First of all, we would like to thank the editor for considering our paper worth of interest. Second, the constructive comments of the three reviewers helped us to improve the original manuscript. We tried to address their questions and comments as accurately as possible to improve the manuscript. We hope our answers meet their expectations. Please note that all page and line numbers refer to the track-change version of the manuscript.

1 Response to Reviewer 1

General comments

In this research article, the authors present a detailed study on the predictive skill of hydrological ensemble forecasts in 10 watersheds in Québec, Canada. Different methods are employed which are subject to different degrees of complexity. Among these methods, a simple application of historical streamflow data is seen as benchmark for more complex approaches. The second method, the ESP approach is based on historical meteorological data and accounts for initial conditions in each forecast. The initialization using known system states is also relevant in the third approach, a dynamical seasonal forecast method, in which meteorological forcing is obtained from bias-corrected climate model forecasts (ECMWF's System4). Given a lead time of 1 month, the dynamical approach provides improved skill in terms of Continuous Ranked Probability Score (CRPS), while for longer lead times the predictive skill is similar to the corresponding ensemble forecasts using ESP. For the period of snowmelt in spring, the CRPS is lowest (best) in the case of ESP and the dynamical approach. In some watersheds, however, the first method which provides forecasts using historical streamflow data performs best. This comparison highlights the fact that the predictability is low in some watersheds. The study is interesting, the results are promising and the paper fits very well into both the special issue on "Sub-seasonal to seasonal hydrological forecasting" in particular and HESS in general. The methodology is comprehensively presented and the results are discussed in a balanced way. Related work and relevant references are mentioned and acknowledged. Especially the assessment of added value provided by each increased level of complexity (using streamflow data only $->$ ESP $->$ dynamical forecasts) is very useful. Another important point is that the paper presents a specific case study in which operational forecasts have already been issued and new methods are going to be implemented. This might be relevant for other forecasting centers. However, in my opinion, the paper needs a few minor revisions and technical corrections. It's my impression that the section on reliability seems to be detached to a certain degree given that the findings from this analysis are not really considered in the summary. Moreover, the paper would benefit from some additional explanations that might improve comprehensibility. Please find my suggestions below.

Response :

Thank you once again for your detailed review and constructive comments. CRPS is used to assess the overall performance of the forecasting system. Reliability diagram and PIT histogram have been used as diagnostic tools to understand the lack of performance of the ensemble forecasts. The conclusion is that even if the reliability is quite good, some under-dispersive behavior or bias are observed in some cases. As the number of event are limited for the verification, these punctual problems can penalize strongly the CRPS of corr-DSP. Some part of the summary has been rewritten to take into account your remark as follows at Page 1, line 18-20 : "Corr-DSP appears quite reliable but, in some cases, under-dispersion or bias is observed. A more complex bias correction method should be further investigated to remedy this weakness and take more advantage of the ensemble forecasts produced by the climate model." Answers and clarifications for the other specific comments are detailed below.

Specific comments :

1. **Page 1, line 9: the abbreviation "corr-DSP" is not explained in this context and might be omitted here**

   Page 1, line 9 : The explicit meaning of the abbreviation is not explained here, as well as ESP at page 1, line 8. However, these abbreviations are useful here as we refer to them in the rest of the abstract. The word "corr-DSP" has been kept in the manuscript.

2. **Page 1, line 9: Would it make sense to point out that "Simulated streamflow computed using observed meteorological data is used as benchmark."?**

   Page 1, lines 10-11: The sentence "Simulated streamflows are used as observations" has been replaced by "Simulated streamflow computed using observed meteorological data is used as benchmark".

3. **Page 2, lines 12-28: In this section, historical streamflow prediction (HSP) and extended ensemble streamflow prediction (ESP) are presented. In my opinion, some additional explanations might be helpful in this context. You could explain that using HSP is in general possible without using a hydrological model, even though, in particular, you involve the output of a model in your specific case study. ESP, in contrast, does require a hydrological model in order to improve forecasts through explicitly incorporating initial states in the forecasts. The relevance of using hydrological models, as already pointed out, might be helpful in the process of understanding the different methods you apply.**

   Page 2, lines 27-28 : The sentence "Contrary to HSP, ESP require the use of a calibrated hydrological model to produce hydrological forecasts" has been added.

4. **Page 3, line 2: I am not sure whether "questioning" is the appropriate verb in this context. As far as I know it would make sense if you have reason to doubt the usefulness. Instead, using "assessing" might be a better option.**

   Page 3, line 14 : We agree and the word "questioning" has been replaced by "assessing".

5. **Page 3, line 14: Please add appropriate references of the DEMETER project and also explain the project's acronym.**

   Page 3, line 26-27 : The signification of the DEMETER acronym "Development of a European Multimodel Ensemble system for seasonal to inTERannual prediction" has been added.

   Also on page 3, line 27, the following reference has been cited according to your suggestion: "Palmer, T., Doblas-Reyes, F., Hagedorn, R., Alessandri, A., Gualdi, S., Andersen, U., Feddersen, H., Cantelaube, P., Terres, J., Davey, M., et al.: Development of a European multimodel ensemble system for seasonal-to-interannual prediction (DEMETER), Bulletin of the American Meteorological Society, 85, 853–872, 2004"

6. **Page 5, Table 1: Please add mean temperature and mean streamflow if easily available. As mean precipitation is indicated, averages of temperature and streamflow might gain insight into the climate characteristics.**

   Page 6, Table 1 : Information about the mean temperature and mean streamflow have been added. Also, the labels of the columns for temperature data have been changed to clarify the content. The content have also been updated. Consequently, small changes can appear compared to the previous version of the manuscript.

7. **Page 5, line 13-14: Does this mean that short-term forecasts are extended by the ensembles generated using ESP? Please consider rephrasing.**

   Page 6, line 1-10 and page 7, line 1-20 : The first paragraph of the operational system description has been reformulated as follows: "The current operational streamflow forecasting system at Hydro-Québec relies on Extended Streamflow Prediction (ESP, Day, 1985) and can be divided into three distinct stages. In the first stage, an analog approach (e.g. Marty et al., 2012) based upon deterministic meteorological forecasts from Environment and Climate Change Canada is used to produce short-term meteorological forecasts. The definition of "short-term" is not fixed. The lead-time depends on watersheds and meteorological events. On average, five to seven days ahead forecasts are produced using this method. The second stage aims to produce seasonal forecasts. Observed precipitation and temperature for previous years are considered as plausible future scenarios. Hence, archived observed meteorological conditions for all previous years (since 1950) form an ensemble. These analog-based meteorological ensembles are used to extend the short-term forecasts obtained in the first stage. These scenarios are then fed to a lumped conceptual hydrological model (described below) to obtain hydrological ensemble forecasts. Lastly, the third and last stage begins when the influence of initial conditions becomes negligible. Observed streamflow for the same Julian day of each available year in the database are then considered as equiprobable long-term forecasts (Historical Streamflow Predictions, see Introduction). The appropriate moment to shift from ESP to HSP is fixed by the forecaster and varies between watersheds. Note that Hydro-Quebec is currently improving its forecasting system by integrating ensemble weather forecasts with statistical post-processing for short-term forecasting, and by developing a weather generator for medium-term forecasting. This new system is expected to become operational in 2018."

