feb 2010</td>
<td>39</td>
</tr>
<tr>
<td>8</td>
<td>J7483010</td>
<td>Seiche at Bruz</td>
<td>811</td>
<td>70</td>
<td>1 Dec 1967</td>
<td>11 Mar 2010</td>
<td>42</td>
</tr>
<tr>
<td>9</td>
<td>K1321810</td>
<td>Arroux at Etang-sur-Arroux</td>
<td>1798</td>
<td>431</td>
<td>1 Jan 1971</td>
<td>28 Mar 2010</td>
<td>39</td>
</tr>
<tr>
<td>10</td>
<td>K6402520</td>
<td>Sauldres at Salbris</td>
<td>1200</td>
<td>70</td>
<td>1 Jan 1971</td>
<td>28 Mar 2010</td>
<td>39</td>
</tr>
<tr>
<td>11</td>
<td>L0563010</td>
<td>Briance at Condat-sur-Vienne</td>
<td>597</td>
<td>386</td>
<td>1 Jan 1966</td>
<td>28 Mar 2010</td>
<td>44</td>
</tr>
<tr>
<td>12</td>
<td>L4411710</td>
<td>Petite Creuse at Fresselines</td>
<td>850</td>
<td>393</td>
<td>1 Jan 1958</td>
<td>28 Mar 2010</td>
<td>52</td>
</tr>
<tr>
<td>13</td>
<td>M0243010</td>
<td>Orne Saosnoise at Montbizon</td>
<td>510</td>
<td>103</td>
<td>1 Dec 1967</td>
<td>4 Mar 2010</td>
<td>43</td>
</tr>
<tr>
<td>14</td>
<td>M7112410</td>
<td>Sèvre Nantaise at Tiffauges</td>
<td>817</td>
<td>170</td>
<td>1 Nov 1967</td>
<td>4 Mar 2010</td>
<td>43</td>
</tr>
<tr>
<td>15</td>
<td>O0592510</td>
<td>Salat at Roquefort-sur-Garonne</td>
<td>1570</td>
<td>986</td>
<td>1 Jan 1913</td>
<td>22 Mar 2010</td>
<td>97</td>
</tr>
<tr>
<td>16</td>
<td>O3121010</td>
<td>Tarn at Montbrun</td>
<td>588</td>
<td>1020</td>
<td>1 Jan 1961</td>
<td>31 Dec 2009</td>
<td>38</td>
</tr>
<tr>
<td>17</td>
<td>Q5501010</td>
<td>Gave de Pau at Berenx</td>
<td>2575</td>
<td>916</td>
<td>1 Jul 1923</td>
<td>28 Mar 2010</td>
<td>87</td>
</tr>
<tr>
<td>18</td>
<td>S2242510</td>
<td>Eyre at Salies</td>
<td>1650</td>
<td>78</td>
<td>1 Jan 1967</td>
<td>19 Mar 2010</td>
<td>43</td>
</tr>
<tr>
<td>19</td>
<td>U4644010</td>
<td>Azergues at Lozanne</td>
<td>798</td>
<td>517</td>
<td>1 Jan 1965</td>
<td>28 Mar 2010</td>
<td>43</td>
</tr>
<tr>
<td>20</td>
<td>V4264010</td>
<td>Drôme at Saillans</td>
<td>936</td>
<td>936</td>
<td>1 Jan 1910</td>
<td>28 Mar 2010</td>
<td>46</td>
</tr>
<tr>
<td>21</td>
<td>Y4624010</td>
<td>Gapeau at Hyères</td>
<td>517</td>
<td>316</td>
<td>1 Feb 1961</td>
<td>1 Mar 2010</td>
<td>49</td>
</tr>
</tbody>
</table>

of low-flow situations to obtain robust and significant model evaluations (see also Giuntoli et al., 2013, for a discussion on low-flow thresholds).

Table 2 illustrates the range of low-flow thresholds and low-flow conditions on the catchment set, using two descriptors, namely the base-flow index (BFI) and the $Q_{90}/Q_{50}$ ratio (where $Q_{90}$ and $Q_{50}$ are the 90th and 50th percentiles of the flow duration curve, respectively). BFI represents the part of base-flow in the total flow volume (Lvovitch, 1972). Low BFI values indicate a catchment with a flashy flow regime and limited groundwater contribution, while high values are an indication of large storage capacity and groundwater-fed rivers (Gustard and Demuth, 2009). The catchment set examined provides a wide range of BFI values, ranging from 11.7 to 93.5%. The $Q_{90}/Q_{50}$ ratio represents the difference between low flows and medium flows, thus indicating the severity of low flows. It shows a similar variability, with values between 7 and 67% and half of the catchments set between 18 and 38%.

2.2 Models

Before presenting the models used in this work in detail, we found it useful to reiterate here the context of their developments in France, to show that these models are the results of an already quite long experience on low-flow simulation and forecasting within the hydrological modelling community.

2.2.1 Modelling background

Among the first attempts to use conceptual hydrological models in France for low-flow forecasting, CTGREF (Centre technique du génie rural des eaux et des forêts, 1977) developed a simple storage-type model on the Durance basin to improve irrigation water management in low-flow conditions. Then a few hydrological models were developed to better take into account low-flow dynamics and are now used in operational conditions. The French Geological Survey (BRGM) first worked on aquifer level forecasts (Thiéry, 1982, 1988b). Subsequently, Thiéry (1988a) reported the application of a conceptual model to forecast low flows on four catchments with various characteristics in France. These studies yielded the hydrological model GARDENIA, which is now used in operational conditions (Thiéry, 2013). EDF, the French national electricity company, was also active in the development of operational tools and they implemented a forecasting system based on a hydrological model (MORDOR) in the 1990s to better manage the reservoirs in the Durance River basin (Garçon, 1996; Garçon et al., 1999). This system was later extended to other river basins in the mountainous regions where EDF manages reservoirs, including the Loire River basin (Mathevet et al., 2010). Using similar methods, Perrin et al. (2001), Staub (2008) and Pushpalatha (2013) evaluated the performance of the GR4J model (or modified version of this model, see Pushpalatha et al., 2011) for low-flow forecasting on a large set of French catchments. Lang et al. (2006a, b) also developed a platform for low-flow analysis and forecasting based on a conceptual hydrological model and implemented it in northeastern France (Meuse, Moselle and Rhine basins). Last, Soubeyroux et al. (2010) discussed the implementation of tools developed by Météo-France for long-term forecasting, especially using the Safran-Isba-Modec (SIM) modelling suite running throughout France in operational conditions.

2.2.2 Selected models

Table 3 shows the five models used in this study. Four of them (called here GARD, GR6J, MORD and PRES) are lumped storage-type models, with various conceptualisations of the rainfall-runoff transformation. The fifth model (SIM) is distributed and more physically oriented. These models have all already been applied in various conditions in France. SIM is implemented throughout France, and the other models were tested in various basins or regions for different purposes (e.g. low-flow or flood simulation and forecasting). The simulation of low flows in these models is governed by different stores and functions. In forecasting mode, the models use assimilation schemes and/or statistical correction procedures (see Table 3).

