
Reply to the Associate Editor

Dear Authors,
your replies address all the points raised by the three referees (who did indeed an excellent and very constructive job!) and you propose to implement modifications that should adequately solve their main concerns, and therefore I warmly invite you to submit a revised version, amending the paper as you already suggest in your Replies.

AR: We thank the Associate Editor for having accepted to evaluate this paper and for his positive feedbacks and useful comments on our manuscript.

AEC: 1) The major issue is certainly that of the fairness of the comparison in forecasting mode as far as data assimilation/updating is concerned. As the Authors themselves propose, in their reply to Ref1, it is indeed necessary to replace the current forecasting experiment with one where the actual observed outputs are not used in any way.

AR: We added results when no assimilation/updating method is used as requested. Analysis can be found at:
Lines 423 to 431 and figure 8 and 9 (results when no assimilation/post-processing methods have been added).
Lines 496 to 505 and figure 13

AEC: 2) Secondly, the calibration procedures should be better described, with details in the text, in addition to Table 3), with particular attention paid to the use of different Objective Functions, especially in order to guarantee a fairness in the models comparison: the use of the same Obj Function(s) would be strongly preferable.

AR: We agree that using the same objective function would guarantee fairness in the models comparison. Here, the spirit of the project was to benefit of the experience of each modellers concerning his model to provide the best performances. This has been better explain, as calibration procedures (optimization method and objective function), in the text at lines 225 to 235.

AEC: 3) Also the proposed deeper analysis/interpretation of the performances (maybe using the integrated criterion you suggest in reply to Ref3) between catchments and between models (and in particular putting them in relation to model structure), even if certainly not easy, would further improve the paper significance.

AR: An integrated criterion has been added in Table 6 and 7, and a short paragraph better supports the interpretation of the performances between catchments and models in the text at lines 393 to 401 (with figure 6).

Additional remarks:
- I agree with you that it is not necessary here to make new analyses with actual forecast ensembles and extending your

justification (as already described in your reply to Ref1, RC1-2) should be enough.

AR: No tests with actual forecast ensemble were done. We extended the justification by adding a short paragraph at lines 292 to 295.

- I agree with Ref2 that section 1.3 should be removed and replaced by a shorter paragraph citing only the papers that are closely related to your work (same models or/and same catchments) and relating their results to yours in the final discussion (when, as you propose in the reply to Ref2-RC1, you will refer to literature results for corroborating your findings.

AR: Section 1.3 has been removed and a short paragraph citing the works related to the models has been kept in section 1.2 (lines 79 to 98). Our results have been replaced in the literature when possible (sentence line 584, paragraph lines 637 to 641).

- The reply to Ref3-RC2 is incomplete.

AR: We wonder if this comment is not about Ref3-RC3 (instead of RC2). In this case, we added additional reply elements (see response to Ref3-RC3). But the case study illustration has been removed.

Reply to the Anonymous Reviewer #1

We thank Reviewer #1 for his careful reading and evaluation of our manuscript and his detailed suggestions, which will help improving the manuscript.
In the following, we explain how we will account for his comments. Each time, the comment is repeated and our reply is given.

Reviewer's comment (RC): The authors present the results of a large experiment on low flow simulation and forecast. They compare different hydrological models (with different complexities) for their performance and try to answer interesting research questions. The very recent low flow literature is included and referred in an appropriate way. Overall the article is well written and quite clear for the reader although there is room for some improvements.

Authors' reply (AR): We thank Reviewer #1 for his positive feedbacks and useful comments on our manuscript.

RC: 1) Section 2.3.3: Instead of using real forecast inputs, long term meteorological archive was used. The justification of long-term archieve is somewhat surprising. Was long term data necessary? It would be nice to have a short test period but with real forecast meteorological ensemble forcings (e.g. the period of 2002-2005 as in Demirel 2013b) to see the effect of input uncertainty due to the different ensembles.
Could you explain/justify (a bit more) the link between possible future conditions based on the historical dataset?

AR: We agree with the reviewer that testing the models with actual series of past meteorological ensemble forecasts would have been better to account for the actual uncertainty linked to meteorological forecasts, especially for short lead times. However, there were several reasons for running the models using archives of past observations instead of actual ensemble meteorological forecasts in the context of the PREMHYCE project:

- *First we wanted to test models on long series to get general results, i.e. including a few key drought events that occurred in France in the past decades, that date back to the 1970s. Such long archives of past forecasts do not exist to our knowledge.*
- *Second, the lead times targeted in the project were up to a few weeks, i.e. much longer than the medium-range forecasts of about two weeks that are available today. Running models up to a few weeks therefore means that medium-range ensemble forecasts should have been extended with other information, basically based on climatic series. Since the objective of the project was not to build scenarios but rather to concentrate on hydrological models, this was not an option we considered. We think that using past observed series provides a representative ensemble of likely conditions for the period of the year, even though the ensemble is probably too large for the short lead times. However, since the target is on low flows, the catchment response to meteorological inputs is much more smoothed than in high flow conditions, which makes this problem probably less essential.*
- *Third, the use of past observed series is one option that was chosen to run one of the tested models in operational conditions, and which provides interesting results.*

For these reasons, it was chosen not to use actual meteorological forecasts.
It will be difficult to include results with actual forecasts in the article. Indeed, we think that building scenarios combining medium-range forecasts and climatic archives to reach the targeted lead times may correspond to various options that should be considered. Actually, in

a separate ongoing work at Irstea (PhD of Louise Crochemore), we are doing tests to investigate this issue and we intend to report it shortly.
So, to answer the reviewer's comment, we added a short paragraph line 302 to 307.

RC: 2) Section 2.3.3: Using historical SAFRAN data is more straightforward than downscaling the ECMWF forecast data. I find it an interesting, pragmatic and sound approach. This approach also avoids different errors due to downscaling. But representativeness of historical data for future scenarios should be better described. This can be in a subbasin for a short period of data, just to see if the two input dataset (51/39 ECMWF ensembles and 51 SAFRAN ensemble) are compatible.

AR: As explained above, this aspect was a bit out of the scope of the PREMHYCE project. The preliminary tests we did in a separate work to compare the use of ECMWF forecasts with the SAFRAN archive option or even combined versions of these two sources of information showed that very little information is brought by the medium-range forecast in terms of reduction of uncertainty for low-flow forecasts. Our interpretation is that the smoothing effect of the catchment is much stronger than in high flow conditions. We plan to finalize these analyses and publish this work in due course. We also added a short paragraph at lines 292 to 295.

RC: 3) Section 4.1 concludes as "a better model in simulation does not systematically mean a better model in forecasting". The reader can be curious why? May be it is the model sophistication handling the input uncertainty (behavior during wetter or dryer inputs)? Is there a similar situation in Demirel 2013b to support this result? For example, in Demirel 2013b while GR4J (NSlow: 0.65) outperforms HBV (NSlow: 0.52) for calibration period, the model output uncertainty of the HBV (the grey range in Fig 3) was lower than GR4J.

AR: The differences in relative performance between simulation and forecasting modes can have several origins. We think that one key aspect is the way models assimilate observed flows and/or use post-processing techniques in forecasting mode. Tests have been done without assimilation and/or post-processing techniques. Results are presented in figure 13 and detailed at lines 496 to 505.

RC: 4) The second part of the sentence "... which strengthens the need for an evaluation relative to specific modeling objectives." is unclear to me. What do you mean? There was a specific modelling objective in this study i.e. low flows. What else?

AR: This is indeed not fully clear. By specific modeling objectives, we meant 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). We added the following sentence:
Line 492: 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).

RC: 5) Another unclear sentence: "These differences in performance in simulation and forecasting can be explained by the specific tools used in forecasting, which assimilate streamflow and/or correct model outputs (see Table 3)."

What kind of specific tools?

AR: By specific tools, we mean the different methods used by modelers to improve the forecasts quality, i.e. streamflow assimilation or post-processing methods. A better model performance in forecasting mode can result from these methods instead of the model himself. As mentioned in our answer to reviewer's comment #3, we added additional insights on this aspect at lines 496 to 505 and lines 423 to 431 (with figure 8 and 9 which have been modified):

RC: 6) Another unclear sentence: "However, given the variety of assimilation and correction methods applied in this study, it is difficult to conclude on the relative advantages of each of them and more systematic tests would be needed."
..the relative advantages of each of them (of What?) Please can you explain?

AR: Linked to comments #3 and 5 above, this has been be clarified at lines 423 to 431 and 496 to 505. Here, we refer to the assimilation and correction methods and meant that testing the reliability of these methods would require systematically applying them to each hydrological model and comparing the performance. Here the spirit of the project was to consider modelling tools, i.e. hydrological models and the associated assimilation and/or post-processing methods selected by the modellers. Therefore, we did not distinguish the two aspects and did not investigate the sensitivity of results to each of them. This sentence has been removed.

RC: 7) Section 4.3: variable-weight average forecast model seems similar to Bayesian model averaging. If so I would recommend the authors to include relevant references from bayesian model averaging literature e.g.
Parrish, M., H. Moradkhani, C.M. DeChant (2012), Toward Reduction of Model Uncertainty: Integration of Bayesian Model Averaging and Data Assimilation, Water Resources Research,48, W03519, doi:10.1029/2011WR011116.

AR: Here, the method's principle looks similar to the BMA from Parrish et al., but is different because we do not use the probability density of forecast for each model to combine them. As for the other methods, each member of the multi-model corresponds to the weighted mean of the forecasts issued by the five models using the same meteorological scenario.

RC: 8) The authors' effort on presenting the catchment characteristics to explain the relations to model performance is very much appreciated although the strength of relations was not significant to reveal a pattern.

AR: We thank the Reviewer for this comment since the choice to include or not these results was a bit difficult, given the lack of clear relationship. But we agree that it is something important since such relationships could be expected. The sentence has been rephrased:
Line 611: 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.

RC: 9) Page 14004, line 4 "The relative gain compared to the benchmark (daily average streamflow) is very high and showed the usefulness of hydrological simulation for low flows."
What do you mean by relative gain?

AR: The term relative may be confusing here. Actually, we meant the performance gain relatively to the benchmark. The term relative has been removed and the sentence has been rephrased:

Line 613: Models are much better than the benchmark (daily average streamflow) and showed the usefulness of hydrological simulation for low flows.

Reply to the Anonymous Reviewer #2

We thank Reviewer #2 for his careful reading and evaluation of our manuscript and his detailed suggestions, which will help improving the manuscript.

In the following, we explain how we will account for his comments. Each time, the comment is repeated and our reply is given.

Reviewer’s comment (RC): The manuscript reports on a model intercomparison project (MIP) on low flow simulation and forecasting in a number of rivers in France. It is well written, and methods are generally clear, with some exceptions (see below). The study is certainly relevant and could fill a niche as no such experiments have yet focused on low flows. Reading a paper with this topic I expected more specific conclusions on what needs to be done to better simulate low flow though – this information, i.e. the relation to of the results back to the actual model differences with respect to low flow modelling are still a bit weak. In order for the paper to make a convincing and useful contribution to wider hydrological sciences, therefore, it needs to be more focused, preferably shifted from a wide reporting of the (provocatively put: rather boring and not directly transferrable) very detailed results on all individual models and score measures to e.g. the (commendably: really interesting and useful) questions already now addressed in the discussion section. The manuscript would then have the potential for a much wider impact.

Authors’ reply (AR): We thank Reviewer #2 for this positive general comment and for his proposal to extend the discussion section. As mentioned in our answer to Reviewer #1, we added the results of model testing in forecasting mode without assimilation/post-processing methods, which give useful insights on the added value of this key aspect of the forecasting methodology:
Line 423 to 431 and figure 8 and 9 (results when no assimilation/post-processing methods have been added)
Line 496 to 505 and figure 13
Figure 8 and 9: results when no assimilation/post-processing methods have been added
Since the article is already quite long and we do not wish to make it much longer, we made the result section a bit shorter, even though we think that this section is essential and represent a valuable output of our work. We removed the case study illustrative applications (section 3.3), which are probably less essential to understand results.

Major comments:

RC: 1) As said above, the interesting aspects are currently hidden in the discussion section. On the other hand a proper discussion relating the results to the literature is missing. The analysis done is largely sufficient for example to make these discussion questions into the main objectives of the manuscript and thus make it visibly more useful to a wider community than the French participants of this specific MIP.

AR: To answer the Reviewer’s comment and better stress the key aspects in the discussion, we propose to (1) better emphasize these questions in the introduction section (sentences *line 62, 98, 143), (2) increase the discussion by referring to results already published in the literature to explain how our results corroborate or not past findings (sentence line 584, paragraph lines 637 to 641) and (3) better stress in the conclusion the key directions we think useful to investigate in the future (sentences line 612, lines 619 to 623).*

RC: 2) The frequently repeated conclusion that “model performance depends more on the catchment than on the choice of model”, which is presented as the main outcome of the results as they are presented now, needs to be better supported. The analysis doesn’t prove this consistently e.g. by testing the degree of systematic ranking of models versus the degree of systematic ranking of catchments with the same approach.

AR: This aspect is indeed not clearly shown in the article. We added a short paragraph where performances mean variability of models is compared to performances mean variability of catchments, at lines 393 to 401 and figure 6.

RC: 3) Related to the previous point is the highlighted but unsupported not-found link to catchment characteristics. For example: p13995 lines 18/19 What is the basis of this statement “satisfactory streamflow simulation seems to depend more on catchment characteristics than on the model”? It contradicts the next sentence and the sentence in the abstract that “all attempts to relate model efficiency to catchment characteristics remained inconclusive”. Two of the three predictors tested are not catchment characteristics, but streamflow characteristics. Only drainage density is a catchment characteristic. In any case what are the hypotheses that these three should be influencing model performance? If influences are tested statistically there should be a hypothesis. Besides characteristics, also processes not represented by the models will hopefully influence the performance. What these might be could also be the basis for such tests and lead to more transferrable results. Here the approach needs to be clearer and more focussed.

AR: We agree that this is confusing. To avoid confusion, the term “characteristics” has been removed:
Line 392: Overall, obtaining satisfactory streamflow simulation seems to depend more on catchment than on the model itself.
This sentence is supported by the added paragraph line 393 to 401 with figure 6 as explained in comments #2.
The hypotheses that the low-flow indices or drainage density influence model performances are given by a short paragraph at lines 402 to 406.

RC: 4) One methodological aspect the functioning of which and the relevance for the results needs to be better explained and elucidated are the post-correction methods and the assimilation methods. As these are related to model application rather than to the models themselves they can distort results. It would have been useful to test these somehow separately. But at least their effect needs to be discussed in detail with reference to literature (see comment on lacking discussion in general).

AR: We agree with the Reviewer that this is a key aspect to interpret results and differences between models in forecasting mode. As explained in our answer to Reviewer #1, we added results showing model performance in forecasting mode without assimilation/post-processing, to better emphasize the role of these methods on differences between models (lines 423 to 431 with figure 8 and 9, and lines 496 to 505 with figure 13, see reviewer #1, RC 3, 5 and 6).

RC: 5) The approach of constructing scenarios appears not well suited to look at the two very severe events in forecasting mode, as this results in most ‘spaghettis’ above the actual low flow. The value of this sub-analysis should be reconsidered or it should be assessed more in terms of the models themselves. What do those that can model these extreme events well have that others don’t?

We agree that the approach of scenario selection has limitations, as acknowledged in our answer to Reviewer #1 RC2 (lines 292 to 295). One scenario that could have been added is the “no rainfall” scenario, although it is very pessimistic and corresponds to very rare events. Actually, the year 1976 in our record is close to this scenario, which is therefore accounted for to some extent. The illustration case studies we had introduced on severe low-flow events partly intended to illustrate the differences between models in such conditions. But the Reviewer is right in stressing that the differences in model performance could be related to model structures. Although it is difficult to draw conclusions on this aspect without a close investigation of model behavior and internal states, which was a bit beyond the objectives of the project, we removed the case study part as proposed earlier, which is a less essential part to understand results and added a short paragraph to explain limits of using such scenarios at lines 296 to 301.

Specific and technical comments:

RC: 6) The abbreviations and variable names of the skill scores are very variable with some being words, some three or four letter abbreviations, some one letter variables. This makes the manuscript difficult to read. Perhaps they could be homogenized in their presentation or less could be selected (see comment above on “too detailed”)

AR: We homogenized all notations to clarify the text as suggested (see Table 4 and 5 and appendix for the names of scores).

RC: 7) Catchment yield (Table 2), Runoff Yield p. 13994, line 4. Choose one term. “Runoff Ratio” may be the more common term, anyway.

AR: We replaced the term catchment yield in table 2 and the term runoff yield lines 192 by runoff ratio as suggested.

RC: 8) Last sentence of the abstract was unclear to me without having read the paper.

AR: We agree and the sentence has been rephrased/extended to improve clarity (lines 31 to 34).

RC: 9) Last sentence of 1.1 is a bit disconnected and surprising and would require some further information of why this is mentioned at all.

AR: We had written this sentence to avoid confusion on the actual focus of the article. But we agree that it is not well placed here. We moved it to the method section when discussing lead times (lines 307 to 310).

RC: 10) p. 13983 line 8ff This sound a bit complicated. Why can it not be called ensemble low flow forecasting (similar to ensemble flood forecasting)?

AR: The sentence has been rephrased as suggested (lines 72 to 78).

RC: 11) 1.3 The list of models or forecasting systems in France is of little use to the paper

AR: We think it is useful to explain the context of this research in France and acknowledge the previous efforts to develop forecasting tools since these efforts lead to the development of most of the models that were tested in the project. However, this link may be unclear and we removed section 1.3. Presentation of the development of the five models has been keeping in section 1.2 (line 79 to 99).

Reply to the Anonymous Reviewer #3

We thank Reviewer #3 for his careful reading and evaluation of our manuscript and his detailed suggestions, which will help improving the manuscript.

In the following, we explain how we will account for his comments. Each time, the comment is repeated and our reply is given.

Reviewer’s comment (RC): The manuscript compares the performances of 5 different hydrological models used to forecast low flows of 21 French watersheds based on a large variety of criteria. The text is well written and structured, clear, referring to the recent literature on low flow forecasting and will certainly be of interest for the readers of HESS. The work could nevertheless benefit from a more in-depth analysis of the obtained low-flow forecasts and their limits. The whole approach remains a little too empirical and descriptive at this stage with no clear conclusion or open perspectives for future improvements. Important questions, some mentioned in the manuscript, could be discussed in more detail:

Authors’ reply (AR): We thank the Reviewer for his constructive comments. We tried to account for his suggestions as detailed below.

RC: 1) Most of the tested models have not been specifically developed for the purpose of simulating low-flows. Have their calibration procedures been adapted to better simulate the low-flow periods? Some information on the calibration procedures of the model, the possible influence on their parameter values, recession dynamics, would be useful here as well as some suggestions.

AR: The calibration method adopted by each modeller was indeed adapted to better simulate low flows: the objective functions used are generally specifically adapted to low-flow simulation (e.g. Nash-Sutcliffe Criteria calculated with Q0.2 for PRESAGES, or mean of the Kling-Gupta criteria calculated both on Q and 1/Q for Mordor and GR6J). The calibration method and criteria are described in Table 3. We better explained these aspects in the text (short paragraph lines 225 to 235).

RC: 2) Beyond the quantitative criteria, the analysis of the simulated discharge series could be a little more developed. Are for instance the forecasts in fig. 11 realistic? Is it really likely that the discharge increases within a few days to exceed the Q80 during a marked low-flow period in mid-August for a significant number of rainfall scenarios as suggested by some tested models? I have some doubts. Most of the tested models seem too sensitive to rainfall during low flows for the Meuse river.

AR: The models were tested using an ensemble of likely rainfall scenarios, some of which are quite wet, explaining the sudden reaction of models. Actually, the spaghetti representation visually emphasizes outlier scenarios, while the 80% confidence intervals would be more narrow. What is reassuring on the actual capacity of models to represent catchment behaviour is that the control run (with the observed meteorological scenario) is close to the observed values. We removed the case study illustrative applications (section 3.3), which are probably less essential to understand results as proposed to reviewer #2.

But this aspect has been better explained, in link with the choice of meteorological scenarios (lines 296 to 301).

RC: 3) It appears that the discharge lies significantly under the average inter-annual discharge already in May for the 3 selected severe low-flow periods and the 2 selected watersheds in figures 10 to 12. This leads to a question: what is the relative importance of the initial conditions and of the summer rainfall scenarios in the determination of the discharge evolution during low-flows? Is this relative weight the same in the observed and simulated series? In other words, are the models representing the correct low-flow dynamics? This is a tricky question that cannot be answered based on aggregated criteria only. By the way, the selected NVQ benchmark could have been improved: distribution of available streamflows in the other years for the considered day, but selecting only the years where the baseflow at the date of the forecast lie in similar ranges as in the considered year. This would probably be less in favor of the tested models. Could the authors test this?

AR: This is indeed a tricky question, and it is difficult to provide an answer as the relative importance of the initial conditions and of rainfall scenarios probably depends on catchment reactivity, which differ from a catchment to another in this study. We think that models are overall able to well reproduce low-flow dynamics in most cases. Moreover, the case study illustration part (section 3.3) has been removed as proposed to reviewer #2.
We thank the Reviewer for this interesting suggestion about the choice of benchmark. We agree that using more demanding benchmarks helps better emphasizing the limits of the tested models. We did not investigate the added value of such a benchmark to better analyse the behaviour of the tested models as model performances will not change, but we added a sentence explaining how useful lead-time could be affected by the selected benchmark (line 515).

RC: 4) The differences between simulation and forecasting performances deserve some more explanation.

AR: We agree that this is a key aspect that deserves clear explanation. We think that performance differences between forecasting and simulation are relying to the use of assimilation/post-processing methods. We better explained the differences by adding short paragraphs at line 423 to 431 and figure 8 and 9 (results when no assimilation/post-processing methods have been added), and lines 496 to 505 and figure 13. Moreover, we introduced a schematic diagram (figure 2) which explains the difference between simulation and forecasting.

RC: 5) Beyond the relative performances of the models, could the authors comment on the absolute values obtained for the various tested criteria? Are the performances of the models really sufficient for decision making (what decisions) on the tested rivers?

AR: Given the feedbacks from operational forecasters on the use of such models, we think they are indeed useful, even though some of their performance criteria remain modest compared to the benchmarks. There is clearly a significant margin of progress. This has been better commented conclusion section (lines 619 to 623).

RC: 6) The figures and tables could also be improved. I am not convinced that the rankings are the most useful peace of information. I would prefer to see the average values of the criterions in tables 6 and 7. Comments on the ranks in the text are sufficient.

AR: The objective of the ranking was to give some index of relative reliability of the tested models. However, we acknowledge the limitations of ranks. We provided average performances in table 6 and 7, and added an integrated criterion based on the mean of the non-dimensional criteria we used.

RC: 7) Many figures and legends are too small. Figures 4 and 8 are for instance attractive, but difficult to read and interpret. They have moreover little added value if compared to tables 6 and 7 (with values of criteria) and figures 14 to 16. Fig 10 is impossible to read because the contrast between the different curves is not sufficiently marked. Colours but also line types should be varied.

AR: We agree that some figures should be improved.
Figure 5, 7, 10, 11 have been modified
Performances in forecasting mode when no assimilation/post-processing method is used have been added in figure 8 and 9.
Hydrograph have been removed due to the suppression of case study section (see reviewer #2 comment).

