The authors would like to thank the two anonymous referees for their careful reading and the interesting comments they provided that will contribute to the quality of the paper.

Shortly after initial submission, while working on the same database, we realize that the streamflow measurements of two of the 20 catchments where of dubious quality. Several clues led us conclude that they should not be longer included in the catchment pool. We sincerely apologize for any inconvenience this may have caused.

The two catchment that have been withdrawn from the initial submission are the ones that often behaved like outliers and were the most unreliable. Thus, the new results are more homogeneous. The two problematic catchments have been substituted by two new. Also, Figures 3, 4, 5, 6, 8, 9, 10, 11, 12, 13 were updated and are provided along this answer. Even if this does not affect any of the conclusion, the author suggest several modifications of the text:

- page 7193, line 17-19: Exceptions can be occasionally observed for catchment 3, 17, and 20 where only one or two models outperform the ensemble. should be replaced by Exceptions can be occasionally observed for catchment 3 and 17 where only one or two models outperform the ensemble.

- page 7194, line 5-6: we suggest to modify For the first lead time, most of the catchments are close to reliability while there is a clear outlier for which accuracy skills do not match its corresponding spread. In fact, this low performing catchment exhibits a constant hydrological wet bias partially explained by a meteorological forecast wet bias that over-forecasts precipitations by 15% that is not captured by any of the models even if the global tendency is respected. to For the first lead time, most of the catchments are close to reliability while there are two outliers for which accuracy skills do not match their corresponding spread. In fact, these catchments exhibits a constant hydrological bias partially explained by an inaccurate meteorological forcing that is not captured by any of the models even if the global tendency is respected

- page 7194, line 20: (outlier) should be deleted.

- page 7194, line 20-23: Data assimilation is particularly effective on catchments that present a systematic bias. For example, catchment number 11 that was problematic from the first lead time lies among the other catchments in terms of performance. should be deleted, even if we think that DA is particularly effective on catchment that have a systematic bias, but this assertion is no longer explicitly supported by the new Figure.

- page 7196, line 12-13: values should be replaced from While they are almost identical with a value of 0.55 and 0.57 mm day-1 respectively for the day 3, G spread drops to 0.44 mm day-1 for day 9 while the use of the MEPS maintains the spread to 0.55 mm day-1 to While they are almost identical with a value of 0.58 mm day-1 and 0.59 mm day-1 respectively

for the day 3, G spread drops to 0.45 mm day-1 for day 9 while the use of the MEPS maintains the spread to 0.59 mm day-1.

- page 7197, line 18-19: and only model 1 and 5 perform should be replace to and only models 1, 5, and 17 perform
- page 7198, line 3: catchment 19, should be replaced to catchment 20.
- page 7199, line 1: reducing the overdispersion with a sensible decrease in the ensemble spread from 0.65 to 0.54 mm day-1 should be replaced with reducing the overdispersion with a sensible decrease in the ensemble spread from 0.72 to 0.57 mm day-1
- page 7199, line 8-11: The two outlier catchments that exhibit poorer reliability present an underdispersed forecast that is a bit more pronounced for the H system than the H system (see Fig. 9). This indicates that uncertainties used to define the EnKF perturbations are under-estimated. should be suppressed. We also suggest to replace by As a matter of fact by Finally.

This manuscript investigated impacts of different uncertainty sources on streamflow forecasting comparing various combinations including multi-model ensemble, data assimilation, and meteorological ensembles. It fits well the scope of Hydrology and Earth System Sciences and the topic is of interest to a broad ranges of the scientific and engineering community. Their research questions and methodologies are of importance to better improve understanding on prediction uncertainty. However, for some study materials, description and information are not enough to convince general readers of their results. Especially, I have concerns on excessively simplified application of hydrologic models in terms of spatial and temporal scales and interpretation of contribution of different uncertainty sources. Therefore, revisions should be required to clarify several issues shown below before possible publication:

Major comments:

1. Multimodel ensemble:
Abstract: One of main findings of this manuscript is that the multimodel approach to take into account structural uncertainty supports the streamflow forecasts to maintain the required dispersion throughout the entire forecast horizon. However, such a statement might mislead a conclusion as if structural uncertainty is a dominating factor rather than forcing uncertainty, which could not convince readers with given results of this study. The fact that contribution of the meteorological ensemble forcing was negligible compared to deterministic one (Fig. 8) could strengthen such misinterpretation. In this study, as I understood correctly, input uncertainty (e.g. forecast

forcing) seemed to be compensated by structural uncertainty (e.g. multimodel) to enhance performance metrics. In addition, when we recall one of aims of this study is to decipher the traditional hydrometeorological sources of uncertainty (Page 7183), it is a bit doubtful if their aim was achieved and demonstrated successfully. Please clarify your findings and opinions on hydrologic prediction uncertainty which can be concluded from your study results.

The authors do not claim that the meteorological uncertainty should be neglected, far from it, and we agree that this source of uncertainty is among dominant ones. Nevertheless, the authors also believe and show that the uncertainty arising from the structure of hydrological model have to be taken into account. If not, system outputs will clearly underestimate total predictive uncertainty.

In this article, rather than trying to compensate for unaddressed sources of uncertainty (like overestimating structural uncertainty to balance a lack located in inputs), the authors try to decipher the different sources of uncertainty explicitly and coherently with dedicated tools that are meant to prevent overlapping in their respective action.

The contribution of meteorological ensemble forecasting may appear smaller than it is in reality for two reasons:

- The superiority of MEPS over deterministic NWP systems has already been demonstrated and MEPS are recognized particularly useful to estimate uncertainties in rainfall prediction. Here, results are assessed on the whole time period and this may contribute to dilute the importance of rainy days. We attached to this answer a complementary plot to Figure 8. The same comparison between system G and H is carried out but with 5 sub-periods of assessment (spring, summer, fall, winter, and days for which streamflow is higher than the yearly median streamflow). One can notice that the benefits of ensemble forcing are clearer during time of the year where streamflows are the higher (spring, $Q > Q_{50}$).

- The meteorological ensemble was not post-processed and this may contribute to underestimate its value. It is likely that HEPS performance could have been enhanced with improved MEPS that benefited from a suitable processing. Yet, we also would like to emphasize that the contribution of meteorological ensemble forcing is not only about a subtitle gain in the MCRPS metric but it provides a substantial improvement in reliability for longer lead times (page 7196, line 6-9). Nevertheless, we agree that removing the bias and correcting the dispersion of the meteorological ensemble could improve the reliability further. Thus MEPS forcing is a piece that cannot be overlooked for medium range forecasting.

2. Specification of individual model:

- **Line 7-8 at page 7186: It is not clear whether or not spatial discretization is considered to construct catchment applications by 20 conceptual lumped models. There are lots of ways and examples to apply lumped hydrologic models considering spatial heterogeneity. Please clarify this sentence and relevant comments below:**

- **If spatial discretization IS considered related to 1st comment, please clarify what spatial resolution was used. Additionally, how spatial heterogeneity was resolved in parameterization using lumped models?**

- **If spatial discretization is NOT considered, please clarify how large catchments (¿10,000 square kilometers) were conceptualized and parameterized.**

- **Regardless of spatial discretization, please clarify which flow routing methods were used in each model.**

To clarify line 7-8: The 20 models were derived from pre-existing models found in the literature. The models, in their original form, are either lumped (GR4J, GARDENIA, HBV, MOHYSE,...) or use a spatial discretization of the watershed (CEQUEAU, TOPMODEL, SACRAMENTO,...). For the models that are initially semi-distributed, they have been converted into lumped models. This has been done in order to facilitate their integration in a common framework (which is the multimodel framework that is used in this study) and for computational requirement. This is also why we emphasize line 4-5 that they are not the original models but only derived from the original models.

The 20 conceptual lumped models are applied in a traditional way. No spatial discretization has been done and hydrological processes are computed at the catchment scale. Consequently, the parameterization is uniform over the entire catchment, even for the largest one that spans over 15000 km.

We recognize that model spatial discretization can be useful and/or necessary for many purposes in hydrological sciences. However, conceptual lumped models are still very competitive, especially in the case where they are combined. We recently published a paper where we compare the multimodel hydrological forecast performance with a semi-distributed and physically based model. Results shown clearly that in our case, the multimodel was superior (Assessment of a multimodel ensemble against an operational hydrological forecasting system. A. Thiboult , F. Anctil Canadian Water Resources Journal / Revue canadienne des ressources hydriques Vol. 40, Iss. 3, 2015).

We dedicated a particular attention to the structural diversity of the lumped conceptual models, including the diversity of representation of flow routing as it is a driving process in hydrology. This ensures, or at least maximizes the chance to encompass the most effective way to achieve routing for a catchment by providing an ensemble of likely descriptions of the process. In depth description of

the routing of each model may be too long for a discussion but these information can be found in G. Seiller's Ph.D. thesis annex (valuation de la sensibilit des projections hydrologiques au choix des outils hydro-mtorologiques globaux conceptuels, http://theses.ulaval.ca/archimede/) from page 288 to 312 (in English).

3. Meteorological ensemble

- **In page 7185, rainfall forecasts seemed to be aggregated in space (one point per catchment) and time (daily). Please clarify possible impacts of excessive aggregation of rainfall forecasts on study results.**
- **For evaluating contribution of different uncertain sources on forecasts, it is essential to check bias of input forcing forecasts for varying lead times, while the manuscript only showed MCRPS. Please clarify the detailed analysis on it.**
- **In Line 8-9 Page 7186, please clarify the sentence such as modifications include their spatial discretization if they were initially distributed and their evapotranspiration formulation.**

It is expected that excessive spatial aggregation deteriorate the results with the possibility to miss local but driving events. However, in the article, several meteorological grid points are systematically situated within the catchment boundaries. Consequently, it is unlikely that the contribution of local event to the streamflow out the outlet is neglected, in the case that particular meteorological event was predicted by the MEPS. Concerning temporal aggregation, it is possible that there is some loss of information, especially for small catchments, but the hydrological models used in this study are designed to work with such time step. We are currently working on the conversion from a daily to a hourly time step multimodel framework.

MEPS performance are mentioned (indeed in terms of CRPS) but their performance respect to meteorological observations are not detailed in the paper. Other scores have been evaluated (NSE, RMSE, MAE, Normalized Root-mean-square error Ratio) and are in agreement with the CRPS values. The choice of the CRPS metric relies on the fact that it is a popular score that is strictly proper. Moreover, it has the advantage that it can be reduced to the MAE and thus allows a straightforward comparison of deterministic (system A) and probabilistic forecasts (system B to H'). Also, to prevent too inaccurate forcing, a selection in the catchments has been carried out. As explained line 15-18 of page 7185, the meteorological forcing quality was too poor for 18 of the 38 catchments. Thus they have been withdrawn from the catchment pool. It is likely that this could have been enhanced with meteorological post-processing but this was deemed out of scope of this study.

Line 8-9. This belongs to the modification of hydrological model part. For the same reason as previously mentioned, models are not the original models

but were derived from these models. Only the hydrological component of the model is kept. For the models that originally included a module to compute PET or snowmelt, this module has been omitted and replaced by a common one, CemaNeige for snow, and the formulation proposed by Oudin for the PET.

4. Information and analysis on catchments

- **In Section 2.1, information on catchments is limited. A new table showing information of each catchment such as catchment size, river length, low and high flow, typical time of concentration of flood, and etc, is required.**
- **Please clarify whether there are critical human intervention facilities such as dam reservoir, water gate, or weir in catchments. If there are, please clarify how such intervention was considered or affected in model configuration or results.**
- **Analysis on catchments in Results (e.g. Fig. 3, 5, 8, 10, 11, and 12) should be revised with additional analysis or figures regarding catchment characteristics such as catchment size or human intervention (e.g. Rakovec et al. 2015). References: Rakovec, O., Kumar, R., Mai, J., Cuntz, M., Thober, S., Zink, M., Attinger, S., Schafer, D., Schron, M., Samaniego, L. (2015): Multiscale and multivariate evaluation of water fluxes and states over European river basins, J. Hydrometeorol., in press,doi:10.1175/JHM-D-15-0054.1.**

The required table that describes some catchment characteristics is provided. A criterion in the selection of the catchments was the absence or the very deem influence of human intervention on streamflows. Considerable efforts have been paid to link estimated time of concentration, size of catchments and other variables without any clear results. We were not able to relate any catchment feature with particular results. This is a reason why catchments are presented anonymously in the paper.

5. DA

- **Please clarify how observation uncertainty was considered in EnKF. In conventional EnKF, noise for observation is commonly added to each ensemble which may lead to increase additional uncertainty. Otherwise, square root formulation can be used to avoid instability coming from observation noise.**
- **In Section 3.5, please clarify how EnKF perturbation was optimized in details in the case of H. Please remind that authors already mentioned that the optimal setting may use unrealistically high perturbations that compensate partially for the structural error.**

- **In Conclusion, authors described quick decrease of reliability is found in EnKF. However, it might be accelerated by coarse spatial and temporal resolution of models and input. Please clarify this issue.**

The formulation that was used is indeed the conventional EnKF. The additional noise to each member of the ensemble is not meant to increase uncertainty but rather to take it into account, in particular the uncertainty related to the catchment state. For the system H', streamflow observations are perturbed with random sampling from a normal law with a standard deviation equal to 10% of the observed streamflow and 0 mean. Added perturbations for precipitation are sampled from a gamma law with a standard deviation of 25% of the initial precipitation and finally, temperatures with a normal law with a 2C standard deviation, still with 0 mean (page 7198). For the system H, details on the perturbations added to streamflow, temperature and precipitation, but also about coupling between individual model and the EnKF can be found in a recently published article that was still under review at the time of initial submission of this article (On the difficulty to optimally implement the Ensemble Kalman filter: An experiment based on many hydrological models and catchments, Journal of Hydrology, Volume 529, Part 3, October 2015, Pages 1147-1160 A. Thiboult, F. Anctil).