The models include different numbers of free parameters (Table 3). Participant were free to choose the optimisation method best suited to parameter estimation, but all opted for automatic calibration, using either global (SCE-UA (Shuffled Complex Evolution method – University of Arizona) method for MORD, multistart simplex method for PRES) or local (gradient-type “step-by-step” method for GR6J, Rosenbrock method for GARD) optimisation algorithms (Table 3). The objective functions were generally chosen to put more weight on low-flow (e.g. Nash and Sutcliffe, 1970 (NS) criterion calculated on transformed streamflow ($Q^{0.2}$) for PRES, Root Mean Square Error (RMSE) calculated with $\ln(Q)$ for GARD, or mean of Kling–Gupta efficiency (KGE, Gupta et al., 2009) criteria calculated on $Q$ (KGE) and $1/Q$ (KGE$_{v}$) for MORD and GR6J, see Table 3). Even though this variety of choices may make the comparison of results less straightforward, this was a means to account for the variety of modelling approaches and for the experience of model developers. Note that SIM was the only model for which no calibration against observed flow data at the catchment outlet was performed. The spatially distributed parameters used in this model were estimated regionally. This should be kept in mind when interpreting the results. Moreover, this version of SIM includes a detailed simulation of the aquifers only on a few parts of France (Seine and Rhône catchments). This may impact the efficiency of the model outside these zones. Moreover, the larger computing requirements of SIM only allowed a limited number of tests (see Sect. 2.3.3).

The models were fed with the same meteorological inputs derived from SAFRAN. For the lumped models, the
Table 2. Percentiles of the distribution of a few climatic and hydrological characteristics of the 21 selected catchments. Interannual variability values correspond to coefficients of variation calculated on the 1974–2009 period. $Q_{50}$, $Q_{80}$ and $Q_{90}$ are respectively the 50th, 80th and 90th exceedance percentiles of the flow duration curve.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Min</th>
<th>25%</th>
<th>Median</th>
<th>75%</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean annual precipitation $P_A$ (mm)</td>
<td>656</td>
<td>842</td>
<td>931</td>
<td>1039</td>
<td>1400</td>
</tr>
<tr>
<td>Interannual variability of $P_A$</td>
<td>0.13</td>
<td>0.15</td>
<td>0.17</td>
<td>0.17</td>
<td>0.26</td>
</tr>
<tr>
<td>Mean annual potential evapotranspiration $PE_A$ (mm)</td>
<td>606</td>
<td>683</td>
<td>698</td>
<td>717</td>
<td>1031</td>
</tr>
<tr>
<td>Interannual variability of $PE_A$</td>
<td>0.05</td>
<td>0.06</td>
<td>0.08</td>
<td>0.09</td>
<td>0.11</td>
</tr>
<tr>
<td>Mean annual streamflow $Q_A$ (mm year$^{-1}$)</td>
<td>135</td>
<td>255</td>
<td>325</td>
<td>437</td>
<td>1033</td>
</tr>
<tr>
<td>Interannual variability of $Q_A$</td>
<td>0.23</td>
<td>0.28</td>
<td>0.33</td>
<td>0.38</td>
<td>0.62</td>
</tr>
<tr>
<td>Runoff ratio $Q_A/P_A$ (%)</td>
<td>21</td>
<td>31</td>
<td>37</td>
<td>41</td>
<td>76</td>
</tr>
<tr>
<td>Base-flow index (BFI) (%)</td>
<td>11.7</td>
<td>35</td>
<td>45.3</td>
<td>51.1</td>
<td>93.5</td>
</tr>
<tr>
<td>$Q_{90}/Q_{50}$ (%)</td>
<td>7</td>
<td>18</td>
<td>28</td>
<td>38</td>
<td>67</td>
</tr>
<tr>
<td>$Q_{80}$ (mm day$^{-1}$)</td>
<td>0.03</td>
<td>0.13</td>
<td>0.19</td>
<td>0.31</td>
<td>1.21</td>
</tr>
</tbody>
</table>

Figure 2. Schematic representation of the difference between (a) simulation and (b) forecasting modes ($L$: lead time).

SAFRAN variables were first aggregated at the catchment scale by simple averaging.

2.3 Testing protocol and evaluation methodology

A common testing and evaluation framework was set up to make the results comparable. It was jointly elaborated by all project participants in the first phase of the project, so that most of the models’ requirements and constraints could be accounted for.

2.3.1 Testing scheme

Model evaluation was based on a classical split-sample test approach (Klemeš, 1986). Streamflow records were divided into two approximately equal sub-periods. Each period was alternately used for calibration and validation, i.e. calibration on period 1 (noted C1) with validation on period 2 (V2), and then calibration on period 2 (C2) with validation on period 1 (V1). Thus the models could be evaluated in validation on all available data. The 1974–1991 and 1992–2009 periods based on calendar years were chosen for periods 1 and 2, respectively. A 3-year warm-up period was used at the beginning of each test period (1971–1973 and 1989–1991 for periods 1 and 2, respectively) to initialise the internal states of the models.

2.3.2 Differences between forecast and simulation tests

As underlined above, the simulation and forecasting exercises differ, which has clear implications in the way models were tested here (see illustration in Fig. 2).

In simulation mode, models are expected to simulate streamflow at time step $t$, knowing observed meteorological inputs until this time step. Observed streamflow values remain unknown at all time steps. The simulation mode shows the models’ ability to reproduce the catchments’ hydrological behaviour without uncertainties due to unknown future
Table 3. Overview of the characteristics of the five models tested.

<table>
<thead>
<tr>
<th>Short name used here</th>
<th>GARD</th>
<th>GR6J</th>
<th>MORD</th>
<th>PRES</th>
<th>SIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full name</td>
<td>GARDENIA</td>
<td>GR6J</td>
<td>MORDOR</td>
<td>PRESAGES</td>
<td>SIM</td>
</tr>
<tr>
<td>Type</td>
<td>Conceptual</td>
<td>Conceptual</td>
<td>Conceptual</td>
<td>Conceptual</td>
<td>Physically based</td>
</tr>
<tr>
<td>Spatial distribution</td>
<td>Semi-distributed</td>
<td>Lumped</td>
<td>Lumped</td>
<td>Lumped</td>
<td>Distributed</td>
</tr>
<tr>
<td>Number of free parameters</td>
<td>4 to 9 (+2 to 4 for snowmelt)</td>
<td>6 (+2; snow routine)</td>
<td>11 (+4; snow routine)</td>
<td>7 (+3; snow routine)</td>
<td>0</td>
</tr>
<tr>
<td>Calibration criteria</td>
<td>RMSE with ln(Q)</td>
<td>(KGE + KGE_i)^2</td>
<td>(KGE + KGE_i)^2</td>
<td>Nash–Sutcliffe with Q^0.2</td>
<td></td>
</tr>
<tr>
<td>Post-correction method (simulation)</td>
<td>Not used</td>
<td>Not used</td>
<td>Not used</td>
<td>Empirical method (Berthier, 2005)</td>
<td>Quantile/quantile post-treatment</td>
</tr>
<tr>
<td>Assimilation method (forecast)</td>
<td>When a flow discrepancy appears, the model tanks are updated proportionally according to their variance</td>
<td>Correction based on error at first time step before forecast, with decreasing effect when lead time increases</td>
<td>Correction based on errors at previous time steps before forecast, with decreasing effect when lead time increases. No update of model stores.</td>
<td>Update of gravitational routing store</td>
<td>No assimilation method but a quantile/quantile post-treatment</td>
</tr>
<tr>
<td>Structure overview: production</td>
<td>Actual evapotranspiration is computed using a non-linear soil capacity. GW (Groundwater) exchange is a proportion of the GW flow</td>
<td>A rainfall interception by PE, a non-linear soil moisture accounting store, an interception GW exchange function</td>
<td>A rainfall excess/soil moisture accounting store; an evaporating reservoir; an intermediate store and a deep store</td>
<td>A soil store, rainfall interception by PE</td>
<td></td>
</tr>
<tr>
<td>Structure overview: transfer</td>
<td>A non-linear tank distributes the effective rainfall into runoff and GW recharge. The aquifer is represented by a linear tank.</td>
<td>Two unit hydrographs, two parallel nonlinear routing stores</td>
<td>Direct, indirect and base flow components are routed using a unit hydrograph (Weibull law)</td>
<td>Two unit hydrographs, two linear routing stores: one for streamflow recession, one for interflow</td>
<td></td>
</tr>
<tr>
<td>References on simulation applications in France</td>
<td>800 to 1000 rivers simulated in France</td>
<td>Pushpalatha et al. (2011)</td>
<td>Garavaglia (2011); Paquet et al. (2013)</td>
<td>Lang et al. (2006a, b)</td>
<td>Vidal et al. (2010b)</td>
</tr>
</tbody>
</table>
conditions (input scenarios) and without the information contributed by external data (typically observed flows) that could be assimilated to adjust the model.