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

Pierre Nicolle[1], Raji Pushpalatha[1], Charles Perrin[1], Didier François[2], Dominique Thiéry[3], Thibault Mathevet[4], Matthieu Le Lay[4], François Besson[5], Jean-Michel Soubeyroux[5], Christian Viel[5], Fabienne Regimbeau[5], Vazken Andréassian[1], Pascal Maugis[6], Bénédicte Augeard[7], Emmanuel Morice[8]

[1] Irstea, HBAN, Antony, France (pierre.nicolle@irstea.fr)

[2] Université de Lorraine, LOTERR, Metz, France

[3] BRGM, Orléans, France

[4] EDF-DTG, Grenoble, France

[5] Météo-France, Direction de la Climatologie, Toulouse, France

[6] IPSL, LSCE, Gif-sur-Yvette, France

[7] ONEMA, Vincennes, France

[8] Direction de l'eau et de la Biodiversité, Ministère de l'écologie, du développement durable et de l'énergie, La Défense, France

Abstract

Low-flow simulation and forecasting remains a difficult issue for hydrological modellers, and intercomparisons can be extremely instructive ~~are needed~~ to assess existing low-flow prediction models and to develop more efficient operational tools. This research ~~study~~ 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 and one distributed physically-oriented model) ~~with different characteristics and conceptualizations~~ were applied within ~~following~~ 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

Mis en fo
Automatic

all catchments and criteria~~in all circumstances~~, a few models appeared 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 ~~contribute~~ useful information compared to the benchmark. This maximum useful lead time logically varies between catchments, but also depends on the model used. ~~Preliminary attempts to implement simple~~Simple multi-model approaches that combine the outputs of the five hydrological models were tested to improve simulation and forecasting efficiency. We find ~~showed~~ that ~~additional efficiency gains~~the multi-model approach was more robust and can ~~be expected from such approaches~~provide better performance than individual models on average.

Keywords

Hydrological modelling, Low flow, Long-term forecast, Evaluation criteria, Comparison

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~~study was carried out~~, 81% of the 33 km^3 of total water withdrawals in 2009 came from rivers ~~(CGDD, 2012)~~(CGDD, 2012). Municipal water supply, irrigation, navigation, hydropower and ~~nuclear~~ thermal power plant cooling are highly dependent on streamflow~~surface water resources~~ and can be strongly affected by water shortages in rivers ~~(Bousquet et al., 2003)~~(Bousquet et al., 2003). Increasing efforts to maintain minimum environmental flows in rivers make the issue even more acute ~~(Saunders and Lewis, 2003; García de Jalón, 2003)~~(García de Jalón, 2003; Saunders and Lewis, 2003).

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 ~~(Chiew and McMahon, 2002; Hamlet et al., 2002; Karamouz and Araghinejad, 2008)~~(Chiew and McMahon, 2002; Hamlet et al., 2002; Karamouz and Araghinejad, 2008). Extreme droughts, which occurred in Western Europe in 1921 ~~(Duband et al., 2004)~~(Duband et al., 2004), 1949 ~~(Duband, 2010)~~(Duband, 2010), 1976 ~~(Gazelle, 1979)~~(Gazelle, 1979) ~~and more recently in 2003 (Moreau, 2004), 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~~ 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 ~~(Feyen and Dankers, 2009; Manoha et al., 2008; Svensson et al., 2005)~~(Svensson et al., 2005; Manoha et al., 2008; Feyen and Dankers, 2009). Operational tools to forecast river low flows are still ~~quite~~ limited in many basins and much less developed than those dedicated to flood forecasting.

In spite of early attempts to develop models ~~(Singh and Stall, 1971; Riggs, 1953; Popov, 1964)~~(Riggs, 1953; Bernier, 1964; Popov, 1964; Singh and Stall, 1971; Larras, 1972; Oberlin and Michel 1978), low-flow forecasting has received only limited attention in the literature compared to flood forecasting. 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. ~~Note that we will not investigate here seasonal forecasting with typical forecast horizons of several months (Singla et al., 2012) and the possible role played by teleconnections (Chiew and McMahon, 2002; Rutten et al., 2008; Céron et al., 2010; Mosley, 2000).~~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

~~Hydrological models are essential tools~~Several simple modelling approaches have been proposed for low-flow forecasting~~. The first models to be used for low-flow forecasting included~~, including linear ARMA-type models, propagation models and recession curves ~~(Stravs and Brilly, 2007; Rivera-Ramirez et al., 2002; Girard, 1977; Yates and Snyder, 1975)~~(Lefèvre, 1974; Yates and Snyder, 1975; Avalos Lingan, 1976; Guilbot et al., 1976; Girard, 1977; Miquel and Roche, 1985; Rivera-Ramirez et al., 2002; Stravs and Brilly, 2007). ~~Campolo et al. (1999)~~Campolo et al. (1999) also proposed a neural network modelling approach.

~~However, t~~These methods generally make the assumption of no-rainfall future conditions, which is the most pessimistic case, but often a not entirely realistic one when lead times of a few weeks are considered. To make more ~~informed~~ reliable forecasts and extend to longer lead times, it is necessary to account for future meteorological conditions~~. Due to~~ and rainfall-runoff models are thus much relevant for low-flow forecasting. To account for the uncertainty in ~~these~~the future conditions (mainly in terms of temperature and precipitation), the typical methodology ~~used to issue~~

consists in simulating an ensemble of low-flow forecasts ~~is to feed~~(similar to ensemble flood forecasts), using a hydrological model ~~with~~fed by an ensemble of meteorological scenarios ~~describing the range of likely future conditions, and to~~. These forecasts are then statistically ~~analyse model outputs~~analysed for the target time period ~~(see e.g. Demirel et al., 2013b; Perrin et al., 2001; Garçon et al., 1999)~~(see e.g. Garçon et al., 1999; Perrin et al., 2001; Demirel et al., 2013b)~~. Rainfall-runoff models are therefore relevant for low-flow forecasting.~~.

~~1.3 Experience in low-flow forecasting in France~~

~~In France, the first initiatives to develop models for operational low-flow forecasting date back to the 1960s and 1970s, with the use of simple methods based on the statistical analysis of flow characteristics and recession curves (Bernier, 1964; Oberlin and Michel 1978; Larras, 1972). This coincided with the increase in hydroelectricity production capacities in mountainous regions and the development of a dense network of nuclear power plants in lowland areas, which needed reliable cooling water. In this perspective, investigations on low flows were made to develop strategies for the management of artificial reservoirs for low-flow augmentation (Lefèvre, 1974; Miquel and Roche, 1985). These authors applied linear models based on upstream information on the Loire and Seine basins. Avalos Lingan (1976) and Guilbot et al. (1976) compared several simple linear or recession-curve methods on the Oise basin (a tributary of the Seine) and also mentioned the possible use of conceptual rainfall-runoff models to overcome the limitations of the simple regression-based methods.~~

~~Among the first attempts to use conceptual models for river low-flow forecasting, CTGREF (1977) developed a simple storage-type model on the Durance basin to improve irrigation water management in low-flow conditions. This model accounted for snow influence on this basin. The French Geological Survey (BRGM) first worked on aquifer level forecasts (Thiéry, 1988b, 1982). 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). Moreover, 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, 2006b) also developed a platform for low-flow analysis and forecasting based on a conceptual hydrological model and implemented it on north-eastern 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-Modcou modelling suite running throughout France in operational conditions.~~

In France, among the first attempts to use conceptual models for low-flow forecasting, CTGREF (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; 2006b) also developed a platform for low-flow analysis and forecasting based on a conceptual hydrological model and implemented it in north-eastern 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-Modcou modelling suite running throughout France in operational conditions. One objective of this research will be to evaluate the strengths and weaknesses of these existing models.

~~1.4~~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. (2013a)~~Demirel et al. (2013a) 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)~~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)~~(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)~~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., 2013b)~~(see e.g. the discussion by Demirel et al., 2013b)~~.~~, as well as multi-model approaches (Georgakakos et al., 2004; Velazquez et al., 2011).

Our literature review showed that there are very few studies comparing the performance of existing hydrological models so that 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. (2013b)~~Demirel et al. (2013b), 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.

~~1.5~~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 ~~decided in 2010 to~~ launched in 2010 a comparative study for evaluating existing operational (or pre-operational) low-flow forecasting models on basins ~~within~~ 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., 2013)~~(Andréassian et al., 2009; Gupta et al., 2013). 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)~~Beven and Young (2013), 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 ~~study~~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 provides a discussion of the main lessons and perspectives of this work.

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)~~(1975, 1986, 1992) was among the first to organize such experiments to evaluate model running for simulation, snowmelt or flood forecasting purposes. More recently, the DMIP experiments ~~(Smith et al., 2012; Smith et al., 2004)~~(Smith et al., 2004; Smith et al., 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 ~~spread~~ 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 human influences considered negligible in low-flow conditions; (3) a sufficient number of stations to reach general conclusions, but not too many to keep tests feasible for all participants. Fourteen of these catchments are part of the national low-flow reference network of near-natural catchments established by Giuntoli et al. (2013).

The catchment set is well distributed over France (see Figure 1), with hydrological regimes ranging from oceanic to Mediterranean. Table 1 lists the set of 21 catchments, showing catchment sizes ranging from 379 km² to 4316 km², median elevations ranging from 70 m to 1020 m and streamflow data covering periods ranging from 36 to 97 years.

2.1.2 Data

Daily streamflow records were retrieved from the French HYDRO database (www.hydro.eaufrance.fr). Daily precipitation, temperature and potential evapotranspiration (PE) data originate from the gridded (8 × 8 ~~-~~km) SAFRAN climate reanalysis developed by Météo-France ~~(Vidal et al., 2010).~~(Vidal et al., 2010a). PE was computed using the Penman-Monteith formula ~~(Monteith, 1965; Penman, 1948).~~(Penman, 1948; Monteith, 1965). The climatic series are continuously available on the 1959–2010 period over France. To treat all catchments as uniformly as possible in the tests, the common 1974–2009 period was selected for model testing. This period includes severe low-flow conditions (e.g. in summers 1976, 1989, 2003 and 2005).

Table 2 displays the ranges of climate characteristics of the catchment set. Climate conditions in France are quite variable in terms of mean annual precipitation, PE and streamflow. Variations in rainfall, PE and streamflow can also be significant between years, as shown by interannual variability, especially for streamflow. On average, 36% of rainfall becomes runoff for the catchment set, but this ratio varies ~~yield can vary~~ between 21% and 76%.

2.1.3 Characteristics of low flows

In France, low flows mostly occur in summer and at the beginning of autumn (except in snow-influenced conditions). However, the duration and intensity of low flows as well as the beginning and ending dates of low-flow periods vary substantially between years and catchments.

For the operational purposes, low-flow periods are often defined using a streamflow threshold, under which specific management measures must be taken to face water shortages. In this study, it was difficult to choose operational low-flow thresholds, because they do not represent the same level of severity in all catchments since managers did not use the same methods to define these thresholds in all catchments. ~~So we~~We thus considered low flows as periods when observed streamflow falls below the threshold defined by the 80th percentiles of the flow duration curve, noted Q_{80}, i.e. the flow exceeded 80% of the time. ~~This was chosen as a compromise between focusing on specific low-flow periods and having a sufficient number of low-flow situations to obtain robust and significant model evaluations.~~This was chosen as a compromise between focusing on specific low-flow periods and having a sufficient number 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)~~(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)~~(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

Table 3 shows the five models used in this study. Four of them (GARD, GR6J, MORD and PRES) are lumped storage-type models, with various conceptualizations 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). ~~Each participant was free to choose the optimization method best suited to parameter estimation. Note that SIM was the only model where~~Participant were free to choose the optimization method best suited to parameter estimation, but all opted for automatic calibration, using either global (SCE-UA 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 flows (e.g. Nash-Sutcliffe (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) criteria calculated on Q and 1/Q for MORD and GR6J, see Table 3). Even though this variety of choices may make the comparison of results less straightforward, this was a mean 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 section 2.3.3).

The models were fed with the same meteorological inputs derived from SAFRAN. For the lumped models, the 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)~~(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 initialize 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 Figure 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 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 they 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 P, PE and T. The following procedure was applied. For a given catchment, let us consider that *N* years of meteorological inputs are available. One wishes to make a forecast on a calendar day *t* of a year *Y* within the test period, i.e. to forecast flows between calendar days *t*+1 and *t*+L. The observed meteorological data available between days *t*+1 and *t*+L in the years 1,…,*Y*−1,*Y*+1,…,N (i.e. *N*−1 scenarios) were used as input scenarios to the model, considering that they are likely meteorological conditions for this period of the year. Here, 51 years (1959–2009) of daily climate data from the SAFRAN reanalysis were available, thus 50 scenarios (for rainfall, temperature and PE) could be used each time. ~~We assumed that this number of scenarios was sufficient for a good representation of the variability of possible future climate~~

~~conditions. Obviously, such scenarios are likely to be less accurate than actual ensemble forecasts from meteorological models, at least for short to medium lead times.~~ The observed meteorological inputs of year *Y* were used as a control forecast, to estimate forecasting efficiency in the idealized case of perfect foreknowledge of future meteorological conditions.

Following this procedure, models were run to issue an ensemble of 50 streamflow forecasts for each day *t*, over a time window of 90 days (from *t*+1 to *t*+90). Due to computing time constraints, SIM only provided forecasts every 5 days, from ~~*d*~~*t*+1 to ~~*d*~~*t*+30 (and ~~*d*~~*t*+90 for each first day of the month), over a period limited to May 1st to October 26th (the low-flow period) and on the second validation period only (1992–2009).

In this study, we assumed that this number of scenarios (50) was sufficient for a good representation of the variability of possible future climate conditions. Obviously, historical scenarios are likely to be less accurate than actual ensemble forecasts from meteorological models, at least for short to medium lead times, since the spread of these scenarios may be too large for short lead-times. However, the catchment response to meteorological inputs is much more smoothed in low-flow than in high-flow conditions, which makes the catchment less sensitive to the spread of the ensemble. This approach may also find some limitations for forecasting the most extreme low-flow events, since most scenarios from the historical archive are likely to be wetter than the conditions actually observed for these extreme events. This can result in an overestimation of low flows forecasted by the models. In operational conditions, adding a “no-rainfall” scenario to the historical ones, i.e. running the model in pure recession, may be a way to overcome this problem and have an estimate of the “worst” low-flow forecast.

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 (Singla et al., 2012) and the possible role played by teleconnections (Mosley, 2000; Chiew and McMahon, 2002; Rutten et al., 2008; Céron et al., 2010).

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)~~(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 ~~(Perrin et al., 2006; Seibert, 2001)~~(Seibert, 2001; Perrin et al., 2006), results in simulation mode were compared to the daily average streamflow curve (noted DAQ). This benchmark was advocated by ~~Martinec and Rango (1989)~~Martinec and Rango (1989). In forecasting mode, the probabilistic forecasts were compared to a benchmark describing the streamflow natural variability (noted 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 lie 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. section 4.2) but it will not impact the evaluation of their respective merits when placed in a comparative framework.

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 choose 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 interval]-∞;1] (1 being the optimum value) to ease interpretation.

Note that the forecasting results presented hereafter were limited in order to adapt to the availability of streamflow forecasts from SIM.

2.3.5 Presentation of results

~~Since t~~The project produced a very large number of results, and it is obviously not possible to detail them all here. Instead, we chose to present summary evaluations using tables and graphical representations. Radial plots, as exemplified in Figure 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. We also used performance maps to investigate the possible regional trend in results. These maps were drawn for three criteria only (~~C2MiQ~~$C2M_i$, CSI and Vdef in simulation; $RMSE_{ut}$, ~~BSutvig~~BS_{vig} and Vdef in forecasting). They were found to be complementary, thus providing an overall picture of model performance in low-flow conditions.

Mis en fo

~~Last, two catchments were selected to illustrate the results using hydrographs: the Meuse River at St-Mihiel (B2220010) and the Orge River at Morsang-sur-Orge (H4252010). These two catchments are quite different in terms of mean annual precipitation (937 mm and 656 mm, respectively) and mean runoff yield (41% and 21%) and in terms of low-flow conditions described by BFI values (25.5% and 65.7%) and the Q_{90}/Q_{50} ratio (0.21 and 0.57).~~

Mis en fo
Gras, Cou

3 RESULTS

3.1 Simulation mode

Figure 4 summarizes 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 (~~C2MiQ, C2MQ~~C2M$_i$, 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 (~~C2MiQ~~C2M$_i$, 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 ~~C2MiQ~~C2M$_i$ 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 ~~C2MiQ~~C2M$_i$, 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 ~~C2MiQ~~C2M$_i$ criterion. Interestingly, most models tend to underestimate the volume deficit (Vdef < 1), i.e. they tend to overestimate low flows below the Q_{80} 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.

~~For~~Table 6 presents the results based on the mean performance in validation on the 21 catchments. An integrated criterion provides an ~~overall evaluation~~overview of the ~~models, we ranked them by decreasing performance for each~~overall performances. It is based on the transformed values of the ~~11 criteria and computed their mean ranks for the~~ nine criteria directly related to low flows (i.e. not considering ~~C2MQ~~C2M and KGE) between 0 and ~~KGEQ). Table 6 presents the results based on the mean performance in validation on the 21 catchments.~~ 1 (where 1 is the best performance), and represents the blue area of Figure 4. It can be observed that GARD performs best for four criteria, PRES ~~for four,~~and MORD for ~~two~~three and GR6J with one. PRES ~~appears~~performs best the most consistently ~~ranked~~ among the best models on ~~average~~the integrated criterion, followed by ~~GARD,~~ GR6J, GARD and MORD, ~~which~~ even if these four models are quite similar, and then SIM. 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_{80} threshold well simulated seven times out of ten). SIM performs less satisfactorily than the four other models for 9 out of 11 criteria, but all the models greatly improve performances relative to the benchmark NVQ (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 (~~C2MQ~~C2M and ~~KGEQ~~KGE): the way PRES was implemented within 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 ~~characteristics than on the model itself. We~~than on the model itself. Figure 6 presents the performance variability between models against the performance variability between catchments for the 11 selected criteria. For each criterion, standard deviation of performances for a model is calculated for all catchments, the average standard deviation for the five models represents the variability of performances between models. For each criterion, standard deviation of performances for a catchment is calculated for all models, the average standard deviation for the 21 catchments represents the variability of performances between catchments. The graph shows that performance varies more between catchments than between models for all criteria (except for $C2M_i$), 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), ~~but it did not show significant trends, as illustrated in~~ 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 Figure 7~~.~~, the relation did not show significant trends.

Mis en fo

Mis en fo

3.2 Forecasting mode

Figure 8 and Figure 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 a forecast interval as possible (excessively spaced forecasts do not contribute information). Performance even becomes close to the benchmark performance NVQ, 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). Assimilation method of GARD (reservoir updating) seems to be 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 error correction at previous time-step. The quantile/quantile post-correction method seems less efficient than streamflow assimilation methods, as performances are not improved for a few criteria ($RMSE_{ut}$, POD, CSI and sharpness (Sharp)).

As in simulation mode, model performance based on several criteria strongly varies among the catchments. Figure 10 and Figure 11 show the performance maps on validation period 2 for $RMSE_{ut}$ (normalized by mean flow under the Q_{80} threshold), ~~BSutvig~~BS_{vig} and Vdef, and for each model on the 21 catchments, for forecasting 7-day (Figure 10) and 30-day (Figure 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 Vascoeuil 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 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 ~~BSutvig~~BS$_{vig}$ or Vdef, compared to the other catchments.

Mis en fo

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. The analyses to link model performance to low-flow indices (BFI or Q_{90}/Q_{50} ratio) did not show significant trends, as had already be shown in simulation mode in Figure 7.

Table 7 presents the ~~rank~~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 ~~is~~ and GARD perform best the most consistently ~~ranked~~ among the best models on average~~, followed~~ as shown by ~~GARD~~the integrated criterion. Then come PRES and MORD~~ which are quite similar, and ~~, 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 (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, 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 result in severe drought situation. The use of other scenarios based on meteorological forecast may help limiting this problem, but further test would be needed to check this point.

~~3.3 Illustration of two case studies~~

~~Here, we present the results in simulation and forecasting modes for two catchments: the Meuse River at St-Mihiel, where the models perform well, and the Orge River at Morsang-sur-Orge, where they perform less satisfactorily.~~

~~Figure 10 shows the observed and simulated hydrographs in the logarithmic scale for two years where severe low-flow events occurred: 1976 and 1996. For the Meuse at St-Mihiel, GARD, GR6J, MORD and PRES simulated the low flows well, even if they overestimate streamflows from October to December in 1976 and in August in 1996. SIM does not adequately reproduce the low-flow dynamic with quite erratic streamflow simulations. For the Orge River at Morsang-sur-Orge, the models tend to substantially overestimate low flows for the two years, except SIM, which accurately simulates the low-flow event in 1996. Interestingly, this catchment benefits from a detailed simulation of the aquifer within SIM while most others do not.~~

~~Figure 11 and Figure 12 present the observed and forecasted hydrographs in the logarithmic scale for the Meuse River at St-Mihiel in 2003 and the Orge River at-Morsang-sur-Orge in 1996 (the most severe low-flow events for validation period 2). Forecast ensembles over the next 15 days are represented for a forecast produced every 20 days, together with the control run in red (i.e. streamflow forecast obtained with a posteriori observed P, PE and T as the future scenario). For the Meuse River, GARD and GR6J tend to be less dispersed than the other models. The control run shows that SIM is overly reactive to precipitation, while PRES tends to underestimate streamflow. Therefore, the ensemble forecasted by PRES surrounds the observation well, while MORD and SIM tend to overestimate streamflow when lead time increases. In these cases of severe low flows, the added value of models compared to the benchmark is clear: given its definition, the benchmark~~

~~consistently overestimates severe low flows, whereas models issue forecast ensembles that are better centred on observation and less dispersed.~~

~~For the Orge River, the low-flow event is poorly forecasted by all models, with a general tendency to overestimation. This is confirmed by the control run, especially for GARD, GR6J and MORD. SIM and PRES surround the observation better and forecast low flows better despite a few missed forecasts for PRES (July and August). In this case and for other low-flow events for the Orge River, the added value of hydrological models compared to the benchmark is limited.~~

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 DISCUSSION

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 0 showed the results of the comparison between hydrological models in simulation and forecasting modes. The mean model ranks show several differences between simulation (Table 6) and forecasting (Table 7) modes. This is further illustrated in Figure 12, which 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 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 ~~modeling objectives.~~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~~, which assimilate~~ (streamflow assimilation and/or ~~correct model outputs (~~output correction methods, see Table 3~~). However, given the variety of assimilation and correction methods applied in this study, it is difficult to conclude on the relative advantages of each of them and more systematic tests would be needed.~~). Figure 13 presents, for each model, the performance difference in CSI for each catchment between 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 indicates 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)~~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 of 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 ~~little~~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 (RMSE$_{ut}$, 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 characteristics (BFI or Q_{90}/Q_{50} ratio). It was also noted that UFL estimates vary between models and/or test periods~~ (see Figure 8).~~. For example, for the

Mis en fo

Briance River at Condat-sur-Vienne, the best mean UFL is provided by PRES and reaches 60 days for validation period 2 versus 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; Velazquez et al., 2011)~~Palmer et al., 2004; Velazquez et al., 2011). Here we chose to focus on three simple methods:

1. Average multi-model forecast (noted 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 model outputs using weights based on model performance. The model weight W_m given to each model is:

$$W_m = \frac{Crit_m}{\sum_{m=1}^{M} Crit_m} \quad \text{Eq. (1)}$$

where *m* is the hydrological model, *M* the number of hydrological models, *Crit* the value of the criterion on the calibration period. Better performing models obtain higher weights. In ensemble forecasting mode, each member of the multi-model corresponds to the 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 (noted VMM): The third method tested is inspired from ~~Loumagne et al. (1995)~~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:

$$W_{m,d} = \frac{\sum_{s=d-p}^{d} \sqrt{(Qfor_{m,s} - Qobs_s)^2}}{\sum_{m=1}^{M} \sum_{s=d-p}^{d} \sqrt{(Qfor_{m,s} - Qobs_s)^2}} \qquad \text{Eq. (2)}$$

where *m* is the hydrological model, *M* the number of hydrological models, *d* the day when the forecast is issued, $Qfor_{m,s}$ the streamflow forecasted by model m at date *s*−1 for *s*, $Qobs_s$ the observed streamflow at date *s*, *p* 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 ~~Figure 16~~Figure 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)~~Loumagne et al. (1995), other approaches accounting better 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~~. The relative gain compared to~~, 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)~~ is very high~~ 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 criteria, 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 remain 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. Here, the five tested models are runoff-rainfall models. Demirel and Booij (2009) compared three low-flow forecast models (a multivariate ARMAX model, a linear regression model and an Artificial Neural Network (ANN) model) for the Meuse River. Results are difficult to compare but comparing ANN and hydrological rainfall-runoff models should be interesting in low-flow forecasting.