We do not think that the decrease of reliability is due to spatial or temporal resolution but rather to the resilience of the models. Same results were found with a semi-distributed models with a 3h time step (Abaza, M.; Anctil, F.; Fortin, V. & Turcotte, R. Sequential streamflow assimilation for short-term hydrological ensemble forecasting. Journal of Hydrology, 2014, 519, 2692-2706). In the aforementioned article, the EnKF spread decreases quickly leading to unreliable forecast from day 2 but has been compensated by direct perturbations of states variable and meteorological ensemble forcing.

6. Figures and analysis

- **Model diagnostic metrics were drawn by aggregating results of all simulation periods. Additional analysis and description on conditional statistics of different flow regimes and seasons are highly recommended.**
- **Similar figures on reliability and catchment comparison are suggested to be removed or merged together.**

We suggest to add the figure that assesses the MCRPS according to the time of the year to the supplementary material of the article.

We would prefer to keep them separate. Even if the same metrics are presented, they represent the evolution and complexification of the systems and follow the article progression.

Minor comments:

1. Please use a consistent term between catchment and watershed throughout the manuscript.

This will be corrected.

2. In Fig. 7, the legend of a dotted line is not shown.

We are not sure to understand the remark. The dotted line corresponds to the RMSE of the ensemble as stated in the legend.

The authors would like to thank the two anonymous referees for their careful reading and the interesting comments they provided that will contribute to the quality of the paper.

Shortly after initial submission, while working on the same database, we realize that the streamflow measurements of two of the 20 catchments where of dubious quality. Several clues led us conclude that they should not be longer included in the catchment pool. We sincerely apologize for any inconvenience this may have caused.

The two catchment that have been withdrawn from the initial submission are the ones that often behaved like outliers and were the most unreliable. Thus, the new results are more homogeneous. The two problematic catchments have been substituted by two new. Also, Figures 3, 4, 5, 6, 8, 9, 10, 11, 12, 13 were updated and are provided along this answer. Even if this does not affect any of the conclusion, the author suggest several modifications of the text:

- page 7193, line 17-19: Exceptions can be occasionally observed for catchment 3, 17, and 20 where only one or two models outperform the ensemble. should be replaced by Exceptions can be occasionally observed for catchment 3 and 17 where only one or two models outperform the ensemble.

- page 7194, line 5-6: we suggest to modify For the first lead time, most of the catchments are close to reliability while there is a clear outlier for which accuracy skills do not match its corresponding spread. In fact, this low performing catchment exhibits a constant hydrological wet bias partially explained by a meteorological forecast wet bias that over-forecasts precipitations by 15% that is not captured by any of the models even if the global tendency is respected. to For the first lead time, most of the catchments are close to reliability while there are two outliers for which accuracy skills do not match their corresponding spread. In fact, these catchments exhibits a constant hydrological bias partially explained by an inaccurate meteorological forcing that is not captured by any of the models even if the global tendency is respected

- page 7194, line 20: (outlier) should be deleted.

- page 7194, line 20-23: Data assimilation is particularly effective on catchments that present a systematic bias. For example, catchment number 11 that was problematic from the first lead time lies among the other catchments in terms of performance. should be deleted, even if we think that DA is particularly effective on catchment that have a systematic bias, but this assertion is no longer explicitly supported by the new Figure.

- page 7196, line 12-13: values should be replaced from While they are almost identical with a value of 0.55 and 0.57 mm day-1 respectively for the day 3, G spread drops to 0.44 mm day-1 for day 9 while the use of the MEPS maintains the spread to 0.55 mm day-1 to While they are almost identical with a value of 0.58 mm day-1 and 0.59 mm day-1 respectively

for the day 3, G spread drops to 0.45 mm day-1 for day 9 while the use of the MEPS maintains the spread to 0.59 mm day-1.

- page 7197, line 18-19: and only model 1 and 5 perform should be replace to and only models 1, 5, and 17 perform
- page 7198, line 3: catchment 19, should be replaced to catchment 20.
- page 7199, line 1: reducing the overdispersion with a sensible decrease in the ensemble spread from 0.65 to 0.54 mm day-1 should be replaced with reducing the overdispersion with a sensible decrease in the ensemble spread from 0.72 to 0.57 mm day-1
- page 7199, line 8-11: The two outlier catchments that exhibit poorer reliability present an underdispersed forecast that is a bit more pronounced for the H system than the H system (see Fig. 9). This indicates that uncertainties used to define the EnKF perturbations are under-estimated. should be suppressed. We also suggest to replace by As a matter of fact by Finally.

The paper analyses different descriptions of uncertainty for hydrological ensemble forecasting and discusses their relative merit. It provides a valuable contribution to the research on probabilistic hydrological forecasting. However, different assumptions are made that may have a significant impact on the results and the general conclusions of the study. More elaborate discussions of the impact of these assumptions are needed (see specific comments below).

Specific comments

Page 7183, line 15. The term open loop scheme may not be familiar to all readers. It is explained later in Section 2.

Indeed, a short definition will be added in a future version.

Page 7185, line 7-8. Not clear why conversion to local time reduces the forecast horizon?

The meteorological forecast retrieved from the database is issued from 12am UTC up to ten days ahead with a 6-hour time step. Since the hydrometeorological day starts at 6am EST (or 12 UTC) for the catchments of the study, the first 12 hours of forecast are not used. Consequently, only 9 and half days of meteorological forecast are available. Lastly, because the time step is daily, the remaining 12 hours of forecast of the last day have been discarded.

Page 7185, line 10-14. Why first downscale and then aggregate to catchment rainfall? You could derive catchment rainfall directly from the ECMWF forecast.

We believe that this is something that should be avoided. The raw resolution of the ECMWF is too coarse for this application and do not match

systematically catchment size, as 6 of them are smaller than 1000km2. Without interpolation, it is possible that only one meteorological forecast grid point would fall within catchment boundaries, if any. Also, as most catchment straddle several initial ECMWF grid points, interpolation allow to take into account the contribution of each of these grid point. Finally, considering the influence of more than one meteorological grid point allows for smoothing out individual members occasional instability.

Page 7185, line 14-18. Pre-processing of meteorological forecasts is widely used in hydrological forecasting systems to improve forecast accuracy and reliability. Since this is not done in the study, the value of the rainfall forecast will most probably be underestimated.

We agree on that point but it was not intended to investigate such pre-processing in this article as we deemed that it is a step that is sufficiently complex that it may require specific investigation. However, one can note that recent attempts at pre-processing meteorological forecasts have been performed without much success at improving streamflow forecasts, although the improvement of the meteorological forcing was indeed successful (e.g. Verkade, J. S.; Brown, J. D.; Reggiani, P. & Weerts, A. H. Post-processing ECMWF precipitation and temperature ensemble reforecasts for operational hydrologic forecasting at various spatial scales Journal of Hydrology, 2013, 501, 73-91). In the light of the comments of both reviewers, we realize that we should remind the reader and emphasize that no pre-processing is used and that the interpretation of the results should be done accordingly, in particular in Section 3.3. It is expected that a successful post-processing would enhance the MEPS capabilities to decipher meteorological uncertainty and would be eventually cascade these benefits through the hydrological systems, thus leading to better accuracy and reliability.

Page 7187, line 20. The H operator has an index t in the equation. I would not expect H to be time varying.

Indeed, the index will be removed in the future version.

Page 7188, line 5-6. Different variants of the EnKF have been proposed in literature. Which method is applied here, and why?

The EnKF has been implemented in its traditional form (Evensen, G. The Ensemble Kalman Filter: theoretical formulation and practical implementation Ocean Dynamics, 2003, 53, 343-367). We had the opportunity to developed our expertise by studying in detail the interactions between the filter and the different models (On the difficulty to optimally implement the Ensemble Kalman filter: An experiment based on many hydrological models and catchments, Journal of Hydrology, Volume 529, Part 3, October 2015, Pages 1147-1160 A. Thiboult, F. Anctil). This variant of the EnKF, if properly set, proved to be able to efficiently reduce the initial condition uncertainty and thus to fulfill the expectations we have from it, in regards with to the hydrometeorological setup that is presented here.

Page 7188, line 21-22. How is reliability and accuracy evaluated in the tuning of the EnKF?

The accuracy is assessed with the NSE that is computed on the median of the ensemble and the Normalized Root-mean-square error Ratio (NRR) is used for reliability assessment. Then, the 2-step criterion described page 7189 line 1-3 is applied. Results concerning the tuning of the EnKF can be found in the article cited in the previous answer.

Page 7188, line 22-27. Only uncertainty in model forcing is assumed, and hence this uncertainty should compensate also for other model uncertainties such as parameter uncertainty. Model parameter uncertainty could be included in the EnKF. This would most likely improve the reliability of the EnKF since this would add uncertainty in the forecast period by propagation of parameter uncertainty.

It is practically hard to untangle uncertainties through the use of EnKF only. EnKF, it its traditional form, can decipher the overall predictive uncertainty but does not distinguish between input-output, structural, and parameter uncertainty. By artificially and deliberately overestimating the input uncertainty, it is possible to compensate for the other uncertainties and achieve reliability for simulation and possibly for the first (and sometimes second) forecast day. This may be desirable only in a case where there is no other tool available to handle the other sources.

We did not consider dual parameter-state variable updating since the multimodel approach allows to take into account parameter uncertainty without the need to modify (update) time invariant values. Thus, model parameter uncertainty is treated outside of the EnKF through the multimodel approach. Moreover, it is shown that several dissimilar hydrological models bring much more diversity than traditional parameter uncertainty estimations, thus indicating that structural uncertainty, in some ways, encompasses parameter uncertainty (Poulin, A., Brissette, F., Leconte, R., Arsenault, R., and Malo, J. S.: Uncertainty of hydrological modelling in climate change impact studies in a Canadian, snow-dominated river basin, J. Hydrol., 409, 626636, doi:10.1016/j.jhydrol.2011.08.057, 2011. 7183)

Page 7188, line 27-28. The definition of the state vector is not clearly described. The state vector is uniquely defined by the system model. Typically, for lumped, conceptual rainfall-runoff models it will consist of storages of the different conceptual reservoirs.

By state vector we meant the ensemble of state variables that are updated. It is indeed a mistake as the definition of state vector is the one you gave. This will be changed.

Page 7194, line 7-8. This illustrates the problem of not pre-processing rainfall forecasts cf. comment above.

We are aware of the limitations that are induced by not pre-processing rainfall forecast and we are currently carrying out research on the subject. The

sentence line 7-8 was initially written to explain the behavior of one of the catchment that has been withdrawn from the database but your remark about pre-processing remains valid.

Page 7195. The results for the EnKF are due to an incomplete description of the uncertainty. It provides a good description of the initial uncertainty but this is quickly washed out of the system as forecast lead time increases. Use of a more elaborate description of the uncertainty in the EnKF would improve the reliability, e.g. by including model parameter uncertainty cf. comment above.

This framework should be regarded as a possible way toward accurate and reliable forecast but we strongly concur that it is not the only one. More sophisticated version of the EnKF may indeed improve the description of uncertainties for longer lead times. In a different study, we tested direct state variable perturbations (Abaza, M.; Anctil, F.; Fortin, V. & Turcotte, R. Sequential streamflow assimilation for short-term hydrological ensemble forecasting. Journal of Hydrology, 2014, 519, 2692-2706). It contributes to maintain the dispersion a little longer (about 1 day longer) but the spread is at the end also maintained by the MEPS forcing spread. However, these results should not be compared in a very strict way with this paper since the hydrological models are different. A secondary objective of the article is also to show that with the traditional formulation of EnKF the spread (and the corresponding description of uncertainty) is not sufficient but can be compensated by the use of multimodel. Also, the combination of the multimodel and the traditional EnKF makes that is not necessary to resort to dual state variable-parameter updating. By keeping the parameters time invariant, the inner model dynamic is better preserved.

Page 7196. I think the lack of pre-processing of the rainfall forecast ensemble can explain the small impact observed of using a probabilistic rainfall forecast.

Despite the absence of pre-processing, the reliability is still better with probabilistic forecast. The hydrological ensemble spread is substantially larger and one could expect this spread to contribute more actively to reliability if the bias of the forcing would be removed.

Page 7198, line 20-23. Not clear why this would correspond to an optimal EnKF?

Traditional EnKF accounts for input and output uncertainty explicitly. It could be optimal in a perfectly controlled environment where only the input and output are subject to uncertainty (in a synthetic experiment for example). Thus there shouldn't be any uncertainty in the structure / parameter / conceptualization. It is suboptimal in real cases as it has to account for other sources of uncertainty if used with a single model.

Page 7198, line 26-28. Not clear.

This assertion is closely related to the previous question. In the case where

the different sources are not explicitly accounted for by dedicated tools, the EnKF has to compensate for them. One way to achieve reliability is to add perturbations to input. However, there is no obvious way to know by which amount the uncertainty on input should be overestimated to compensate for the other uncertainties. Thus, to ensure hydrological reliability, one needs to perform a calibration of EnKF hyper parameters (research of required noise magnitude) which is a fastidious step.

Page 7198. Figure 12 is not referred in Section 3.5.

Indeed, it was originally meant to be in the article but we finally decided to put it only as supplementary material as the main changes concern reliability and forgot to remove the corresponding figure from the manuscript.

Page 7200, line 5-10. There seems to a contradiction here. First, it is stated that the EnKF does not provide a satisfactory uncertainty propagation. And then it is stated that the EnKF is the component that provides the most dispersion.