In forecasting mode, models are expected to forecast streamflow from time steps $t + 1$ to $t + L$ (with $L$ the lead time), knowing both observed meteorological inputs and streamflow until time step $t$ and making assumptions (i.e. choosing scenarios) for the future meteorological inputs from $t + 1$ to $t + L$. Streamflow data can be used within an assimilation scheme and/or a statistical correction procedure. Models were actually tested in hindcasting mode, i.e. retrospectively running the models at each time step of the available test periods and making forecasts as if models were used in real time.

2.3.3 Choice of scenarios in forecasting mode

An ensemble of scenarios of future meteorological inputs must be chosen for the forecasting mode. Usually, real-time ensemble forecasts from meteorological models are used to forecast streamflow. Here, since no long-term archive of actual forecasts was available over the test period, the meteorological archive was used as possible scenarios for the future meteorological inputs from $t + 1$ to $t + L$. Streamflow data can be used within an assimilation scheme and/or a statistical correction procedure. Models were actually tested in hindcasting mode, i.e. retrospectively running the models at each time step of the available test periods and making forecasts as if models were used in real time.

Although models provided streamflow simulations or forecasts at a daily time step, we chose to evaluate models on the streamflow averaged over a 3-day sliding window. This aimed at smoothing the low-flow series and avoiding putting too much emphasis on isolated streamflow variations (Henny, 2010). Note that this target variable is quite commonly used in France for regulation purposes.

Since long archives of ensemble meteorological forecasts from an ensemble prediction system (EPS) were not available for this study, using long archives of observed meteorological data gave the advantage to get general results and also included severe drought conditions observed in the past decades. Moreover, the targeted lead time in the study is up to a few weeks, i.e. longer than medium-range forecasts of about two weeks which are currently available. Extending medium-range forecasts with other information (i.e. climatic series) was out of the scope of this study. Note that we did not investigate here seasonal forecasting, with typical forecast horizons of several months (see e.g. Céron et al., 2010; Singla et al., 2012).

2.3.4 Benchmarks and evaluation criteria

Although models provided streamflow simulations or forecasts at a daily time step, we chose to evaluate models on the streamflow averaged over a 3-day sliding window. This aimed at smoothing the low-flow series and avoiding putting too much emphasis on isolated streamflow variations (Henny, 2010). Note that this target variable is quite commonly used in France for regulation purposes.

Since the use of benchmarks is important to evaluate the relative advantages of model predictions (Seibert, 2001; Perrin et al., 2006), results in simulation mode were compared to the daily average streamflow curve (DAQ). This benchmark was advocated by Martinec and Rango (1989). In forecasting mode, the probabilistic forecasts were compared to a benchmark describing the streamflow natural variability (NVQ). NVQ is defined for a given calendar day $d$ of year $Y$ as the distribution of available streamflows in the other years for this day. Obviously, more demanding benchmarks could have been chosen to raise the level of expected performance. For example, in forecasting mode, one may use a constrained version of NVQ by selecting the years for which flow at the day of forecast lies in similar ranges as the observed flow for the current year. Here NVQ benchmark has been chosen to keep a more uniform evaluation among years. Note that the choice of the benchmark may change interpretations when comparing the models with the benchmark (see e.g. Sect. 4.2) but it will not impact the evaluation of their respective merits when placed in a comparative framework.
### Table 4. List of efficiency criteria used for model evaluation in simulation mode (see details in Appendix).

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Quadratic criteria</strong></td>
<td></td>
</tr>
<tr>
<td>KGE</td>
<td>Kling-Gupta efficiency</td>
</tr>
<tr>
<td>C2M</td>
<td>Nash–Sutcliffe efficiency bounded in $[-1; 1]$</td>
</tr>
<tr>
<td><strong>Low-flow quadratic criteria</strong></td>
<td></td>
</tr>
<tr>
<td>C2M_{\text{LFD}}</td>
<td>Nash–Sutcliffe efficiency calculated with $1/Q$ and bounded in $[-1; 1]$</td>
</tr>
<tr>
<td>RMSE\text{\footnotesize{ref}}</td>
<td>Root mean square error calculated when observed streamflow is less than $Q_{80}$ threshold</td>
</tr>
<tr>
<td><strong>Volume-based and temporal criteria</strong></td>
<td></td>
</tr>
<tr>
<td>Vdef</td>
<td>Ratio of observed and simulated cumulative annual volume deficits</td>
</tr>
<tr>
<td>LFD</td>
<td>Ratio of observed and simulated cumulative low-flow duration</td>
</tr>
<tr>
<td>DatSt</td>
<td>Relative difference between observed and simulated start of annual low-flow period</td>
</tr>
<tr>
<td>DatEn</td>
<td>Relative difference between observed and simulated end of annual low-flow period</td>
</tr>
<tr>
<td><strong>Threshold criteria</strong></td>
<td></td>
</tr>
<tr>
<td>POD</td>
<td>Probability of detection, based on contingency table</td>
</tr>
<tr>
<td>FAR</td>
<td>False alarm rate, based on contingency table</td>
</tr>
<tr>
<td>CSI</td>
<td>Critical success index, based on contingency table</td>
</tr>
</tbody>
</table>

### Table 5. List of efficiency criteria used for model evaluation in forecasting mode (see details in Appendix).

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Continuous low-flow quadratic and probabilistic criteria</strong></td>
<td></td>
</tr>
<tr>
<td>RMSE\text{\footnotesize{ref}}</td>
<td>Root mean square error calculated when observed streamflow is less than $Q_{80}$ threshold</td>
</tr>
<tr>
<td>DRPS</td>
<td>Discrete ranked probability score</td>
</tr>
<tr>
<td><strong>Volume-based and temporal criteria</strong></td>
<td></td>
</tr>
<tr>
<td>Vdef</td>
<td>Ratio of observed and simulated cumulative annual volume deficits</td>
</tr>
<tr>
<td>LFD</td>
<td>Ratio of observed and simulated cumulative low-flow duration</td>
</tr>
<tr>
<td><strong>Sharpness/reliability</strong></td>
<td></td>
</tr>
<tr>
<td>Sharp</td>
<td>Mean width of interval defined by 10 and 90 % percentiles of forecast distribution when observed streamflow is less than $Q_{80}$ threshold</td>
</tr>
<tr>
<td>Cont\text{\footnotesize{ratio}}</td>
<td>Percentage of observation in the 80 % forecasted confidence interval when observed streamflow is less than $Q_{80}$ threshold (80 % of observed streamflow should be included in the interval)</td>
</tr>
<tr>
<td><strong>Threshold criteria</strong></td>
<td></td>
</tr>
<tr>
<td>POD</td>
<td>Probability of detection, based on contingency table</td>
</tr>
<tr>
<td>FAR</td>
<td>False alarm rate, based on contingency table</td>
</tr>
<tr>
<td>CSI</td>
<td>Critical success index, based on contingency table</td>
</tr>
<tr>
<td>BS\text{\footnotesize{vig}}, BS\text{\footnotesize{cri}}</td>
<td>Brier score with vigilance threshold ($Q_{80}$) or crisis threshold ($Q_{95}$)</td>
</tr>
</tbody>
</table>

We used two sets of evaluation criteria for model evaluation in simulation (see list in Table 4) and forecasting (see Table 5) modes. They were chosen to assess various modelling skills expected in low-flow conditions for different objectives, after discussions with stakeholders. The detailed mathematical formulation of the criteria is given in the Appendix.