8. **Page 6, lines 27-28: Do the Nash-Sutcliffe values are computed using daily time series?**

Page 8, line 7 : The sentence "The Nash-Sutcliffe efficiencies (NSE) ranges from 0.30 to 0.86 for the 1981-2015 period." has been changed for "The Nash-Sutcliffe efficiencies (NSE) based on daily streamflow data range from 0.30 to 0.86 for the 1981-2015 period" in order to be more specific.

9. **Page 7, line 18: Forecasts are also computed using one day time steps?**

   Response : Yes. Meteorological ensemble forecasts are really computed for 6-hour time steps. However, for this study, forecasts were only available at daily time steps from 0Z to 0Z. Page 8, lines 32-33 : To clarify, the sentence "Ensemble forecasts are computed at finer time-steps than one day but are available only at daily time-step from 00Z to 00Z for this study." has been added. Furthermore, hydrological forecasts are computed at daily time steps. However, hydrological observations are only available at daily time step between 05Z and 05Z as mentioned in Section 2. In addition, the following sentence has been added at page 9, lines 19-22: "Both meteorological and hydrological forecasts are available at daily time steps. However, as mentioned in section 2 and 3, a lag exists between daily forecasts and observations. A monthly aggregation of the different variables is performed for verification, in order to limit the impact of the lag between forecasts (meteorological and hydrological) and observations."

10. **Page 8, line 6: Please indicate why the number of forecasts amounts to 420. 35 years x 3 months x 4 seasons?**

    Page 9, lines 32-33 : The sentence "Because both meteorological and streamflow observations are not available after the 31th of December 2015, 2 to 7-months lead-time have 419 to 413 forecast-observation pairs for the verification, respectively." has been added. In addition, on page 10, line 7-9, the sentence "For one season and one lead-time, each set of verification comprises 105 monthly ensemble forecast-observation pairs" was modified for "For one season and one lead-time, each set of verification comprises around between 100 and 105 monthly ensemble forecast-observation pairs."

11. **Page 8, line 23: Is the term "confidence interval" really correct in this context? As far as I understand, say we consider a reliable forecast of a specific event, a probability of 95% should at best also refer to 95 out of 100 occasions in the observed dataset. Please also define the terms "nominal" and "effective".**

    Page 10, lines 19-22: In order to define the terms 'nominal' and 'effective' as well as to clarify the use of the term 'confidence interval', the sentence "Confidence intervals computed from reliable forecasts should be in agreement with their definition: the 95% confidence interval, for instance, must include on average 95 observations out of 100. For each nominal confidence level probability from 0.1 to 0.9, the effective frequency of the observation occurrence in the given nominal interval is calculated. Then, the effective frequencies are plotted against the nominal confidence level probability." now reads "The reliability diagram diagnostic tool compares the observed coverage frequency (effective, $1 - \hat{\alpha}$) with the corresponding theoretical confidence levels (nominal, $1 - \alpha$) of predictive confidence intervals calculated from ensemble forecasts. Of course, if forecasts are reliable, these values $1 - \hat{\alpha}$ and $1 - \alpha$ should be equal for any confidence level."

12. **Page 10, lines 13-14: This phrase is hard to understand. Please consider rephrasing.**

    Page 12, lines 21-23 : The sentence "A leave-one-year-out procedure is used, which consists in excluding the forecast to correct from the bias evaluation process" was modified to improve clarity and now reads "A leave-one-year-out procedure is used to calculate bias and correct the forecasts. This consists in calculating the bias based on available forecasts issued on the same month, excluding the month under correction".

13. **Page 12, line 13: "evolution" might be more appropriate that "maturation" in this context.**

    Page 14, line 5: We agree. The word "maturation" has been replaced by "evolution".

14. **Page 13, line 1: corr-DSP forecasts**

    Page 14, line 10: Thank you for pointing this out. The word "ensemble forecasts" has been replaced by "corr-DSP".

15. **Page 14, line 1: Here, you state that Fig. 6 presents a reliability diagram while in the figure's caption it is labeled as PIT diagram. This is a little bit misleading and might cause confusions even if the type of information is similar to a certain degree. Please confirm or specify the type of figure more detailed.**

Page 17, Figure 6 : We apologize for this typo. The words "PIT histograms" have been replaced by "Reliability diagrams" in the title of the figure.

16. **Page 14, line 11: Here, you state that Fig. 7 presents a PIT histogram while in the figure's caption it is labeled as rank histogram. Is this in line with your explanations in Sect. 4?**

    Page 17, Figure 7 : You are right and "Rank histograms" has been replaced by "PIT histograms" in the title of the figure. As mentioned in section 4, PIT histograms and rank histograms are equivalent in terms of interpretation. However, Fig. 7 really represents a PIT histogram.

17. **Page 14, lines 11-12: Further explanations might improve comprehensibility (e.g., by stating that an equal distribution indicate accurate ensemble forecasts).**

    Page 16, lines 26-31 : Further explanations about the interpretation (flat, bias and over/under-dispersive cases) of the PIT histogram has been added as follows: "A flat PIT histogram corresponds to an accurate forecasting system whereas a higher effective frequency on one side of the histogram indicates the presence of bias (asymetric shape). Higher effective frequency in the middle of the PIT histogram is linked with too much dispersion of the ensemble (bell-shape) and, on the contrary, higher effective frequencies on both sides of the histogram is the sign of an under-dispersive behavior (U-shape, the spread of the ensemble is too small for most forecasts)."

18. **Page 14, Figure 6: In my opinion, labeling each row of the diagram by stating the watershed's numbers might be more intuitive (see, e.g., Figure 2). This is also relevant in the case of Figures 7, 8, 9, and 10.**

    Pages 17, 19, 20, 21, Figures 6, 7, 8, 9 and 10 : The watershed's numbers have been added directly on the rows of the figures 6 and 7 and the lead-time on the rows of the figures 8, 9 and 10 to your suggestion.

19. **Page 15, line 16: The bias correction is applied for each month. Single events at time scales smaller than one month might be subject to biases different to the monthly values.**

    Response: We completely agree. This is certainly a limit of the chosen bias correction method, which has the advantage of being simple, but it is possible that in this case a more sophisticated bias correction method would be worth the additional complexity. Hence, the sentence "Moreover, single events at time scales smaller than one month might by be subject to biases different than the monthly values used for bias correction" has been added on page 18, lines 19-20.

20. **Page 16, line 11: By the way, the term dispersion is often used throughout the manuscript if the variability is overestimated (or underestimated). Variability might be more appropriate as mentioned in line 6 on page 15.**

    Response : The same issue has been raised by Reviewer #2. Throughout the manuscript, *Dispersion* refers directly to the spread of a single ensemble forecast, namely the variability of the members. To clarify the meaning of the word *dispersion*, the following sentence has been added on page 10, line 18: "In the following, the term 'dispersion' refers to the spread of the ensemble forecasts" The term 'variability' could also be used but is less frequent than the term 'dispersion.

21. **Page 17, line 11: Do you mean corr-DSP when discussing the results of ensemble meteorological forecasts?**

    Page 11, lines 18-19 : To clarify, the sentence "raw and bias-corrected meteorological forecasts will refer to the ensemble meteorological forecasts of the ECMWF System4" has been added.