It requires indeed some clarifications. It should have been specified that it is the component that provides the most dispersion, but only for the first lead times.

Manuscript prepared for Hydrol. Earth Syst. Sci.
with version 2014/09/16 7.15 Copernicus papers of the LaTeX class copernicus.cls.
Date: 25 January 2016

Accounting for three sources of uncertainty in ensemble hydrological forecasting

A. Thiboult[1], F. Anctil[1], and M.-A. Boucher[2]

[1]Dept. of Civil and Water Engineering, Université Laval, 1065 avenue de la Médecine, Quebec, Canada
[2]Dept. of Applied Sciences, Université du Québec à Chicoutimi, 555, boulevard de l'Université, Chicoutimi, Canada

Correspondence to: Antoine Thiboult, antoine.thiboult.1@ulaval.ca

Abstract. Seeking for more accuracy and reliability, the hydrometeorological community has developed several tools to decipher the different sources of uncertainty in relevant modeling processes. Among them, the Ensemble Kalman Filter, multimodel approaches and meteorological ensemble forecasting proved to have the capability to improve upon deterministic hydrological forecast. This study aims at untangling the sources of uncertainty by studying the combination of these tools and assessing their contribution to the overall forecast quality. Each of these components is able to capture a certain aspect of the total uncertainty and improve the forecast at different stage in the forecasting process by using different means. Their combination outperforms any of the tool used solely. The EnKF is shown to contribute largely to the ensemble accuracy and dispersion, indicating that the initial condition uncertainty is dominant. However, it fails to maintain the required dispersion throughout the entire forecast horizon and needs to be supported by a multimodel approach to take into account structural uncertainty. Moreover, the multimodel approach contributes to improve the general forecasting performance and prevents from falling into the model selection pitfall since models differ strongly in their ability. Finally, the use of probabilistic meteorological forcing was found to contribute mostly to long lead time reliability. Particular attention needs to be paid to the combination of the tools, especially in the Ensemble Kalman Filter tuning to avoid overlapping in error deciphering.

1 Introduction

The complexity of hydrometeorological systems is such that it is not possible to perfectly represent their "true" descriptive physical processes, and even less to integrate them forward in time with

mathematical models. These models are only an approximation of varying quality to represent and predict variables of interest, yet they proved to be skilful and useful for water resource management and hazard prevention (e.g. Bartholmes et al., 2009; Pagano et al., 2014; Demargne et al., 2014).

Inadequacies between simulation or predictions and observations can be largely attributed to the many sources of uncertainty that are located along the meteorological chain (e.g. Walker et al., 2003; Beven and Binley, 2014). Hence, it is admitted that improvement of the forecast ought to go through understanding and reducing the sources of uncertainty (e.g. Liu and Gupta, 2007). These sources have different nature that range from epistemic uncertainty due to the imperfection of our knowledge to variability uncertainty where the imperfections are due to the inherent system variability (e.g. Walker et al., 2003; Beven, 2008). They also differ in location, i.e. where they lay in the hydrometeorological modeling process: meteorological forcing, model parameter and structure, hydrological initial conditions, and, to a lesser extent, observations (Walker et al., 2003; Vrugt and Robinson, 2007; Ajami et al., 2007; Salamon and Feyen, 2010).

As all models are exposed to these sources of uncertainty, they necessarily lead to forecasts with imperfections. It is thus possible – and frequent – that several models can simulate the process of interest with the same accuracy. These simulation are equally likely in the mathematical sense; it is referred as the principle of equifinality (Beven and Binley, 1992).

Ensembles provide a probabilistic answer to the equifinality problem. They are a collection of deterministic predictions issued by different models to simulate the same event and attempt to produce a representative sample of the future. They can be built by a suitable method wherever a source of uncertainty needs to be put under scrutiny. Additionally, the ensemble mean generally have better skills than deterministic systems and offer a better ability to forecast extreme events (e.g. Wetterhall et al., 2013).

As the sources of uncertainty differ in their location, nature and statistical properties, they need specific tools to be deciphered efficiently (Liu and Gupta, 2007). A wide range of methods have been developed in the past year to cater hydrological forecast needs.

At the beginning of the 90s, meteorologists pioneered the operational use of ensembles by constructing Meteorological Ensemble Prediction Systems (MEPS), mostly to take into account imperfect initial conditions that is a prime importance uncertainty source in view of the chaotic nature of the atmospheric physics. Several methods have been proposed to tackle this issue. For instance, to define the initial condition uncertainty, the European Center for Medium-Range Weather Forecasts (ECMWF) generates an ensemble by initiating their model with singular vectors (Molteni et al.,

1996) to which a stochastic scheme is added to deal with the model physical parametrisation uncertainty (Buizza et al., 1999).

The increasing accessibility of MEPS benefited to the hydrology community to issue probabilistic hydrological forecasts that take into account meteorological uncertainty forcing with Hydrological Ensemble Prediction Systems (HEPS, e.g. Cloke and Pappenberger, 2009; Brochero et al., 2011; Boucher et al., 2012; Abaza et al., 2014). Since 2007, The Observing System Research and Predictability Experiment (THORPEX) Interactive Grand Global Ensemble (TIGGE) allows free access to meteorological ensemble forecasts for hydrologists and other researchers. This database regroups the outputs from nine operational atmospheric models around the world, which can be downloaded in grib2 format.

A lot of attention has been paid to the identification of hydrological model parameters and the non uniqueness of the solutions. Among other technique, Vrugt et al. (2003) proposed the Shuffled Complex Evolution Metropolis Algorithm (SCEM-UA), a calibration technique that retains several sets of parameters instead of a single one for a more realistic assessment of parameter uncertainty. Beven and Binley (1992) suggested a more comprehensive approach for model acceptance or rejection with the Generalized Likelihood Uncertainty Estimation (GLUE) that allows to include different forms of competing models.

Gourley and Vieux (2006) assert that dealing only with input and parameter uncertainty is likely to issue unreliable forecast and that hydrological model structural uncertainty should be deciphered explicitly. This statement is substantiated by Clark et al. (2008) who compares 79 unique model structures and concludes that a single structure is unlikely to perform better than the others in all situations. Poulin et al. (2011) adds that the structural uncertainty is larger than the parameter estimation uncertainty and provides more diverse outputs. Combining dissimilar hydrological model structures proved to possess a great potential (Breuer et al., 2009) even with simple combination patterns (Ajami et al., 2006; Velázquez et al., 2011; Seiller et al., 2012).

Initial condition uncertainty has also aroused scientific interest. Many studies using various data assimilation techniques to incorporate observations within the simulation processes demonstrated that the specification of catchment descriptive states is a crucial aspect of short and medium range forecasts (DeChant and Moradkhani, 2011; Lee et al., 2011). Among them, sequential data assimilation technique such as the Particle Filter (e.g. DeChant and Moradkhani, 2012; Thirel et al., 2013), the Ensemble Kalman Filter (e.g. Weerts and El Serafy, 2006; Rakovec et al., 2012) and variants (Noh et al., 2013; Chen et al., 2013; McMillan et al., 2013; Noh et al., 2014) substantially improve forecast over the open loop scheme (i.e. no data assimilation is performed), by reducing and charac-

terizing the uncertainty in initial conditions.

Considerable efforts have been made in the development of these sophisticated techniques and this gave rise to many tools that have been individually tested useful. As Bourdin et al. (2012) points out, "To date, applications of ensemble methods in streamflow forecasting have typically focused on only one or two error sources [...] A challenge will be to develop ensemble streamflow forecasts that sample a wider range of predictive uncertainty". As underlined, the forecasting tools frequently tackle different sources of uncertainty and therefore do not exclude each other but can be seen as complementary, combining their assets to compose an overall better system.

The present study identifies three efficient tools, namely a hydrological multimodel approach, Ensemble Kalman Filter, and MEPS forcing that are used together to decipher the traditional hydrometeorological sources of uncertainty. The paper scope is to identify how they are complementary to each other, to assess their individual contribution to the hydrological forecast reliability and accuracy, and to eventually evaluate the possibility of achieving reliability without resorting to post-processing.

This is achieved by issuing a hindcast on 20 catchments using the aforementioned techniques, either individually or combined, to investigate their specific role in the forecasting process. Each of them produces an ensemble that can be cascaded through the next ensemble technique in order to produce a larger ensemble that possesses a more comprehensive error handling. Finally, if all sources of error are accounted for, the ensemble should generate a forecast that is reliable (Bourdin et al., 2012).

This paper is organized as follow: section 2 presents the catchments, models, the Ensemble Kalman Filter basics and scores, section 3 sums up the systems specificities and their respective performances followed by a conclusion in section 4

2 Material and methodology

2.1 Catchments and hydrometeorological data

20 catchments situated in the south of the Province of Québec have been selected for this study (Fig. 1). The catchments experience a mixed hydrological regime with a spring freshet resulting from the important winter snow cover and a lesser second peak in autumn. There is little or no human intervention on the catchments.

Table 1. Main characteristics of the 20 catchments

River name	Area (km2)	River length (km)	Average slope (%)	Mean ann. Q (m3/s)	Coeff. of variation of Q	Mean ann. P (mm)	Mean ann. snow (cm)
Trois Pistoles	923	52	0.52	18	1.81	1109	382
Du Loup	512	45	0.78	10	1.47	1050	378
Gatineau	6796	190	0.12	127	1.08	1023	332
Dumoine	3743	145	0.13	50	0.81	968	297
Kinojévis	2572	83	0.12	39	1.12	921	324
Matawin	1383	68	0.29	24	1.11	1025	328
Croche	1551	102	0.33	29	1.24	996	360
Vermillon	2650	145	0.20	39	1.10	957	312
Batiscan	4483	167	0.45	96	1.03	1162	381
Saint-Anne	1539	84	0.81	51	1.20	1412	502
Bras du Nord	643	77	0.82	19	1.21	1385	499
Du loup	767	57	0.78	12	1.27	1020	332
Aux Ecorces	1107	54	1.04	28	1.09	1236	450
Métabetchouane	2202	155	0.43	48	1.19	1168	420
Péribonka	1010	101	0.50	19	1.16	1000	376
Ashuapmushuan	15342	342	0.16	300	0.92	984	379
Ashuapmushuan	11200	232	0.12	227	0.88	1001	394
Au Saumon	586	69	0.65	8	1.36	877	334
Mistassini	9534	278	0.20	200	1.08	1004	409
Valin	761	59	1.06	24	1.13	1123	453

~~The climatology of the catchments is varied, with a mean annual snow fall ranging from 3 meters to 5 meters and total precipitation fluctuates between 877 mm to 1412 mm. The size of the watersheds extends from 512 km^2 to 15342 km^2 and annual mean streamflow from 8 m^3/s to 300 m^3/s.~~

The climatology of the catchments is varied (Table 1), in particular in terms of annual snow fall and annual total precipitation. The differences in the catchment physical characteristics (area, length, slope,...) and in climatology are reflected in their streamflow statistics (e.g. average streamflow, coefficient of variation).

Daily total precipitation, maximum and minimum temperature and streamflows are provided by the Centre d'Expertise Hydrique du Québec. They performed kriging on the observations over a 0.1° resolution grid to which a temperature correction with an elevation gradient of -0.005°C/m is added. The data base is split into three periods: 1990-2000 for the calibration of the models, Octo-

ber 2005-October 2008 for the spin up, while November 2008-December 2010 is committed to the hydrological forecast assessment.

The MEPS used as inputs to the hydrological model were retrieved from the TIGGE database. The temperatures and precipitation forecasts from the European Center for Medium range Weather Forecasts (ECMWF) were chosen for this study. They are formed by 50 exchangeable members (Fraley et al., 2010) with a 6 hours time-step and a 10 day horizon. However, after conversion from Greenwich time to local Quebec time, the horizon reduces to 9 days. For the sake of the study and to match the common framework of the hydrological models, weather forecast is aggregated at a daily time step. ~~The forecast is provided on a regular grid with a 0.5° resolution (N200 Gaussian grid) that is downscaled to a 0.1° resolution during data retrieval by using bilinear interpolation.~~ The ECMWF raw forecast is provided on a regular grid with a 0.5° resolution (N200 Gaussian grid), which is too coarse for this application, especially for the smallest catchments. To ensure to have several representative grid points situated within the catchment boundaries, meteorological forecast is downscaled to a 0.1° resolution during data retrieval by using bilinear interpolation (e.g. Gaborit et al., 2013). Also, the interpolation allows to take into account the contribution of the grid points that are close but not directly situated within catchment boundaries and thus allows a better description of the catchment meteorological conditions. As the rainfall-runoff models are lumped, a single representative point forecast is obtained for each MEPS member by averaging the downscalled grid points situated within the catchment boundaries.

The weather forecast displays acceptable performance over the 20 selected catchments. In fact, in the initial group of 38 catchments, 18 displayed unsatisfactory performances so they where withdrawn from the experiment from the beginning, as pre-processing the meteorological inputs falls outside the scope of the project. When compared to the meteorological observations, rainfall and temperature $MCRPS$ over the 9 days (see sect. 2.4) remain below 3 mm and 3°C respectively for selected catchments. Other scores have been evaluated (Nash-Sutcliffe efficiency, root-mean-square error, mean absolute error, normalized root-mean- square error ratio) and are in agreement with the $MCRPS$ values, confirming the exclusion of the catchments.

An alternative to the ECMWF ensemble is used to simulate a deterministic meteorological forcing with equivalent theoretical skill. For this purpose, a single member is drawn randomly among the 50 exchangeable members.