In forecasting mode, the models were expected to produce forecasts over a future time window of 90 days. Therefore, model forecasting performance could be investigated for all lead times between 1 and 90 days. To simplify the presentation of results, we chose to focus on two specific lead times: a short one (7 days) and a longer one (30 days). This choice was made in agreement with stakeholders since those are the typical horizons useful for water managers. The longer lead time was limited to 30 days given the computation constraints of the SIM model.
In some cases, the mathematical form of the criteria was changed to have all of them vary within the $]-\infty; 1]$ interval (1 being the optimum value) to ease interpretation. Note that the forecasting results presented hereafter were analysed only on the time steps where streamflow forecasts from SIM were available.

### 2.3.5 Presentation of results

The project produced a very large number of results, and it is obviously not possible to elaborate them all here. Instead, we chose to present summary evaluations using tables and graphical representations. Radial plots, as exemplified in Fig. 3, were used to present mean model performance on the set of 21 catchments for all selected criteria. Visually, the larger the polygon linking the performance values, the better the model. On these graphs, criteria focusing on similar aspects were grouped together (the groups are those defined in Tables 4 and 5). We also used performance maps to investigate the possible regional trend in results. These maps were drawn for three criteria only (C2M, CSI and Vdef in simulation; RMSECan, BSvig and Vdef in forecasting). They were found to be complementary, thus providing an overall picture of model performance in low-flow conditions.

### 3 Results

#### 3.1 Simulation mode

Figure 4 summarises the mean performance obtained by the five models tested in validation on the 21 catchments and the two test periods. Quite similar results can be observed for four lumped models on average. The performance of the SIM model was lower for a few criteria (C2M, C2M, POD, FAR and CSI). However, no model seemed able to outperform all the other models for all criteria.

Performance on some criteria can vary substantially between catchments. Figure 5 presents the maps of mean performance on the two validation periods for three criteria (C2M, Vdef and CSI). A few catchments (e.g. the Meuse at St-Mihiel) are properly simulated by more or less all models. However, performance can be much more variable between models on other catchments: e.g. the PRES model performs well on the Gapeau at Hyères for the C2M and Vdef criteria, while the performance of the other models is significantly lower. The relative advantages of one model may also depend on the criteria selected. For the Gapeau at Hyères, PRES performs better than GARD in terms of C2M, while the reverse is true for Vdef. Although it achieves lower performance than the other models on average, SIM can prove better on some catchments, e.g. the Orge at Morsang-sur-Orge for the C2M criterion. Interestingly, most models tend to underestimate the volume deficit ($V_{\text{def}} < 1$), i.e. they tend to overestimate low flows below the $Q_{50}$ threshold. GR6J is the only model which tends to underestimate low flows. The models clearly outperform the benchmark (DAQ) for all criteria. Note that the DAQ model is by definition perfect for the DatSt and DatEn criteria (see the Appendix), so comparison with the other models on these criteria is pointless.

Table 6 presents the results based on the mean performance in validation on the 21 catchments. An integrated criterion provides an overview of the overall performances. It is based on the nine criteria directly related to low flows (i.e. not considering C2M and KGE) with transformed values ranging between 0 and 1 (where 1 is the best performance). It represents the blue area in the radial plots shown in Fig. 4. It can be observed that GARD performs best for four criteria, PRES and MORD for three and GR6J with one. When looking at the integrated criterion, PRES performs best on average, followed by GR6J, GARD and MORD. However these four models are quite similar compared to SIM which obtained comparatively lower performance. DAQ performs poorly for most criteria. Mean performances and performance variability (standard deviation) on all catchments for GARD, GR6J, PRES and MORD are quite similar; the models provide good performance (e.g. at least 0.79 for KGE, and 0.7 for POD, which indicates an event under the $Q_{50}$ threshold well simulated 7 times out of 10). SIM performs less satisfactorily than the four other models for 9 out of 11 criteria, but all the models obtained greatly improve performances relative to the benchmark DAQ (except SIM for false alarm rate FAR). Interestingly, PRES performs a bit less well than the three other conceptual models on the two criteria focusing on high flows (C2M and KGE); the way PRES was implemented within...
Table 6. Models’ mean performances (with standard deviation in brackets) in validation on the 21 catchments for the simulation mode. The integrated criterion is calculated in the nine low-flow criteria (i.e. not considering C2M and KGE) and on transformed values of criteria. Bold values indicate the best model.

<table>
<thead>
<tr>
<th>Criterion</th>
<th>GARD</th>
<th>GR6J</th>
<th>MORD</th>
<th>PRES</th>
<th>SIM</th>
<th>DAQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>C2M</td>
<td>0.73 (0.09)</td>
<td>0.69 (0.10)</td>
<td>0.69 (0.11)</td>
<td>0.67 (0.11)</td>
<td>0.53 (0.13)</td>
<td>0.13 (0.05)</td>
</tr>
<tr>
<td>KGE</td>
<td>0.81 (0.09)</td>
<td>0.83 (0.09)</td>
<td>0.86 (0.06)</td>
<td>0.79 (0.10)</td>
<td>0.80 (0.07)</td>
<td>0.27 (0.11)</td>
</tr>
<tr>
<td>C2M_up</td>
<td>0.57 (0.12)</td>
<td>0.53 (0.14)</td>
<td>0.48 (0.22)</td>
<td>0.56 (0.13)</td>
<td>0.23 (0.19)</td>
<td>0.11 (0.06)</td>
</tr>
<tr>
<td>RMSEout</td>
<td>0.52 (0.29)</td>
<td>0.61 (0.52)</td>
<td>0.81 (0.80)</td>
<td>0.55 (0.35)</td>
<td>1.23 (1.06)</td>
<td>3.48 (2.66)</td>
</tr>
<tr>
<td>FAR</td>
<td>0.21 (0.12)</td>
<td>0.25 (0.13)</td>
<td>0.24 (0.12)</td>
<td>0.22 (0.12)</td>
<td>0.37 (0.12)</td>
<td>0.34 (0.12)</td>
</tr>
<tr>
<td>CSI</td>
<td>0.58 (0.15)</td>
<td>0.60 (0.11)</td>
<td>0.58 (0.14)</td>
<td>0.61 (0.11)</td>
<td>0.42 (0.10)</td>
<td>0.18 (0.12)</td>
</tr>
<tr>
<td>POD</td>
<td>0.70 (0.19)</td>
<td>0.78 (0.14)</td>
<td>0.72 (0.17)</td>
<td>0.75 (0.14)</td>
<td>0.57 (0.13)</td>
<td>0.21 (0.14)</td>
</tr>
<tr>
<td>Vdef</td>
<td>0.89 (0.50)</td>
<td>1.21 (0.64)</td>
<td>0.99 (0.44)</td>
<td>0.95 (0.46)</td>
<td>0.90 (0.38)</td>
<td>0.13 (0.14)</td>
</tr>
<tr>
<td>LFD</td>
<td>0.92 (0.33)</td>
<td>1.10 (0.35)</td>
<td>0.98 (0.26)</td>
<td>0.99 (0.29)</td>
<td>0.92 (0.24)</td>
<td>0.32 (0.21)</td>
</tr>
<tr>
<td>DatSt</td>
<td>4.67 (5.64)</td>
<td>−0.55 (8.83)</td>
<td>0.14 (9.88)</td>
<td>2.43 (5.71)</td>
<td>−13.31 (12.07)</td>
<td>NA NA</td>
</tr>
<tr>
<td>DatEn</td>
<td>1.57 (4.00)</td>
<td>−1.93 (6.38)</td>
<td>1.31 (15.31)</td>
<td>0.40 (4.08)</td>
<td>−7.83 (8.73)</td>
<td>NA NA</td>
</tr>
<tr>
<td>Integrated criterion</td>
<td>0.734</td>
<td>0.735</td>
<td>0.721</td>
<td>0.747</td>
<td>0.617</td>
<td>0.422</td>
</tr>
</tbody>
</table>