As far as perspectives are concerned, we would like to mention (i) that 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.

6 ACKNOWLEDGMENTS

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 for their comments, which helped improving the manuscript.

7 REFERENCES

Abrahart, R. J., and See, L.: Multi-model data fusion for river flow forecasting: an evaluation of six alternative methods based on two contrasting catchments, Hydrology and Earth System Sciences, 6(4), 655-670, 2002.

Andréassian, V., Bergström, S., Chahinian, N., Duan, Q., Gusev, Y. M., Littlewood, I., Mathevet, T., Michel , C., Montanari, A., Moretti, G., Moussa, R., Nasonova, O. N., O'Connor, K. M., Paquet, E., Perrin, C., Rousseau, A., Schaake, J., Wagener, T., and Xie, Z.: Catalogue of the models used in MOPEX 2004/2005, in: Large sample basin experiments for hydrological model parameterization: Results of the Model Parameter Experiment - MOPEX, edited by: Andréassian, V., Hall, A., Chahinian, N., and Schaake, J., IAHS Publication 307, 41-93, 2006.

Andréassian, V., Perrin, C., Berthet, L., Le Moine, N., Lerat, J., Loumagne, C., Oudin, L., Mathevet, T., Ramos, M. H., and Valéry, A.: Crash tests for a standardized evaluation of hydrological models, Hydrol. Earth. Syst. Sci., 13, 1757-1764, ~~doi:~~1710.5194/hess-1713-1757-2009, 2009.

Avalos Lingan, R. F.: Essais méthodologiques de prévision des débits d'étiages : application au bassin de l'Oise à Sempigny (Methodological attempts for low-flow forecasting: application on the Oise River basin at Sempigny) Université des Sciences et Techniques du Languedoc, 145 pp., 1976.

Bernier, J.: La prévision statistique de bas débits (Statistical forecast of low flows), in: IAHS Publication n°63, 1964, 340-351,

Berthier, C. H.: Quantification des incertitudes des débits calculés par un modèle pluie-débit empirique, Université Paris Sud XI, 52 pp., 2005.

Beven, K., and Young, P.: A guide to good practice in modeling semantics for authors and referees, Water Resour. Res., 49(8), 1-7, 10.1002/wrcr.20393, 2013.

Bousquet, S., Gaume, E., and Lancelot, B.: Évaluation des enjeux socio-économiques liés aux étiages de la Seine (Evaluation of the socio-economical impacts of the Seine river low water periods), La Houille Blanche(3), 145-149, 2003.

Brier, G. W.: Verification of forecasts expressed in terms of probability, Mon. Weather Rev., 78(1), 1-3, 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2, 1950.

Campolo, M., Soldati, A., and Andreussi, P.: Forecasting river flow rate during low-flow periods using neural networks, Water Resour. Res., 35(11), 3547-3552, 1999.

Céron, J. P., Tanguy, G., Franchistéguy, L., Martin, E., Regimbeau, F., and Vidal, J. P.: Hydrological seasonal forecast over France: feasibility and prospects, Atmos. Sci. Lett., 11(2), 78-82, 2010.

CGDD: Les prélèvements d’eau en France en 2009 et leurs évolutions depuis dix ans (Water withdrawals in France in 2009 and their evolution over the last ten years), Commissariat général au développement durable, 8, 2012.

Chiew, F. H. S., and McMahon, T. A.: Global ENSO-streamflow teleconnection, streamflow forecasting and interannual variability, Hydrol. Sci. J.-J. Sci. Hydrol., 47(3), 505-522, 10.1080/02626660209492950, 2002.

CTGREF: Prévision d'étiages pour la gestion de réserves agricoles du barrage de Serre-Ponçon (Low-flow forecasting for the management of the storage for agriculture in the Serre-Ponçon reservoir), CTGREF Aix-en-Provence, SRAE PACA, Le Tholonet, Rapport d'étude, 56, 1977.

Demirel, M. C., and Booij, M. J.: Identification of an appropriate low flow forecast model for the Meuse River IAHS Publ. 331, 296-303, 2009.

Demirel, M. C., Booij, M. J., and Hoekstra, A. Y.: Identification of appropriate lags and temporal resolutions for low flow indicators in the River Rhine to forecast low flows with different lead times, Hydrol. Processes, 27(19), 2742-2758, 10.1002/hyp.9402, 2013a.

Demirel, M. C., Booij, M. J., and Hoekstra, A. Y.: Effect of different uncertainty sources on the skill of 10 day ensemble low flow forecasts for two hydrological models, Water Resour. Res., 49(7), 4035–4053, 10.1002/wrcr.20294, 2013b.

Duband, D., Schoeneich, P., and Stanescu, V. A.: Exemple de l'étiage 1921 en Europe (Italie, France, Roumanie, Suisse,…) : climatologie et hydrologie (The example of 1921 drought in Europe (Italy, France, Rumania, Swiss...): climatology and hydrology), La Houille Blanche(5), 18-29, 2004.

Duband, D.: Rétrospective hydro-pluviométrique des étiages rares depuis 140 ans, dans l'ouest de l'Europe (bassins Loire, Seine, Rhin, Rhône, Pô) (Rainfall-run-off retrospective of extremes droughts since 1860 in Europe (Germany, Italia, France, Rumania, Spain, Switzerland)), La Houille Blanche(4), 51-59, 2010.

Feyen, L., and Dankers, R.: Impact of global warming on streamflow drought in Europe, J. Geophys. Res.-Atmos., 114, D17116, 10.1029/2008JD011438, 2009.

Franz, K. J., and Hogue, T. S.: Evaluating uncertainty estimates in hydrologic models: borrowing measures from the forecast verification community, Hydrol. Earth Syst. Sci., 15(11), 3367-3382, 10.5194/hess-15-3367-2011, 2011.

Garavaglia, F.: Méthode SCHADEX de prédétermination des crues extrêmes (SCHADEX method of determination of extreme floods), Université de Grenoble, 2011.

García de Jalón, D.: The Spanish Experience in Determining Minimum Flow Regimes in Regulated Streams, Canadian Water Resources Journal, 28(2), 185-198, 10.4296/cwrj2802185, 2003.

Garçon, R.: Prévision opérationnelle des apports de la Durance à Serre-Ponçon à l'aide du modèle MORDOR. Bilan de l'année 1994-1995 (Operational forecast of the inputs from the Durance in Serre-Ponçon reservoir using the MORDOR model. Assessment of the year 1994-1995) La Houille Blanche, 5, 71-76, 1996.

Garçon, R., Carre, C., and Lyaudet, P.: An example of forecasting and operating simulation of low water flows, Houille Blanche-Revue Internationale De L Eau, 54(6), 37-42, 1999.

Gazelle, F.: Les répercussions des étiages de 1976 sur l'hydroélectricité languedocienne (Impacts of the 1976 low flows on the hydroelectricty in Languedoc), La Houille Blanche(1), 59-61, 1979.

Georgakakos, K. P., Seo, D. J., Gupta, H., Schaake, J., and Butts, M. B.: Towards the characterization of streamflow simulation uncertainty through multimodel ensembles, J. Hydrol., 298(1-4), 222-241, 2004.

Girard, G.: Etude de l'efficacité relative et du domaine d'application de différents modèles hydrologiques pour prévoir les étiages (Study of the relative efficiency and the application domain of various models for low-flow forecasting), ORSTOM, Paris, Contrat de recherche 75/98, 64, 1977.

Giuntoli, I., Renard, B., Vidal, J. P., and Bard, A.: Low flows in France and their relationship to large-scale climate indices, J. Hydrol., 482(0), 105-118, 10.1016/j.jhydrol.2012.12.038, 2013.

Guilbot, A., Masson, J.-M., Bédiot, G., and Ducastelle, C.: Essai de prévision des étiages de l'Oise à Sempigny (Attempts of low-flow forecasts on the Oise River at Sempigny), La Houille Blanche(6-7), 549-568, 1976.

Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F.: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, J. Hydrol., 377(1-2), 80-91, 2009.

Gupta, H. V., Perrin, C., Kumar, R., Blöschl, G., Clark, M., Montanari, A., and Andréassian, V.: Large-sample hydrology: a need to balance depth with breadth, Hydrol. Earth Syst. Sci. Discuss., 10(7), 9147-9189, 10.5194/hessd-10-9147-2013, 2013.

Mis en fo

Mis en fo

Mis en fo
Police :(Pa
Français (
et la gram

Mis en fo

Gustard, A., and Demuth, S.: Manual on low-flow estimation and prediction, Geneva, Operational Hydrology report No. 50, WMO-No 1029, 2009.

Habets, F., Boone, A., Champeaux, J. L., Etchevers, P., Franchisteguy, L., Leblois, E., Ledoux, E., Le Moigne, P., Martin, E., Morel, S., Noilhan, J., Quintana Seguí, P., Rousset-Regimbeau, F., and Viennot, P.: The SAFRAN-ISBA-MODCOU hydrometeorological model applied over France, J. Geophys. Res.-Atmos., 113(D6), D06113, 10.1029/2007jd008548, 2008.

Hamlet, A. F., Huppert, D., and Lettenmaier, D. P.: Economic value of long-lead streamflow forecasts for Columbia River hydropower, J. Water Resour. Plan. Manage.-ASCE, 128(2), 91-101, 10.1061/(~~asce~~ASCE)0733-9496(~~2001~~2002)128:2(91), 2002.

Henny, F.: Gestion de la ressource en eau en Alsace : Révision des arrêtés cadre « sécheresse » (Water ressources management in Alsace : update of the drought framework order), ENGEES, Strasbourg, 118 pp., 2010.

Karamouz, M., and Araghinejad, S.: Drought mitigation through long-term operation of reservoirs: Case study, Journal of Irrigation and Drainage Engineering-Asce, 134(4), 471-478, 10.1061/(asce)0733-9437(2008)134:4(471), 2008.

Klemeš, V.: Operational testing of hydrological simulation models, Hydrol. Sci. J., 31(1), 13-24, 1986.

Lang, C., Freyermuth, A., Gille, E., and François, D.: Le dispositif PRESAGES (PREvisions et Simulations pour l'Annonce et la Gestion des Etiages Sévères) : des outils pour évaluer et prévoir les étiages (The PRESAGES system (Forecast and simulation for warning and management of severe low flows): tools for evaluating and predicting low-flows), Géocarrefour, 81(1), 15-24, 2006a.

Lang, C., Gille, E., Francois, D., and Auer, J.-C.: PRESAGES: A collection of tools for predicting low flows, in: IAHS Publication, Climate Variability and Change - Hydrological Impacts, 2006b, WOS:000249093700024, 145-150,

Larras, J.: Prévision et prédétermination des étiages et des crues (Forecast and estimation of low flows and floods), Collection du BCEOM, Eyrolles, Paris, 159 pp., 1972.

Lefèvre, J.: Le soutien des étiages en Loire à l'aide de réservoirs situés dans le haut bassin.Application au barrage de Naussac (Augmenting low-flows in the Loire River using reservoirs in the upper basin. Application to the Naussac dam), La Houille Blanche(4/5), 271-278, 1974.

Loumagne, C., Vidal, J. J., Feliu, C., Torterotot, J. P., and Roche, P. A.: Procédure de décision multimodèle pour une prévision des crues en temps réel. Application au bassin supérieur de la Garonne (A multimodel weighting decision process for real time flood forecasting: Application to the upper Garonne watershed), Revue des Sciences de l'Eau, 8(4), 539-561, 1995.

Lvovitch, M. I.: Hydrologic budget of continents and estimate of the balance of global fresh water resources, Soviet Hydrology(N 4), 349-360, 1972.

Manoha, B., Hendrickx, F., Dupeyrat, A., Bertier, C., and Parey, S.: Impact des évolutions climatiques sur les activités d'EDF (projet impec) (Climate change impact on the activities of Electricite de France), La Houille Blanche(2), 55-60, 2008.

Martinec, J., and Rango, A.: Merits of statistical criteria for the performance of hydrological models, Water Resources Bulletin, 25(2), 421-432, 1989.

Mathevet, T., Michel, C., Andréassian, V., and Perrin, C.: A bounded version of the Nash-Sutcliffe criterion for better model assessment on large sets of basins, in: Large sample basin experiments for hydrological model parameterisation: Results of the Model Parameter Experiment - MOPEX, edited by: Andréassian, V., Hall, A., Chahinian, N., and Schaake, J., IAHS Red Books Series n°307, 211-219, 2006.

Mathevet, T., Perret, C., Garçon, R., Periers, P., Goutx, D., Gibey, J.-M., Oudin, R., Xhaard, H., and Roy, J.-L.: Modèles de prévision et prise de décision pour le soutien d'étiage de la Loire (Drought forecasts and decision support on Loire river), La Houille Blanche(5), 40-51, 2010.

Miquel, J., and Roche, P. A.: La gestion d'un réservoir de soutien d'étiage peut-elle être optimale en cas de prévisions imparfaites? (Can an optimal release policy for a reservoir be obtained in the case of imperfect low flow forecasts?), in: IAHS Publication n° 147, 1985, 301-320,

Monteith, J. L.: Evaporation and the environment, Proceedings of the XIXth Symposium of the Soc. for Exp. Biol., The state and Movement of water in living organisms, Swansea, 1965, 205-234, Moreau, F.: Gestion des étiages sévères : l'exemple de la Loire (Drought crisis management : Loire basin example), La Houille Blanche(4), 70-76, 2004.

Mosley, M. P.: Regional differences in the effects of El Niño and La Niña on low flows and floods, Hydrol. Sci. J., 45(2), 249-267, 10.1080/02626660009492323, 2000.

Nash, J. E., and Sutcliffe, J. V.: River flow forecasting through conceptual models. Part I - A discussion of principles., J. Hydrol., 10(3), 282-290., 1970.

Oberlin, G., and Michel , C.: Eléments de méthodes de secours pour une prévision improvisée des étiages (Principles of an emergency method for an improvised forecast of low flows), Bulletin du BRGM, Section III(n° 3), 203-214, 1978.

Palmer, T. N., Alessandri, A., Andersen, U., Cantelaube, P., Davey, M., Delecluse, P., Deque, M., Diez, E., Doblas-Reyes, F. J., Feddersen, H., Graham, R., Gualdi, S., Gueremy, J. F., Hagedorn, R., Hoshen, M., Keenlyside, N., Latif, M., Lazar, A., Maisonnave, E., Marletto, V., Morse, A. P., Orfila, B., Rogel, P., Terres, J. M., and Thomson, M. C.: Development of a European multimodel ensemble system for seasonal-to-interannual prediction (DEMETER), Bulletin of the American Meteorological Society, 85(6), 853-872, 2004.

Paquet, E., Garavaglia, F., Garçon, R., and Gailhard, J.: The SCHADEX method: A semi-continuous rainfall–runoff simulation for extreme flood estimation, J. Hydrol., 495(0), 23-37, 10.1016/j.jhydrol.2013.04.045, 2013.

Penman, H. L.: Natural evaporation from open water, bare soil and grass, Proc. R. Soc. London, A193, 120-145, 1948.

Perrin, C., Michel, C., and Andréassian, V.: Long-term low flow forecasting for French rivers by continuous rainfall-runoff modelling, in: BHS Occasional Paper n° 13, Meeting of the British Hydrological Society on Continuous River Flow Simulation, Wallingford, UK, 5th July 2001, 2001, 21-29,

Perrin, C., Andréassian, V., and Michel, C.: Simple benchmark models as a basis for criteria of model efficiency, Archiv für Hydrobiologie Supplement 161/1-2, Large Rivers, 17(1-2), 221-244, ~~DOI:~~ 10.1127/lr/17/2006/221, 2006.

Pointet, T., Amraoui, N., Golaz, C., Mardhel, V., Negrel, P., Pennequin, D., and Pinault, J. L.: The contribution of groundwaters to the exceptional flood of the Somme River in 2001 - Observations, assumptions, modelling, Houille Blanche-Revue Internationale De L Eau(6), 112-122, 2003.

Popov, E. G.: Long-term river flow forecasting in the low-water period, in: IAHS Publication n°63, 1964, 63-67,

Pushpalatha, R., Perrin, C., Le Moine, N., Mathevet, T., and Andréassian, V.: A downward structural sensitivity analysis of hydrological models to improve low-flow simulation, J. Hydrol., 411(1-2), 66-76, ~~doi:~~10.1016/j.jhydrol.2011.1009.1034, 2011.

Pushpalatha, R., Perrin, C., Le Moine, N., and Andréassian, V.: A review of efficiency criteria suitable for evaluating low-flow simulations, J. Hydrol., 420-421, 171-182, ~~doi:~~ 10.1016/j.jhydrol.2011.11.055, 2012.

Pushpalatha, R.: Low-flow simulation and forecasting on French river basins: a hydrological modelling approach, AgroParisTech (Paris), Irstea (Antony), 230 pp., 2013.

Rassam, D. W.: A conceptual framework for incorporating surface-groundwater interactions into a river operation-planning model, Environmental Modelling & Software, 26(12), 1554-1567, 10.1016/j.envsoft.2011.07.019, 2011.

Riggs, H. C.: A method of forecasting low flow of streams, Transactions, American Geophysical Union, 34(3), 427-434, 1953.

Rivera-Ramirez, H. D., Warner, G. S., and Scatena, F. N.: Prediction of master recession curves and baseflow recessions in the Luquillo mountains of Puerto Rico, J. Am. Water Resour. Assoc., 38(3), 693-704, 10.1111/j.1752-1688.2002.tb00990.x, 2002.

Rutten, M., van de Giesen, N., Baptist, M., Icke, J., and Uijttewaal, W.: Seasonal forecast of cooling water problems in the River Rhine, Hydrol. Processes, 22(7), 1037-1045, 10.1002/hyp.6988, 2008.

Mis en fo
Police :(Pa
Vérifier l'o

Saunders, J. F., and Lewis, W. M.: Implications of climatic variability for regulatory low flows in the South Platte River basin, Colorado, J. Am. Water Resour. Assoc., 39(1), 33-45, 10.1111/j.1752-1688.2003.tb01559.x, 2003.

Schäfer, J. T.: The Critical Success Index as an Indicator of Warning Skill, Weather Forecast., 5(4), 570-575, 10.1175/1520-0434(1990)005<0570:TCSIAA>2.0.CO;2, 1990.

Seibert, J.: On the need for benchmarks in hydrological modelling, Hydrol. Processes, 15(6), 1063-1064, 2001.

Singh, K. P., and Stall, J. B.: Derivation of Base Flow Recession Curves and Parameters, Water Resour. Res., 7(2), 292-303, 10.1029/WR007i002p00292, 1971.

Singla, S., Céron, J. P., Martin, E., Regimbeau, F., Déqué, M., Habets, F., and Vidal, J. P.: Predictability of soil moisture and river flows over France for the spring season, Hydrology and Earth System Sciences, 16(1), 201-216, 2012.

Smith, M. B., Seo, D. J., Koren, V. I., Reed, S. M., Zhang, Z., Duan, Q., Moreda, F., and Cong, S.: The distributed model intercomparison project (DMIP): motivation and experiment design, J. Hydrol., 298(1-4), 4-26, 2004.

Smith, M. B., Koren, V., Reed, S., Zhang, Z., Zhang, Y., Moreda, F., Cui, Z., Mizukami, N., Anderson, E. A., and Cosgrove, B. A.: The distributed model intercomparison project - Phase 2: Motivation and design of the Oklahoma experiments, J. Hydrol., 418-419, 3-16, 2012.

Soubeyroux, J.-M., Vidal, J.-P., Baillon, M., Blanchard, M., Ceron, J.-P., Franchisteguy, L., Regimbeau, F., Martin, E., and Vincendon, J.-C.: Characterizing and forecasting droughts and low-flows in France with the Safran-Isba-Modcou hydrometeorological suite, Houille Blanche-Revue Internationale De L Eau(5), 30-39, 2010.

Staub, P. F.: Prévision d'étiage par modélisation hydrologique : mise au point d'une méthode d'évaluation (Low-flow forecasting through rainfall-runoff modelling: development of an evaluation method), Master Géo-Hydrosystèmes Continentaux en Europe, Université de Tours, UFR Sciences et Techniques, Cemagref (Antony), 88 pp., 2008.

Staudinger, M., Stahl, K., Seibert, J., Clark, M. P., and Tallaksen, L. M.: Comparison of hydrological model structures based on recession and low flow simulations, Hydrol. Earth Syst. Sci., 15(11), 3447-3459, 10.5194/hess-15-3447-2011, 2011.

Stewart, I. T., Cayan, D. R., and Dettinger, M. D.: Changes toward Earlier Streamflow Timing across Western North America, Journal of Climate, 18(8), 1136-1155, 10.1175/JCLI3321.1, 2005.

Stravs, L., and Brilly, M.: Development of a low-flow forecasting model using the M5 machine learning method, Hydrol. Sci. J.-J. Sci. Hydrol., 52(3), 466-477, 10.1623/hysj.52.3.466, 2007.

Svensson, C., Kundzewicz, W. Z., and Maurer, T.: Trend detection in river flow series: 2. Flood and low-flow index series / Détection de tendance dans des séries de débit fluvial: 2. Séries d'indices de crue et d'étiage, Hydrol. Sci. J., 50(5), null-824, 10.1623/hysj.2005.50.5.811, 2005.

Tajjar, M. H.: Modélisation de l'hydrodynamique des échanges nappe-rivière. Simulation d'une lâchure expérimentale en Seine en période d'étiage (Modelling the hydrodynamics of aquifer-river exchanges. Simulation of an experimental release in the Seine River in low-flow period), Ecole nationale supérieure des Mines de Paris, Paris, 183 pp., 1993.

Thiéry, D.: Utilisation d'un modèle global pour identifier sur un niveau piézométrique des influences multiples dues à diverses activités humaines (Use of a lumped model to identifyon a piezometric level multiple influences due to human activities), IAHS Publication n° 136, 71-77, 1982.

Thiéry, D.: Application à quatre bassins hydrologique des méthodes de prévision des étiages par convolution (Application of low-flow forecasting methods by convolution on four hydrological basins), BRGM, Orléans, 207, 1988a.

Thiéry, D.: Forecast of changes in piezometric levels by a lumped hydrological model, J. Hydrol., 97, 129-148, 1988b.