2.2 Models, snow module and evapotranspiration

The multimodel ensemble is composed of 20 conceptual lumped models. In this study, their outputs are pooled together with equal weights or studied individually. Models have been initially selected

by Perrin (2000) for their conceptual and structural diversity and revised by Seiller et al. (2012). They present various degrees of complexity: 4 to 10 calibrated parameters and 2 to 7 reservoirs to describe the main hydrological processes (Table 2). The model selection is a key element for an efficient multimodel ensemble as the diversity of them contributes to encompass the error in model conceptualization and structure (Viney et al., 2009). A close attention has been paid to the diversity of the different components of the models, in particular to the representation of the different storages and flows. This ensures, or at least maximizes the chance to encompass the most effective way to describe storage and routing by providing an ensemble of likely descriptions of the processes. All models were derived from existing ones, keeping their main specificities but adapting them to match a common framework where every snow module-model sets share the same inputs, namely precipitation and potential evapotranspiration. The models, in their original form, are either lumped (GR4J, GARDENIA, HBV, MOHYSE,...) or use a spatial discretization of the catchment (CEQUEAU, TOPMODEL, SACRAMENTO,...). For the models that were initially semi-distributed, they have been converted into lumped models (Perrin, 2000). This has been done in order to facilitate their integration in the common framework used in this study and for computational requirement.

The 20 conceptual lumped models are applied in a traditional way, i.e. no subsequent spatial discretization has been done, hydrological processes are computed at the catchment scale and the parameterization is uniform over the entire catchment. Despite their simplicity and the approximations they rely on, they have shown to perform well and be competitive with more complex ones, especially when combined together (Thiboult and Anctil, 2015). ~~Modifications include their spatial discretization if they were initially distributed and their evapotranspiration formulation. The snow accumulation and melt module have been also omitted in the case they had their own to be replaced by Cemaneige.~~

The snow accumulation and melt module, as well as the evapotranspiration formulation, have been also omitted in the case they had their own to be replaced by Cemaneige and Oudin's potential evapotranspiration formulation respectively. A detailed description of the models structure can be found in Perrin (2000).

Cemaneige, a degree day snow accounting routine, is used to model the catchment snow processes (Valery et al., 2014). It divides the catchment into 5 elevation bands and requires 2 parameter to be calibrated: a snowmelt and a cold-content factor. As it is calibrated conjointly with individual models and according to an objective function based on streamflow observations, its parameter values depend on the hydrological model with which it is coupled. The 20 hydrological models have therefore precipitation inputs that are driven by the same snow accounting routine but differently parametrized. Thus, part of the uncertainty related to the snowmelt module is taken into account

Table 2. Main characteristics of the 20 lumped models (Seiller et al., 2012)

Model acronym	Number of optimized. parameters	Number of reservoirs	Derived from
M01	6	3	BUCKET (Thornthwaite and Mather, 1955)
M02	9	2	CEQUEAU (Girard et al., 1972)
M03	6	3	CREC (Cormary and Guilbot)
M04	6	3	GARDENIA (Thiery, 1982)
M05	4	2	GR4J (Perrin et al., 2003)
M06	9	3	HBV (Bergström and Forsman, 1973)
M07	6	5	HYMOD (Wagener et al., 2001)
M08	7	3	IHACRES (Jakeman et al., 1990)
M09	7	4	MARTINE (Mazenc et al., 1984)
M10	7	2	MOHYSE (Fortin and Turcotte, 2007)
M11	6	4	MORDOR (Garçon, 1999)
M12	10	7	NAM (Nielsen and Hansen, 1973)
M13	8	4	PDM (Moore and Clarke, 1981)
M14	9	5	SACRAMENTO (Burnash et al., 1973)
M15	8	3	SIMHYD (Chiew et al., 2002)
M16	8	3	SMAR (O'Connell et al., 1970)
M17	7	4	TANK (Sugawara, 1979)
M18	7	3	TOPMODEL (Beven et al., 1984)
M19	8	3	WAGENINGEN (Warmerdam et al., 1997)
M20	8	4	XINANJIANG (Zhao et al., 1980)

through dissimilar parameter sets that drives the state of the snow pack accumulation and melting.

All models were given the same input potential evapotranspiration which is computed following Oudin et al. (2005) formula that relies on the mean air temperature and the calculated extraterrestrial radiation.

2.3 Forecasting approaches

Two approaches are used and compared for forecasting, the open loop and the Ensemble Kalman Filter. Regardless of the method used, the meteorological observations over the three years preceding the forecast period are used for model spin up to bring models states to values that estimates the catchment conditions.

2.3.1 Open loop forecasting

When the open loop forecast is activated, the state variables are obtained in simulation mode and used as starting point to initiate the hydrological forecast. The simulation and forecast steps then alternate as follow: 1) the models are forced with observations up to the first day t of the forecast and 2) the models are next forced with the meteorological forecast to issue the hydrological prediction until $t+9$. The procedure is repeated as the models are brought forward in time with the observations from t.

2.3.2 Ensemble Kalman Filter

The Ensemble Kalman Filter (EnKF) is a sequential data assimilation technique that uses a recursive Bayesian estimation scheme to provide an ensemble of possible model reinitializations. The model state variable vector $\boldsymbol{X}$ is updated according to its likelihood probability density function that is inferred by the observations z, $p(\boldsymbol{X}_t|\boldsymbol{z}_{1:t})$ with the indices t referring to the time.

When an observation becomes available, model states are updated ($\boldsymbol{X}^+$, the a posteriori estimation) as a combination of the predicted ($\boldsymbol{X}^-$, also called the a priori states) and the difference between the prior estimate of the variable of interest $\mathbf{H}\boldsymbol{X}^-$ and the corresponding observation z_t.

$$\boldsymbol{X}_t^+ = \boldsymbol{X}_t^- + \mathbf{K}_t(z_t - \mathbf{H}\boldsymbol{X}_t^-) \tag{1}$$

where $\mathbf{H}$ is the observation model that relates the state vectors and observations, and $\mathbf{K}$ is the Kalman gain matrix that defines the relative importance given to the output error respect to the prior state estimate.

The Kalman gain is defined with the model error covariance matrix $\mathbf{P}_t$ and the covariance of observation noise $\mathbf{R}_t$ as:

$$\mathbf{K}_t = \mathbf{P}_t\mathbf{H}^T(\mathbf{H}\mathbf{P}_t\mathbf{H}^T + \mathbf{R}_t)^{-1} \tag{2}$$

A detailed explanation of the EnKF mathematical background and concepts can be found in Evensen (2003). In this study, the filter has been implemented in its traditional form following Mandel (2006).

The EnKF is able to decipher catchment initial condition as it acts on variables after the spin up time, i.e. at the very start of the hydrological forecast. Thus, it is frequently presented as a tool that describes catchment descriptive states uncertainty such as soil moisture but it also implicitly takes into account model parameter and structural uncertainty as these are reflected in the model states and outputs errors. The forecast system comprises inaccuracies at several levels and consequently

the error statistics that the EnKF uses to update state variables are not only intrinsic variability but also epistemic uncertainty that lay also in the value of the state variables.

The EnKF performance is highly influenced by its setting, in particular by the required noise specification of inputs and outputs (Noh et al., 2014) and also by the choice of the updated state variables ~~vector~~ (Li et al.). This affects directly the spread of the ensemble and the corresponding uncertainty description (Thiboult and Anctil, 2015a). As the level of uncertainty varies from the model used and the simulated catchment, the optimal EnKF implementation also depends to a great extent on these aspects (Thiboult and Anctil, 2015a).

In practice, it is complex to untangle uncertainties through the use of the EnKF. The filter, in its traditional form, can decipher the overall predictive uncertainty but does not distinguish between input-output, structural, and parameter uncertainty. By artificially and deliberately overestimating the input uncertainty, it is possible to compensate for uncertainties that are not explicitly addressed and achieve reliability in simulation and possibly during forecast for the first lead times.

In this study, the EnKF is tuned to optimize reliability and accuracy per catchment and per model. The retained specification are identified after extensive testing has been carried out. More precisely, two or three noise levels for each input and output were tested (a 25-50-75% standard deviation of the mean value with a gamma law for precipitation, 10-25-50% standard deviation of the mean value with the normal law for streamflow observations and 2-5° standard deviation with a normal law for the temperature). Additionally, as the choice of updated state variables is also a key component of the EnKF, all possible combinations of ~~the state vector~~updated state variables were tested with the 12 noise combinations described above. The retained EnKF setting were based on a two-step criterion; firstly the 3 settings that presented the best reliability are kept and then the one among them that led to the lowest bias. Therefore, the optimal setting may use unrealistically high perturbations that compensate partially for the structural error. A detailed description of the EnKF optimization with the 20 models is provided in Thiboult and Anctil (2015a)

In this study where the EnKF is meant to be combined with the multimodel approach, dual state-parameter updating was not considered since it is expected that the multimodel takes simultaneously into account structural and parameter uncertainty (Poulin et al., 2011), releasing the need to modify (update) model time-invariant parameters.

2.4 Scores

The continuous ranked probability score ($CRPS$, Matheson and Winkler, 1976) is a common verification tool for probabilistic forecasts that assesses accuracy and resolution. A cumulative distribution

function is built based on the raw predictive ensemble, i.e. the collection of deterministic forecasts and then compared to the observation.

$$CRPS(F_t, x_{obs}) = \int_{-\infty}^{+\infty} \left(F_t(x) - H(x \geq x_{obs})\right)^2 dx \tag{3}$$

where $F_t(x)$ is the cumulative distribution function at time t, x the predicted variable, and x_{obs} is the corresponding observed value. The function H is the Heaviside function which equals O for for predicted values smaller than the observed value, 1 otherwise. The $CRPS$ shares the same unit as the predicted variable x.

As the $CRPS$ assesses the forecast for a single time step, the $MCRPS$ is defined as the average $CRPS$ over the entire period. The $MCRPS$ can reduce to the Mean Absolute Error (MAE) if a single member is considered and thus it allows to compare deterministic and probabilistic forecasts (Hersbach, 2000; Gneiting and Raftery, 2007). Finally, ~~a 0 value~~a value of 0 indicates a perfect forecast and there is no upper bound.

The reliability diagram (Stanski et al., 1989) is a graphical method to assess the reliability of a predictive ensemble by plotting forecasted against observed event frequencies. A perfectly reliable forecast is represented by a 45° line that indicates that forecasted and observed frequencies are equal. If the joint distribution curve differs from the perfect reliability lines, it indicates that the spread of the ensemble does not perfectly match its predictive skills. If the curve is situated above the perfect reliability line, this denotes an overdispersion of the ensemble, and an underdispersion in the opposite case.

The reliability is twofold. Since the reliability curve assesses the dispersion regarding the predictive skills of the ensemble, it is possible to have a perfectly reliable system with a low predictive capability in the case the dispersion is very high. For disambiguation, the ensemble spread is added to the plots.

Practically, one can define the deviation from perfect reliability by estimating a measure of distance between the forecast reliability curve and the perfect reliability line by computing the Mean Absolute Error (MAE) or Mean Square Error (MSE, Brochero et al., 2013). This dimensionless score allows to reduce the measure of reliability to a scalar. In the case where the MAE is used, it can be easily interpreted as the average distance between forecasted frequencies and the observed frequencies over all quantiles of interest. This verification score is henceforth referred as Mean absolute error of the Realiability Diagram, $MaeRD$.

Additional information about reliability can be obtained from the Spread Skill Plot (SSP, Fortin et al., 2014). It compares the Root Mean Square Error $RMSE$ and the square root of average ensemble variance that is a measure of the ensemble spread. The reliability is thus somehow decomposed into an accuracy error part and a spread component. Ideally, the spread should match the $RMSE$.

3 Results

Table 3 summarizes the specificities of the nine variants of the hydrometeorological forecast framework according to the three "forecasting tools": multimodel, EnKF, and ensemble meteorological forcing. Each of these switch may be activated or not and are marked as on/off in the table.

The multimodel switch dictates if the members issued by the 20 individual models are pooled together to create a single probabilistic forecast. In the case where the multimodel approach is not used, the models outputs are kept individually and 20 distinct ensembles – one per model – are considered.

The EnKF switch indicates if sequential data assimilation or the open loop procedure is applied. When EnKF updating is used, an ensemble of 50 members is created from 50 likely initial conditions sets identified by the filter. Otherwise, a single set of state variable values determined from the simulation is provided to the forecasting step. Note that the H and H' system differ by the EnKF perturbations magnitude, where H uses perturbations that aim at optimizing the combined criterion while H' uses lower perturbations that are deemed to be more realistic.

Lastly, the meteorological forcing employed during the forecast step can be either deterministic or probabilistic, using one randomly picked member or all 50 MEPS members.

These tools can be used alternatively or combined. For instance, if the EnKF and the meteorological ensemble forcing are used collectively, each of the 50 initial conditions sets will serve as starting point for each of the 50 meteorological forecast member creating a larger hydrometeorological ensemble that contains 2500 members.

We chose to disregard more complex or "hybrid" cases in this study, where for example, the final ensemble is composed with some models that benefit EnKF state updating while others are used in an open loop forecasting mode as these setups do not add additional information about the role of the tools, increase the degree of freedom for the system optimization and would shoot up computational costs.

Table 3. Description of the nine systems

Systems	A	B	C	D	E	F	G	H	H'
Multimodel	Off	Off	Off	Off	On	On	On	On	On
EnKF	Off	Off	On	On	Off	Off	On	On	On
Met. ensemble	Off	On	Off	On	Off	On	Off	On	On
Nb of members	(20x)1	(20x)50	(20x)50	(20x)2500	20	1000	1000	50000	50000

The results for each of the nine systems applied to every catchment, lead time and possibly every model are not systematically detailed and compared to each other. The following graphs are deemed sufficient to interpret the role and benefits that the system components play on the forecast quality. Additional graphs representing the resolution and reliability of each system are provided online for readers who are interested in a specific set up.