Figure 4. Radial plots showing the mean validation results obtained by the five tested models and the benchmark (DAQ) for the selected criteria over the 21 catchments and the two test periods, in simulation mode.
Figure 5. Maps of the mean performance on the two validation periods for the C2M, Vdef and CSI criteria, obtained by the five tested models and the benchmark (DAQ) on the 21 catchments, in simulation mode. The letters on the top right map show the catchments (first two letters of the HYDRO code, see Table 1) whose results are commented in more detail in the text (B2: Meuse; H4: Orge; Y4: Gepeau).
Figure 6. Comparison of the variability (expressed by the standard deviation) in mean performance on all models per catchment (x axis) and the variability in mean performance on all catchments per model (y axis), in simulation for the 21 selected performance criteria.

This study makes it more low-flow-oriented than the other models.

These results indicate that differences are quite limited between the lumped conceptual models for low-flow simulations. A more detailed analysis (not shown here) indicated that performance can vary considerably between validation periods. Overall, obtaining satisfactory streamflow simulation seems to depend more on catchment than on the model itself. Figure 6 compares the mean variability (standard deviation) of performance between models (y axis) against the mean variability of performance between catchments (x axis) for each of the 11 selected criteria. The variability between models was calculated by first computing the standard deviation of performances of the five models for each catchment and then computing the mean of these standard deviation values. The variability between catchments was calculated by first computing the standard deviation of performances on the 21 catchments for each model and then computing the mean of these standard deviation values. The graph shows that performance varies more between catchments than between models for all criteria (except for $C_2M_i$, for which the variability between models is greater than the variability between catchments), which supports that streamflow simulation depends more on catchments than on models.

Given this result, we analysed the relation between model performance and low-flow indices (BFI or $Q_{90}/Q_{50}$ ratio) or catchment characteristic (drainage density here), as they are closely related to low-flow dynamic and could explain in which case models show more difficulties to simulate low flows; BFI values indicate the level of groundwater contribution, the $Q_{90}/Q_{50}$ ratio represents the severity of low flows and drainage density informs on soil permeability. Unfortunately, as illustrated in Fig. 7, the relation did not show significant trends.

3.2 Forecasting mode

Figures 8 and 9 present the radial plots of all criteria for each model, for 7-day and 30-day lead times, respectively. Here, red lines represent the radial plot in forecasting mode when no observed streamflow is used (i.e. without using assimilation or output correction methods). The performance of the benchmark model, NVQ, was also included. Here, the differences between models seem more significant than in simulation mode for a few criteria (e.g. containing ratio (Cont_ratio), sharpness (Sharp), Vdef or low-flow duration (LFD)), especially for the 7-day lead time. However, it is still difficult to identify a single best model. We can only confirm that SIM performs a bit less well, even if the differences with the other models appear to be more limited for the 30 day lead time. One of the expected results is the loss of performance with increasing lead time for all models (and all catchments). This loss is significant for all criteria, except for the containing ratio, which is better; members of the ensemble forecast are more dispersed. Containing ratio (Cont_ratio) and sharpness (Sharp) are two complementary scores that should be evaluated together: a model should first be as reliable as possible and then provide as narrow
forecast intervals as possible (excessively spaced forecasts do not contribute information). Performance even becomes close to the benchmark performance NVQ for the 30-day lead time, but still remains better. The comparison with performance when no observed streamflow is used shows that assimilation or output correction methods improve performances for all the models (average improvement of 14.2% for GARD, 10.7% for GR6J, 12.0% for MORD, 11.3% for PRES and 7.3% for SIM for the 7-day lead-time). The assimilation method of GARD (reservoir updating) seems the most efficient. However PRES assimilation method (similar to GARD) provides similar improvement compared to GR6J and MORD, which use a correction method based on the error made at previous time-step. The quantile/quantile post-correction method, only used in the SIM model, seems less efficient than streamflow assimilation methods, since performances deteriorate for a few criteria (RMSE$_{ut}$, POD, CSI and sharpness – Sharp). Here, SIM tends to underestimate low-flows when the post correction method is not used. The quantile/quantile method for SIM in low-flows tends to increase each forecast member in low flows. Sharpness decreases ($Q_{10}/Q_{90}$ interval of ensemble forecast is larger) because the method is multiplicative. POD decreases when the quantile/quantile method is used because the decrease in the number of hits is larger than the increase in the number of correct misses.

As in simulation mode, model performance based on several criteria strongly varies among the catchments. Figures 10 and 11 show the performance maps on validation period 2 for RMSE$_{ut}$ (normalised by mean flow under the $Q_{80}$ threshold), BS$_{vig}$ and Vdef, and for each model on the 21 catchments, for forecasting 7 day (Fig. 10) and 30-day (Fig. 11) lead times, respectively. We reach the same conclusions as in simulation mode: even if for some catchments the models satisfactorily forecast low flows (e.g. the Andelle at Vascœuil and the Oise at Sempigny in RMSE$_{ut}$, whatever the forecast lead time), performance is quite variable in other catchments: e.g. the Petite Creuse at Fresselines in RMSE$_{ut}$ is properly modelled by GARD but less satisfactorily by the other models. Performance also depends on the criteria.
Figure 9. Radial plot of the mean model results for the 21 catchments for the selected criteria in validation period 2, for a $d+30$ forecasting lead time. Red lines represent the results when no assimilation or post correction method is used.

considered: for the Orge at Morsang-sur-Orge, model performance is quite good in RMSE$_{ut}$ for the two forecasting lead times but decreases significantly in BS$_{vig}$ or Vdef, compared to the other catchments.

The fact that models remain better than the benchmark model indicates that they contribute information, even for a long forecasting lead time. An analysis on the two validation periods has shown that performance can vary greatly between periods. Overall, it appears that a satisfactory streamflow forecast depends more on the catchments and their specificities than on the model, as already noted in the case of simulation results. However, the analyses to link model performance to low-flow indices (BFI or $Q_{90}/Q_{50}$ ratio) did not show significant trends, as already shown in simulation mode in Fig. 7.