22. **Page 17, line 12: Is it possible to prove if the skill is significant or not significant from your analyses? The term significant should be proved by providing statistical measures.**

    Response : We agree. We computed approximate confidence intervals for the CRPS of corr-DSP and ESP using a bootstrap procedure. These intervals could then be used to add some more formal indications about the significance of the verification results. This has been done for the revised version of the manuscript. We found that the difference between the CRPS of corr-DSP and ESP are not significant for almost all cases. The remaining of the manuscript has also been verified to make sure that there are no other instances. Page 19, line 12 : The word "significant" has been deleted.

Page 21, line 8 : The word 'significant' has been replaced by 'notable'.
Page 21, line 12 : The word 'significant' has been replaced by 'important'.

23. **Page 18, line 5-6: Please explain in brief why corr-DSP is less reliable. Is this finding relevant for winter or all seasons? Maybe you can refer to the reliability diagram?**

    Page 20, lines 5-10 : The following sentences have been added: "see Figure 7 for an example, in both meteorological and hydrological forecasts. The lack of skill of corr-DSP can originate from different sources. First, linear scaling is a rather simple bias correction method. It was performed using monthly bias and thus, there is a possibility that biases at smaller temporal scales can remain. Second, as precipitation was originally over-predicted in most cases by System4 (see Figure 2), bias correction results in a reduction of the ensemble spread for precipitation forecasts, and possibly for streamflow forecasts also."

Technical corrections :
Page 1, line 7: Please add the CRPS (abbreviation) here as it is mentioned later without explanation (cf. line 14)
Page 1-line 6: The word "CRPS" has been added in parenthesis.
Page 2, line 32: I would suggest using the singular form of precipitation
Page 3, line 9 : Done!
Page 8, line 20: distributions (plural)
Page 10, line 16: Also done.
Page 10, Figure 2: Please add the dimension of the precipitation bias in the color bar.
Page 13, Figure 2 has been modified according to your comment.
Page 15, line 15: remain
Page 18, line 18: It has been corrected.
Page 17, line 11: "is predictable" instead of "are predictable"
Page 19, line 11: The sentence was corrected according to your suggestion.
Page 18, line 6: originates
Page 20, line 5: Done!

2 Response to Reviewer 2

General comment: The paper presents study of potential skill of different meteorological forcing for seasonal forecasting over 10 basins in Quebec that are operationally (short-term) forecasted and economically used for hydro-power production. For these basins in particular, a seasonal forecasting system delivering streamflow volume forecast might be of great potential economic benefits resulting from more effective operation planning. The aim of the study is to compare three methods of seasonal forecasting, namely: a) hydroclimatology (based on simulated streamflow); b) ESP (streamflow simulation based on known initial conditions of the basin and ensemble of historical precipitation and temperature observations); and 3) dynamic hydrological modelling using ECMWF seasonal forecasts of precipitation and temperature. Topic of the paper is fully appropriate for the HESS. Authors present solid introduction and literature review. They use correct methodology that is generally well explained. Results are presented in a clear and understandable manner. The results are probably less optimistic than one might expect when a complex dynamic modelling approach is implemented, especially for a lead times longer than 1 month, however even negative (or not clearly positive) results are worth of publication (I suspect the limited resolution of aggregated observed meteorological data to be one of the factors that contributed to bit fuzzy results.). I recommend accepting the paper after some minor revisions to the paper as proposed bellow. Authors presents results in more detail for 3 of 10 researched, as they are referred as representing different behaviour of evaluation statistics. For readers, I believe, some more explanation (e.g. on how basins are clustered in this aspect to groups represented by selected basins) would be beneficial. This should also be reflected in the discussion of results (could some physical geographical characteristics be the underlying reason? Do the verification results correlate or not with N-S performance of the hydrological model for these basins?). Authors use simple linear bias correction of ECMWF System4 Forecasts based on differences between forecast mean and observation on a monthly time scale. This method doesn't reflect the ensemble spread of the forecast or the temporal variability of precipitation and temperature within individual months. It would be valuable if authors shortly discuss this issue, in particular, if the bias corrected precipitation and temperature forecasts exhibit ensemble spread over-prediction or under-prediction behaviour (it might have a consequence for interpretation of stream flow and volume forecast results). In general, I would suggest that reasons of a failure of corr-DSP to outperform the ESP beyond 1 month lead time are further investigated

and discussed.

Response :

Thank you very much for reviewing the manuscript and providing comments. We agree that the result of this study are not as clearly positive as expected for the performance of ECMWF's System4, especially for lead-times longer than one month. As you mentioned, different reasons can limit the skill of seasonal forecasts, such as the spatial resolution of both observations and forecasts (grid), as well as the choice of a particular bias correction method.

The watersheds chosen for results presentation are linked with their CRPS performance. Indeed, watersheds 3 and 5 present opposite behaviors of the CRPSS for the 1-month lead-time during winter and spring. Actually, during our study we did not find any clear relationship or pattern between geographical location or characteristics of the watersheds and forecasts' performance. Moreover, no clear link between N-S performance and the forecasts performance has been noted. Consequently, no mention of these points has been added to the manuscript.

Concerning the time scale in the application of the bias correction method, Reviewer #1 also pointed out this element. Page 20, lines 5-10 : The following sentences have been added: "see Figure 7 for an example, in both meteorological and hydrological forecasts. The lack of skill of corr-DSP can originate from different sources. First, linear scaling is a rather simple bias correction method. It was performed using monthly bias and thus, there is a possibility that biases at smaller temporal scales can remain. Second, as precipitation was originally over-predicted in most cases by System4 (see Figure 2), bias correction results in a reduction of the ensemble spread for precipitation forecasts, and possibly for streamflow forecasts also."

Through the different comments and remarks, we have addressed some of the explanations related to the lack of performance for lead-times longer than 1-month. Further studies are needed to explore which aspects of these potential explanations are the most relevant. Possible future studies could include the comparison of other bias correction methods, the comparison with other products characterized by different resolutions as well as the comparison with gridded meteorological observations instead of watershed average forecasts, among others.

Finally, we would like to thank you for highlighting some technical issues. Answers and modifications regarding your other comments and suggestions are detailed below.

Specific comments:

1. **p. 1 lines 13 to 16 – I am afraid that the wording of abstract doesn't reflect properly results presented in the paper itself.**

   Response: Though we agree that some nuances can be added, we do not fully agree with your recommendation.
   Page 1, lines 14-15 : The 1st sentence of the highlighted lines is "For the 1-month lead-time, a gain exists for almost all watersheds during winter, summer and fall." This sentence is based on the results presented in the bottom row of Figure 5, where the performance of corr-DSP is compared with that of ESP. Except for watershed number 5 during winter, all skill scores are indeed positive for winter, summer and fall, which indicate a gain in performance when using corr-DSP instead of ESP. We exclude spring because the results are too contrasted over the different watersheds. This first sentence has been kept as is.
   Page 1, lines 15-16 : The second sentence is "However, volume forecasts performance for spring is close to the performance of ESP." This sentence is still based on the bottom row of Figure 5. It reflects the fact that for watersheds 2, 4 and 10, the CRPSS is really close to 0. Then, for watersheds 1, 3, 6, 8 and 9, the performance is still close to 0 in favor of ESP or DSP. There is skill only for watersheds 5 and 7 during this season (spring). The second sentence was replaced by "However, volume forecasts performance for spring varies from one watershed to another. For most of them, the performance is close to the performance of ESP."
   Page 1, lines 16-18 : The third sentence "For longer lead-times, results are mixed and the CRPS skill score is close to 0 in most cases." is still based on the same figure. Even if the CRPSS is close to 0, the color scale shows that if a preference is given, it is, in most cases, in favor of ESP. This precision could easily be added in the abstract, according to your comment. So, the third sentence has been replaced by "For longer lead-times, the CRPS skill score is mostly in favor of ESP, even if for many watersheds, ESP and corr-DSP have comparable skill."
   Page 1, line 18-20 : The sentence has been added "Corr-DSP appears quite reliable but, in some cases, under-dispersion or bias is observed. A more complex bias correction method should be further investigated to remedy this weakness and take more advantage of the ensemble forecasts produced by the climate model."
   Page 1, lines 20-22 : The last sentence "Bias-corrected ensemble meteorological forecasts appear to be an interesting source of information for hydrological forecasting." could indeed benefit from a reformulation. Corr-DSP is interesting compared to the use of streamflow climatology. Moreover, compared to ESP, the added value of Corr-DSP is mostly visible for the 1st month. The last sentence has been replaced by "Overall, in this study, bias-corrected ensemble meteorological forecasts appear to be an interesting source of informa-