To picture an overview of the results, Figure 2 represents the accuracy in terms of $MCRPS$ (or MAE for system A that is fully deterministic) and $MaeRD$. For graphical convenience, the full distribution of performance according to various factors is not displayed but only a single representative value. To reduce the whole of the results to a single scalars, the median performance has been considered. In the case where a multimodel approach is used, the median performance over the 20 catchments is displayed on the figure. Otherwise, when individual models are considered, firstly the median performing model is identified and then the median performance over the catchment is represented. This implies that the performance of individual models systems (A, B, C, and D) may refer to a different model for each lead time.

The four radar plots situated on the top of the figure illustrate the $MCRPS$ performance. As a reference, the center of the disk consist of the the median $MCRPS$ value of the climatology over the 20 catchments while the perimeter represent a perfect $MCRPS$ equals to 0. The radius lines represent the nine systems described in Table 3 and are referred by their corresponding letter.

The nine systems present varying performance but all decrease logically with lead time. System A, which is deterministic, undoubtedly performs worse for every lead time. It is challenged from the 3rd day and is outperformed for medium range forecast by the hydrological climatology. System B presents a quite similar behaviour to A but with a lower decrease of accuracy with lead time. System C may be considered as competitive for shorter lead times but looses quickly its edge. These preliminary results tend to indicate that simpler HEPS may not be appropriate to accurately forecast streamflows over a nine day horizon. However, all versions including the simpler version except system A are more informative than the climatology for all lead times. Systems G, H and H' stand

out from the others for all lead times.

The second row in Figure 2 illustrates the reliability of each system. The center of the disk corresponds to a $MaeRD$ equals to 0.5. System A is artificially placed at the center of the radar plot to denote that no reliability information is communicated since it is deterministic.

The reliability results shares similarities with the accuracy assessment. Simpler systems face difficulties to provide a reliable forecast. Despite the use of meteorological ensemble forcing, system B is far from providing the right dispersion. Systems C and D provide some information for short lead times but experiences a substantial loss with increasing lead time. Once again, G, H and H' are performing best.

3.1 Multimodel approach and structural uncertainty

To assess the gain related to the multimodel approach, Figure 3 presents a comparison of the individual model MAE (A) and the $MCRPS$ that pools all model output together (E). At this step, only the structural uncertainty is taken into account as the meteorological forcing is kept deterministic and no initial condition uncertainty estimation is provided for both cases. These systems are computationally cheap as they contain either 20x1 member or 20 members.

In Figure 3, each boxplot represents the distribution of performance (minimum, quantiles 0.25, 0.5, and 0.75, and maximum) of the 20 models while the curve details the multimodel accuracy. On the x axis, the 20 test catchments are sorted according to increasing multimodel $MCRPS$ for the first lead time. This allows to notice that certain catchments exhibit a faster growing error.

The multimodel performs consistently better than the median performance of the model but also better than any model in the large majority of cases. Exceptions can be occasionally observed for catchment ~~3, 17, and 20~~3 and 17 where only one or two models outpeform the ensemble. However, the best performing models differ from a catchment to another while the multimodel presents the advantage of being more robust than any of the models. This is explained by the varied individual model behaviours. Each model may grasp different specificities of the hydrograph by focussing more specifically on different (conceptual) hydrological processes. Consequently, the ensemble members – the models – have disparate errors. Whenever the mismatch between forecast members and observation is poorly correlated, their errors tend to cancel out each other.

Figure 4 presents the reliability of the system E. Each curves refers to one of the 20 catchments. As mentioned, the structural uncertainty of the hydrological models is solely explicitly taken into

account by the combination of the models.

System E is generally slightly over confident for all lead times and this trend becomes more apparent as the lead time increases. This is expected as the meteorological forcing uncertainty increases with time while the deterministic forcing do not support that aspect. One can notice that the reliability also depends on the catchments. ~~For the first lead time, most of the catchments are close to reliability while there is a clear outlier for which accuracy skills do not match its corresponding spread. In fact, this low performing catchment exhibits a constant hydrological wet bias – partially explained by a meteorological forecast wet bias that over-forecasts precipitations by 15% – that is not captured by any of the models even if the global tendency is respected.~~ For the first lead time, most of the catchments are close to reliability while there are two outliers for which accuracy skills do not match their corresponding spread. In fact, these catchments exhibits a constant hydrological bias partially explained by an inaccurate meteorological forcing that is not captured by any of the models even if the global tendency is respected. Consequently, the models errors are highly correlated and this prevents the members to form a performing ensemble. This bias indicates that the aggregation of the other sources of uncertainty drive the system toward an inaccurate state.

3.2 Data assimilation and initial condition uncertainty

Figure 5 illustrates the increase of performance related to the data assimilation by comparing systems E and G. System G improves upon E as it benefits from the EnKF data assimilation to handle the initial condition uncertainty. The models states are updated according to the last available observations and an ensemble is created for each model based on the probabilistic estimation of best initial conditions.

The EnKF provides considerable gain over open loop forecasts for all catchments and reduces the number of lower performance ~~(outlier)~~ catchments. ~~Data assimilation is particularly effective on catchments that present a systematic bias. For example, catchment number 11 that was problematic from the first lead time lies among the other catchments in terms of performance.~~ This indicates that inaccuracies accumulated and stored during the spin up period in the state variable as the results of structural and forcing errors can be significantly reduced by providing adequate model reinitialization.

As the EnKF acts on model state variables right after the spin up period, it is not surprising to see its efficiency decreasing with lead time. This clarifies why the EnKF is beneficial for all lead times but that its skill decreases faster than the open loop scheme one. Moreover, the EnKF provides satisfactory initial condition distribution to minimize the error at the time the observation becomes

available but does not sample the posterior states to be optimally integrated through time.

Figure 6 details the reliability of system G. There is a considerable increase of spread in comparison to system E for shorter lead time that goes beyond adequate dispersion and lead to a slightly overdispersed forecast for the first lead time. This was expected as the EnKF was initially implemented to maximize individual model reliability for system G (see section 2.3.2). As the EnKF also takes into account the parameter and structural uncertainties and is combined with a multimodel approach, there may be a redundancy in the error deciphering. The structural error and the corresponding ensemble spread that it should describe may be somewhat accounted twice in that particular case. However, the overestimation of the ideal spread diminishes as the EnKF influence fades away quickly and the system goes back toward a better reliability for medium range forecast and underdispersion from days 4-5.

To explain the rapid decrease of reliability, Figure 7 displays the ensemble mean $RMSE$ and the square root of average ensemble variance. This individual spread skill plot (one model and one catchment) is typical. The spread and the $RMSE$ are close to a perfect match for the first day indicating an appropriate dispersion, yet, they diverge rapidly. The reliability deterioration of the system is twofold: the increase of the ensemble mean bias and the decrease of the spread. The loss of hydrological predictive skill is coherent regarding that the meteorological accuracy diminishes with increasing lead time. Concerning the second point, in most cases, the ensemble of initial conditions that EnKF provides often differ little from each other – few percent – indicating that the posterior distribution of each parameter is rather narrow (DeChant and Moradkhani, 2012; Abaza et al., 2015). These dissimilarities are not large enough to provoke a divergence in the behaviour of EnKF members during the forecasting step as the model are resilient. The different initial conditions thus tend to merge toward a certain value – often close the open loop one – which may not be accurate. This behavior is attributed to the EnKF rather than to the model structures as it has been also observed by others, for example with a three-hour time step and spatially distributed model in Abaza et al. (2014). Alternatives to the traditional EnKF (e.g. dual state-parameter, additional direct perturbations of state variables) may possibly contribute to slightly maintain the spread for longer lead times but they may not be consistent with the use of the multimodel, as it may imply to take into account the same source of uncertainty twice.

3.3 Contribution of the meteorological ensemble forcing

One step further in the system complexity is taken as the MEPS forcing is introduced. In this study, meteorological forcing was not processed as the investigation of such technique was deemed out of scope. It is expected that a successful pre-processing would enhance the MEPS forecast and that these improvements could be possibly cascaded through the hydrological components to the final

hydrological forecast. Counter-intuitively, recent attempts demonstrated that no or minor improvements were obtained in the hydrological forecast (Kang et al., 2010; Verkade et al., 2013; Zalachori et al., 2012; Roulin and Vannitsem, 2015).

Figure 8 compares the $MCRPS$ of systems G and H. They differ only in their meteorological forcing as the latter uses the 50 member probabilistic forecast. Difference between them is negligible until the 7th or 8th day where an improvement in performance can be noticed on some catchments. For these longer lead times, the probabilistic forcing is slightly more efficient for the $MCRPS$ but the main difference lies in the reliability (Figure 9). In fact, the reliability is substantially improved for the longest lead times when the meteorological uncertainty is provided to the system. A comparison of these systems with respect to the seasonality is provided in the supplementary material.

The ECMWF MEPS dispersion grows with lead time and logically contributes to the HEPS spread accordingly. This is confirmed by comparing the spread of the G and H systems as they decrease at a different pace. While they are almost identical with a value of ~~0.55 mm day^{-1} and 0.57 mm day^{-1}~~0.58 mm day^{-1} and 0.59 mm day^{-1} respectively for the day 3, G spread drops to ~~0.44 mm day^{-1}~~0.45 mm day^{-1} for day 9 while the use of the MEPS maintains the spread to ~~0.55 mm day^{-1}~~0.59 mm day^{-1}. This also indicates that the tool that contributes the most to the HEPS dispersion is the EnKF since the raw MEPS forcing is not able to fully balance the decrease of the spread induced by the EnKF. Further improvement in the reliability could perhaps be achieved through bias removal and suitable pre-processing technique.

The main sources of uncertainty – structure, initial conditions, and meteorological forcing – are cascaded through the different components of the forecasting system to provide better forecast than any of the systems previously described. Yet the system reliability is not perfect as the forecast for day 1 and day 9 are slightly overdispersive and underdispersive in addition to present sensitivity to the catchments. To realistically represent the uncertainty of the system, the spread should grow with lead time as the future is more uncertain. This suggest that further improvement of this setup and particular application could be obtained with a more dispersed meteorological forcing.

3.4 Simplification of the framework

A potential drawback for operational use of such system is that it is computationally expensive as 50000 members are exploited to build it. The efficiency of a simpler system is assessed on Figure 10. Eight typical catchments are displayed in the sub plots to illustrate the conclusion. The box plots represent the $MCRPS$ distribution of the 20 models results from system D that benefits EnKF state updating and MEPS forcing. Each of these models can be considered as a sub-ensemble of the

large ensemble H driven by a single model instead of using a multimodel approach. This is a more consistent approach with the EnKF individual optimization that is carried out to aim for reliability for each model one at a time. The numbers at the top of the sub-plots refer to the model number that are better than the multimodel for each lead time.

In Figure 10, sub-ensembles are more skilful than the hydrological climatology for all lead times but rarely outperform the multimodel forecast. More precisely, the median performing sub-ensemble is always poorer than the multimodel and only the best models among the 20 occasionnally exhibit lower $MCRPS$. Individual models that outperform the multimodel frequently differ from a catchment to another and from a lead time to another. This emphasizes the difficulty to chose a priori a single model as half of the 20 models never behave better than the multimodel and only model ~~1 and 5~~1, 5, and 17 perform better than the multimodel for several catchments. Choosing a sub-ensemble doubtlessly enhances the system computational requirements and eases operational implementation but relying on a single model may be misleading or, at least, minimize the expectation that one can have from the HEPS.

Figure 11 assesses the reliability of the same system with the $MaeRD$ score. Like for the previous plots, the box plots contain the 20 ensembles that correspond to the 20 models and are sorted by catchment with increasing multimodel $MaeRD$. Note that the $MaeRD$ does not provides precise information about dispersion but only about the distance from perfect reliability. Nevertheless, individual model ensemble may be either slightly over or underdispersive for the first lead time but are systematically underdispersive for longer lead times. On the other hand, system H can be either over or underdispersive depending on the catchment. Overdispersive forecasts, like for the catchment ~~19~~20, can be recognized as they tend to become more reliable for longer lead time.

For the first lead time, the best individual model ensembles may be competitive with the multimodel but are already less efficient from day 3 and are drastically underdispersive for day 9. Even if the EnKF takes into account the structural uncertainty at $t = 0$, it loses its efficiency during the forecast. The information that the updated state sets contain about the structural uncertainty vanishes when the sets converge toward a common value. The multimodel approach, by its nature, allows to take over the role of the EnKF by dynamically preserving the required diversity.

3.5 Required EnKF perturbations

If the different sources of uncertainty along the hydrometeorological modeling chain are not explicitly accounted for by dedicated tools, the EnKF has to compensate for them. One way to achieve reliability is to increase the level of perturbations to the input. However, there is no obvious way to

know by which amount the uncertainty on input should be overestimated to compensate for the other uncertainties. Thus, to ensure hydrological reliability, one needs to perform a fastidious calibration of the EnKF hyper-parameters to identify the required noise magnitude (Thiboult and Anctil, 2015a).

H' is identical to system H except that it relies on a different optimization of the EnKF. Instead of maximizing the combined criterion for individual models (see section 2.3.2), the EnKF noise specification is set lower to values that are more consistent with real uncertainties estimations of observed climatological and streamflow observations at catchment scale. Namely, precipitation is perturbed with a gamma law with a standard deviation of 25% of the mean value, temperatures with a normal law with a 2° standard deviation and streamflow observations with normal law with a 10% standard deviation.