Table 7 presents the results of the models on each criterion for the two selected lead times, based on the mean performance and standard deviation on the 21 catchments for validation period 2, and the mean rank on all criteria. For the short lead time (7 days), GARD and GR6J perform best on four criteria and MORD and PRES on one. GR6J and GARD are most often among the best models on average, as shown by the integrated criterion. Then come PRES and MORD, followed by SIM. The benchmark remains the poorest model, which shows that all models contribute information compared to this reference. The ranking is a bit different for the longer lead time (30 days). It changes for some criteria, which modifies the mean ranks; GARD appears to be the most highly ranked model, followed by GR6J, PRES and MORD, which are similar. SIM does not seem to contribute information on average compared to the benchmark for this lead time. Interestingly, SIM shows a lower performance loss than the four other models on the integrated criterion when the lead time increases (only 10% against 21 to 23% for the other models). We observe that models tend to underestimate low-flow characteristics, as shown by Vdef and LFD values; while the models are well balanced in simulation (Vdef and LFD around 1), all models obtain Vdef and LFD values lower than 1 in forecasting mode, indicating that they forecast lower deficit of volume and low-flow duration.
i.e. they overestimate low flows. This may be partly related to the use of historical input scenarios, since only a few of them allow representing the climatic situations that resulted in severe drought situation. The use of other scenarios based on meteorological forecasts may help limiting this problem, but further test would be needed to check this point.

This overestimation is more important for all models when the lead time increases. This is due to the attenuation of the effect of post-correction or streamflow assimilation methods. These methods should be improved to better take into account this attenuation with increasing lead-time, especially in the case of low-flow forecasting where long forecast lead-time is expected.

4 Discussions

This intercomparison experiment shows that hydrological models can provide useful information for low-flow simulation and forecasting. Here, we wished to further discuss three main issues raised in the introduction, relative to (1) the relation between simulation and forecasting performance, (2) the lead times achievable on the test catchments for low-flow forecasting and (3) whether models can collaborate to enhance overall performance. In each case, a few additional tests/analyses are presented. Here our intention is solely to provide complementary insights on these results to open clear perspectives based on this work, rather than propose new methodologies.

4.1 Within a set of models, is a better low-flow simulation model also a better forecasting model?

Section 3 showed the results of the comparison between hydrological models in simulation and forecasting modes. The hierarchy based on the integrated criterion shows several differences between simulation (Table 6) and forecasting (Table 7) modes. This is illustrated in Fig. 12. It presents the mean rank of each model in forecasting (for the 7 day lead time) for the models ranked in 1st, 2nd, …, 5th position in simulation for the 21 catchments. The hierarchy of the models between simulation and forecasting differs; the best model in the simulation (mean rank in simulation equal to 1) is also the best model in forecasting for only nine catchments. Overall for all the ranks, the hierarchy between models is the same in only 33 % of cases. Therefore, a better model in simulation does not systematically mean a better model in forecasting, which strengthens the need for an evaluation relative to specific modelling objectives. By modelling objective, we mean simulation or forecasting, which are used for different operational applications (e.g. low-flow estimation for simulation, operational real-time hydrological drought management for forecasting). These differences in performance in simulation and forecasting can be explained by the specific tools used in forecasting (streamflow assimilation and/or output correction methods, see Table 3). Figure 13 presents, for each model, the performance difference in CSI for each catchment between 7-day forecast when observed streamflow assimilation or post-correction is done (FAP) or not (For), versus the performance difference between simulation (Sim) and forecast when assimilation or post-correction is done (FAP). Positive values for the CSI difference between FAP and For indicate that the model provides better performances when using assimilation or post-correction method in forecasting. Positive values for the CSI difference between FAP and Sim indicate that the model provides better performances when the model is used in forecasting mode. We observe that CSI differences between FAP and For, and FAP and Sim are well correlated; performance differences between simulation and forecasting are closely related to the use of assimilation or post-correction methods.

4.2 Which maximum useful lead time can be expected in low-flow forecasting?

The results obtained in forecasting mode were presented for two specific lead times (7 and 30 days). As expected, model performance decreased when lead time increased, which means that the added value of the information provided by the models compared to the benchmark decreases. Therefore, there should be a maximum lead time beyond which the model cannot provide useful information compared to the benchmark. This lead time will be called “useful forecasting lead time” (noted UFL) hereafter, as proposed by Staub (2008). For each catchment and each model, the UFL can be determined by comparing the performance of the model tested and the benchmark (NVQ) when lead time increases. Note that the definition of UFL strongly depends on the benchmark used; a more demanding benchmark would tend to yield lower UFL values. Here UFL was arbitrarily chosen as the lead time beyond which model performance is not at least 20 % better than benchmark performance. We considered that beyond this limit, the operational added value would be too small. Obviously, UFL depends on the criteria chosen and benchmark. The variability of UFL values when considering a given criteria will be an indication of model capacity to represent the corresponding low-flow characteristics, and the more demanding the benchmark, the shorter the UFL.

Figure 14 presents maps of mean UFL values obtained using three efficiency criteria (RMSEa, CSI and Vdef) for the 21 catchments. The symbol indicates the model which provides the best UFL. Note that SIM was not considered here because it was run to issue 90-day forecasts on too few time steps to allow robust conclusions. The results logically depend on the catchments. For some of them, it is not possible to usefully anticipate low flows beyond 1 week, while others seem to have longer inertia and hydrological memory, with forecasts still dependent on initial conditions after several weeks. However, we could not link UFL to low-flow
Figure 10. Model performance in forecasting mode on validation period 2 for the RMSE_{ut}, BS_{vig} and Vdef criteria for each model on the 21 catchments for a 7 day forecasting lead time. The letters on the top right map show the catchments (first two letters of the HYDRO code, see Table 1) whose results are commented in more details in the text (H4: Orge; H7: Oise; H8: Andelle; L4: Petite Creuse).
Figure 11. Model performance if forecasting mode on validation period 2 for the RMSE_{ut}, BS_{vlg} and V_{def} criteria for each model on the 21 catchments for a 30 day forecasting lead time.
characteristics (BFI or \(Q_{90}/Q_{50}\) ratio). It was also noted that UFL estimates vary between models and/or test periods. For example, for the Briance River at Condat-sur-Vienne, the best mean UFL is provided by PRES and reaches 60 days for validation period 2 versus only 21 days for period 1 provided by MORD. The variability in model efficiency may partly explain these results.

The UFL estimation is very useful operationally when adapted to specific criteria/objectives defined by the water manager. The level of improvement over the benchmark, here set to 20 %, could be raised if one wishes to reach a higher level of reliability or could even replace an absolute criterion under specific circumstances.

### 4.3 Could models be efficiently combined in a multi-model approach?

Since it was not possible to identify a single model which would outperform the others for all catchments, validation periods or evaluation criteria, we attempted to investigate the possible complementarity between models via model output combinations in simulation and forecasting modes. Many multi-model approaches exist to combine the outputs of several models (see e.g. Abrahart and See, 2002; Palmer et al., 2004; Velázquez et al., 2011). Here we chose to focus on three simple methods:

1. Average multi-model forecast (AMM): this is the simplest method and consists in averaging the outputs of the five hydrological models at each time step. In ensemble forecasting mode, each multi-model member corresponds to the mean of the forecasts issued by the models using the same scenario. This multi-model approach is applicable in simulation and forecasting modes.

2. Fixed-weight average multi-model forecast (noted FMM): this consists in averaging the outputs of the five hydrological models at each time step. In ensemble forecasting mode, each multi-model member corresponds to the mean of the forecasts issued by the models using the same scenario. This multi-model approach is applicable in simulation and forecasting modes.

### Table 7. Models’ mean performances (with standard deviation in brackets) on the 21 catchments for validation period 2 and for the two selected forecasting lead times. Bold values indicate the best model.