tion for hydrological forecasting for lead-times up to 1-month. They could also complement ESP for longer lead-times."

2. **p. 11 line 4 "... of bias corrected forecasts. The raw ensemble ..."**

   Response: We apologize for this typo.
   Page 12, line 27 : The word "with" has been deleted.

3. **p. 13 line 16 Authors state that "in general, corr-DSP outperforms ESP for the 1-month lead-time for watershed 5 and 7." Just bye eye control of figure 5, I haven't that intention especially as for basin 5 the ESP performs much better for winter period.**

   Response : This is a mistake as we clearly see that the watershed 5 is not a good example in this sentence.
   Page 16, lines 8-10 : The sentence "In general, corr-DSP outperforms ESP for the 1-month lead-time for watershed 5 and 7." has been replaced by "In general, corr-DSP outperforms ESP for the 1-month lead-time, with some exceptions such as watershed number 5 in winter or watersheds number 3 and 9 during spring."

4. **p. 15 line 9 "...(a) ESP and (b) corr-DSP ..."**

   Page 18, line 9 : The word "and" has been deleted.

5. **p. 16-19 figures 8 to 10 present 1, 2 and 3 months lead times of spring freshet forecasts. This is defined as (for majority of basins) period from April 1st to June 30th. Does it mean that the 1-month lead-time is forecast issued on March 1st (etc.). Please note that in fig. 11 this is obviously the case as the 0 months lead time is also included. More description of graphical symbols in fig. 8 to 11 should be provided too.**

   Response: In figures 8 to 10, the 1-month lead-time corresponds to the forecast issued on the 1st day of the spring freshet (namely on the 1st of April for the majority of basins), for the following month. This is obviously incoherent with Figure 11. We thank you for pointing out this mistake. Lead-time is the delay between the date of emission of the forecast and the end of the validity period. We have added this definition in the manuscript.
   Page 11, lines 22-26 : The sentence "Moreover, the lead-time refers to the time lag between the emission date of the forecast and the time at which the forecast is valid. For instance, a skill score for the January-February-March season for the 5-month lead-time correspond to the performance of the forecasts issued 5 months earlier, in August, September and October." has been replaced by the sentence "The lead-time is defined herein as the time between the date of emission of the forecast and the end of the validity period of the forecasts. For instance, the 1-month lead-time of the forecast issued on the 1st of January is the monthly volume or temperature of January."
   Page 19, 20, 21, 22, respectively Figure 8, 9, 10, and 11 : The lead-times in Figure 7 to 11, as well as their titles, have been changed accordingly to the previous definition.
   Page 18 , line 9 : Concerning the last sentence of the comment, we have modified the original sentence, which now reads "Figure 8 to 10 present the boxplots of ensemble forecasts". We assume that boxplots are well-known tools for all readers.
   Page 18, lines 10-13 : This sentence has been added "In those figures, the 3-month lead-time forecast corresponds to the forecasts issued at the beginning of the spring freshet period with a 3-month validity period. The 4-month and the 5-month lead-time forecasts are issued respectively 1 and 2 months before the beginning of the spring freshet period, with corresponding 4- and 5-month validity period."

6. **p. 18, line 2-3 consider to use "monthly flow volume" instead of "monthly volume"**

   Page 20, line 2 : The word "flow" has been added.

7. **p. 18, line 6 Authors use term "dispersion" throughout the paper, e.g. "this possibly originate from bias propagation or dispersion issues." However, I am afraid that the meaning of "dispersion" is not clear and needs some correction (e.g. ensemble spread of meteorological**

inputs, variability of ...).

Response: The same issue has been raised by Reviewer #1. Throughout the manuscript, *Dispersion* refers directly to the spread of a single ensemble forecast, namely the variability of the members in a single forecast. Page 10, line 18 : To clarify the definition of the term "dispersion", the following sentence has been added "In the following, the term 'dispersion' refers to the spread of the ensemble forecasts."

3 Response to Reviewer 3

General comments
The paper is well written, and technically and scientifically sound. Applied methods and data used are well described, and results are presented in a concise and clear way. The paper uses methodologies and results from previous research. The main contri- bution is the verification of bias-corrected ECMWF System 4 forecasts for hydrological forecasting in Quebec, Canada. This supplements, and to a large degree confirms, previous verification studies in other regions.

Response : We would like to thank you for your review and comments. Answers and modifications for the detailed comments are detailed below.

Detailed comments
1. Page 7, line 28-30. The procedure for deriving catchment average precipitation and temperature is not that clear. Why is it necessary to first downscale ECMWF forecasts and then aggregate over a catchment?

Page 9, line 9-16: The original paragraph "Those original resolutions are both too coarse for hydrological applications, as only very few grid points fall inside the watersheds delineations. The original grid was thus downscaled to a 0.1 degree grid through linear interpolation in order to obtain multiple grid points for each watershed. Then, since HSAMI is a lumped model, grid points were averaged to aggregate the information at the watershed scale." has been modified as follows "Since HSAMI is a lumped model, meteorological forecasts has to be a single point representative of the meteorological conditions over the watershed. The original resolution is too coarse for hydrological applications, as only very few grid points fall inside the watersheds delineations. The original grid was thus downscaled to a 0.1 degree grid through linear interpolation in order to obtain more points inside the watersheds boundaries. This allows to ensure that points close to the watersheds boundaries contribute to more accurate meteorological forecasts over the watershed. Then, grid points were averaged to aggregate the information at the watershed scale."

2. Page 9, line 18-20. Repetition. Described earlier.

Response : Although we agree that the information is repeated, it was made on purpose and we would prefer to leave the sentence there. Indeed, the denominations 'DSP' and 'HSP' are not conventional and we wanted to remind the reader about their meaning at this point in the manuscript. Reviewer 1's 21st specific comment (about Page 17, line 11) also indicates that the denominations used in the manuscript can perhaps be confusing, so we would really prefer to keep this repetition.

3. Page 10, line 12-13. Both precipitation and temperature are bias-corrected.

Response : Yes, they are both bias-corrected.
Page 12, lines 19-20 : The original sentence has been completed by "as well as minimal and maximal temperatures".
Page 12, lines 17-19 : Moreover, we found that the previous sentence was not clear. It was thus modified and now reads "As mentioned above, Crochemore et al. (2016) have shown that the simple linear scaling method provides results comparable to the more complex distribution mapping to correct the bias in precipitation ensemble forecasts".