These noise magnitudes are therefore meant to describe the real uncertainties in forcing and observations in the EnKF but do not implicitly account for model error any longer. Also, in a perfect-model environment, i.e. without any model error, it has been shown that the EnKF spread is representative of the ensemble mean error with respect to a truth integration (Houtekamer et al., 2009). In other words, the implementation of the EnKF with realistic input and output perturbations corresponds to a potential 'perfect' EnKF implementation if the total uncertainty could be summarized to the input and output error and were perfectly identified, i.e. in a perfectly controlled environment with a negligible model structural error. ~~This would corresponds to a potential optimal EnKF implementation if the total uncertainty could be summarized to the input and output error and were perfectly identified, i.e. in a perfectly controlled environment with a negligible model structural error~~. Consequently, with the system H', the structural error is theoretically only deciphered through the multimodel pooling. Yet this needs to be qualified as it is practically hard to untangle the source of uncertainty within the actual configuration of the EnKF but it reduces the risk that the tools effects overlap. By choosing these perturbations, the user also gets rid of a fastidious EnKF tuning by screening adequate perturbation (e.g. Moradkhani et al., 2005; Thiboult and Anctil, 2015a) and hence simplifies the system implementation.

In Figure 12, system H' improves reliability for first lead times by reducing the overdispersion with a sensible decrease in the ensemble spread from ~~0.65 mm day^{-1} to 0.54 mm day^{-1}~~0.72 mm day^{-1} to 0.57 mm day^{-1} for day 1 without any degradation of the $MCRPS$ (except for 2 catchments; all results are shown on additional figures online). System H' maintains a more constant spread and reliability with increasing lead time as the main sources of uncertainty are more accurately deciphered specifically by their corresponding tool, leading to an overall better forecast.

~~The two outlier catchments that exhibit poorer reliability present an underdispersed forecast that is a bit more pronounced for the H' system than the H system (see Figure 9). This indicates that uncertainties used to define the EnKF perturbations are under-estimated. As a matter of fact~~Finally, it is unreasonable to assume that uncertainties are invariant from one catchment to another. The comparison of the MEPS forecast and meteorological observations shown that the quality over the 20 catchments remains close and indicates that the misfit probably originates from the structures composing the multimodel ensemble that can be maladapted to simulate this particular catchments or from doubtful streamflow measurements. This lead us think that further improvements in very uncertain environments are limited by a preliminary accurate quantification of error.

Also, considerable efforts have been paid to link performance with estimated time of concentration, size of catchments, and river slope without any clear results. The authors were not able to relate any catchment feature to particular results.

4 Conclusions

This work investigates the contribution of three different probabilistic tools commonly used in hydrometeorological sciences. They are used conjointly and alternatively to identify their effect on the hydrological predictive ensemble and to untangle sources of uncertainty that are aggregated in the outputs.

Each of these tools is dedicated to capture a certain aspect of the total uncertainty. A multimodel approach is used to quantify and reduce explicitly the hydrological model error, the Ensemble Kalman Filter to decipher the uncertainty related to initial conditions and the meteorological ensemble to account for the forcing uncertainty.

The experiment shows that important gain may be achieve in term of accuracy and reliability by adequately using these techniques. Their action differ substantially by their mean and range of action.

The EnKF provides accurate quantification of initial error but fails to maintain reliability as its effect fades out quickly after model spin up. The information about the structural uncertainty deciphered by the EnKF, which is contained in the state variable posterior distribution, is not propagated with time integration during the forecast step. However, the EnKF remains a key component of the system as it is the one that provides the most dispersion for the first lead times. This also indicates that the accumulation of past errors in the initial conditions is a dominant source of uncertainty.

The multimodel approach is able to partially compensate for the EnKF decreasing action by taking over the structural uncertainty. Moreover, the combination of independent models improve accuracy as their errors may cancel each other. Lastly, the use of ensemble meteorological forecast contributes to the reliability of medium range forecast by representing the meteorological forcing errors.

Their action are complementary as they decipher different nature of uncertainty at different locations by acting at particular stages in the forecasting process. When combined, they need to be set according to the tools they are juxtaposed with to prevent overlapping actions. This is particularly the case for the EnKF that has important degree of freedom in its implementation. It can eventually be tuned with more realistic input perturbations by coupling with the multimodel ensemble and therefore, facilitate its implementation by relaxing the constraints of optimal perturbation screening.

Possible avenues for further improvements may be achieved through a multimodel state updating rather than individual model updating, i.e. by treating initial condition in a single step as a whole. Lastly, the meteorological forecast shown to be a little underdispersed for this application and could be possibly improved by applying suitable pre-processing techniques.

Acknowledgements. The authors acknowledge the Centre d'Expertise Hydrique du Québec for providing hydrometeorological data. They also acknowledge financial support from the Chaire de recherche EDS en prévisions et actions hydrologiques and from the Natural Sciences and Engineering Research Council of Canada. Finally, we would like to thank Florian Pappenberger for advices and two anonymous reviewers from fruitful suggestions.

References

Abaza, M., Anctil, F., Fortin, V., and Turcotte, R.: A comparison of the Canadian global and regional meteorological ensemble prediction systems for short-term hydrological forecasting (vol 141, pg 3462, 2013), Monthly Weather Review, 142, 2561–2562, doi:10.1175/mwr-d-14-00018.1, 2014.

Abaza, M., Anctil, F., Fortin, V., and Turcotte, R.: Exploration of sequential streamflow assimilation in snow dominated watersheds, Advances in Water Resources, 80, 79–89, doi:http://dx.doi.org/10.1016/j.advwatres.2015.03.011, 2015.

Ajami, N. K., Duan, Q., Gao, X., and Sorooshian, S.: Multimodel combination techniques for analysis of hydrological simulations: Application to Distributed Model Intercomparison Project results, Journal of Hydrometeorology, 7, 755–768, doi:10.1175/jhm519.1, 2006.

Ajami, N. K., Duan, Q. Y., and Sorooshian, S.: An integrated hydrologic Bayesian multimodel combination framework: Confronting input, parameter, and model structural uncertainty in hydrologic prediction, Water Resources Research, 43, doi:W01403 10.1029/2005wr004745, 2007.

Bartholmes, J. C., Thielen, J., Ramos, M. H., and Gentilini, S.: The european flood alert system EFAS - Part 2: Statistical skill assessment of probabilistic and deterministic operational forecasts, Hydrology and Earth System Sciences, 13, 141–153, 2009.

Bergström, S. and Forsman, A.: Development of a conceptual deterministic rainfall-runoff model, Nordic Hydrology, 4, 147–170, 1973.

Beven, K. and Binley, A.: The future of distributed models - Model calibration and uncertainty prediction, Hydrological Processes, 6, 279–298, doi:10.1002/hyp.3360060305, 1992.

Beven, K.: On doing better hydrological science, Hydrological Processes, 22, 3549–3553, doi:10.1002/hyp.7108, 2008.

Beven, K. and Binley, A.: GLUE: 20 years on, Hydrological Processes, 28, 5897–5918, doi:10.1002/hyp.10082, 2014.

Beven, K. J., Kirkby, M. J., Schofield, N., and Tagg, A. F.: Testing a physically-based flood forecasting model (TOPMODEL) for 3 UK catchments, Journal of Hydrology, 69, 119–143, doi:10.1016/0022-1694(84)90159-8, 1984.

Boucher, M. A., Tremblay, D., Delorme, L., Perreault, L., and Anctil, F.: Hydro-economic assessment of hydrological forecasting systems, Journal of Hydrology, 416, 133–144, doi:10.1016/j.jhydrol.2011.11.042, 2012.

Bourdin, D. R., Fleming, S. W., and Stull, R. B.: Streamflow Modelling: A Primer on Applications, Approaches and Challenges, Atmosphere-Ocean, 50, 507–536, doi:10.1080/07055900.2012.734276, 2012.

Breuer, L., Huisman, J. A., Willems, P., Bormann, H., Bronstert, A., Croke, B. F. W., Frede, H. G., Graff, T., Hubrechts, L., Jakeman, A. J., Kite, G., Lanini, J., Leavesley, G., Lettenmaier, D. P., Lindstrom, G., Seibert, J., Sivapalan, M., and Viney, N. R.: Assessing the impact of land use change on hydrology by ensemble modeling (LUCHEM). I: Model intercomparison with current land use, Advances in Water Resources, 32, 129–146, doi:10.1016/j.advwatres.2008.10.003, 2009.

Brochero, D., Anctil, F., and Gagne, C.: Simplifying a hydrological ensemble prediction system with a backward greedy selection of members - Part 2: Generalization in time and space, Hydrology and Earth System Sciences, 15, 3327–3341, doi:10.5194/hess-15-3327-2011, 2011.

Brochero, D., Gagne, C., and Anctil, F.: Evolutionary Multiobjective Optimization for Selecting Members of an Ensemble Streamflow Forecasting Model, Gecco'13: Proceedings of the 2013 Genetic and Evolutionary Computation Conference, pp. 1221–1228, 2013.

Buizza, R., Miller, M., and Palmer, N.: Stochastic representation of model uncertainties in the ECMWF Ensemble Prediction System, Quarterly Journal of the Royal Meteorology Society , 125, 2887–2908, 1999.

Burnash, R. J. C., Ferral, R. L., and McGuire, R. A.: A Generalized Streamflow Simulation System – Conceptual Modelling for Digital Computers, Technical Report, Joint Federal and State River Forecast Center, US National Weather Service and California Department of Water Resources, Sacramento, p. 204, 1973.

Charron, M., Pellerin, G., Spacek, L., Houtekamer, P. L., Gagnon, N., Mitchell, H. L., and Michelin, L.: Toward Random Sampling of Model Error in the Canadian Ensemble Prediction System, Monthly Weather Review, 138, 1877–1901, doi:10.1175/2009mwr3187.1, 2010.

Chen, H., Yang, D. W., Hong, Y., Gourley, J. J., and Zhang, Y.: Hydrological data assimilation with the Ensemble Square-Root-Filter: Use of streamflow observations to update model states for real-time flash flood forecasting, Advances in Water Resources, 59, 209–220, doi:10.1016/j.advwatres.2013.06.010, 2013.

Chiew, F. H. S., Peel, M. C., and Western, A. W.: Application and testing of the simple rainfall-runoff model SIMHYD, in Mathematical Models of Small Watershed Hydrology and Applications, Water Resources Publication, Littleton, Colorado, 2002.

Clark, M. P., Rupp, D. E., Woods, R. A., Zheng, X., Ibbitt, R. P., Slater, A. G., Schmidt, J., and Uddstrom, M. J.: Hydrological data assimilation with the ensemble Kalman filter: Use of streamflow observations to update states in a distributed hydrological model, Advances in Water Resources, 31, 1309–1324, doi:10.1016/j.advwatres.2008.06.005, 2008.

Cloke, H. L. and Pappenberger, F.: Ensemble flood forecasting: A review, Journal of Hydrology, 375, 613–626, doi:10.1016/j.jhydrol.2009.06.005, 2009.

Cormary, Y. and Guilbot, A.: Étude des relations pluie-débit sur trois bassins versants d'investigation, in: IAHS Madrid Symposium, edited by IHAS publications, 265–279, 1973

DeChant, C. M. and Moradkhani, H.: Improving the characterization of initial condition for ensemble streamflow prediction using data assimilation, Hydrology and Earth System Sciences, 15, 3399–3410, doi:10.5194/hess-15-3399-2011, 2011.

DeChant, C. M. and Moradkhani, H.: Examining the effectiveness and robustness of sequential data assimilation methods for quantification of uncertainty in hydrologic forecasting, Water Resources Research, 48, doi:W04518 10.1029/2011wr011011, 2012.

Demargne, J., Wu, L. M., Regonda, S. K., Brown, J. D., Lee, H., He, M. X., Seo, D. J., Hartman, R., Herr, H. D., Fresch, M., Schaake, J., and Zhu, Y. J.: The Science of NOAA's Operational Hydrologic Ensemble Forecast Service, Bulletin of the American Meteorological Society, 95, 79–98, doi:10.1175/bams-d-12-00081.1, 2014.

Evensen, G.: The Ensemble Kalman Filter: theoretical formulation and practical implementation, Ocean Dynamics, 53, 343–367, doi:10.1007/s10236-003-0036-9, 2003.

Fortin, V. and Turcotte, R.: Le modèle hydrologique MOHYSE, Note de cours pour SCA7420, Report, Département des sciences de la terre et de l'atmosphère, Université du Québec à Montreal, 2007.

Fortin, V., Abaza, M., Anctil, F., and Turcotte, R.: Why should ensemble spread match the RMSE of the ensemble mean ?, Journal of Hydrometeorology, 15, 1708 – 1713, doi:10.1175/JHM-D-14-0008.1, 2014.

Fraley, C., Raftery, A. E., and Gneiting, T.: Calibrating Multimodel Forecast Ensembles with Exchangeable and Missing Members Using Bayesian Model Averaging, Monthly Weather Review, 138, 190–202, doi:10.1175/2009mwr3046.1, , 2010.

Gaborit É., Anctil F., Fortin V., Pelletier G.: On the reliability of spatially disaggregated global ensemble rainfall forecasts. Hydrological Processes 27, 45–56, doi:10.1002/hyp.9509, 2013

Garçon, R.: Modèle global Pluie-Débit pour la prévision et la prédétermination des crues, La Houille Blanche, 7/8, 88–95, doi:http://dx.doi.org/10.1051/lhb/1999088, 1999.

Girard, G., Morin, G., and Charbonneau, R.: Modèle précipitations-débits à discrétisation spatiale, Cahiers ORSTOM, Série Hydrologie, 9, 35–52, 1972.