Thiéry, D.: Logiciel GARDENIA, version 8.2, Guide d'utilisation (GARDENIA software, version 8.2, User Guide), BRGM, BRGM/RP-62797-FR, 102, 2013.

Toth, Z., Talagrand, O., Candille, G., and Zhu, Y.: Probability and Ensemble Forecasts, in: Forecast Verification: a Practitioner's Guide in Atmospheric Science, edited by: Jolliffe, I. T., and Stephenson, D. B., John Wiley & Sons, Chichester, UK, 137–164, 2003.

Velazquez, J. A., Anctil, F., Ramos, M. H., and Perrin, C.: Can a multi-model approach improve hydrological ensemble forecasting? A study on 29 French catchments using 16 hydrological model structures, Advances in Geosciences, 29, 33-42, ~~doi:~~10.5194/adgeo-29-33-2011, 2011.

Vidal, J.-P., Martin, E., Franchisteguy, L., Baillon, M., and Soubeyroux, J.-M.: A 50-year high-resolution atmospheric reanalysis over France with the Safran system, International Journal of Climatology, 30(11), 1627-1644, 2010a.

Vidal, J. P., Martin, E., Franchisteguy, L., Habets, F., Soubeyroux, J. M., Blanchard, M., and Baillon, M.: Multilevel and multiscale drought reanalysis over France with the Safran-Isba-Modcou hydrometeorological suite, Hydrology and Earth System Sciences, 14(3), 459-478, ~~2010~~2010b.

WMO: Intercomparison of conceptual models used in operational hydrological forecasting, World Meteorological Organization, Geneva, Switzerland, Operational Hydrology Report n° 7, WMO n°429, 1975.

WMO: Intercomparison of models of snowmelt runoff, World Meteorological Organization, Geneva, Switzerland, Operational Hydrology Report n° 23, WMO n°646, 1986.

WMO: Simulated real-time intercomparison of hydrological models, World Meteorological Organization, Geneva, Switzerland, Operational Hydrology Report n° 38, WMO 779, 1992.

Yates, P., and Snyder, W. M.: Predicting recessions through convolution, Water Resour. Res., 11(3), 418-422, 10.1029/WR011i003p00418, 1975.

APPENDIX

Formulation of the numerical criteria selected for simulation evaluation

- **KGE**

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

Code de

~~$KGE = 1-\sqrt{(r-1)^2+(\alpha-1)^2+(\beta-1)^2}$~~ $KGE = 1-\sqrt{(r-1)^2+(\alpha-1)^2+(\beta-1)^2}$ **Eq. (A1)**

Code de

with *r* the correlation coefficient between observed and simulated flows, the ratio of simulated and observed flow standard deviations and β the model bias.

- **~~C2MQ~~C2M**

~~C2MQ~~C2M is a bounded version of the Nash-Sutcliffe efficiency index calculated on streamflow Q (NSE_Q), as proposed by ~~Mathevet et al. (2006)~~Mathevet et al. (2006)

~~$C2MQ = \frac{NSE_Q}{2-NSE_Q}$~~ $C2M = \frac{NSE_Q}{2-NSE_Q}$ **Eq. (A2)**

Code de

- **~~C2MiQ~~$C2M_i$**

This is similar to the previous criterion, but NSE is calculated on inverse flows to more strongly emphasize low flows, as proposed by ~~Pushpalatha et al. (2012)~~Pushpalatha et al. (2012)

- **$RMSE_{ut}$**

Mis en fo

$RMSE_{ut}$ is the root mean square error for flows under the low-flow threshold, normalized by the mean observed flow.

Mis en fo

~~$RMSEut = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Q_{sim_i} - Q_{obs_i}\right)^2}}{\frac{1}{n}\sum_{i=1}^{n} Q_{obs_i}}$~~ $RMSE_{ut} = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Q_{sim_i} - Q_{obs_i}\right)^2}}{\frac{1}{n}\sum_{i=1}^{n} Q_{obs_i}}$ **Eq. (A3)**

Code de

where $Qobs_i$ is the observed streamflow for day *i*, $Qsim_i$ the simulated streamflow for day *i*, and *n* the number of time steps on the validation period where $Qobs_i$ is less than the Q_{80} threshold.

- **Vdef**

Vdef is the ratio of simulated and observed flow deficits under the low-flow threshold:

~~$Vdef = \frac{\sum_{i=1}^{n}\max\left(0; Q_{threshold} - Q_{sim_i}\right)}{\sum_{i=1}^{n}\max\left(0; Q_{threshold} - Q_{obs_i}\right)}$~~ $Vdef = \frac{\sum_{i=1}^{n}\max\left(0; Q_{threshold} - Q_{sim_i}\right)}{\sum_{i=1}^{n}\max\left(0; Q_{threshold} - Q_{obs_i}\right)}$ **Eq. (A4)**

Code de

- **LFD**

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

~~$LFD = \frac{Duration_{sim}}{Duration_{obs}}$~~ $LFD = \frac{Duration_{sim}}{Duration_{obs}}$ **Eq. (A5)**

Code de

where $Duration_{sim}$ is the number of days where the $Qsim_i$ is less than the Q_{80} threshold on the validation period and $Duration_{obs}$ is the number of days where the $Qobs_i$ is less than the Q_{80} threshold on the validation period.

- **DatSt and DatEn**

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

~~$Dat = Date\_sim - Date\_obs$~~ $Dat = Date\_sim - Date\_obs$ **Eq. (A6)**

Code de

where *Date_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. *Date_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.

~~**● 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):~~

Vdef, LFD, and DatSt and DatEn have been adapted from the concept of "centre of mass" proposed by Stewart et al. (2005).

● 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):

~~$FAR=\frac{b}{a+b}$~~ $FAR=\frac{b}{a+b}$ **Eq. (A7)**

~~$POD=\frac{a}{a+c}$~~ $POD=\frac{a}{a+c}$ **Eq. (A8)**

~~$CSI=\frac{a}{a+b+c}$~~ $CSI=\frac{a}{a+b+c}$ **Eq. (A9)**

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.

Numerical criteria for forecasting evaluation

● $RMSE_{ut}$, 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.

● **~~Sharpness~~Sharp**

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

~~$Sharp = \frac{1}{n}\sum_{i=1}^{n} Q90_i - Q10_i$~~ $Sharp = \frac{1}{n}\sum_{i=1}^{n} Q90_i - Q10_i$ **Eq. (A10)**

Code de

where *n* is the number of time steps on the validation period where the *Qobs_i* 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*.

● **~~Reliability~~Cont_ratio**

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

~~$Cont\_ratio = \frac{n}{N}$~~ $Cont\_ratio = \frac{n}{N}$ **Eq. (A11)**

Code de

where *n* is the number of observed streamflows in the 80% forecasted confidence interval when the $Qobs_i$ is less than the Q_{80} threshold, and *N* the number of time steps where the $Qobs_i$ 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:

Code de

~~$BS=\frac{1}{n}\sum_{i=1}^{n}(y_i-o_i)^2$~~ $BS=\frac{1}{n}\sum_{i=1}^{n}(y_i-o_i)^2$ **Eq. (A12)**

Code de

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 ~~BSutvig~~BS_{vig}) or the crisis threshold (Q_{95} ~~threshold~~for BS_{cri}). *n* is the number of time steps where $Qobs_i$ is less than the Q_{50} threshold (~~BSutvig~~BS_{vig}) or the Q_{80} threshold (~~BSutcri~~BS_{cri}).

- **DRPS**

The Discrete Ranked Probability Score (DRPS) ~~(Toth et al., 2003)~~(Toth et al., 2003):

~~$DRPS=\frac{1}{Nthreshold}\sum_{k=1}^{Nthreshold}(BS_k)$~~ $DRPS=\frac{1}{Nthreshold}\sum_{k=1}^{Nthreshold}(BS_k)$ **Eq. (A13)**

Code de

where Nthreshold is the number of thresholds chosen (ten percentiles here, k=Q_{80}, Q_{82}, Q_{84}, … , Q_{96}, Q_{98}).

Table 1: Summary of the 21 selected catchments' characteristics.

N°	HYDRO Code	River at Station	Area (km²)	Median elevation (m)	Starting date for flow series	Ending date for flow series	Flow availability (years)
1	A1080330	Ill at Didenheim	657	390	01/11/1973	02/03/2010	36
2	B2220010	Meuse at Saint-Mihiel	2542	350	01/07/1968	03/01/2010	42
3	H2342020	Serein at Chablis	1121	309	01/08/1954	03/03/2010	56
4	H4252010	Orge at Morsang-sur-Orge	927	133	01/10/1967	07/03/2010	43
5	H7401010	Oise at Sempigny	4316	137	01/01/1955	02/03/2010	55
6	H8212010	Andelle at Vascoeuil	379	159	01/01/1973	27/02/2010	36
7	I5221010	Vire at Saint-Lô	868	159	01/01/1971	03/02/2010	39
8	J7483010	Seiche at Bruz	811	70	01/12/1967	11/03/2010	42
9	K1321810	Arroux at Etang-sur-Arroux	1798	431	01/11/1971	27/03/2010	39
10	K6402520	Sauldres at Salbris	1200	220	01/01/1971	28/03/2010	39
11	L0563010	Briance at Condat-sur-Vienne	597	386	01/01/1966	28/03/2010	44
12	L4411710	Petite Creuse at Fresselines	850	393	01/01/1958	28/03/2010	52
13	M0243010	Orne Saosnoise at Montbizot	510	103	01/12/1967	04/03/2010	43
14	M7112410	Sèvre Nantaise at Tiffauges	817	170	01/11/1967	04/03/2010	43
15	O0592510	Salat at Roquefort-sur-Garonne	1570	986	01/01/1913	22/03/2010	97
16	O3121010	Tarn at Montbrun	588	1020	01/01/1961	31/12/2009	38
17	Q5501010	Gave de Pau at Berenx	2575	916	01/07/1923	28/03/2010	87
18	S2242510	Eyre at Salle	1650	78	01/01/1967	19/03/2010	43
19	U4644010	Azergues at Lozanne	798	517	01/01/1965	28/03/2010	43
20	V4264010	Drôme at Saillans	936	936	01/01/1910	28/03/2010	46
21	Y4624010	Gapeau at Hyères	517	316	01/02/1961	01/03/2010	49

2050 Table 2: Percentiles of the distribution of certain climate and hydrological catchment 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

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

Table 3: Overview of the characteristics of the five models tested

Short name used here	GARD	GR6J	MORD	PRES	SIM
Full name	GARDENIA	GR6J	MORDOR	PRESAGES	SIM
Reference on model structure	Thiéry (2013)	Pushpalatha (2011, 2013)	~~Garçon (1999) ; Andreassian et al. (2006)~~ Garçon et al. (1999) ; Andréassian et al. (2006)	Lang et al. (2006a , 2006b)	
Type	Conceptual	Conceptual	Conceptual	Conceptual	Physically-based
Spatial distribution	Semi-distributed	Lumped	Lumped	Lumped	Distributed
Number of free-parameters	4 to 9 (+2 to 4 for snowmelt)	6 (+2 : snow routine)	11 (+4: snow routine)	7 (+3 : snow routine)	0
Calibration method	Automatic calibration ~~on observed streamflow and groundwater levels~~: Rosenbrock method	Automatic calibration: local research method (step by step)	Automatic calibration: Shuffled Complex Evolution Method and Pareto Front Exploitation	Automatic calibration: simplex method with multistart	No calibration
Calibration criteria	~~User selected :~~ ~~Nash, Nash(Log(flow))~~ ~~Nash(sqrt(flow))~~ ~~+ weighting on bias~~RMSE with ln(Q)	(~~KGEQ + KGEiQ~~KGE + KGE$_i$)/2	(~~KGEQ + KGEiQ~~KGE + KGE$_i$)/2	Nash–Sutcliffe with $Q^{0.2}$	
Post-correction method (simulation)	Not used	Not used	Not used	~~Empirical method (Berthier, 2005)~~Empirical method (Berthier, 2005)	Quantile/quantile post-treatment
Assimilation method (forecast)	When a flow discrepancy appears, the model tanks are updated proportionally to their variance	Correction based on error at first time step before forecast, with decreasing effect when lead time increases	Correction based on errors at previous time steps before forecast, with decreasing effect when lead time increases. No update of model stores.	Update of gravitary routing store	No assimilation method but a quantile/quantile post-treatment
Structure overview: production	Actual evapotranspiration is computed using a non-linear soil capacity. GW exchange is a proportion of the GW flow	A rainfall interception by PE, a non-linear SMA store, an intercatchment GW exchange function	A rainfall excess/soil moisture accounting store ; an evaporating reservoir ; an intermediate store and a deep store	A soil store, rainfall interception by PE	
Structure overview: transfer	A non lineau tank distributes the effective rainfall into runoff and GW recharge. The aquifer is represented by a linear tank.	Two unit hydrograph, two parallel nonlinear routing stores	Direct, indirect and baseflow components are routed using a unit hydrograph (Weibull law)	Two unit hydrographs, two linear routing stores : one for streamflow recession, one for interflow	
References on simulation applications in France	800 to 1000 rivers simulated in France		~~Garavaglia (2011) ; Paquet et al. (2013)~~Garavaglia (2011); Paquet et al. (2013)	Lang et al. (2006a~~-~~, 2006b)	Vidal et al. (2010b) Habets et al. (2008)
References on low-flow forecasting		Pushpalatha (2011, 2013)	Mathevet et al. (2010)	Lang et al. (2006a~~-~~, 2006b)	Céron et al. (2010) Soubeyroux et al. (2010)

applications in France					Singla et al. (2012)

Table 4: List of efficiency criteria used for model evaluation in simulation mode

Name	Description
Quadratic criteria	
~~KGEQ~~KGE	Kling-Gupta Efficiency
~~C2MQ~~C2M	Nash-Sutcliffe Efficiency bounded in]-1 ; 1]
Low-flow quadratic criteria	
~~C2MiQ~~C2M_i	Nash-Sutcliffe Efficiency calculated with 1/Q and bounded in]-1 ; 1]
RMSE_{ut}	Root mean square error calculated when observed streamflow is less than Q_{80} threshold
Volume based criteria	
Vdef	Ratio of observed and simulated cumulative annual volume deficits
Temporal criteria	
LFD	Ratio of observed and simulated cumulative low-flow duration
DatSt	Relative difference between observed and simulated start of annual low-flow period
DatEn	Relative difference between observed and simulated end of annual low-flow period
Threshold criteria	
POD	Probability of detection, based on contingency table
FAR	False alarm rate, based on contingency table
CSI	Critical success index, based on contingency table

Mis en fo

Table 5: List of efficiency criteria used for model evaluation in forecasting mode

Name	Description
Low-flow quadratic criteria	
$RMSE_{ut}$	Root mean square error calculated when observed streamflow is less than Q_{80} threshold
Volume based criteria	
Vdef	Ratio of observed and simulated cumulative annual volume deficits
Temporal criteria	
LFD	Ratio of observed and simulated cumulative low-flow duration
Sharpness/reliability	
~~Sharpness~~Sharp	Mean width of interval defined by 10% and 90% percentiles of forecast distribution when observed streamflow is less than Q_{80} threshold
~~Reliability~~Cont_ratio	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)
Threshold criteria	
POD	Probability of detection, based on contingency table
FAR	False alarm rate, based on contingency table
CSI	Critical success index, based on contingency table
~~BSutvig, BSutcri~~BS_{vig}, BS_{cri}	Brier Score with vigilance threshold (Q_{80}) or crisis threshold (Q_{95})
DRPS	Discrete Ranked Probability Score

Mis en fo

Table 6: ~~Models ranked based on~~Models' mean ~~performance~~performances (standard deviation) in validation on the 21 catchments. The ~~mean rank~~integrated criterion is calculated with the nine low-flow criteria (i.e. not considering C2MQ and KGEQ~~).~~) and on transformed values of criteria. Bold values indicate the best model.

Model's mean performances (standard deviation)												
Criterion	GARD		GR6J		MORD		PRES		SIM		DAQ	
C2M	**0.73**	**(0.09)**	0.69	(0.10)	0.69	(0.11)	0.67	(0.11)	0.53	(0.13)	0.13	(0.05)
KGE	0.81	(0.09)	0.83	(0.09)	**0.86**	**(0.06)**	0.79	(0.10)	0.80	(0.07)	0.27	(0.11)
$C2M_i$	**0.57**	**(0.12)**	0.53	(0.14)	0.48	(0.22)	0.56	(0.13)	0.23	(0.19)	0.11	(0.06)
$RMSE_{ut}$	**0.52**	**(0.29)**	0.61	(0.52)	0.81	(0.80)	0.55	(0.35)	1.23	(1.06)	3.48	(2.66)
FAR	**0.21**	**(0.12)**	0.25	(0.13)	0.24	(0.12)	0.22	(0.12)	0.37	(0.12)	0.34	(0.12)
CSI	0.58	(0.15)	0.60	(0.11)	0.58	(0.14)	**0.61**	**(0.11)**	0.42	(0.10)	0.18	(0.12)
POD	0.70	(0.19)	**0.78**	**(0.14)**	0.72	(0.17)	0.75	(0.14)	0.57	(0.13)	0.21	(0.14)
Vdef	0.89	(0.50)	1.21	(0.64)	**0.99**	**(0.44)**	0.95	(0.46)	0.90	(0.38)	0.13	(0.14)
LFD	0.92	(0.33)	1.10	(0.35)	0.98	(0.26)	**0.99**	**(0.29)**	0.92	(0.24)	0.32	(0.21)
DatSt	4.67	(5.64)	-0.55	(8.83)	**0.14**	**(9.88)**	2.43	(5.71)	-13.31	(12.07)	NA	(7.20)
DatEn	1.57	(4.00)	-1.93	(6.38)	1.31	(15.31)	**0.40**	**(4.08)**	-7.83	(8.73)	NA	(6.47)
Integrated criterion (rank)	0.734	(3)	0.735	(2)	0.721	(4)	**0.747**	**(1)**	0.617	(5)	0.422	(6)

Mis en fo

~~Model's rank~~						
~~Criterion~~	~~GARD~~	~~GR6J~~	~~MORD~~	~~PRES~~	~~SIM~~	~~DAQ~~
~~C2MQ~~	**~~1~~**	~~2~~	~~3~~	~~4~~	~~5~~	~~6~~
~~KGEQ~~	~~3~~	~~2~~	**~~1~~**	~~5~~	~~4~~	~~6~~
~~C2MiQ~~	**~~1~~**	~~3~~	~~4~~	~~2~~	~~5~~	~~6~~
~~RMSEut~~	**~~1~~**	~~3~~	~~4~~	~~2~~	~~5~~	~~6~~
~~FAR~~	**~~1~~**	~~4~~	~~3~~	~~2~~	~~6~~	~~5~~
~~CSI~~	~~4~~	~~2~~	~~3~~	**~~1~~**	~~5~~	~~6~~
~~POD~~	~~4~~	**~~1~~**	~~3~~	~~2~~	~~5~~	~~6~~
~~Vdef~~	~~5~~	~~3~~	~~2~~	**~~1~~**	~~4~~	~~6~~
~~LFD~~	~~5~~	~~4~~	**~~1~~**	~~3~~	~~2~~	~~6~~
~~Date Start~~	~~2~~	~~3~~	~~4~~	**~~1~~**	~~5~~	~~NA~~
~~Date End~~	~~2~~	~~3~~	~~4~~	**~~1~~**	~~5~~	~~NA~~
~~Mean rank~~	~~2.8~~	~~2.9~~	~~3.1~~	**~~1.7~~**	~~4.7~~	~~5.9~~

Table 7: ~~Models ranked based on~~Models' mean ~~performance~~performances (standard deviation) on the 21 catchments for validation period 2 and for the two forecasting lead times selected.

	Model's mean performances (standard deviation)											
	7-day lead time						30-day lead time					
Criterion	GARD	GR6J	MORD	PRES	SIM	NVQ	GARD	GR6J	MORD	PRES	SIM	NVQ
$RMSE_{ut}$	**0.72** **(0.43)**	1.22 (1.13)	1.16 (0.91)	0.99 (0.52)	1.25 (0.83)	2.33 (1.54)	**1.88** **(1.17)**	2.81 (2.13)	2.16 (1.59)	2.02 (1.15)	2.06 (1.41)	2.57 (1.75)
DRPS	0.13 (0.07)	**0.12** **(0.05)**	0.13 (0.05)	0.12 (0.04)	0.18 (0.03)	0.19 (0.02)	0.18 (0.06)	0.18 (0.03)	0.19 (0.04)	**0.17** **(0.03)**	0.20 (0.03)	0.21 (0.02)
POD	0.82 (0.16)	0.85 (0.06)	**0.87** **(0.08)**	0.8 (0.11)	0.55 (0.21)	0.58 (0.16)	0.65 (0.17)	0.68 (0.09)	**0.72** **(0.10)**	0.59 (0.18)	0.52 (0.17)	0.55 (0.16)
FAR	0.23 (0.08)	0.22 (0.06)	0.27 (0.07)	**0.22** **(0.06)**	0.32 (0.11)	0.38 (0.11)	0.31 (0.08)	0.32 (0.08)	0.35 (0.08)	**0.29** **(0.07)**	0.36 (0.11)	0.38 (0.11)
CSI	0.67 (0.14)	**0.69** **(0.08)**	0.66 (0.08)	0.65 (0.10)	0.42 (0.14)	0.41 (0.12)	0.51 (0.13)	0.52 (0.07)	**0.52** **(0.08)**	0.47 (0.14)	0.38 (0.10)	0.40 (0.12)
BS_{vig}	0.09 (0.05)	**0.08** **(0.04)**	0.1 (0.03)	0.09 (0.03)	0.13 (0.03)	0.13 (0.02)	0.12 (0.04)	**0.12** **(0.03)**	0.14 (0.03)	0.12 (0.03)	0.14 (0.03)	0.14 (0.02)
BS_{cri}	0.06 (0.03)	**0.06** **(0.03)**	0.07 (0.03)	0.07 (0.03)	0.09 (0.03)	0.09 (0.02)	**0.08** **(0.03)**	0.08 (0.03)	0.10 (0.04)	0.09 (0.03)	0.10 (0.03)	0.09 (0.03)
Cont_ratio	0.34 (0.13)	0.45 (0.20)	0.52 (0.20)	0.64 (0.08)	0.68 (0.18)	**0.84** **(0.07)**	0.59 (0.16)	0.65 (0.16)	0.63 (0.20)	**0.82** **(0.08)**	0.69 (0.19)	0.84 (0.08)
Sharp	**0.95** **(0.53)**	1.58 (1.30)	1.95 (1.45)	1.92 (0.98)	2.96 (1.92)	4.69 (2.95)	**3.29** **(1.89)**	4.88 (3.48)	4.06 (2.43)	4.30 (2.11)	4.12 (2.43)	5.06 (3.12)
Vdef	**0.73** **(0.22)**	0.7 (0.16)	0.55 (0.23)	0.62 (0.21)	0.18 (0.21)	0.12 (0.12)	**0.41** **(0.19)**	0.38 (0.16)	0.37 (0.20)	0.39 (0.23)	0.15 (0.23)	0.12 (0.13)
LFD	**0.79** **(0.19)**	0.77 (0.15)	0.69 (0.23)	0.67 (0.20)	0.35 (0.22)	0.33 (0.23)	**0.53** **(0.20)**	0.49 (0.16)	0.50 (0.21)	0.45 (0.22)	0.30 (0.25)	0.34 (0.22)
Integrated criterion (rank)	0.673 (2)	**0.674** **(1)**	0.636 (4)	0.652 (3)	0.473 (5)	0.448 (6)	**0.527** **(1)**	0.516 (2)	0.504 (4)	0.514 (3)	0.425 (6)	0.436 (5)