4. Page 13, line 16. General performance of watersheds 5 and 7 described is not clear

Response : This issue was also raised by Reviewer 2 (his or her 3rd specific comment). This is a mistake. In fact, we clearly see that watershed 5 is not a good example in this sentence.
Page 16, lines 8-10 : The sentence "In general, corr-DSP outperforms ESP for the 1-month lead-time for watershed 5 and 7." has been replaced by "In general, corr-DSP outperforms ESP for the 1-month lead-time, with some

exceptions such as watershed number 5 in winter or watersheds number 3 and 9 during spring.”

6. Page 15, Figure 7. PIT histograms and not rank histograms, I expect.

Response : Thank you for highlighting this mistake. Page 17, Figure 7 : The words ”Rank histograms” have been replaced by ”PIT histograms”.

7. Page 16, line 1-2. The problem of underdispersion of the bias-corrected ensemble could be elaborated. There is a general overestimation of precipitation cf. Fig. 2. In this case, linear scaling will produce a bias-corrected ensemble with smaller dispersion than the raw ensemble.

Response : We agree. The linear scaling method modifies the dispersion of precipitation forecasts and can influence the dispersion of streamflow forecasts.
Page 18, line 23-24 : The sentence ”As precipitation is most of the time overpredicted, as shown in Figure 2, bias corrected precipitation forecasts exhibit a lower dispersion than raw forecasts. This can explain the smaller dispersion of the volume forecasts.” has been added.

4 Additional changes in the manuscript

The additional changes include all typo corrections, sentences rephrasing or additional information.

Page 1 , line 1 : Some complementary information have been added, now reads ”Hydro-power production requires optimal dam and reservoir management to prevent flooding damage and avoid operation losses. ”
Page 2 , line 14-16 : The sentence has been reformulated as follows ”Examples include multiple regression type of models treated in a Bayesian perspective (e.g. Wang et al., 2009) or a frequentist framework (e.g. Moradkhani and Meier, 2010; Sveinsson et al., 2008)”.
Page 2 , line 22-23 : The complementary information ”either observed or simulated” has been added.
Page 3, line 7 : The word ”Nordic” has been changed for ”northern”.
Page 3, line 30 : The word ”Mutlimodel” has been corrected by ”Multimodel”.
Page 4, line 10 : The word ”Predicition” has been corrected by ”Prediction”.
Page 4 , line 22-24 : The sentence has been added ”Hydro-Québec is a government owned corporation that produces and distributes electricity in the Province of Québec. The installed hydroelectricity capacity of Hydro-Québec is more than 36 000MW.” and the next sentence adapted, now reads ” Together, the ten watersheds under study represent more than 8750 MW”.
Page 6, Table 1 : Update of the values and footnote.
Page 9 , line 23 : Add-on ”and optimization of electricity production”.
Page 9, line 26-27 : Add-on ”and reservoir operation”.
Page 9 , line 33 to page 10, line 5 : Reformulation, now reads ”The verification set should be as large as possible, in order to ensure statistical significance of the results. It should also be homogeneous. However, in reality, forecasts characteristics change depending on the period of the year and contradictory behaviors can balance each other out.”
Page 10, line 11 : ”tools” switched to ”diagnostic devices”.
Page 10 , line 12 : ”score” switched to ”scoring rule”.
Page 10, line 27-28 : The sentence ”PIT histograms are preferred over rank histograms herein because of the changing number of members (see section 3).” has been deleted.
Page 10 , line 31 : Typo correction, ”of” switched to ”for”.
Page 11 , line 3 : Reformulation change, ”presented in equation 2” switched to ”expressed as”.
Page 12 , Equations 3 and 4 : k index has been added into the equations and the descriptive paragraph line 7.
Page 12 , line 8 : Correction ”considered for the bias estimation” instead of ”in the verification set”.
Page 12, line 8-11 : Reformulation, now reads ”We therefore assume an additive bias for monthly mean minimal and maximal temperature: it is the mean error. For monthly precipitation, bias is defined as the ratio of the forecasts mean to the mean observed accumulation. A multiplicative bias is then assumed for this meteorological variable.”
Page 16 , line 23 : Terminology modification to be in line with technical terms used in section 4 for reliability diagram description. The modified sentence is ”However, small differences between the observed coverage frequency and the theoretical confidence level remain.”
Page 17 , Figure 6 : Modification of the axis labels to be in line with the definition of the reliability description given at Section 4.
Page 18 , line 25 : ”boxplot” changed to plural form ”boxplots”.
Page 19, line 6 : Type, ”analyse” switch to ”analysis”.

Page 20 , line 2 : Typo correction ”CRPSS” instead of ”CRPSSS”.
Page 21 , line 11-12 : Reformulation, now reads ”Moreover, according to our results, it is for the 1-month lead-time that the most important gain could be achieved by dynamical models over climatology (ESP).”
Page 22, Figure 11 title : ”boxplot” changed to plural form ”boxplots”.

Verification of ECMWF System4 for seasonal hydrological forecasting in a northern climate

Rachel Bazile[1], Marie-Amélie Boucher[1], Luc Perreault[2], and Robert Leconte[1]

[1]Département de génie civil, Université de Sherbrooke, 2500 Boul. de l'Université, Sherbrooke, Québec, J1R 2R2, Canada
[2]Institut de Recherche d'Hydro-Québec (IREQ), 1800 boul. Lionel-Boulet, Varennes, Québec, J3X 1S1, Canada

Correspondence to: rachel.bazile@gadz.org

Abstract. Hydro-power production requires optimal dam **and reservoir** management **to prevent flooding damage and avoid operation losses**. In a northern climate, where spring freshet constitutes the main inflow volume, seasonal forecasts can help to establish a yearly strategy. Long-term hydrological forecasts often rely on past observations of streamflow or meteorological data. Another alternative is to use ensemble meteorological forecasts produced by climate models. In this paper, those produced by the ECMWF (European Center for Medium-Range Forecast)'s System 4 are examined and bias is characterized. Bias correction, through the linear scaling method, improves the performance of the raw ensemble meteorological forecasts in terms of Continous Ranked Probability Score **(CRPS)**. Then, three seasonal ensemble hydrological forecasting systems are compared: 1) the climatology of simulated streamflow, 2) the ensemble hydrological forecasts based on climatology (ESP) and 3) the hydrological forecasts based on bias-corrected ensemble meteorological forecasts from System4 (corr-DSP). ~~Simulated streamflows are used as observations.~~ **Simulated streamflow computed using observed meteorological data is used as benchmark.** Accounting for initial conditions is valuable even for long-term forecasts. ESP and corr-DSP both outperform the climatology of simulated streamflow for lead-times from 1-month to 5-month depending on the season and watershed. ~~Corr-DSP appears quite reliable but sometimes suffer from under-dispersion.~~Integrating information about future meteorological conditions also improves monthly volume forecasts. For the 1-month lead-time, a gain exists for almost all watersheds during winter, summer and fall. However, volume forecasts performance for spring **varies from one watershed to another. For most of them, the performance** is close to the performance of ESP. For longer lead-times, ~~results are mixed and the CRPS skill score is close to 0 in most cases.~~ **the CRPS skill score is mostly in favor of ESP, even if for many watersheds, ESP and corr-DSP have comparable skill. Corr-DSP appears quite reliable but, in some cases, under-dispersion or bias is observed. A more complex bias correction method should be further investigated to remedy this weakness and take more advantage of the ensemble forecasts produced by the climate model. Overall, in this study, b**~~B~~ias-corrected ensemble meteorological forecasts appear to be an interesting source of information for hydrological forecasting **for lead-times up to 1-month. They could also complement ESP for longer lead-times**.