<table>
<thead>
<tr>
<th>7-day lead time</th>
<th>30-day lead time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Criterion</td>
<td>GARD</td>
</tr>
<tr>
<td>RMSE(_{\text{int}})</td>
<td>0.72 (0.43)</td>
</tr>
<tr>
<td>DRPS</td>
<td>0.13</td>
</tr>
<tr>
<td>POD</td>
<td>0.82</td>
</tr>
<tr>
<td>FAR</td>
<td>0.23</td>
</tr>
<tr>
<td>CSI</td>
<td>0.67</td>
</tr>
<tr>
<td>BS(_{\text{vig}})</td>
<td>0.09</td>
</tr>
<tr>
<td>BS(_{\text{crit}})</td>
<td>0.06</td>
</tr>
<tr>
<td>Cont(_{\text{ratio}})</td>
<td>0.34</td>
</tr>
<tr>
<td>Sharp</td>
<td>0.95</td>
</tr>
<tr>
<td>Vdef</td>
<td>0.73</td>
</tr>
<tr>
<td>LFD</td>
<td>0.79</td>
</tr>
<tr>
<td>Integrated criterion</td>
<td>0.673</td>
</tr>
</tbody>
</table>

Figure 12. Mean model ranks for the 7 day forecast for the 21 catchments for the models ranked 1st, 2nd, … 5th in simulation.

Figure 13. CSI difference for each model in forecasting mode when streamflow assimilation or output correction method is used (FAP) or not (For), versus CSI difference for each model in forecasting mode when streamflow assimilation or output correction method is used (FAP) and in simulation mode.
weighted mean of the forecasts issued by the five models using the same scenario. This multi-model approach is applicable in simulation and forecasting modes.

3. Variable-weight average forecast (VMM): the third method tested is inspired from Loumagne et al. (1995) and is applicable in forecasting mode only. It is equivalent to the previous method, but here weights are time-dependent and are based on the mean of model errors on the last $p$ time steps. This error is calculated using the control run. For each time step, the weight given to a model is

$$W_{m,d} = \frac{\sum_{s=d-p}^{d} \sqrt{(Q_{\text{form},s} - Q_{\text{obs},s})^2}}{\sum_{m=1}^{M} \sum_{s=d-p}^{d} \sqrt{(Q_{\text{form},s} - Q_{\text{obs},s})^2}},$$

(2)

where $m$ is the hydrological model, $M$ is the number of hydrological models, $d$ is the day when the forecast is issued, $Q_{\text{form},s}$ is the streamflow forecasted by model $m$ at date $s - 1$ for $s$, $Q_{\text{obs},s}$ is the observed streamflow at date $s$, and $p$ is the length of the time window over which previous forecasting errors are considered. This approach could not be applied to the SIM model given limited availability of streamflow forecasts.

Figure 15 presents the maps of the best ranked models in simulation (mean of the models’ ranks by criteria for each catchment) for each evaluation period. The comparison between AMM and FMM (not detailed here) showed very similar results for each catchment and test period and we kept only the FMM approach in the rest of the analysis, since it is slightly better. The multi-model presented in Fig. 15 is FMM, weighted using the POD criteria. It provides better results than individual models on 13 and 12 catchments out of 21 for validation periods 1 and 2, respectively. For a few catchments, the multi-model performs best on one validation period but not on the other. Moreover, since a model that performs best on the calibration period compared to the other models does not systematically perform best on the validation period, the weight given to this model in the
FMM approach may not be optimal. The performance of the multi-model seems not to be impacted by this robustness effect. The multi-model does not drastically change performance compared to the single best models; if all models perform poorly, the multi-model does not produce satisfactory results either, which is not surprising. Interestingly however, the multi-model seems more robust than the individual models in the sense that it limits severe model failures, since it allows compensations between poor and good models. FMM provides overall better performance than the other models (integrated criterion of 0.769 against 0.747 for the best model in simulation). Here, we reach the same conclusion as Georgakakos et al. (2004), where using several distributed models with a variety of structures benefits to mean flow simulation compared to a best single distributed one. Combining several lumped and distributed models overall improve low-flow simulation here.

In forecasting mode, SIM was excluded from the three combination methods since it was not possible to use it in the VMM option. For VMM, the mean error to weight the model was calculated over the six last time steps, which appeared to be a good compromise between performance and length of this backtracking period. Here, as in simulation, the results (not detailed here) are similar between the three options, but VMM is slightly better. Therefore, we kept only the VMM model in the rest of the analysis. Figure 16 presents the maps of the best ranked model in forecasting for a 7 day lead time (mean of the ranks of models by criteria for each catchment) for each evaluation period. The multi-model provides the best results only on six and five catchments out of 21 for validation periods 1 and 2, respectively. GARD and GR6J are also often the best models. The limited efficiency of the multi-model may be due to the overly crude combination approach; even if it proved useful in a flood forecasting context in the study reported by Loumagne et al. (1995), other approaches that better account for the slow dynamics of low flows may be more efficient and should be further investigated.

5 Conclusion and perspectives

In this paper, we presented a comparison between five hydrological models for low-flow simulation and forecasting on 21 French catchments representing a variety of physical and hydro-climatic characteristics. A general evaluation of models was made using several criteria which represent different qualities expected of models. Moreover, the use of benchmarks contributed comparative information on the actual operational utility of these models.

In simulation mode, the comparison showed that calibrated models perform better (GARD, MORD, GR6J and PRES). SIM, the only uncalibrated model included in the comparison, nonetheless performs as well as the other models on a few catchments. It was difficult to define a clear hierarchy between these calibrated models, since the results vary according to the selected criteria, the catchment considered or even the test period. Tests to relate performance to catchment or streamflow characteristics proved unsuccessful, but this is a key aspect to improve low-flow simulation as results depends more on the catchments than on models. Models are much better than the benchmark (daily average streamflow) and showed the usefulness of hydrological simulation for low flows.

In forecasting mode, we reached the same conclusions, with better results for calibrated models. Here, establishing a hierarchy between the models is also difficult, since performance varies according to the criterion, catchment, validation period and lead time. The results are quite good for short lead times, especially compared to the benchmark. As can be expected, this gain decreases as lead time increases, and performance remains modest, especially for longer lead times; there is an important need for further investigation to improve low-flow forecasting. It is difficult to conclude on the actual
usefulness of such models for operational management, as performance can vary much between catchments. But forecast might be improved by using alternative input scenarios (e.g. actual meteorological ensemble). Although models perform differently from one period to another, overall they tend to present the same ability to forecast low flows on a catchment. The rainfall scenarios (historical archive) used here to test models were quite crude and it is likely that using the ensemble forecast from meteorological models would improve results, at least for short lead times, but this would require further investigation.

In forecasting, we presented a simple approach to determine the maximum lead time beyond which models do not add significant information compared to the benchmark. This maximum lead time was variable because models behaved differently with increasing lead time and the results differed according to the criteria and the validation period.

Combining the single models into a multi-model was successful even with simple combination methods, but the performance of the multi-model strongly depends on the performance of individual models; where all the models present difficulties in simulating or forecasting low flows, a model combination cannot compensate for model errors. The main advantage in building a multi-model lies in its robustness, where only one model presents difficulties on a catchment, a multi-model corrects this weakness.