Mis en fo

	~~Model's rank~~											
	~~7-day lead time~~						~~30-day lead time~~					
~~Criterion~~	~~GARD~~	~~GR6J~~	~~MORD~~	~~PRES~~	~~SIM~~	~~NVQ~~	~~GARD~~	~~GR6J~~	~~MORD~~	~~PRES~~	~~SIM~~	~~NVQ~~
~~RMSEut~~	~~1~~	~~2~~	~~3~~	~~4~~	~~5~~	~~6~~	~~1~~	~~5~~	~~2~~	~~4~~	~~3~~	~~6~~
~~DRPS~~	~~3~~	~~1~~	~~4~~	~~2~~	~~5~~	~~6~~	~~3~~	~~1~~	~~4~~	~~2~~	~~6~~	~~5~~
~~POD~~	~~3~~	~~2~~	~~1~~	~~4~~	~~6~~	~~5~~	~~3~~	~~2~~	~~1~~	~~4~~	~~6~~	~~5~~
~~FAR~~	~~3~~	~~2~~	~~4~~	~~1~~	~~5~~	~~6~~	~~2~~	~~3~~	~~4~~	~~1~~	~~5~~	~~6~~
~~CSI~~	~~2~~	~~1~~	~~3~~	~~4~~	~~5~~	~~6~~	~~3~~	~~2~~	~~1~~	~~4~~	~~6~~	~~5~~
~~BSutvig~~	~~3~~	~~1~~	~~4~~	~~2~~	~~5~~	~~6~~	~~3~~	~~2~~	~~4~~	~~1~~	~~5~~	~~6~~
~~BSutcri~~	~~2~~	~~1~~	~~4~~	~~3~~	~~6~~	~~5~~	~~1~~	~~2~~	~~5~~	~~3~~	~~6~~	~~4~~
~~Cont_ratio~~	~~6~~	~~5~~	~~4~~	~~2~~	~~3~~	~~1~~	~~6~~	~~4~~	~~5~~	~~2~~	~~3~~	~~1~~
~~Sharp~~	~~1~~	~~2~~	~~3~~	~~4~~	~~5~~	~~6~~	~~1~~	~~3~~	~~2~~	~~5~~	~~4~~	~~6~~
~~Vdef~~	~~1~~	~~2~~	~~4~~	~~3~~	~~5~~	~~6~~	~~1~~	~~3~~	~~4~~	~~2~~	~~5~~	~~6~~
~~LFD~~	~~1~~	~~2~~	~~3~~	~~4~~	~~5~~	~~6~~	~~1~~	~~3~~	~~2~~	~~4~~	~~6~~	~~5~~
~~Mean rank~~	~~2.4~~	~~1.9~~	~~3.4~~	~~3.0~~	~~5.0~~	~~5.4~~	~~2.3~~	~~2.7~~	~~3.1~~	~~2.9~~	~~5.0~~	~~5.0~~

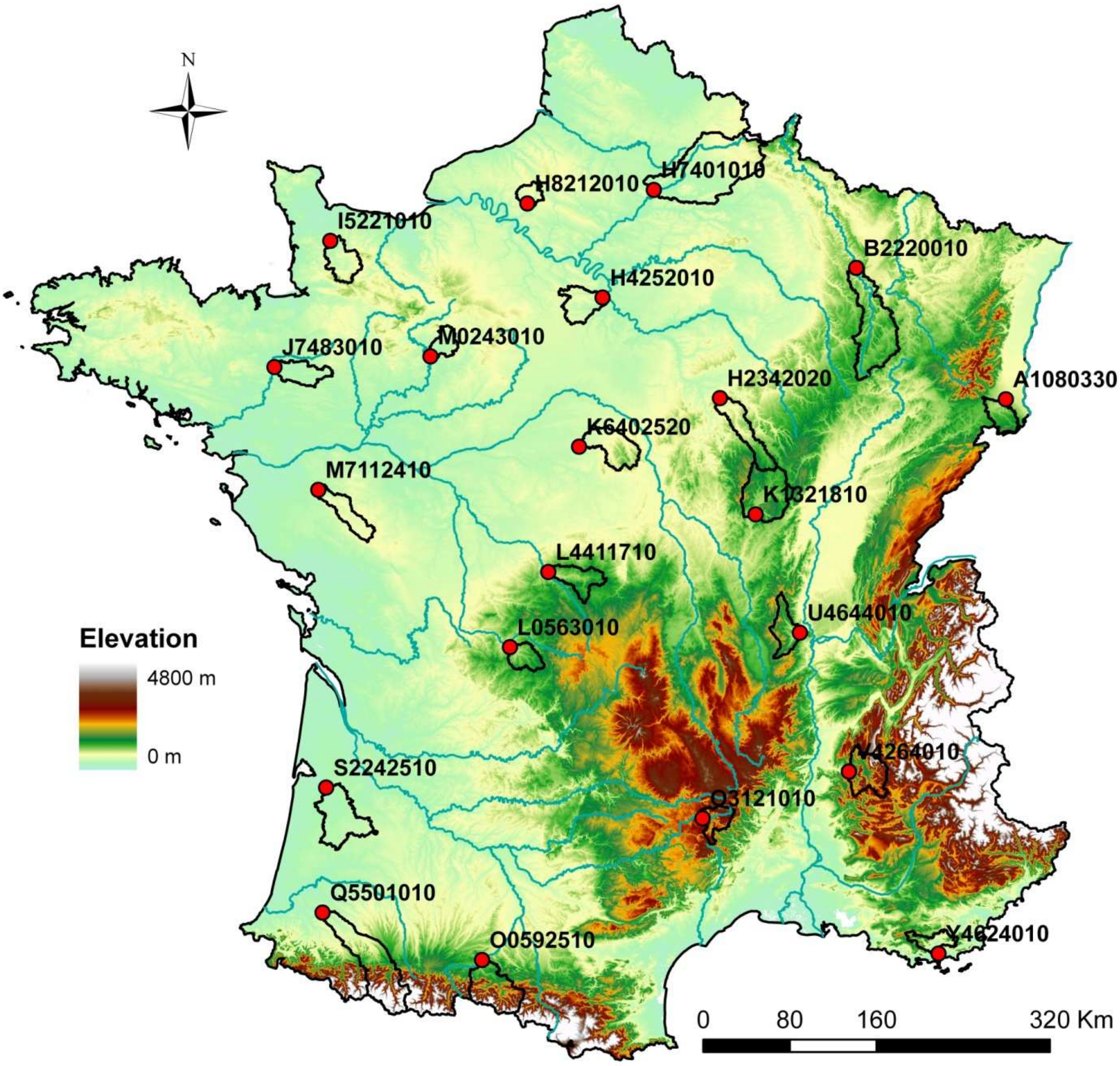
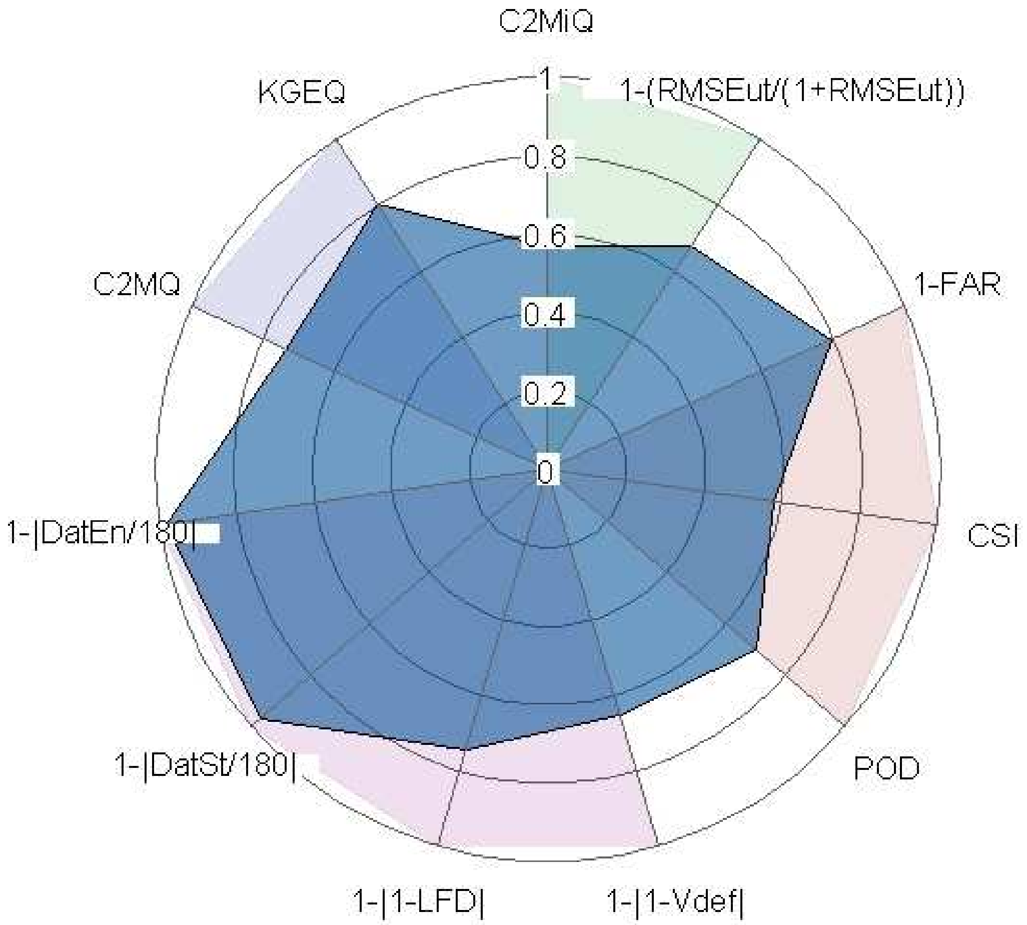


Figure 1: Location of the 21 selected catchments in France



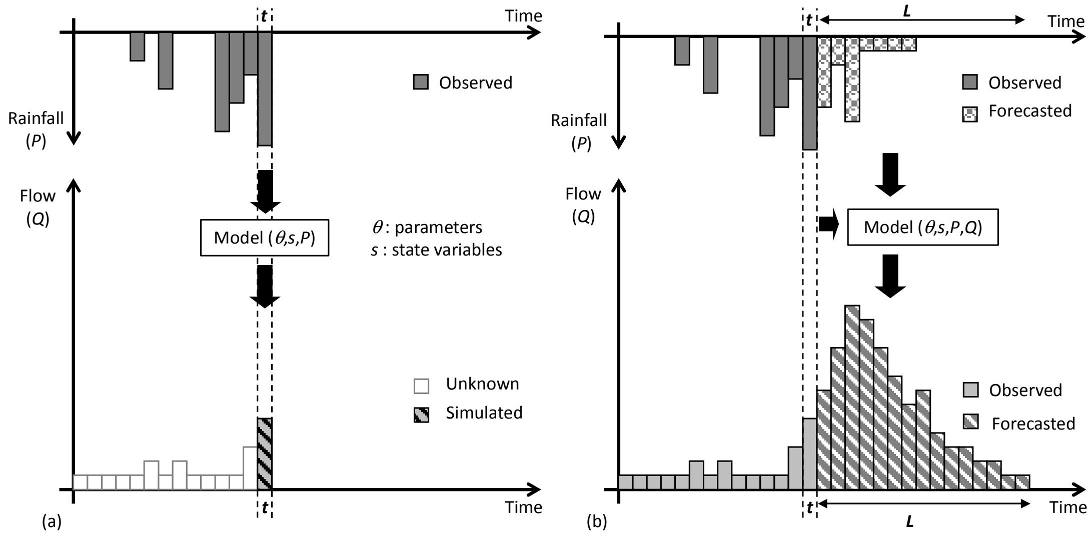


Figure 2: Schematic representation of the difference between (a) simulation and (b) forecasting modes (*L*: lead time)

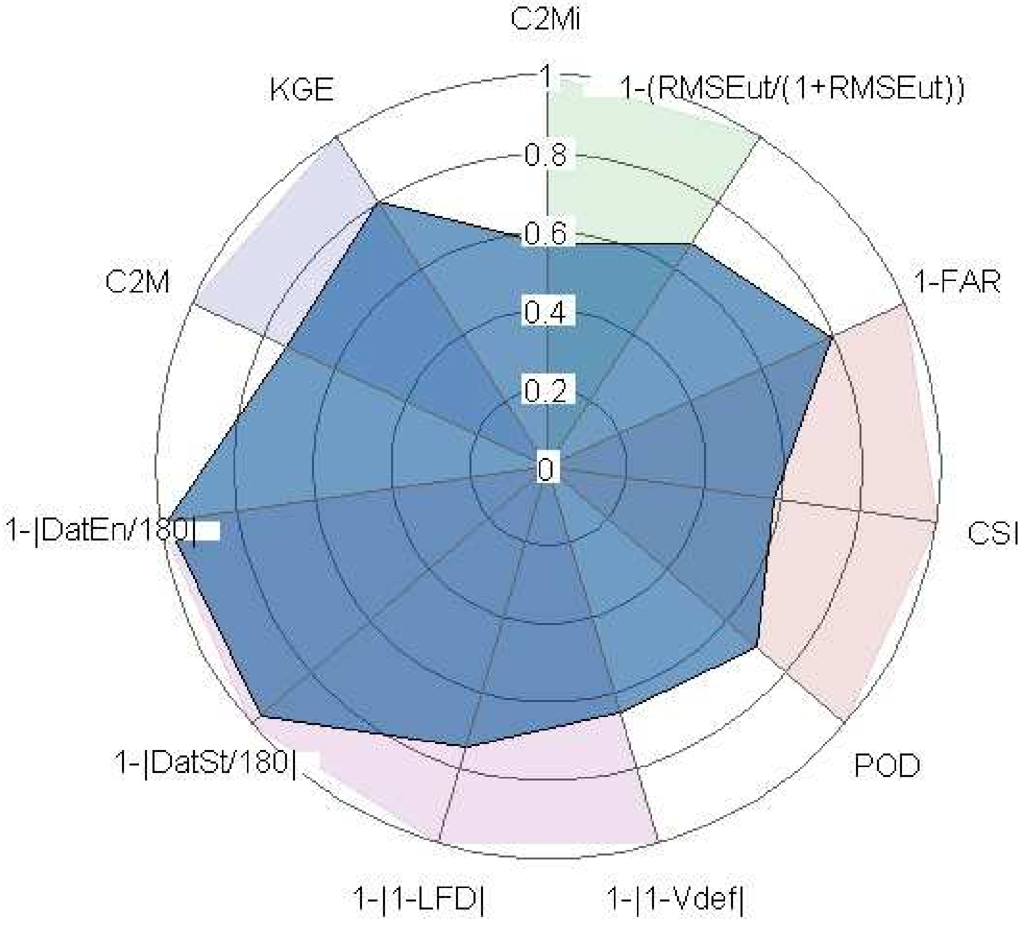
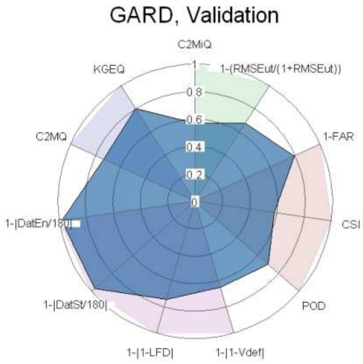
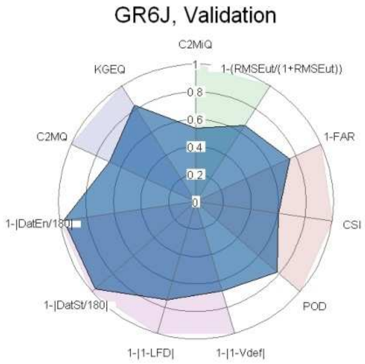
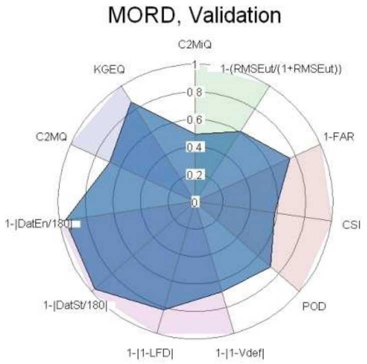
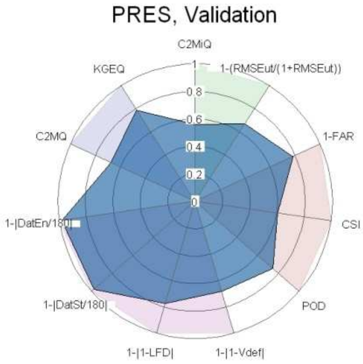


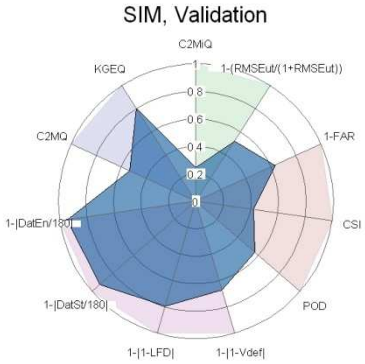
Figure 3: Example of radial plot showing mean model results on the set of 21 catchments for the selected evaluation criteria. The larger the blue surface, the better the model. Background colours link criteria focusing on similar aspects

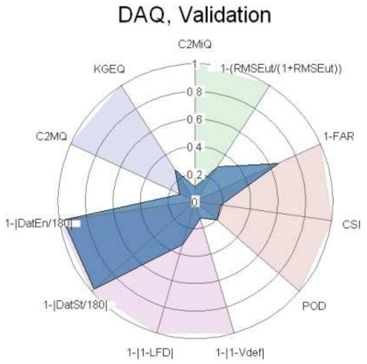

GARD, Validation

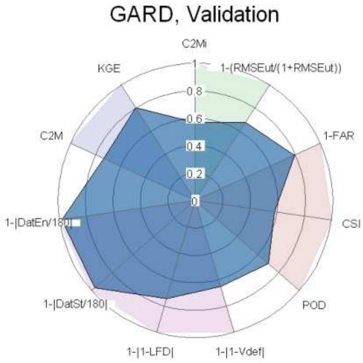

GR6J, Validation

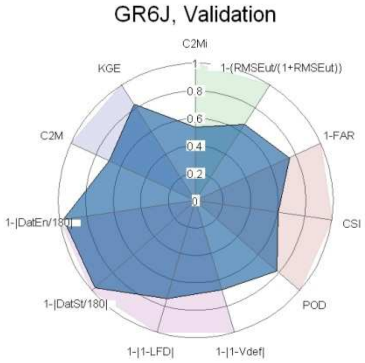

MORD, Validation

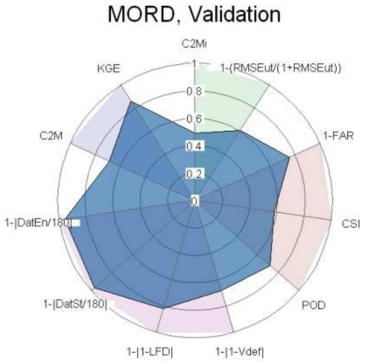

PRES, Validation

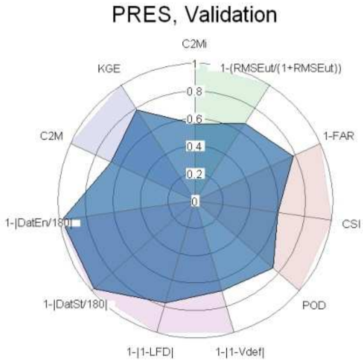

SIM, Validation

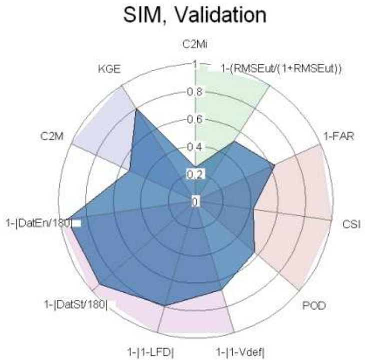

DAQ, Validation


GARD, Validation


GR6J, Validation


MORD, Validation


PRES, Validation


SIM, Validation

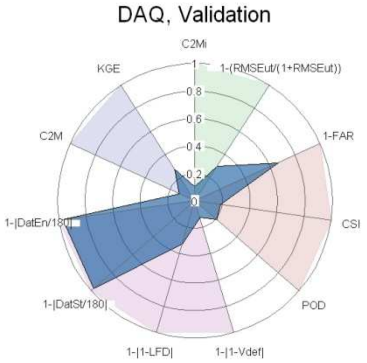
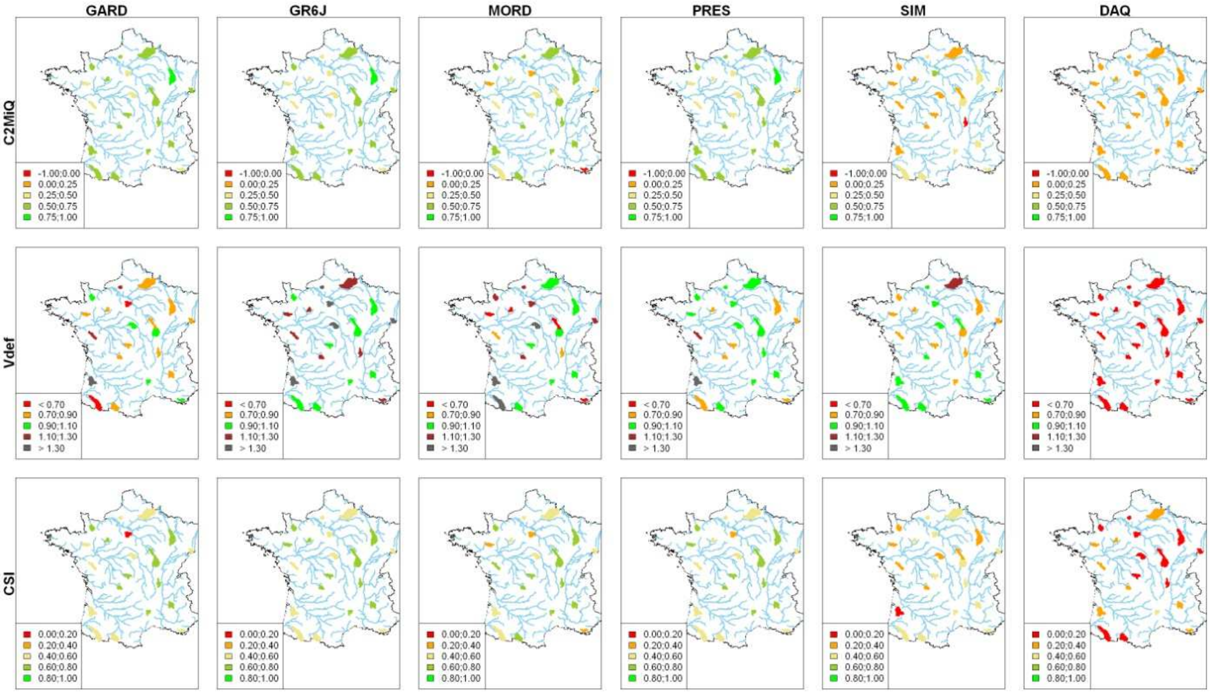

DAQ, Validation

Figure 4: Radial plot showing the mean results for the selected criteria in validation for the 21 catchments and the two periods. Results of the five models tested and the benchmark (DAQ) are shown.