1 Introduction

Hydro-power production planning typically requires inflow forecasts to reservoirs at different lead-times. Whereas short term forecasts are used for day to day planing, sub-seasonal (1 to 3 months) to seasonal (up to 6 month) forecasts are used to establish a yearly strategy. Improving the skill of hydrological forecasts at the sub-seasonal to seasonal scale is thus essential. According to DelSole (2004): "A system is said to be unpredictable if the forecast distribution, which gives the most complete description of the future state based on all available knowledge, is identical to the climatological distribution, which describes the state in the absence of time lag information". Hence, the advent of seasonal meteorological forecasts that are more informative than climatology could support water managers in their decision process.

Probabilistic forecasts are necessary to quantify uncertainty about future hydrological conditions. It is especially true for long-term forecasts, as uncertainty grows with lead-time. Operationally, several methods exist to produce sub-seasonal to seasonal hydrological forecasts. They can be broadly divided into two main categories: statistical forecasting and ensemble-based forecasting (Yuan et al., 2015). However, hybrid methods also exist.

Statistical methods can take advantage of relationships between past and future streamflow persistence, (e.g. Svensson, 2016) or between streamflows and teleconnections indices. Examples include **multiple regression type of models treated in a** Bayesian ~~inference~~**perspective** (e.g. Wang et al., 2009) ~~and multiple regression~~**or a frequentist framework** (e.g. Moradkhani and Meier, 2010; Sveinsson et al., 2008).

Ensemble-based forecasting is a widespread uncertainty assessment technique (Cloke and Pappenberger, 2009). Ensemble forecasts comprise different potential future scenarios also called 'members'. One possible method to obtain hydrological ensembles is to provide ensemble meteorological forecasts as inputs to one or several hydrological model(s). Each meteorological scenario leads to one hydrological scenario (member). From these members, different techniques exist to derive probabilistic forecasts (Bröcker and Smith, 2008).

For long forecasting horizons, the simplest type of ensemble forecasts is the climatology of streamflow, **either observed or simulated,** hereafter called 'Historical Streamflow Prediction' (HSP). This naive forecasting method is by definition a reliable forecasting system but of course its resolution can be improved. Even if this kind of forecasting system does not show any predictability, it accounts for different plausible hydrological scenarios based on the past. A simple alternative method, proposed by Day (1985), is called 'Extended Streamflow Prediction' (ESP). To produce ESP, past meteorological observations are considered as equiprobable potential future meteorological scenarios. **Contrary to HSP, ESP require the use of a calibrated hydrological model to produce hydrological forecasts.** If the historical record is long enough, climatology provides a reliable estimation of the distribution of future meteorological conditions, including some extreme scenarios. The main advantage of ESP relative to HSP is that it allows accounting for the current hydrological initial conditions. Several studies have shown that state variables such as soil moisture or snow water equivalent can provide relevant information to extend predictability for lead-times from 1 to several months ahead (e.g. Wood and Lettenmaier, 2008; Shukla et al., 2013; Yang et al., 2014; Yuan et al., 2016). The influence of initial conditions depends on the period of the year and on the location of watersheds (e.g. Yossef et al., 2013). ESP are intuitively appealing, since they are coherent with a natural tendency of humans to judge actual situations according to their memory of past experiences. Moreover, ESP allow practitioners to condition streamflow scenarios only on selected past meteorological scenarios if they wish, for instance by using only the most extremes historical scenarios. Because of their simplicity and efficiency, both ESP and HSP are popular among operational agencies for forecasts from several days to weeks to months (e.g. García-Morales and Dubus, 2007) and still arouse interest as a forecasting system (e.g. Singh, 2016).

However, given the current context of climate change, some past meteorological and hydrological data might not be representative of plausible future conditions. In **northern**~~Nordic~~ contexts, it is expected that climate change will gradually modify the repartition of rain and snowfall during the year, for instance. For the province of Quebec in Canada, climatic projections anticipate a rise in temperature and precipitation~~s~~ (Ouranos, 2015). It is expected that these changes will modify hydrological conditions both at the annual and intra annual scales. Indeed, higher winter streamflows, earlier spring freshet and longer periods of low streamflow during the summer are expected (Guay et al., 2015). In central Sweden, climate change will also affect the seasonality of streamflow, mostly by decreasing the mean snow water equivalent and the mean annual runoff (Xu, 2000).

During the past decade, sub-seasonal to seasonal ensemble meteorological forecasts produced by dynamic climate models have undergone constant improvements and it is worth ~~questioning~~ **assessing** their usefulness for long lead-times inflow forecasting. A dynamic climate model is an atmospheric model, sometimes coupled with an ocean model. Considering the interactions between the atmosphere and oceans allows for modeling long-term phenomenon such as El-Niño and La Niña phases of the ENSO cycle. For instance, according to Kim et al. (2012), for winter in the northern hemisphere, both the European Centre for Medium-Range Weather Forecasts (ECMWF) System4 and the National Centers for Environmental Prediction (NCEP) Climate Forecast System Version 2 (CFSv2) accurately reproduce El Niño/La Niña phases. Temperature variations are more difficult to capture. Regarding the ECMWF System4, Weisheimer and Palmer (2014) assessed the performance of 2m temperature and precipitation forecasts throughout the world. The reliability of forecasts vary from "perfect" to "dangerous", depending on the month of the year, the variables and the location.

Meteorological forecasts from dynamic climate models can be used to produce hydrological forecasts, hereafter called "Dynamical Streamflow Prediction" (DSP). However, according to previous studies, their potential for hydrological purposes is highly variable, depending on the location and the context. Luo and Wood (2008) compared forecasts from the NCEP Climate Forecast System (CFS), multimodel forecasts from a combination of CFS and seven models from the DEMETER **('Development of a European Multimodel Ensemble system for seasonal to inTERannual prediction', (Palmer et al., 2004))** database and ESP for hydrological forecasting on a watershed in Ohio during summer. They found that the multimodel approach is more efficient than a single climate model. Both outperform the ESP approach in terms of Ranked Probability Scores. ~~Mutlimodel~~ **Multimodel** approaches improve seasonal forecasts significantly for the 1-month and 2-month lead-times, whereas single model improvements are limited. Across the United States, the ESP approach has been compared to hydrological ensemble forecasts based on NCEP CFSv1 and CFSv2 by Yuan et al. (2013). Their results indicate that CFSv2 improves hydrological forecasting performances for the 1-month lead-time, whereas CFSv1-based forecasts are not very efficient. Similarly, He et al. (2016) compared the performance of climatology (ESP) and CFSv2 for a single watershed in the Sierra Nevada. Their results indicate only little improvement when using CFSv2. Some agencies already integrate information from long-term ensemble forecasts into their operational hydrological forecasting systems (e.g. Demargne et al., 2014). One major problem of DSP is that ensemble meteorological forecasts produced by dynamic models suffer from bias. However, in a context of climate change, using dynamical meteorological forecasts seems intuitively valuable, as they are expected to better represent the current climate, compared to methods based on past climatology (such as ESP). Crochemore et al. (2016) compared different strategies for bias correction of daily precipitation forecasts and evaluated their efficiency for hydrological forecasting over 16 watersheds in France. They show that correcting precipitation forecasts does indeed translate into an improvement of hydrological forecasts. However, they also show that simple bias correction methods, such as linear scaling, are as efficient as more sophisticated methods.