Gneiting, T. and Raftery, A. E.: Strictly proper scoring rules, prediction, and estimation, Journal of the American Statistical Association, 102, 359–378, doi:10.1198/016214506000001437, 2007.

Gourley, J. J. and Vieux, B. E.: A method for identifying sources of model uncertainty in rainfall-runoff simulations, Journal of Hydrology, 327, 68–80, doi:10.1016/j.jhydrol.2005.11.036, 2006.

Hersbach, H.: Decomposition of the continuous ranked probability score for ensemble prediction systems, Weather and Forecasting, 15, 559–570, doi:10.1175/1520-0434(2000)015<0559:dotcrp>2.0.co;2, 2000.

Houtekamer, P. L., Mitchell, H. L., and Deng, X. X.: Model Error Representation in an Operational Ensemble Kalman Filter, Monthly Weather Review, 137, 2126–2143, doi:10.1175/2008mwr2737.1, 2009.

Jakeman, A. J., Littlewood, I. G., and Whitehead, P. G.: Computation of the instantaneous unit hydrograph and identifiable component flows with application to two small upland catchments, Journal of Hydrology, 117, 275–300, doi:10.1016/0022-1694(90)90097-h, 1990.

Kang, T. H., Kim, Y. O., and Hong, I. P.: Comparison of pre- and post-processors for ensemble streamflow prediction, Atmospheric Science Letters, 11, 153–159, doi:10.1002/asl.276, 2010.

Lee, H., Seo, D. J., and Koren, V.: Assimilation of streamflow and in situ soil moisture data into operational distributed hydrologic models: Effects of uncertainties in the data and initial model soil moisture states, Advances in Water Resources, 34, 1597–1615, doi:10.1016/j.advwatres.2011.08.012, 2011.

Li, Y., Ryu, D., Wang, Q., Pagano, T., Western, A., Hapuarachchi, P., and Toscas, P.: Assimilation of streamflow discharge into a continuous flood forecasting model, in: The International Union of Geodesy and Geophysics XXV General Assembly, edited by: Bloschl, G., Takeuchi, K., Jain, S., Farnleitner, A., and Schumann, A., vol. 347, Internation Association of Hydrological Sciences Press, Melbourne, 107–113, 2011.

Liu, Y. Q. and Gupta, H. V.: Uncertainty in hydrologic modeling: Toward an integrated data assimilation framework, Water Resources Research, 43, W07401, doi:10.1029/2006wr005756, 2007.

Mandel, J.: Efficient Implementation of the Ensemble Kalman Filter, Report, University of Colorado at Denver and Health Sciences Center, Denber, 2006.

Matheson, J. E. and Winkler, R. L.: Scoring rules for continuous probability distributions, Management Science, 22, 1087–1096, doi:10.1287/mnsc.22.10.1087, 1976.

Mazenc, B., Sanchez, M., and Thiery, D.: Analyse de l'influence de la physiographie d'un bassin versant sur les paramètres d'un modèle hydrologique global et sur les débits caractéristiques à l'exutoire, Journal of Hydrology, 69, 97–188, 1984.

McMillan, H. K., Hreinsson, E. O., Clark, M. P., Singh, S. K., Zammit, C., and Uddstrom, M. J.: Operational hydrological data assimilation with the recursive ensemble Kalman filter, Hydrology and Earth System Sciences, 17, 21–38, doi:10.5194/hess-17-21-2013, 2013.

Molteni, F., Buizza, R., Palmer, T. N., and Petroliagis, T.: The ECMWF ensemble prediction system: Methodology and validation, Quarterly Journal of the Royal Meteorological Society, 122, 73–119, doi:10.1002/qj.49712252905, 1996.

Moore, R. J. and Clarke, R. T.: A distribution function approach to rainfall runoff modeling, Water Resources Research, 17, 1367–1382, doi:10.1029/WR017i005p01367, 1981.

Moradkhani, H., Sorooshian, S., Gupta, H. V., and Houser, P. R.: Dual state-parameter estimation of hydrological models using ensemble Kalman filter, Advances in Water Resources, 28, 135–147, doi:10.1016/j.advwatres.2004.09.002, 2005.

Nielsen, S. A. and Hansen, E.: Numerical simulation of the rainfall-runoff process on a daily basis, Nordic Hydrology, 4, 171–190, 1973.

Noh, S. J., Tachikawa, Y., Shiiba, M., and Kim, S.: Sequential data assimilation for streamflow forecasting using a distributed hydrologic model: particle filtering and ensemble Kalman filtering, in: Floods: from Risk to Opportunity, 357, IAHS Publications, Tokyo, 341–349, 2013.

Noh, S. J., Rakovec, O., Weerts, A. H., and Tachikawa, Y.: On noise specification in data assimilation schemes for improved flood forecasting using distributed hydrological models, Journal of Hydrology, 519, 2707–2721, doi:10.1016/j.jhydrol.2014.07.049, 2014.

O'Connell, P. E., Nash, J. E., and Farrell, J. P.: River flow forecasting through conceptual models, Part II - The Brosna catchment at Ferbane, Journal of Hydrology, 10, 317–329, 1970.

Oudin, L., Hervieu, F., Michel, C., Perrin, C., Andreassian, V., Anctil, F., and Loumagne, C.: Which potential evapotranspiration input for a lumped rainfall-runoff model? Part 2 - Towards a simple and efficient potential evapotranspiration model for rainfall-runoff modelling, Journal of Hydrology, 303, 290–306, doi:10.1016/j.jhydrol.2004.08.026, 2005.

Pagano, T. C., Wood, A. W., Ramos, M. H., Cloke, H. L., Pappenberger, F., Clark, M. P., Cranston, M., Kavetski, D., Mathevet, T., Sorooshian, S., and Verkade, J. S.: Challenges of Operational River Forecasting, Journal of Hydrometeorology, 15, 1692–1707, doi:10.1175/jhm-d-13-0188.1, 2014.

Perrin, C.: Vers une amélioration d'un modèle global pluie-débit, Thesis, Institut National Polytechnique de Grenoble, Grenoble, 2000.

Perrin, C., Michel, C., and Andreassian, V.: Improvement of a parsimonious model for streamflow simulation, Journal of Hydrology, 279, 275–289, doi:10.1016/s0022-1694(03)00225-7, 2003.

Poulin, A., Brissette, F., Leconte, R., Arsenault, R., and Malo, J. S.: Uncertainty of hydrological modelling in climate change impact studies in a Canadian, snow-dominated river basin, Journal of Hydrology, 409, 626–636, doi:10.1016/j.jhydrol.2011.08.057, 2011.

Rakovec, O., Weerts, A. H., Hazenberg, P., Torfs, P., and Uijlenhoet, R.: State updating of a distributed hydrological model with Ensemble Kalman Filtering: effects of updating frequency and observation network density on forecast accuracy, Hydrology and Earth System Sciences, 16, 3435–3449, doi:10.5194/hess-16-3435-2012, 2012.

Roulin, E. and Vannitsem, S.: Post-processing of medium-range probabilistic hydrological forecasting: impact of forcing, initial conditions and model errors, Hydrological Processes, 29, 1434–1449, doi:10.1002/hyp.10259, 2015.

Salamon, P. and Feyen, L.: Disentangling uncertainties in distributed hydrological modeling using multiplicative error models and sequential data assimilation, Water Resources Research, 46, W12501, doi:10.1029/2009wr009022, 2010.

Seiller, G., Anctil, F., and Perrin, C.: Multimodel evaluation of twenty lumped hydrological models under contrasted climate conditions, Hydrology and Earth System Sciences, 16, 1171–1189, doi:10.5194/hess-16-1171-2012, 2012.

Stanski, H. R., Wilson, L. J., and Burrows, W. R.: Survey of common verification methods in meteorology, WMO World Weather Watch Tech Report 8, WMO, Downsview, Ontario, 1989.

Sugawara, M.: Automatic calibration of the tank model, Hydrological Sciences, 24, 375–388, 1979.

Thiboult, A. and Anctil, F.: Assessment of a multimodel ensemble against an operational hydrological forecasting system, Canadian Water Resources Journal, 40, 272–284, 2015.

Thiboult, A. and Anctil, F.: On the difficulty to optimally implement the Ensemble Kalman filter: An experiment based on many hydrological models and catchments, Journal of Hydrology, 529, 1147–1160, 2015a.

Thiery, D.: Utilisation d'un modèle global pour identifier sur un niveau piézométrique des influences multiples dues à diverses activités humaines, Improvement of methods of long term prediction of variations in groundwater resources and regimes due to human activity, IAHS Publications, 136, 71–77, Exeter, 1982.

Thirel, G., Salamon, P., Burek, P., and Kalas, M.: Assimilation of MODIS Snow Cover Area Data in a Distributed Hydrological Model Using the Particle Filter, Remote Sensing, 5, 5825–5850, doi:10.3390/rs5115825, 2013.

Thornthwaite, C. W. and Mather, J. R.: The water balance, Report, Drexel Institute of Climatology, Centerton, New Jersey, 1955.

Valery, A., Andreassian, V., and Perrin, C.: 'As simple as possible but not simpler': What is useful in a temperature-based snow-accounting routine? Part 2 - Sensitivity analysis of the Cemaneige snow accounting routine on 380 catchments, Journal of Hydrology, 517, 1176–1187, doi:10.1016/j.jhydrol.2014.04.058, 2014.

Velázquez, 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.

Verkade, J.S., Brown, J.D., Reggiani, P. and Weerts, A. H.: Post-processing ECMWF precipitation and temperature ensemble reforecasts for operational hydrologic forecasting at various spatial scales, Journal of Hydrology, 501, 73–91, 2013.

Viney, N. R., Bormann, H., Breuer, L., Bronstert, A., Croke, B. F. W., Frede, H., Graeff, T., Hubrechts, L., Huisman, J. A., Jakeman, A. J., Kite, G. W., Lanini, J., Leavesley, G., Lettenmaier, D. P., Lindstroem, G., Seibert, J., Sivapalan, M., and Willems, P.: Assessing the impact of land use change on hydrology by ensemble modelling (LUCHEM) II: Ensemble combinations and predictions, Advances in Water Resources, 32, 147–158, doi:10.1016/j.advwatres.2008.05.006, 2009.

Vrugt, J. A. and Robinson, B. A.: Treatment of uncertainty using ensemble methods: Comparison of sequential data assimilation and Bayesian model averaging, Water Resources Research, 43, doi:W01411 10.1029/2005wr004838, 2007.

Vrugt, J. A., Gupta, H. V., Bouten, W., and Sorooshian, S.: A Shuffled Complex Evolution Metropolis algorithm for optimization and uncertainty assessment of hydrologic model parameters, Water Resources Research, 39, doi:1201 10.1029/2002wr001642, 2003.

Wagener, T., Boyle, D. P., Lees, M. J., Wheater, H. S., Gupta, H. V., and Sorooshian, S.: A framework for development and application of hydrological models, Hydrology and Earth System Sciences, 5, 13–26, 2001.

Walker, W. E., Harremoës, P., Rotmans, J., van der Sluijs, J. P., van Asselt, M. B. A., Janssen, P., and Krayer von Krauss, M. P.: Defining uncertainty : A conceptual basis for uncertainty management in model-based decision support, Integrated Assessment, 4, 5–17, doi:10.1076/iaij.4.1.5.16466, 2003.

Warmerdam, P. M., Kole, J., and Chormanski, J.: Modelling rainfall–runoff processes in the Hupselse Beek research basin, in: IHP-V, Technical Documents in Hydrology, IHP/UNESCO – ERB – CEREG, Strasbourg, 155–160, 1997.

Weerts, A. H. and El Serafy, G. Y. H.: Particle filtering and ensemble Kalman filtering for state updating with hydrological conceptual rainfall-runoff models, Water Resources Research, 42, 17, W09403, doi:10.1029/2005wr004093, 2006.

Wetterhall, F., Pappenberger, F., Alfieri, L., Cloke, H. L., Thielen-del Pozo, J., Balabanova, S., Danhelka, J., Vogelbacher, A., Salamon, P., Carrasco, I., Cabrera-Tordera, A. J., Corzo-Toscano, M., Garcia-Padilla, M., Garcia-Sanchez, R. J., Ardilouze, C., Jurela, S., Terek, B., Csik, A., Casey, J., Stankunavicius, G., Ceres, V., Sprokkereef, E., Stam, J., Anghel, E., Vladikovic, D., Eklund, C. A., Hjerdt, N., Djerv, H., Holmberg, F., Nilsson, J., Nystrom, K., Susnik, M., Hazlinger, M., and Holubecka, M.: HESS Opinions "Forecaster priorities for improving probabilistic flood forecasts", Hydrology and Earth System Sciences, 17, 4389–4399, doi:10.5194/hess-17-4389-2013, 2013.

Zalachori I., Ramos M.-H., Garçon R., Mathevet T., and Gailhard J. : Statistical processing of forecasts for hydrological ensemble prediction: a comparative study of different bias correction strategies , Advances in Science and Research, 8, 135–141, doi:10.5194/asr-8-135-2012, 2012.

Zhao, R. J., Zuang, Y. L., Fang, L. R., Liu, X. R., and Zhang, Q.: The Xinanjiang model, Hydrological Forecasting, IAHS Publications, 129, 351–356, 1980.