As far as perspectives are concerned, we would like to mention that (i) tests were made on two other catchments in a very different climatic context on Reunion Island (Indian Ocean). They were not detailed here for the sake of brevity but yielded similar conclusions; (ii) this study used catchments where human influence was considered negligible, but the use of catchments where anthropogenic pressure on water resources is significant constitutes the second part of the PREMHYCE project, and the results will be reported in due course.
Appendix A: Formulation of the numerical criteria selected for simulation evaluation

KGE

This criterion was proposed by Gupta et al. (2009) as a modification of the Nash–Sutcliffe (1970) efficiency index:

\[ \text{KGE} = 1 - \sqrt{ \frac{1}{n} \sum_{i=1}^{n} (Q_{\text{sim}_i} - Q_{\text{obs}_i})^2 } \]

with \( r \) the correlation coefficient between observed and simulated flows, the ratio of simulated and observed flow standard deviations and \( \beta \) the model bias.

C2M

C2M is a bounded version of the Nash-Sutcliffe efficiency index calculated on streamflow \( Q \) (\( \text{NSE}_Q \)), as proposed by Mathevet et al. (2006)

\[ \text{C2M} = \frac{\text{NSE}_Q}{2 - \text{NSE}_Q}. \]

This is similar to the previous criterion, but \( \text{NSE}_Q \) is calculated on inverse flows to more strongly emphasise low flows, as proposed by Pushpalatha et al. (2012).

\( \text{RMSE}_{\text{ut}} \)

\( \text{RMSE}_{\text{ut}} \) is the root mean square error for flows under the low-flow threshold, normalised by the mean observed flow:

\[ \text{RMSE}_{\text{ut}} = \sqrt{ \frac{1}{n} \sum_{i=1}^{n} (Q_{\text{sim}_i} - Q_{\text{obs}_i})^2 } \]

where \( Q_{\text{obs}_i} \) is the observed streamflow for day \( i \), \( Q_{\text{sim}_i} \) the simulated streamflow for day \( i \), and \( n \) the number of time steps on the validation period where \( Q_{\text{obs}_i} \) is less than the \( Q_{80} \) threshold.

\( \text{Vdef} \)

\( \text{Vdef} \) is the ratio of simulated and observed flow deficits under the low-flow threshold:

\[ \frac{\sum_{i=1}^{n} \max(0; Q_{\text{threshold}} - Q_{\text{sim}_i})}{\sum_{i=1}^{n} \max(0; Q_{\text{threshold}} - Q_{\text{obs}_i})} \]

\( \text{LFD} \)

This is the ratio of simulated and observed low-flow durations:

\[ \frac{\text{Duration}_{\text{sim}}}{\text{Duration}_{\text{obs}}} \]

where \( \text{Duration}_{\text{sim}} \) is the number of days where the \( Q_{\text{sim}_i} \) is less than the \( Q_{80} \) threshold on the validation period and \( \text{Duration}_{\text{obs}} \) is the number of days where the \( Q_{\text{obs}_i} \) is less than the \( Q_{80} \) threshold on the validation period.

\( \text{DatSt} \) and \( \text{DatEn} \)

This is a comparison of observed and simulated dates when low flows start (St) or end (En).

\[ \text{Dat} = \text{Date}_{\text{sim}} - \text{Date}_{\text{obs}}, \]

where \( \text{Date}_{\text{obs}} \) is the Julian day of daily average streamflow when 10% (resp. 90%) of the observed volume deficit is exceeded for DatSt (resp. DatEn). The threshold for the observed volume deficit calculation is the observed \( Q_{80} \) calculated of the daily average streamflow. \( \text{Date}_{\text{sim}} \) is the Julian day of the daily average streamflow, where 10% (resp. 90%) of the simulated volume deficit is exceeded for DatSt (resp. DatEn). The threshold for the simulated volume deficit calculation is the simulated \( Q_{80} \) calculated of the daily average streamflow.

\( \text{Vdef, LFD, and DatSt and DatEn} \) have been adapted from the concept of “centre of mass” proposed by Stewart et al. (2005).

\( \text{False alarm ratio (FAR), probability of detection (POD) and critical success index (CSI)} \)

These are criteria based on the contingency table for low flows considering the \( Q_{80} \) threshold (Schäfer, 1990):

\[ \text{FAR} = \frac{b}{a + b}, \]

\[ \text{POD} = \frac{a}{a + c}, \]

\[ \text{CSI} = \frac{a}{a + b + c}. \]

where \( a \) is the number of hits, \( b \) the number of false alarms, \( c \) the number of correct misses and \( d \) the number of correct rejects.

\( A1 \) Numerical criteria for forecasting evaluation

\( \text{RMSE}_{\text{ut}}, \text{Vdef, LFD} \)

These criteria have the same definition as in the simulation but are calculated using the mean of the ensemble forecasts for the horizon considered.

\( \text{Sharp} \)

This criterion measures the width of the ensemble forecast (Franz and Hogue, 2011):

\[ \text{Sharp} = \frac{1}{n} \sum_{i=1}^{n} Q_{90_i} - Q_{10_i}, \]
where \( n \) is the number of time steps on the validation period where the \( Q_{\text{obs}} \) is less than the \( Q_{80} \) threshold, and \( Q_{90} \) (resp. \( Q_{10} \)) the 90\% (resp. 10\%) percentile of the distribution of forecasts for day \( i \).

**Cont\_ratio**

The containing ratio measures how often the observation lies within the ensemble forecast (Franz and Hogue, 2011):

\[
\text{Cont\_ratio} = \frac{n}{N},
\]

where \( n \) is the number of observed streamflows in the 80\% forecasted confidence interval when the \( Q_{\text{obs}} \) is less than the \( Q_{80} \) threshold, and \( N \) the number of time steps where the \( Q_{\text{obs}} \) is less than the \( Q_{80} \) threshold.

**FAR, POD and CSI**

The same definition as in the simulation is used. Here an event is forecasted if more than 50\% of members are below the low-flow threshold.

**BS**

The Brier Score (BS) (Brier, 1950) compared the observed and forecast probabilities relative to a threshold:

\[
\text{BS} = \frac{1}{n} \sum_{i=1}^{n} (y_i - o_i)^2,
\]

where \( o_i \) is the observation probability, \( y_i \) the forecast probability. An event is observed/forecasted if the observed/forecasted streamflow is less than the vigilance threshold (\( Q_{80} \) for \( \text{BS}_{\text{vig}} \)) or the crisis threshold (\( Q_{95} \) for \( \text{BS}_{\text{cri}} \)). \( n \) is the number of time steps where \( Q_{\text{obs}} \) is less than the \( Q_{50} \) threshold (\( \text{BS}_{\text{vig}} \)) or the \( Q_{80} \) threshold (\( \text{BS}_{\text{cri}} \)).

**DRPS**

The Discrete ranked probability score (DRPS) (Toth et al., 2003):

\[
\text{DRPS} = \frac{1}{N_{\text{threshold}}} \sum_{k=1}^{N_{\text{threshold}}} (\text{BS}_k),
\]

where \( N_{\text{threshold}} \) is the number of thresholds chosen (ten percentiles here, \( k = Q_{80}, Q_{82}, Q_{84}, \ldots, Q_{96}, Q_{98} \)).
Acknowledgements. The authors thank Météo-France for providing meteorological data and the national hydrometeorological forecasting centre (SCHAPI) for providing streamflow data. The Regional Departments for the Environment (DREAL) are also thanked for their advice in selecting catchments and their feedback on the project. The PREMHYCE project was funded by the French National Agency for Water and the Aquatic Environment (ONEMA) and the Department of Fresh Water and Biodiversity of the French Ministry for Ecology (MEDDE). The second author thanks DIM ASTREA of Région Ile-de-France for financial support. The authors thank the three anonymous reviewers and the Associate Editor, Elena Toth, who handled the manuscript, for their comments, which helped improving the manuscript.

Edited by: E. Toth

References