GARD
GR6J
MORD
PRES
SIM
DAQ
C2MiQ
-1.00;0.00
0.00;0.25
0.25;0.50
0.50;0.75
0.75;1.00
Vdef
< 0.70
0.70;0.90
0.90;1.10
1.10;1.30
> 1.30
CSI
0.00;0.20
0.20;0.40
0.40;0.60
0.60;0.80
0.80;1.00

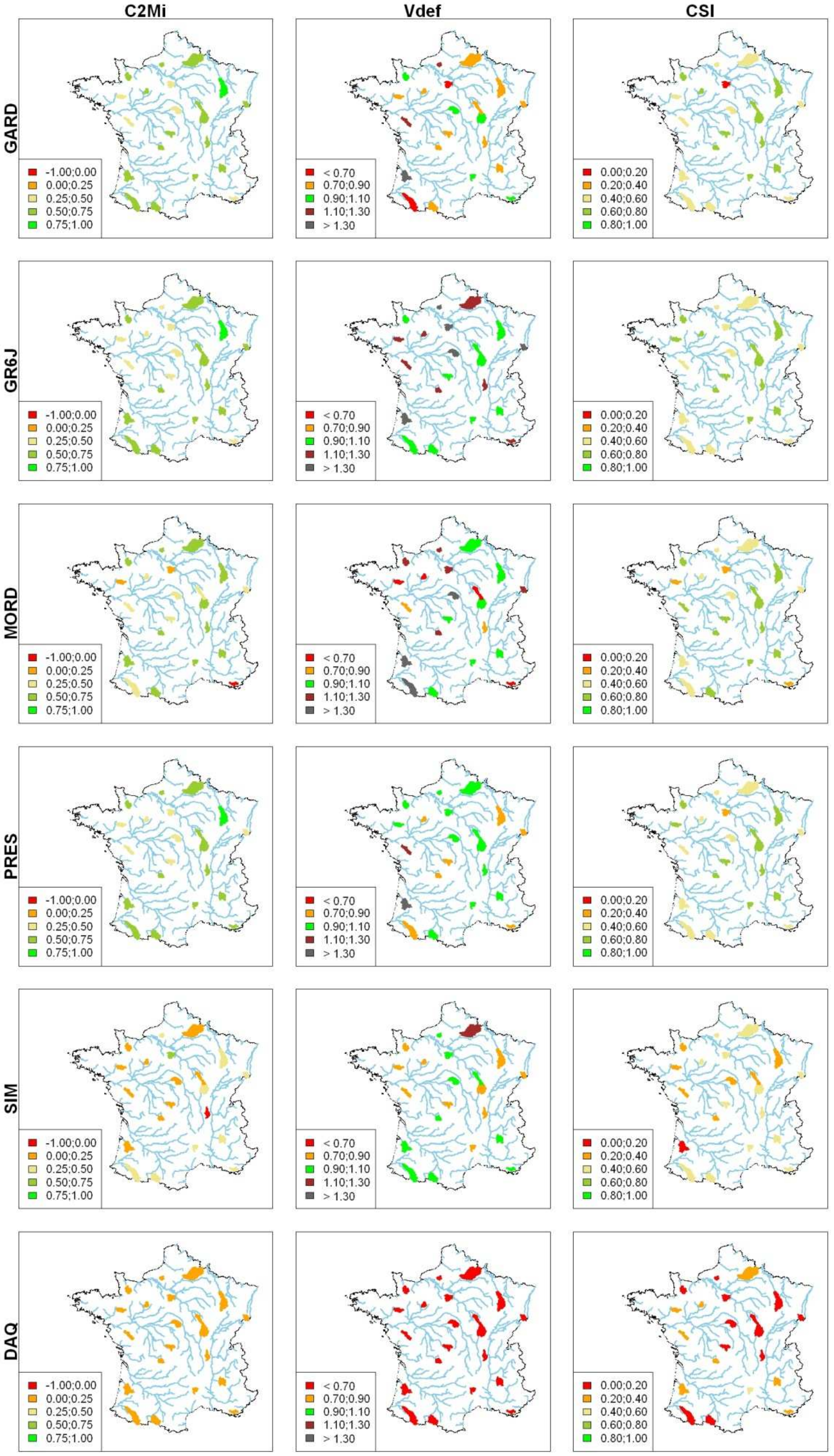
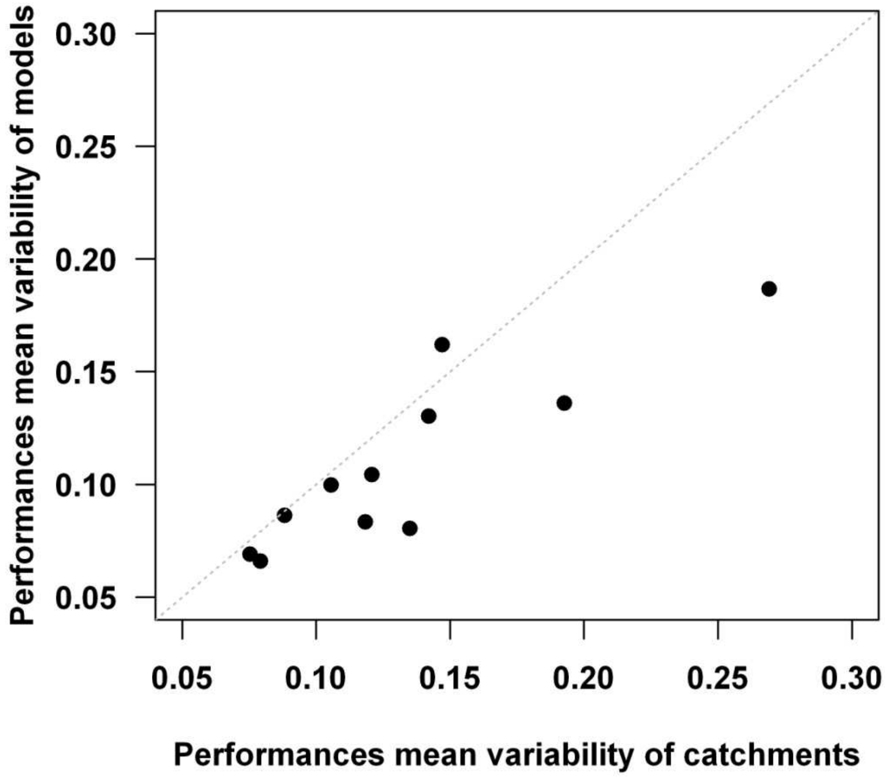


Figure 5: Maps of mean performance on the two validation periods in ~~C2MiQ~~C2M$_i$, Vdef and CSI for the five models tested and the benchmark (DAQ) on the 21 catchments

Mis en fo



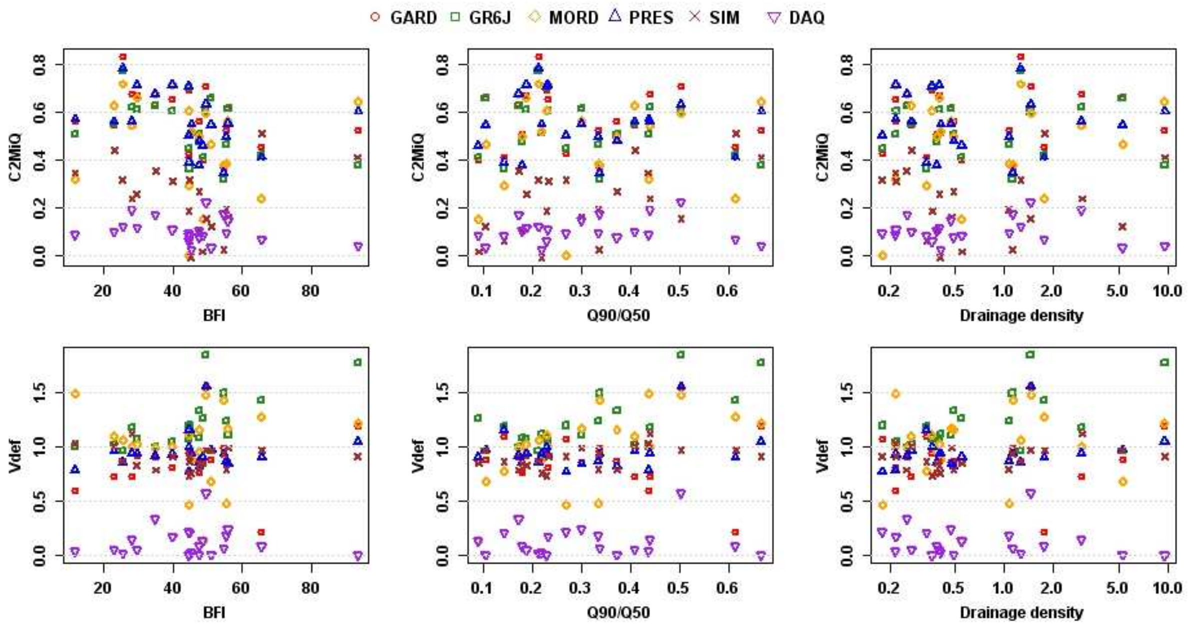


Figure 6~~:~~: Performances mean variability of models versus performances mean variability of catchments in simulation for the 11 selected criteria

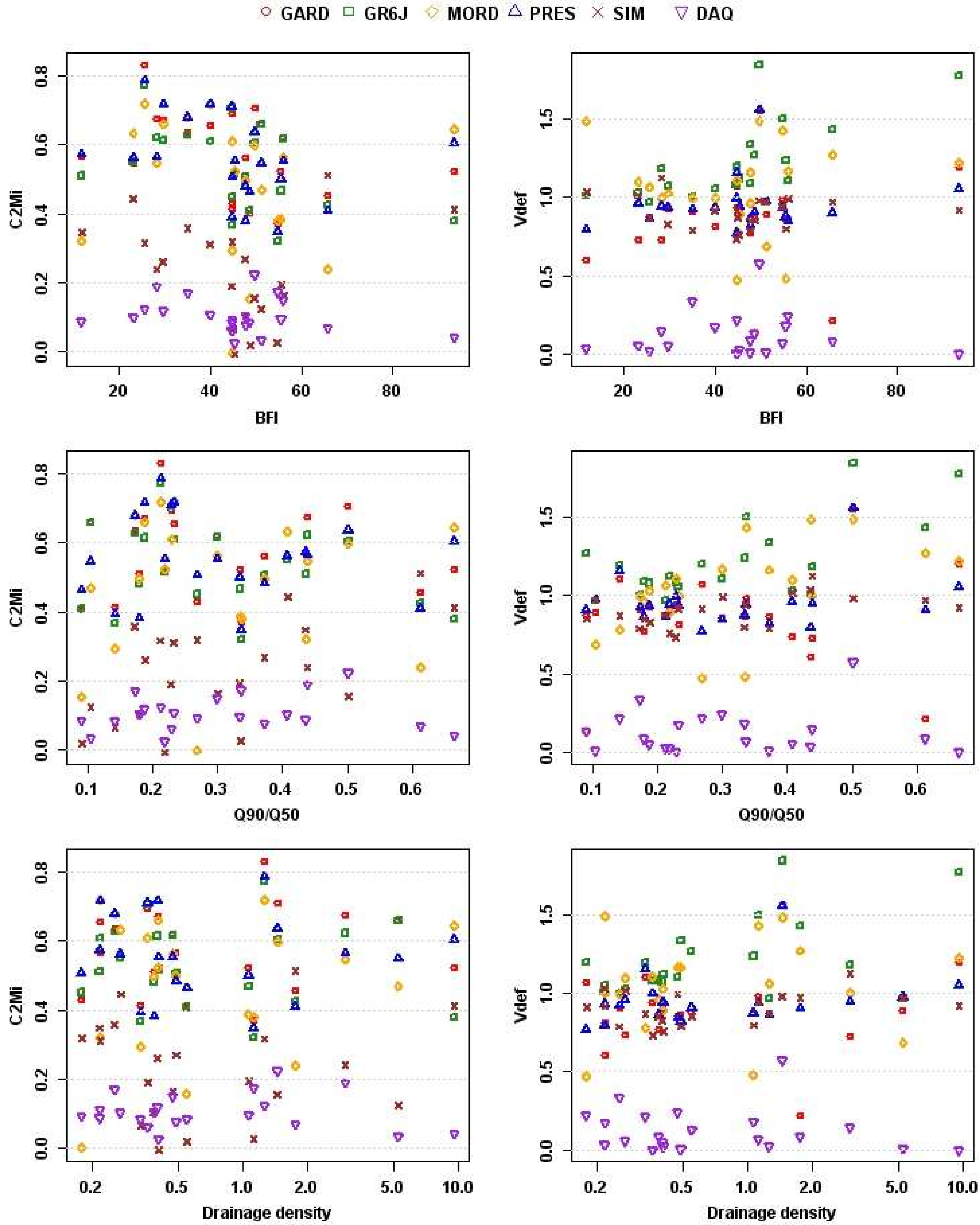


Figure 7: Relation between mean performance on the two validation periods in terms of ~~C2MiQ~~C2M$_i$ (a) and Vdef (b), and catchment or streamflow characteristics (left: Base-Flow Index, centre: Q_{90}/Q_{50} ratio; right: drainage density) for the 21 catchments and the models tested.

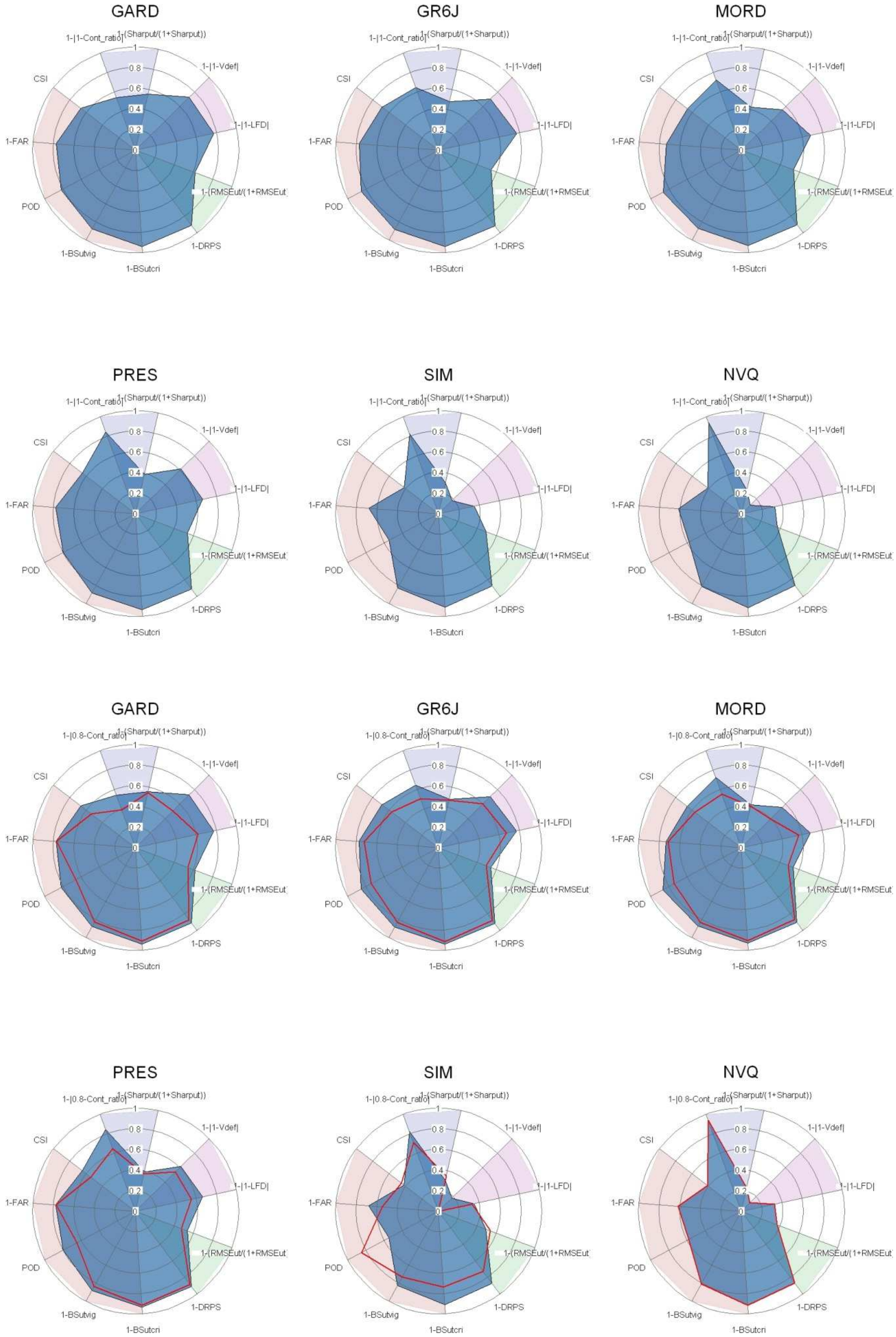
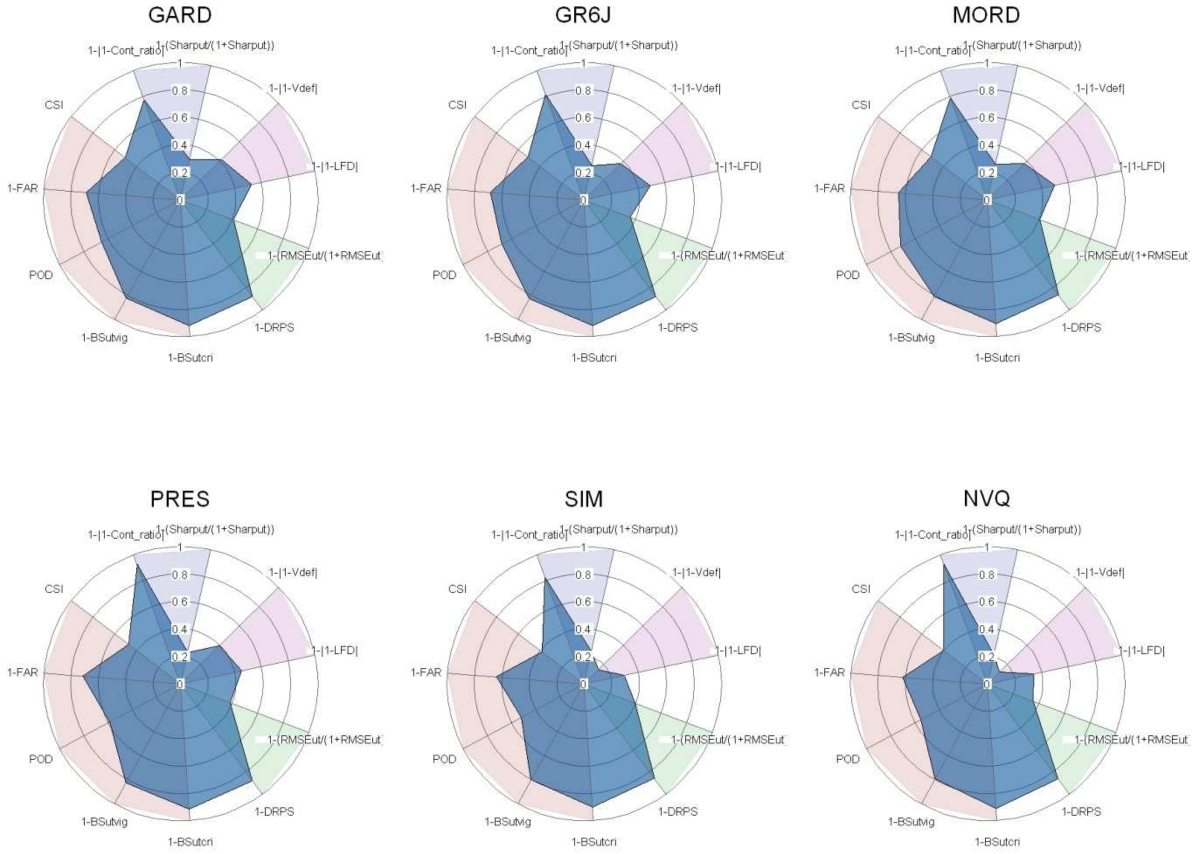

GARD
GR6J
MORD
PRES
SIM
NVQ
GARD
GR6J
MORD
PRES
SIM
NVQ
1-(Sharput/(1+Sharput))
1-|1-Vdef|
1-|1-LFD|
1-(RMSEut/(1+RMSEut
1-DRPS
1-BSutcri
1-BSutvig
POD
1-FAR
CSI
1-|1-Cont_ratio|
1-|0.8-Cont_ratio|
1
0.8
0.6
0.4
0.2
0

Figure 8: Radial plot of the results of the mean selected criteria in validation for the 21 catchments in validation period 2, for a *d*+7 forecasting lead time. Red lines represent the results when no assimilation or post correction method is used.