The goal of this study is to evaluate the potential of Dynamical Streamflow Prediction (DSP) in terms of predictability improvement for long term streamflow forecasting, compared to Historical Streamflow Predic~~t~~tion (HSP) and Extended Streamflow Prediction (ESP). More specifically, long-term meteorological and hydrological forecasts are assessed for 10 northern watersheds in the province of Québec in Canada. Those watersheds are all exploited for hydro-power production. Therefore, skillful long lead-time forecasts are crucial for optimal water management, especially for anticipating and exploiting the large inflow to reservoirs during spring melt.

The paper is organized as follow. After describing the context in section 2, details regarding the case studies and available data will be given in section 3. The forecasts verification methodology is presented in section 4. Results are presented and discussed in section 5 and conclusions are drawn in section 6, which also identifies potential research avenues for future studies.

2 Hydro-meteorological context of the study

2.1 Watersheds

The ten watersheds used as a testbed in this study are all located in the province of Québec and exploited by Hydro-Québec to generate hydro-power. **Hydro-Québec is a government owned corporation that produces and distributes electricity in the Province of Québec. The installed hydroelectricity capacity of Hydro-Québec is more than 36 000MW.** Together, ~~they~~**the ten watersheds under study** represent more than 8750 MW (Hydro-Québec, 2015), as the outlet of each watershed is a hydro-power reservoir. -Figure 1 illustrates the geographical location of the ten watersheds. -Two of them are located in the southern portion of the province, whereas the others are located in the central portion. In addition, some watersheds are part of larger hydro-power production complexes. For instance, watersheds 4, 7, 8 and 9 compose the Manicouagan complex and watersheds 5, 6 and 10 are part of the La Grande complex.

The streamflow regime of the ten watersheds is dominated by a northern climate, which induces snow accumulation and low streamflow during winter (December to February), followed by high streamflow during spring. The exact timing of the spring freshet for a particular watershed is a function on its latitude, its physiographic characteristics, such as slope and orientation, but also of the meteorological conditions that prevail during a particular year. Watersheds located in the southern portion of the province generally produce their highest streamflows in March or April, whereas those located in the central and northern

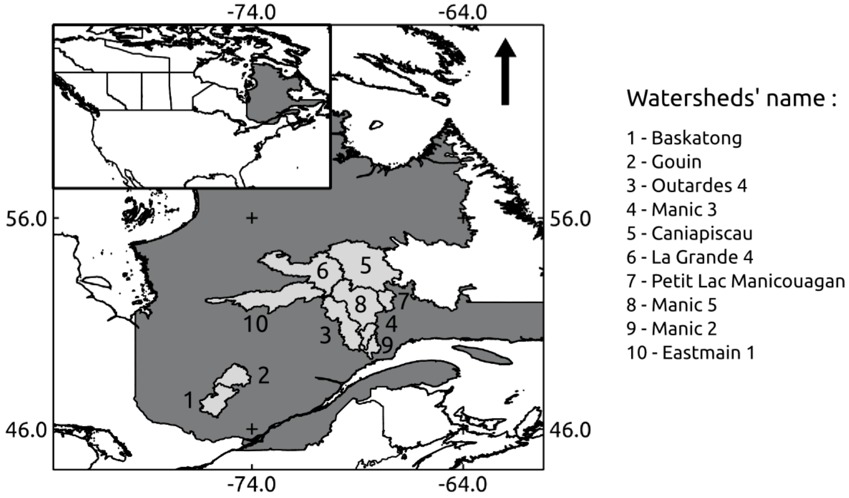

Figure 1. Geographical location of the watersheds used in this study

parts reach their maximum streamflows in May or June. The total volume of runoff associated with the spring freshet obviously depends on the accumulated snow pack during the winter season. Most watersheds also exhibit high streamflow during fall on most years, when evapotranspiration is low and soil is saturated. In fact, whereas temperatures generally show large variability during the winter, the spatio-temporal variability of precipitation is higher during summer and fall. Table 1 presents the hydrometeorological characteristics of the ten watersheds under study.

Table 1. Hydro-meteorological characteristics of the watersheds illustrated on Figure 1

ID (see Figure 1)	Area (km^2)	~~Mean min. temp.~~ **Annual min. temp.*** (°C)	**Annual mean temp.* (°C)**	~~Mean max. temp.~~ **Annual max. temp.*** (°C)	Mean annual precip. (mm)	~~Mean min. streamflow~~ **Annual min. streamflow** (m^3/s)	**Annual mean streamflow** (m^3/s)	~~Mean max. streamflow~~ **Annual max. streamflow** (m^3/s)	Mean date of the max. spring streamflow	Proportion of the spring freshet volume compared to the annual volume (%)
1	13 057	-30	**2.2**	23.9	~~1018~~**1027**	17	**231**	~~1359~~**1306**	28/04	~~46~~**45**
2	9 426	-31.2	**1.3**	23.8	~~971~~**980**	11	**180**	1282	11/05	40
3	17 119	-33.6	**-1.2**	21.7	~~909~~**918**	80	**359**	1759	20/05	44
4	4 245	-31.1	**0.1**	21.6	~~970~~**980**	19	**104**	536	18/05	45
5	37 328	-34.7	**-3.6**	20.5	~~798~~**807**	151	**747**	~~3074~~**3073**	30/05	38
6	28 443	-36.6	**-3.4**	20.9	~~806~~**816**	140	**564**	~~1384~~**1392**	09/06	30
7	4 565	-33.3	**-2.3**	20.9	~~902~~**909**	12	**106**	638	25/05	45
8	24 608	-33.6	**-2.0**	20.8	~~901~~**910**	~~42~~**41**	**526**	~~3275~~**3276**	21/05	48
9	4 100	-28	**1.5**	21.5	~~1005~~**1015**	~~6~~**5**	**77**	537	15/05	51
10	26 944	-35.8	**-2.5**	21.2	~~838~~**848**	109	**572**	1981	23/05	37

* Based on the mean daily temperatures.