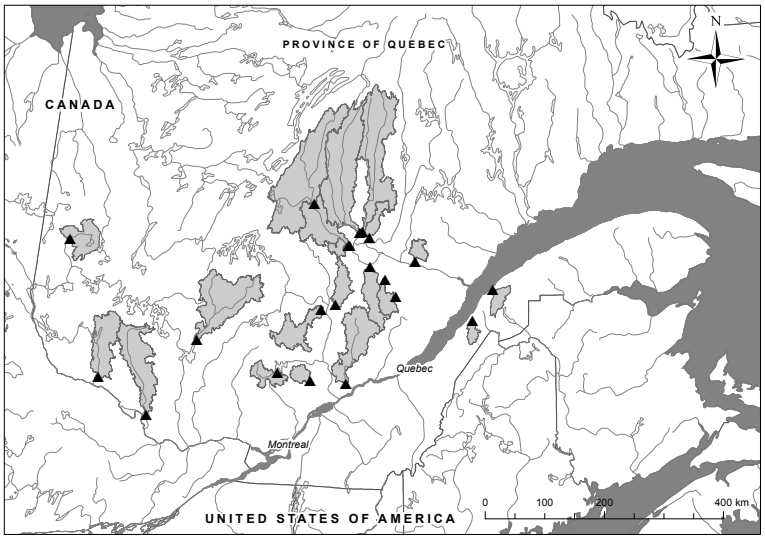

Figure 1. Spatial distribution of the catchments

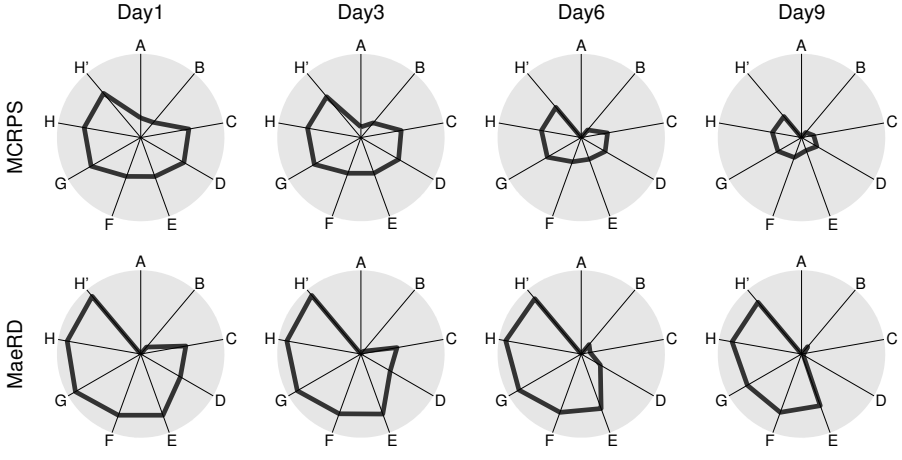

Figure 2. Synthetic results of the 9 systems that are referred by their code letter (see Table 3). The 4 top radar plots illustrate the $MCRPS$ with the center indicating the climatology reference performance, and the perimeter representing a perfectly accurate simulation. The 4 bottom plots describe the measure of distance from perfect reliability, with the center indicating a $MaeRD$=0.5 while the perimeter corresponds to a perfect reliability.

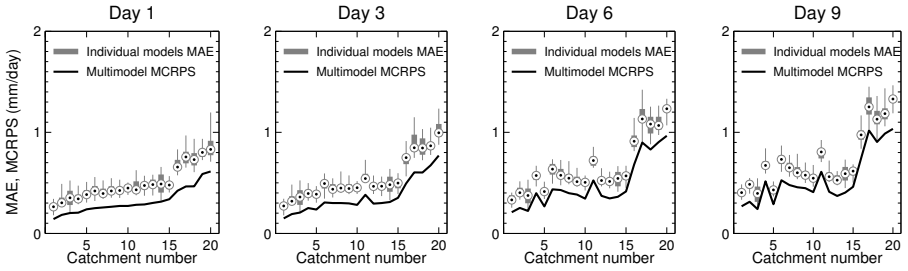

Figure 3. Comparison of individual models MAE and multimodel $MCRPS$ sorted by increasing multimodel $MCRPS$ for the first day (version A vs E)

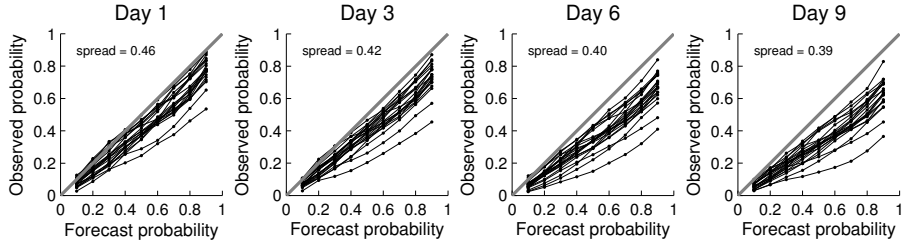

Figure 4. Reliability of the multimodel ensemble (system E) for all individual catchments. The spread represents the square root of mean ensemble variance averaged over all catchments.

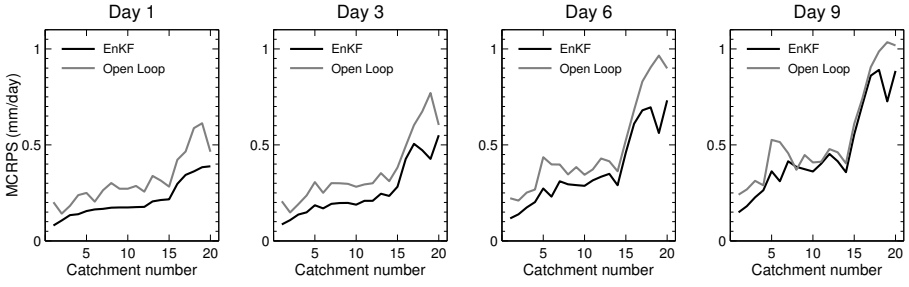

Figure 5. Comparison of open loop and EnKF multimodel $MCRPS$ sorted by increasing EnKF $MCRPS$ (system E vs G)

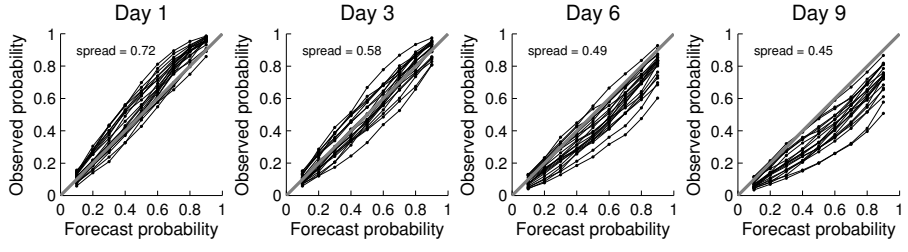

Figure 6. Reliability of the EnKF multimodel ensemble (system G) for all individual catchments. The spread represents the square root of mean ensemble variance averaged over all catchments.

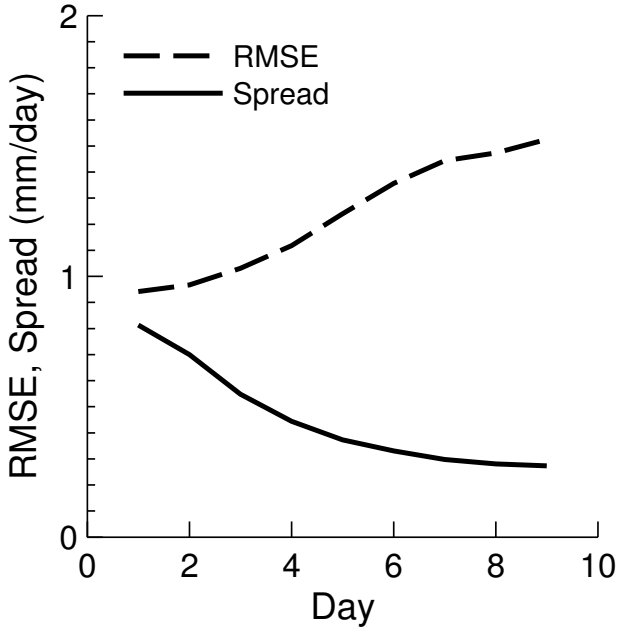

Figure 7. Typical Spread Skill plot of a single model EnKF ensemble

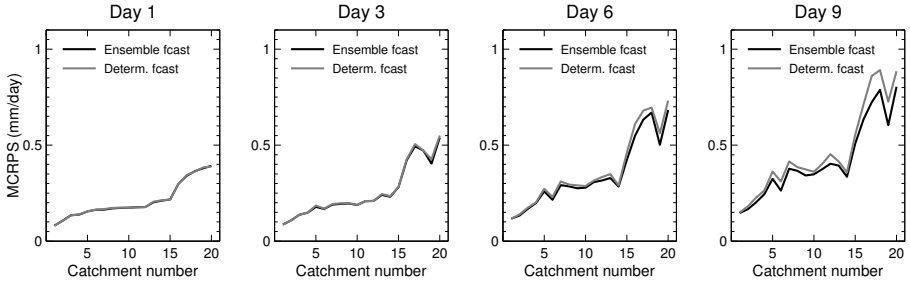

Figure 8. Comparison of EnKF multimodel $MCRPS$ with deterministic and ensemble meteorological forcing (system G vs H)

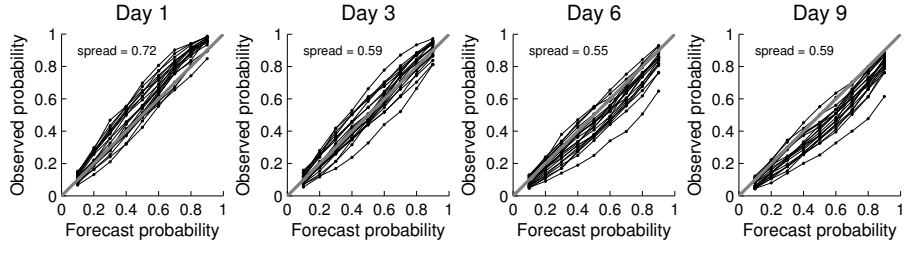

Figure 9. Reliability of the EnKF multimodel ensemble with MEPS forcing (system H)

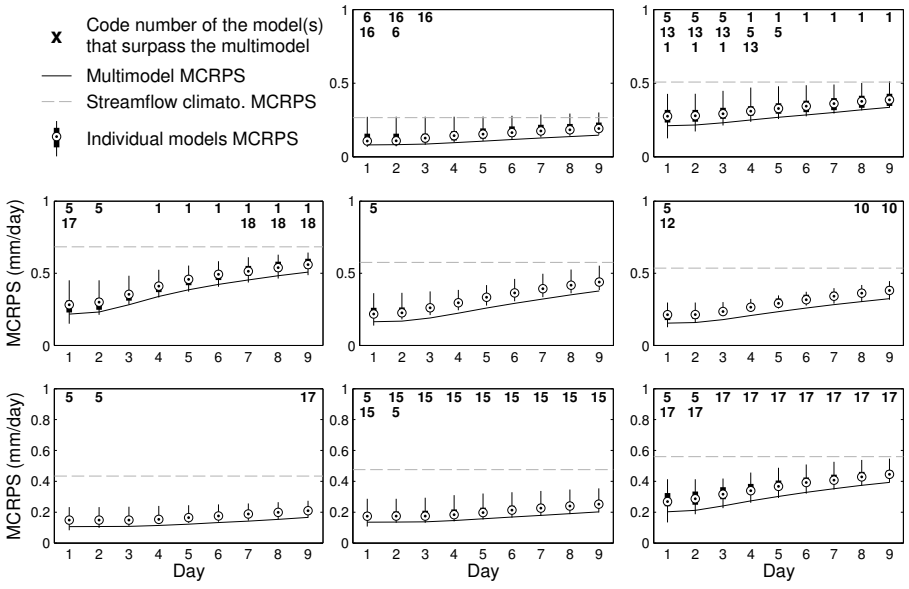

Figure 10. Comparative examples of the $MCRPS$ on 8 catchments of the EnKF individual models and the EnKF multimodel, both using MEPS forcing (system D vs H)

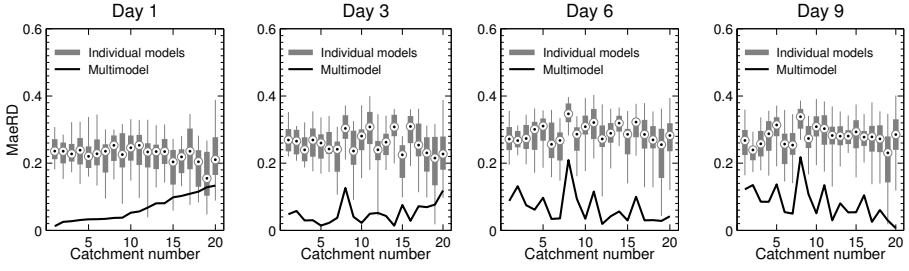

Figure 11. Comparison of the deviation from perfect reliability of EnKF individual models and the EnKF multimodel, both using MEPS forcing sorted by increasing EnKF multimodel $MaeRD$ for the first day (system D vs H)

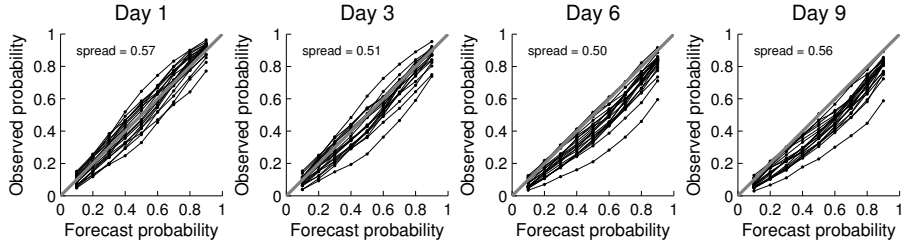

Figure 12. Reliability of the EnKF multimodel ensemble with MEPS forcing and lower input-output perturbations (system H')