GARD
GR6J
MORD
PRES
SIM
NVQ
1-|1-Cont_ratio|
1-(Sharput/(1+Sharput))
1-|1-Vdef|
1-|1-LFD|
1-(RMSEut/(1+RMSEut
1-DRPS
1-BSutcri
1-BSutvig
POD
1-FAR
CSI
0
0.2
0.4
0.6
0.8
1

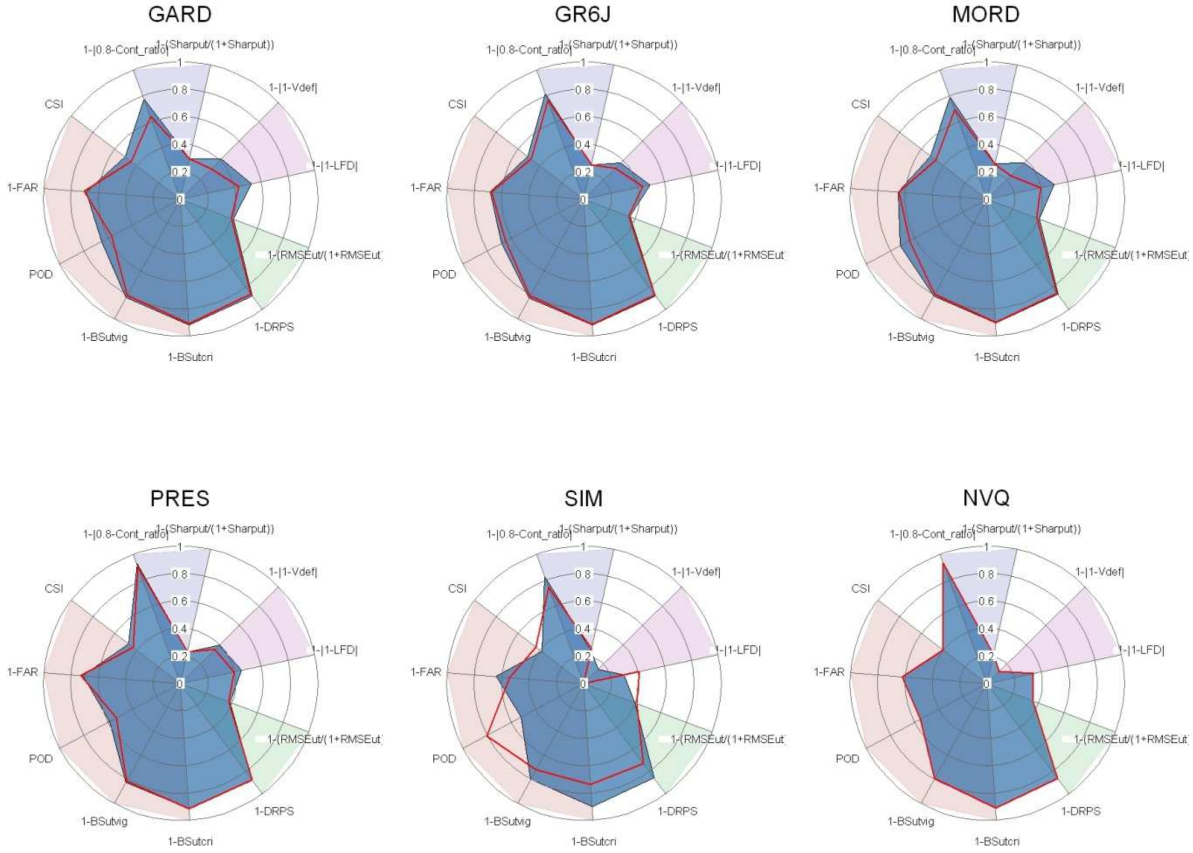
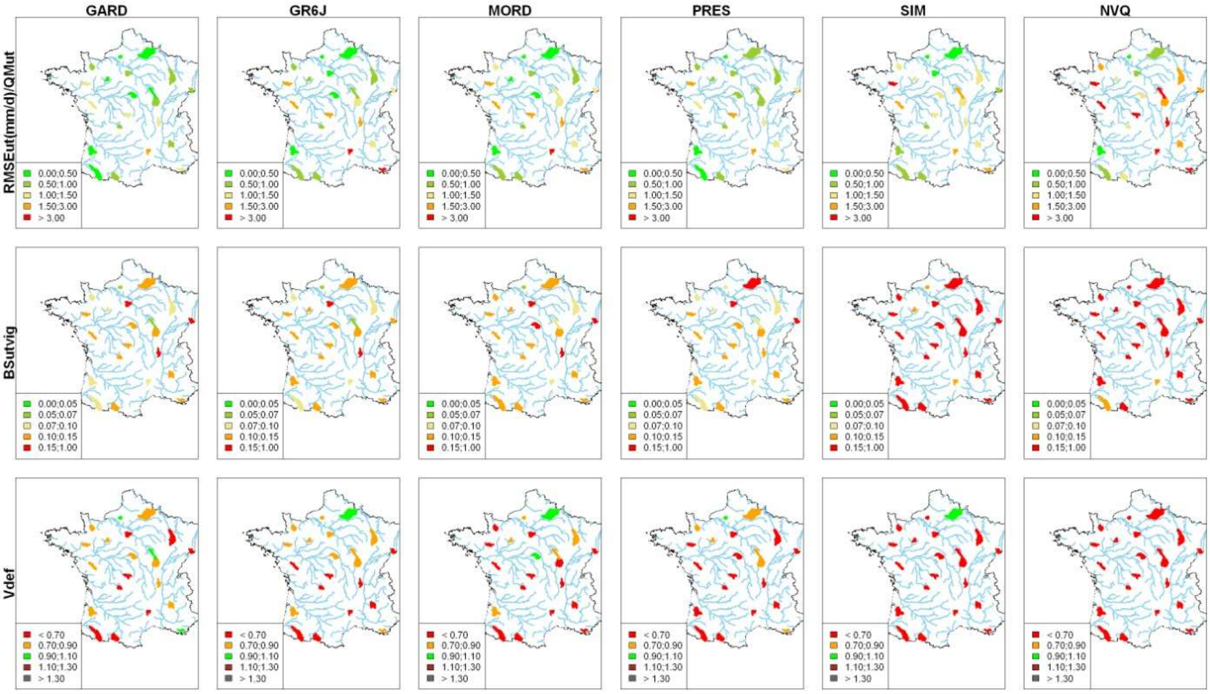

GARD
GR6J
MORD
PRES
SIM
NVQ
1-|0.8-Cont_ratio|
1-(Sharput/(1+Sharput))
1-|1-Vdef|
1-|1-LFD|
1-(RMSEut/(1+RMSEut
1-DRPS
1-BSutcri
1-BSutvig
POD
1-FAR
CSI
0
0.2
0.4
0.6
0.8
1

Figure 9: Radial plot of the results of the mean selected criteria in validation for the 21 catchments 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.

2130


GARD
GR6J
MORD
PRES
SIM
NVQ
RMSEut(mm/d)/QMut
0.00;0.50
0.50;1.00
1.00;1.50
1.50;3.00
> 3.00
BSutvig
0.00;0.05
0.05;0.07
0.07;0.10
0.10;0.15
0.15;1.00
Vdef
< 0.70
0.70;0.90
0.90;1.10
1.10;1.30
> 1.30

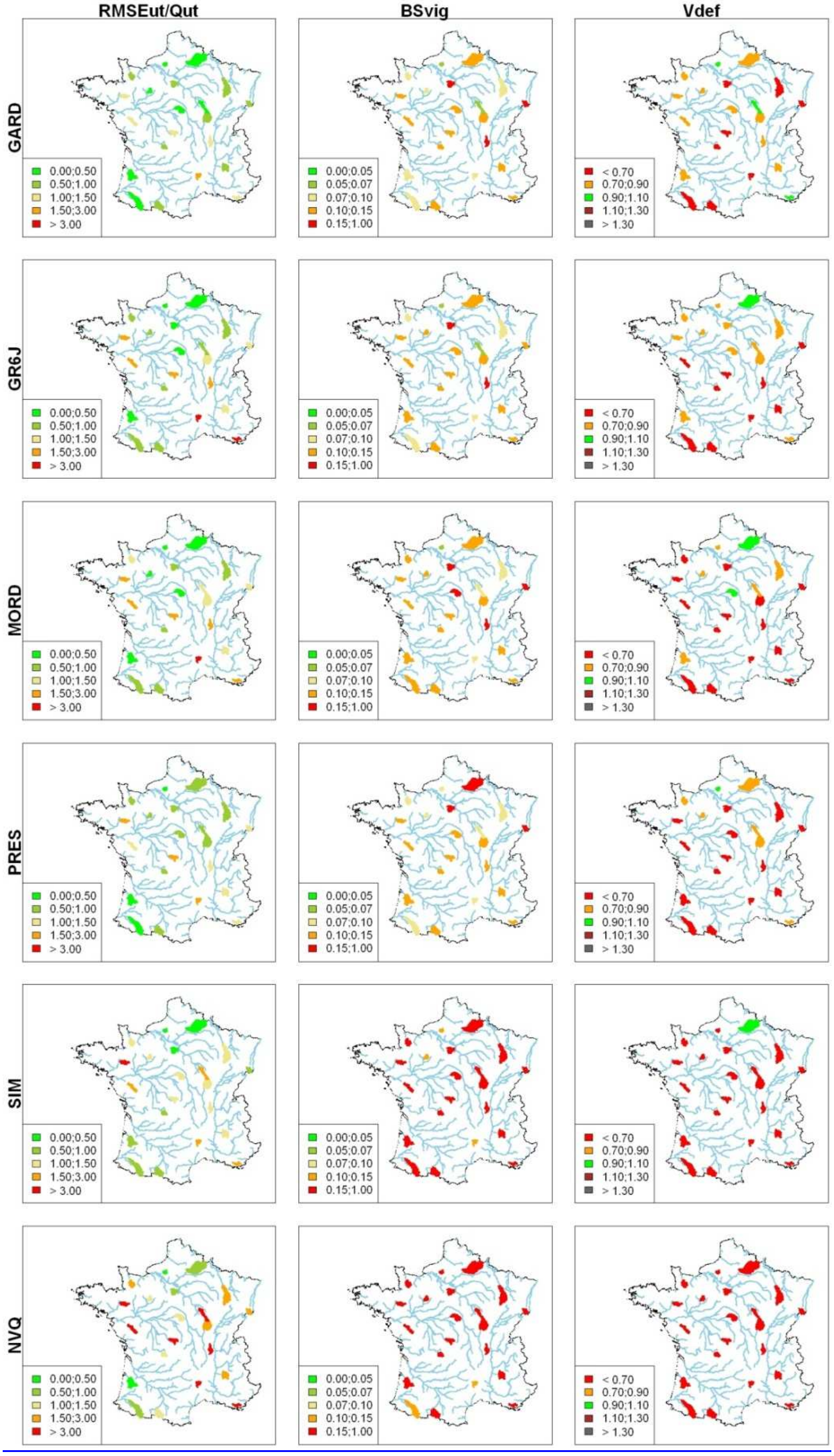
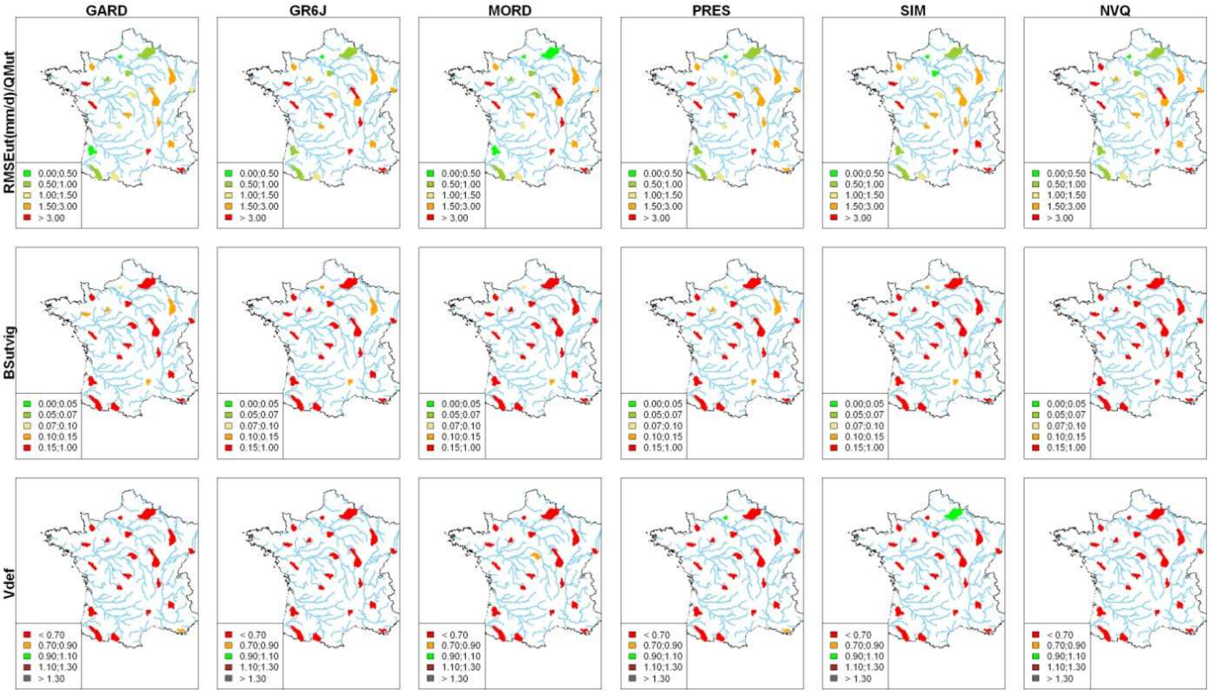


Figure 10: Performance on validation period 2 in RMSE$_{ut}$, ~~BSutvig~~BS$_{vig}$ and Vdef for each model on the 21 catchments for a 7-day forecasting lead time

Mis en fo

Mis en fo

Mis en fo


GARD
GR6J
MORD
PRES
SIM
NVQ
RMSEut(mm/d)/QMut
0.00;0.50
0.50;1.00
1.00;1.50
1.50;3.00
> 3.00
BSutvig
0.00;0.05
0.05;0.07
0.07;0.10
0.10;0.15
0.15;1.00
Vdef
< 0.70
0.70;0.90
0.90;1.10
1.10;1.30
> 1.30

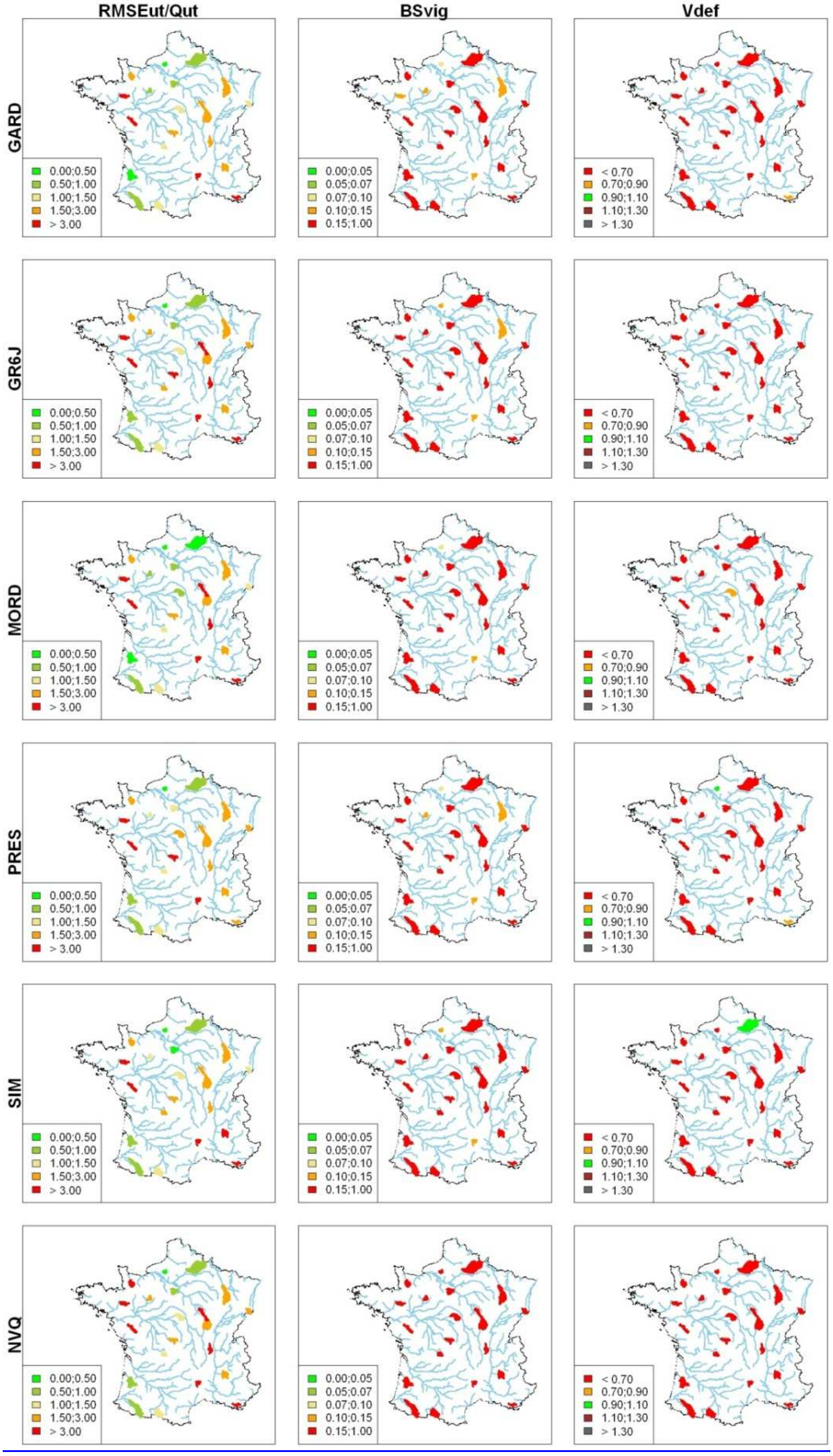
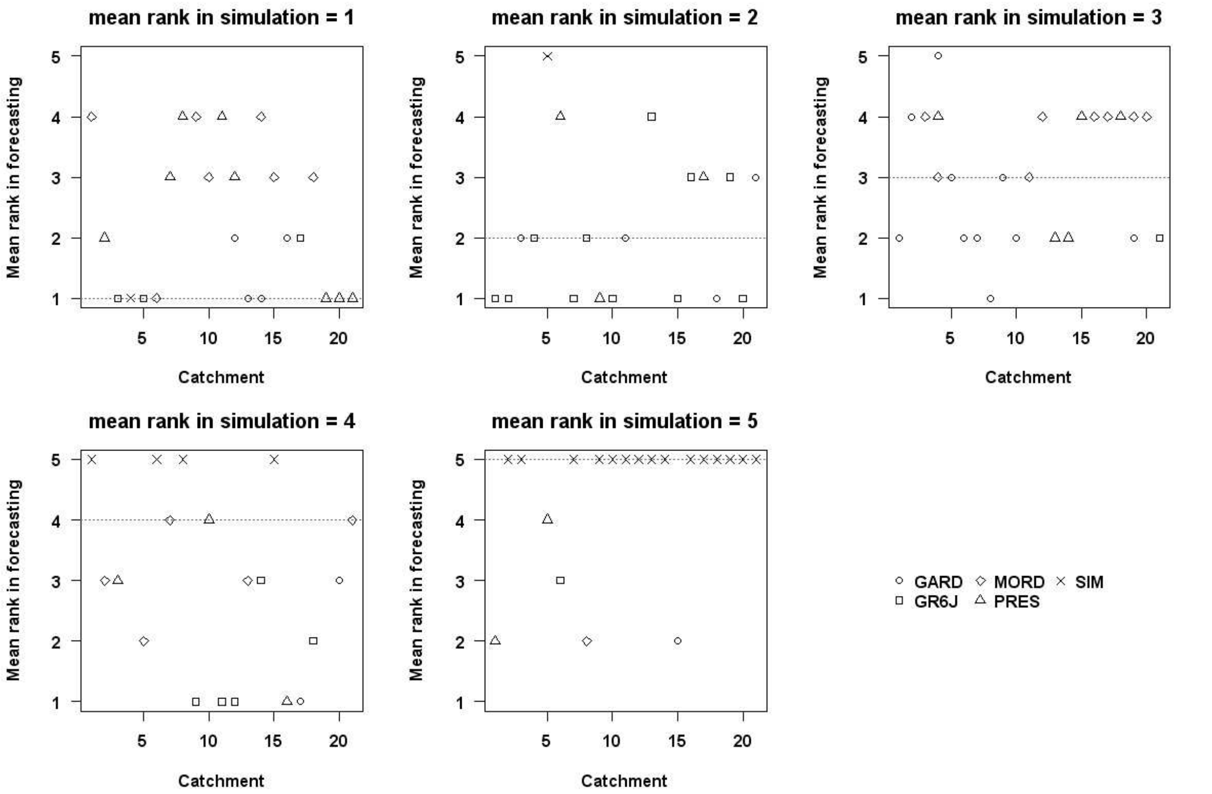
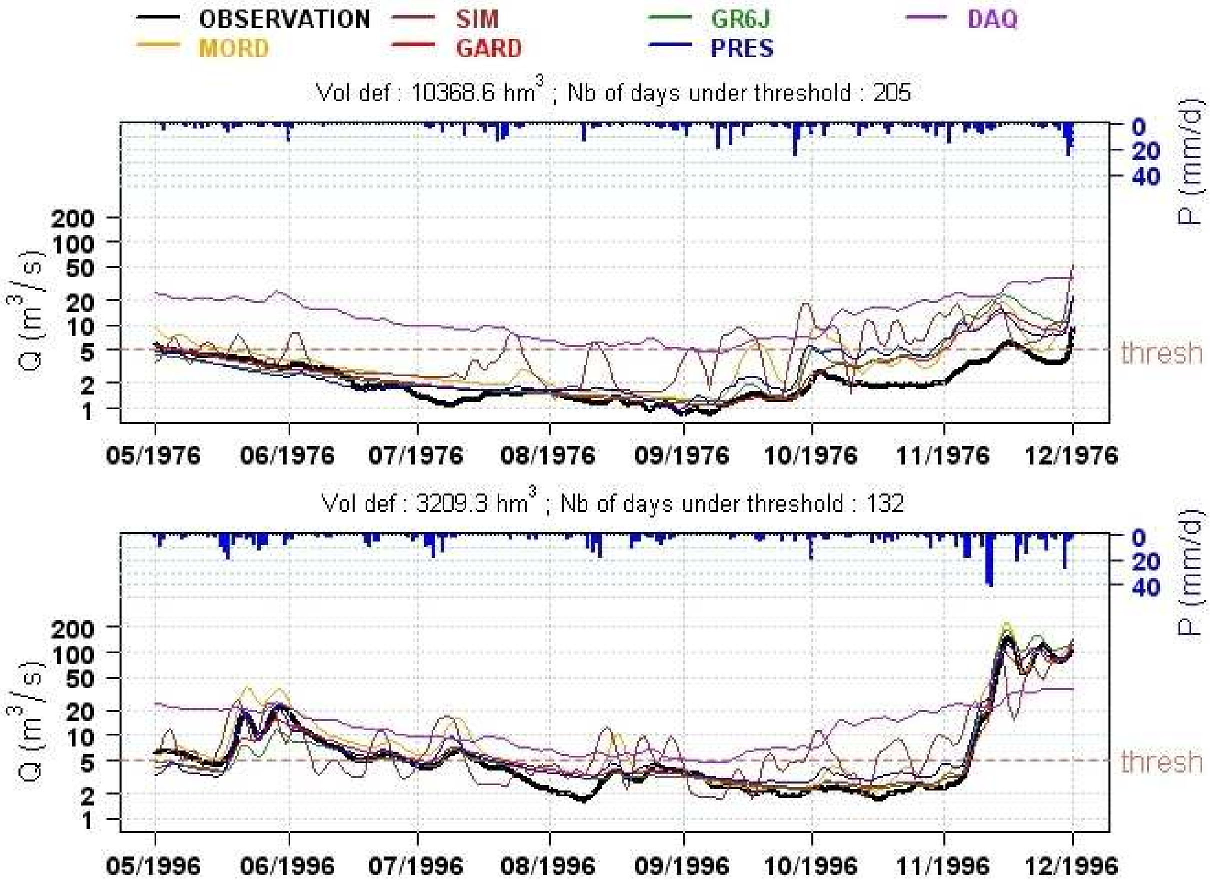


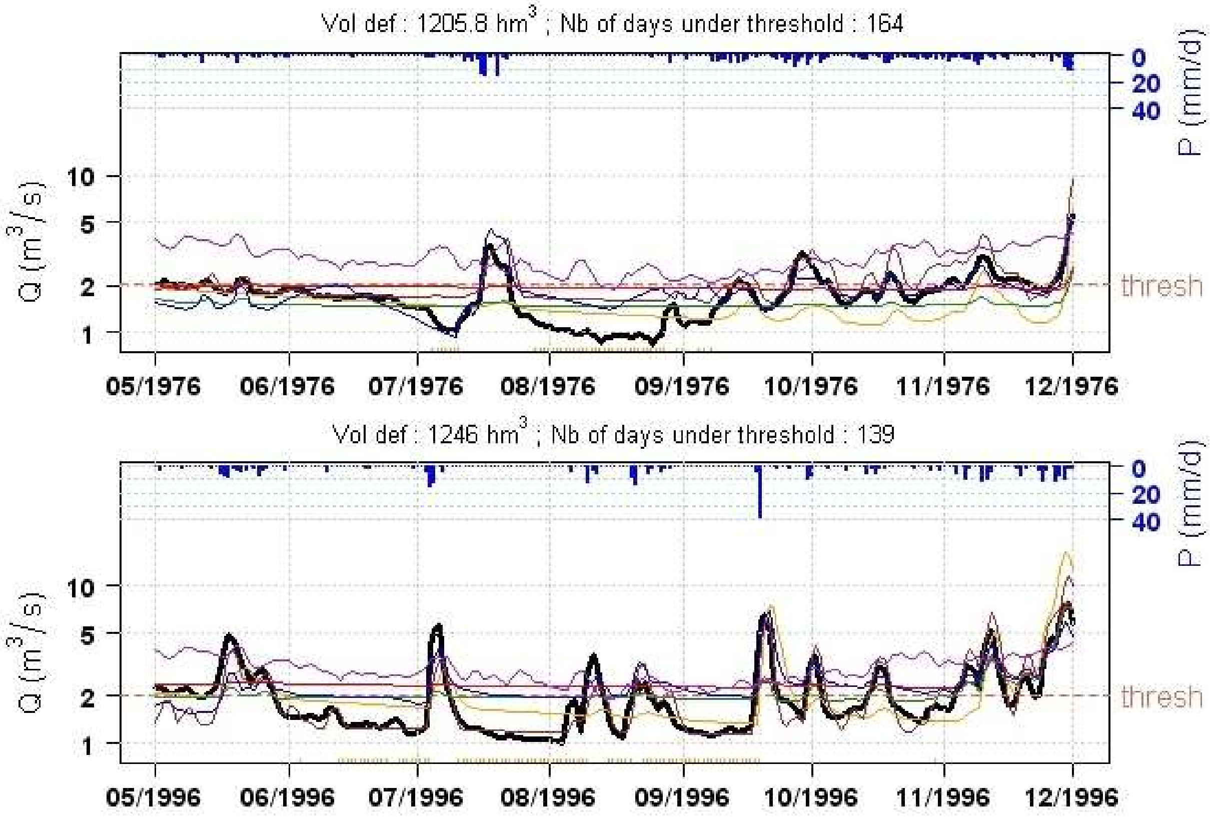
Figure 11: Performance on validation period 2 in RMSE$_{ut}$, ~~BSutvig~~BS$_{vig}$ and Vdef for each model on the 21 catchments for a 30-day forecasting lead time

Mis en fo


mean rank in simulation = 1
mean rank in simulation = 2
mean rank in simulation = 3
mean rank in simulation = 4
mean rank in simulation = 5
Mean rank in forecasting
Catchment
1
2
3
4
5
5
10
15
20
○ GARD ◇ MORD × SIM
□ GR6J △ PRES

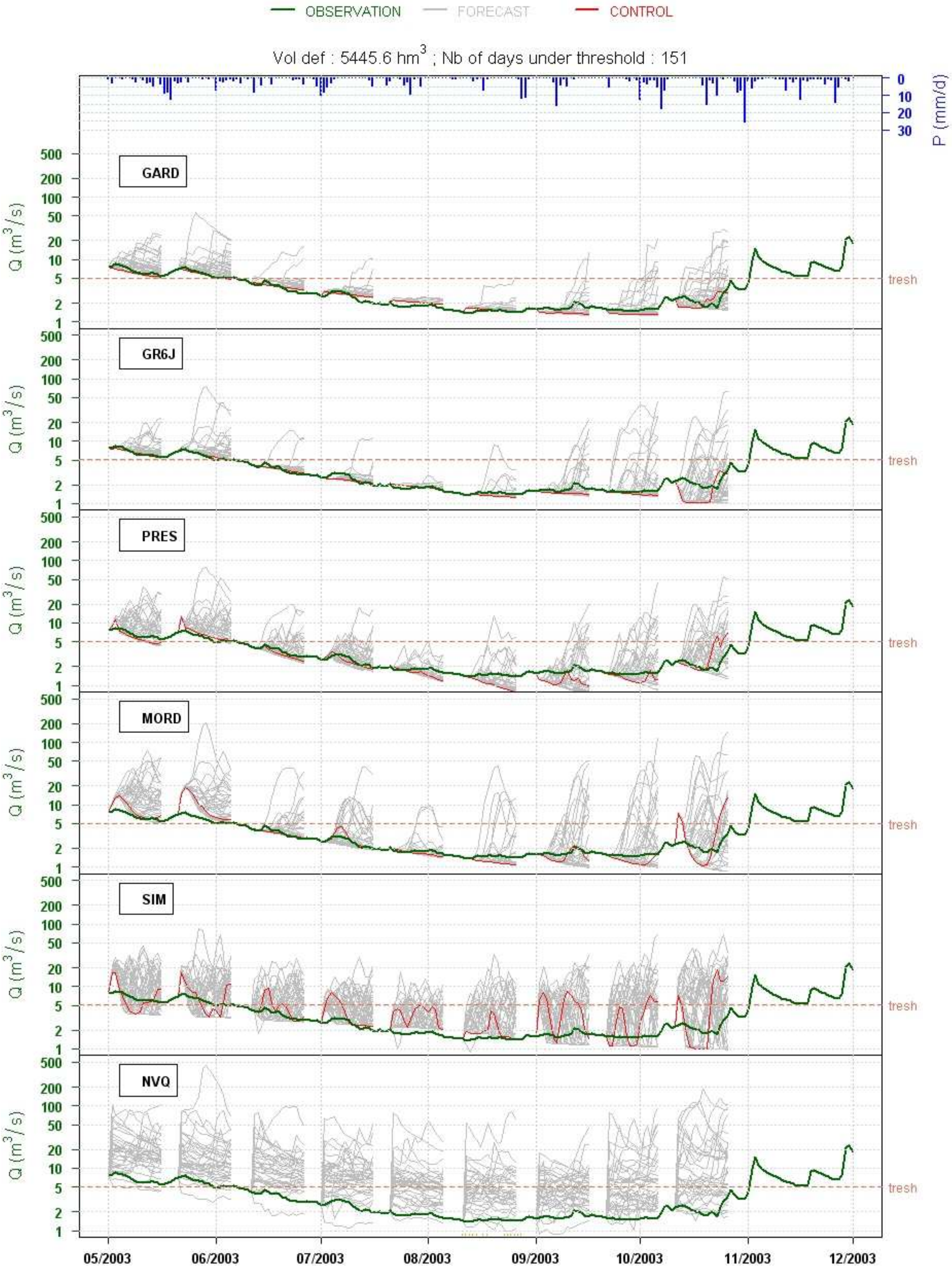

OBSERVATION
SIM
GR6J
DAQ
MORD
GARD
PRES
Vol def : 10368.6 hm3 ; Nb of days under threshold : 205
Q (m3/s)
P (mm/d)
0
20
40
200
100
50
20
10
5
2
1
thresh
05/1976
06/1976
07/1976
08/1976
09/1976
10/1976
11/1976
12/1976
Vol def : 3209.3 hm3 ; Nb of days under threshold : 132
05/1996
06/1996
07/1996
08/1996
09/1996
10/1996
11/1996
12/1996

(a)

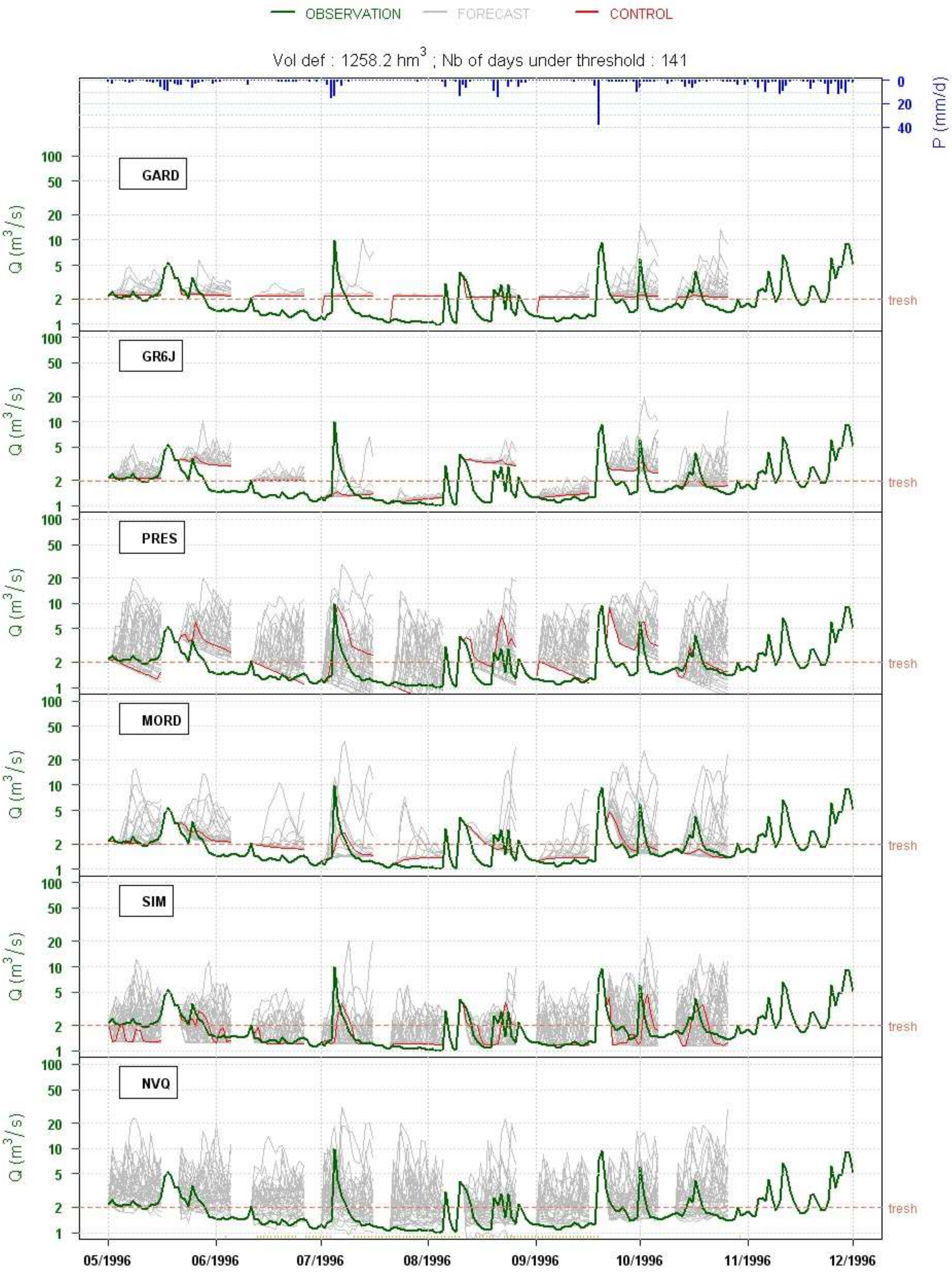

Vol def : 1205.8 hm3 ; Nb of days under threshold : 164
Q (m3/s)
P (mm/d)
0
20
40
10
5
2
1
thresh
05/1976
06/1976
07/1976
08/1976
09/1976
10/1976
11/1976
12/1976
Vol def : 1246 hm3 ; Nb of days under threshold : 139
05/1996
06/1996
07/1996
08/1996
09/1996
10/1996
11/1996
12/1996

(b)

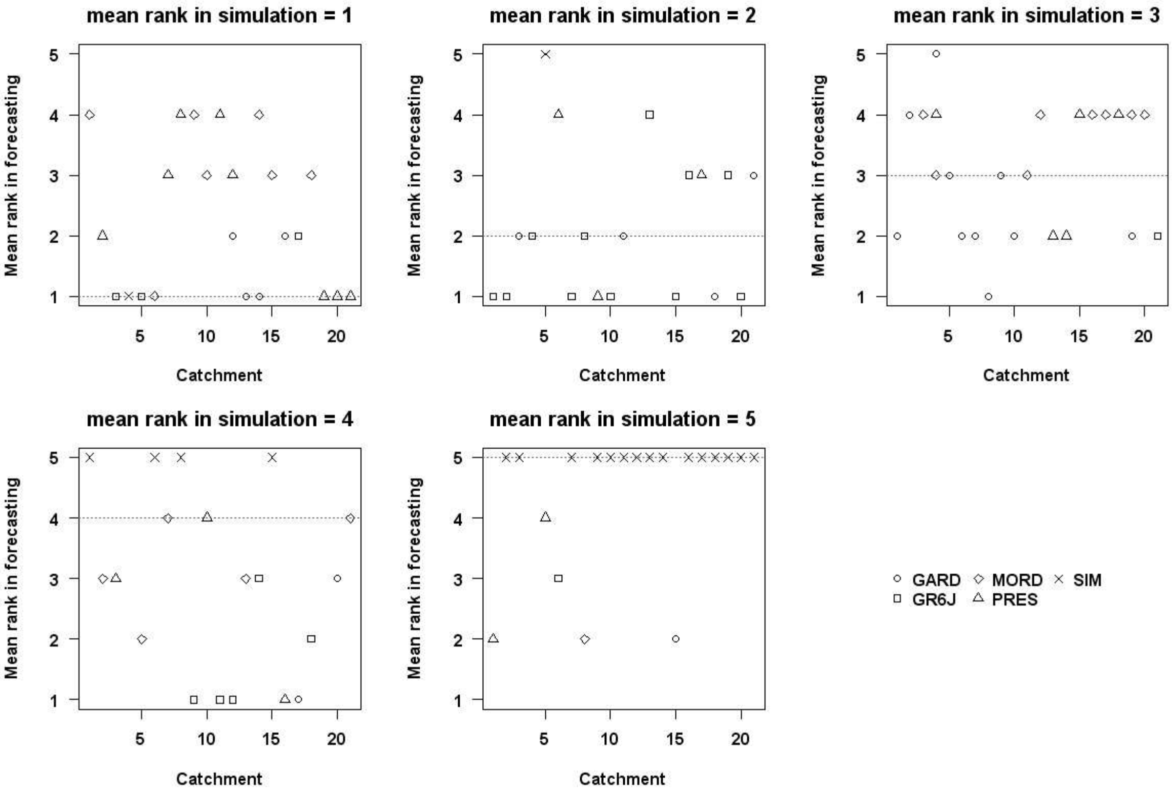
Figure 12~~: Observed and simulated hydrographs for (a) the Meuse River at St-Mihiel and (b) the Orge River at Morsang-sur-Orge for 1976 (top graph) and 1996 (bottom graph). The secondary axis shows rainfall.~~



~~Figure 11: Examples of forecasts issued by the five models tested and the benchmark every 20 days for the next 15 days for the Meuse River at St-Mihiel for 2003~~



~~Figure 12: Examples of forecasts issued by the five tested models and the benchmark every 20 days for the next 15 days for the Orge River at Morsang-sur-Orge for 1996~~



~~Figure 13~~: Mean rank in forecasting at the 7-day lead time for the 21 catchments for the models ranked 1st, 2nd,… 5th in simulation.

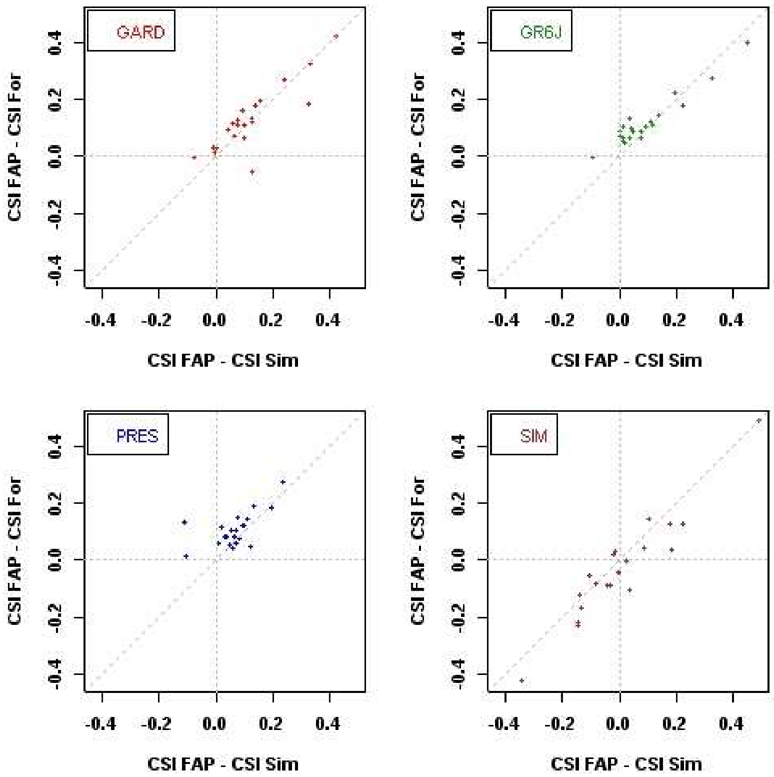


Figure 13: CSI difference for each model in forecasting mode when streamflow assimilation or output correction method is used (FAP) or not (For), versus CSI difference for each model in forecasting mode when streamflow assimilation or output correction method is used (FAP) and in simulation mode.

Mis en fo

Mis en fo

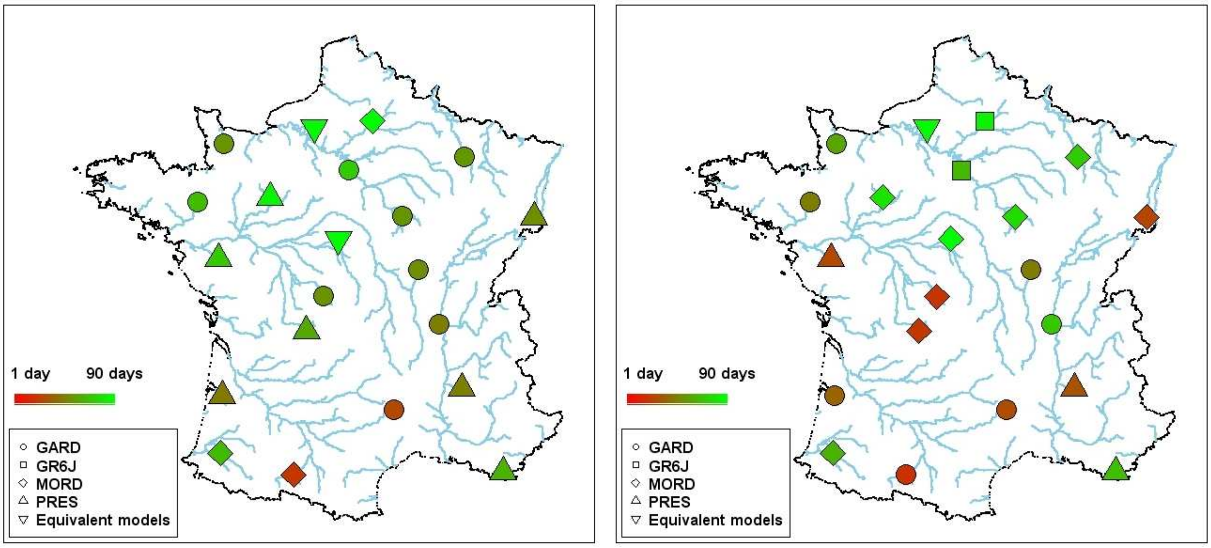


Figure 14: Map of useful forecasting lead time (UFL) for the 21 catchments, for validation periods 1 (left) and 2 (right). Symbols indicate the model which provides the best UFL and the colour scale indicates the value of this UFL.

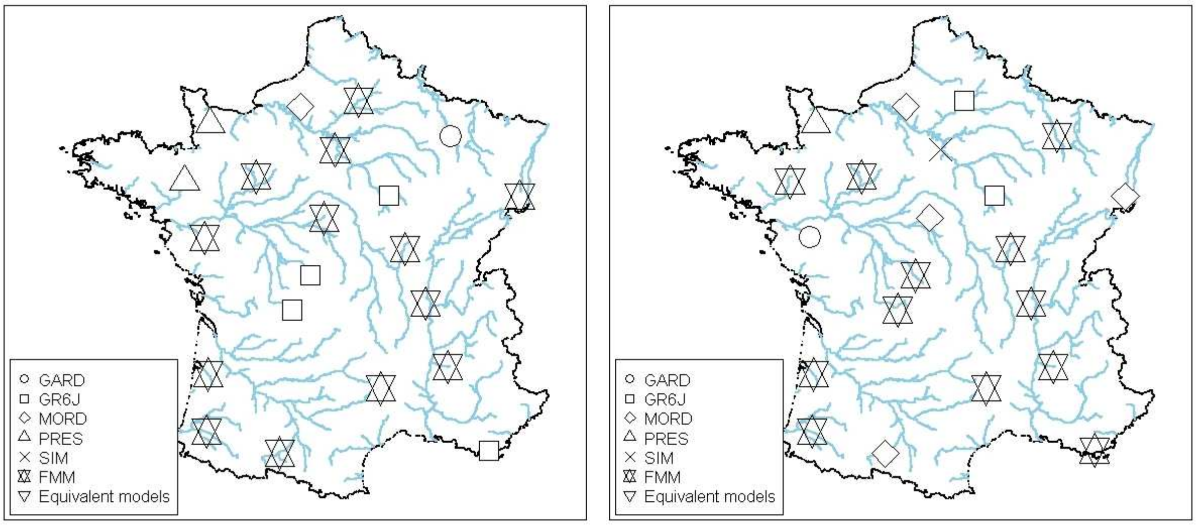


Figure 15: Maps of the model ranked best in simulation for the mean of all criteria and for validation periods 1 (left) and 2 (right), including the multi-model (fixed-weight average approach, FMM)

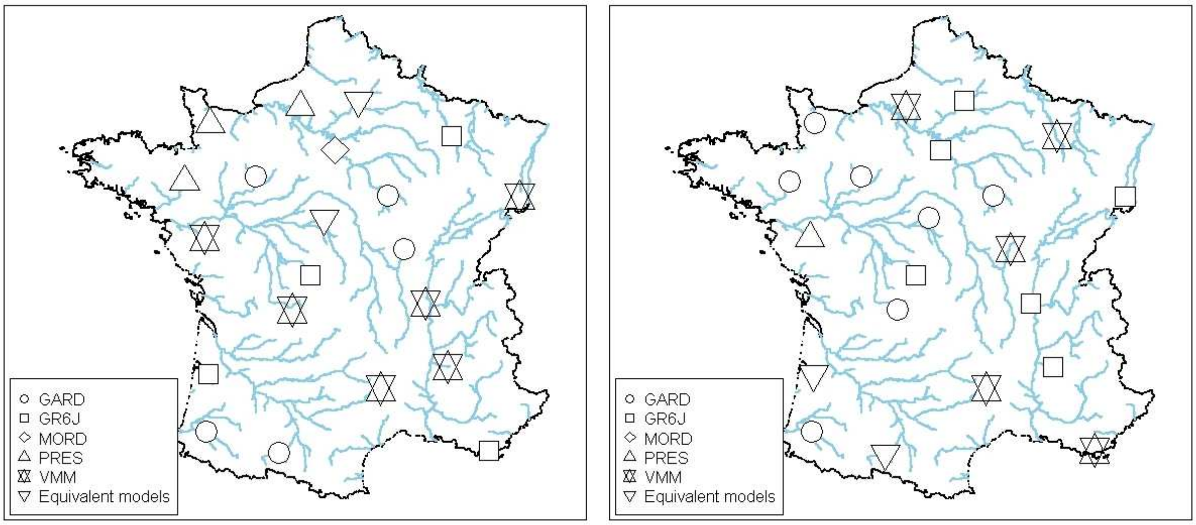


Figure 16: Maps of the model best ranked in forecasting for the mean of all criteria and for validation periods 1 (left) and 2 (right), for a *d*+7 forecasting lead time.

Mis en fo