2.2 Current operational streamflow forecasting system

~~The current operational streamflow forecasting system at Hydro-Québec is divided into three distinct stages. The first stage, for short-term forecasting, is an analog method~~ (e.g. Marty et al., 2012) ~~based on the deterministic meteorological forecast from Environment and Climate Change Canada. The analog-based meteorological ensembles are then fed to a lumped conceptual hydrological model (described below). The definition of "short-term" is not fixed but rather varies with watersheds and events. On average, it varies between five to seven days. The second stage, for seasonal forecasting, relies on Extended Streamflow Prediction~~ (ESP, Day, 1985). ~~Observed precipitation and temperature for previous years are considered as plausible future scenarios. Hence, archived observed meteorological conditions for all previous years (since 1950) form an ensemble. This ensemble is concatenated with short term meteorological forecasts and both are used as inputs for the hydrological model. Lastly, the third and last stage begins when the influence of initial conditions becomes negligible. Observed streamflow for the same~~

~~Julian day of each available year in the database are then considered as equiprobable long-term forecasts (Historical Streamflow Predictions, see Introduction). The appropriate moment to shift from ESP to HSP is fixed by the forecaster and varies between watersheds. Note that Hydro-Quebec is currently improving its forecasting system by integrating ensemble weather forecasts with statistical post-processing for short-term forecasting, and by developing a weather generator for medium-term forecasting. This new system is expected to become operational in 2018.~~ **The current operational streamflow forecasting system at Hydro-Québec relies on Extended Streamflow Prediction (ESP, Day, 1985) and can be divided into three distinct stages. In the first stage, an analog approach (e.g. Marty et al., 2012) based upon deterministic meteorological forecasts from Environment and Climate Change Canada is used to produce short-term meteorological forecasts. The definition of "short-term" is not fixed. The lead-time depends on watersheds and meteorological events. On average, five to seven days ahead forecasts are produced using this method. The second stage aims to produce seasonal forecasts. Observed precipitation and temperature for previous years are considered as plausible future scenarios. Hence, archived observed meteorological conditions for all previous years (since 1950) form an ensemble. These analog-based meteorological ensembles are used to extend the short-term forecasts obtained in the first stage. These scenarios are then fed to a lumped conceptual hydrological model (described below) to obtain hydrological ensemble forecasts. Lastly, the third and last stage begins when the influence of initial conditions becomes negligible. Observed streamflow for the same Julian day of each available year in the database are then considered as equiprobable long-term forecasts (Historical Streamflow Predictions, see Introduction). The appropriate moment to shift from ESP to HSP is fixed by the forecaster and varies between watersheds. Note that Hydro-Quebec is currently improving its forecasting system by integrating ensemble weather forecasts with statistical post-processing for short-term forecasting, and by developing a weather generator for medium-term forecasting. This new system is expected to become operational in 2018.**

The available archive of past meteorological observations covers the 1950-2015 period. Data include daily minimum and maximum temperature as well as daily rainfall and snowfall. Those variables are only available at the watershed scale, meaning that observations from individual weather stations were spatially aggregated before being archived. Those weather stations are part of a province-wide cooperative network called RMCQ (in French *Réseau météorologique coopératif du Québec*, Lepage and Bourgeois, 2011). The aim of this cooperative network is to pool together data from private and public collaborators. Unfortunately, the number of stations and the interpolation method have evolved over time. At the time of writing, it was not possible to obtain detailed information regarding those successive changes. However, meteorological data is generally of good quality and there is no missing day. Daily data are collected from 6 UTC to 6 UTC for precipitation and temperature and from 5 UTC to 5 UTC for streamflow data.

For the purpose of this study, climatology-based ensemble forecasts were built. All available years were used except one, in rotation. For instance, the climatology based forecasts for year 1980 include all years but 1980. This setup is of course different than the operational framework, where the ensemble size grows year after year and the future is unknown. However, the methodology used here allows to maintain a constant ensemble size (64 members). In addition, this produces ensembles that are free of any possible trend in the time series of climate data. Indeed, all information about past meteorological conditions are used as inputs to the hydrological model.

In all cases, meteorological series are used as input to HSAMI, a lumped conceptual hydrological model described below (in French, Fortin, 2000). This model is based on a series of three linear reservoirs which supply two hydrographs. Snow accumulation and melt are based on a degree-day approach. HSAMI uses daily minimum and maximum temperature, as well as rainfall and snowfall to compute the mean streamflow at the outlet of the watershed at a daily time-step. The model has 23 parameters that must be calibrated against previous streamflow observations. The sets of parameters used in this study are provided by Hydro-Québec. Modeling performance varies greatly from one watershed to another. The Nash-Sutcliffe efficiencies (NSE) **based on daily streamflow data** range~~s~~ from 0.30 to 0.86 for the 1981-2015 period. Despite low NSE for some watersheds, it was judged appropriate to use these parameter sets rather than recalibrating the model. First, this variation is attributable mostly of the quality of hydrological data collected before 2000, on which the calibration of the model is based. Second, since the goal of this study is to assess the influence of meteorological forecasts on hydrological forecasts, simulated streamflows are used as pseudo "observations" in the verification process, and therefore a perfectly well calibrated model is not required (also see Section 4 for details).

The next section describes the ECMWF System4 that is explored in this study as a potential replacement for the current operational forecasting system.

3 An alternative system for seasonal forecasting based on long-term dynamical climate modeling : The ECMWF System4

Our hypothesis is that exploiting dynamical meteorological forecasts in the streamflow forecasting chain would improve the latter compared to ESP. The rationale behind this hypothesis is that dynamical meteorological forecasts should be driven by the current state of the atmosphere at their initialization. Eventually, considering the context of an evolving climate, this could also help hydro-power producers to adapt reservoir and dam management to new situations. The long-range ensemble meteorological forecasts used in this study are produced by the ECMWF (European Center for Medium-Range Forecast)'s System4.

System4 (Molteni et al., 2011) is a global coupled ocean-atmosphere model that officially became operational in 2011. It is used to produce reforecasts and real-time forecasts, that are both archived. The atmospheric model component, namely the ECMWF IFS (Integrated Forecast System) model (version 36r4) includes a lake model and also involves ozone, volcanic aerosol and solar cycle action. Sea ice is depicted by initial sea ice condition for short lead-times and by observed conditions for the five previous years. The initialization of the atmospheric model is performed using ERA-Interim for the reforecasts and the operational procedure of the ECMWF for real-time forecasts. The ocean model is initialized by the Nemovar ocean analysis. Ensemble forecasts are produced by perturbing initial conditions. In the current model setup, five members originate from perturbations of ocean wind surface initial conditions, whereas other members originate from sea surface temperature perturbations and stochastic physics. More details can be found on the ECMWF website (ECMWF, 2017).

Real-time forecasts are issued on the first day of each month for the next 215 days (approx. seven months). **Ensemble forecasts are computed at finer time-steps than one day but are available only at daily time-step from 00Z to 00Z for this**

study. They are archived and available from 2012 to 2015. A set of reforecasts is also available, from 1981 to 2011. Reforecasts for the months of February, May, August and November as well as real-time forecasts comprise 51 members. Reforecasts for the other months comprise 15 members. Both archived past real-time forecasts and reforecasts are used in the present study. This allows the extension of the verification data base length, but poses certain challenges in terms of performance assessment, since the number of members vary. In the following, the term "forecasts" will refer indifferently to real-time forecasts and reforecasts.

In the context of this study, the meteorological variables of interest are those that are inputs to HSAMI, namely daily minimum and maximum temperature as well as total daily precipitation. The original output grid of System4 has a 0.7 degree horizontal resolution for the atmospheric model and around 1 degree for the ocean model at mid-latitudes. **Since HSAMI is a lumped model, meteorological forecasts has to be a single point representative of the meteorological conditions over the watershed.** ~~Those original resolutions are both~~ **The original resolution is** too coarse for hydrological applications, as only very few grid points fall inside the watersheds delineations. The original grid was thus downscaled to a 0.1 degree grid through linear interpolation in order to obtain ~~multiple grid points for each watershed~~ **more points inside the watersheds boundaries. This allows to ensure that points close to the watersheds boundaries contribute to more accurate meteorological forecasts over the watershed**. Then, ~~since HSAMI is a lumped model,~~ grid points were averaged to aggregate the information at the watershed scale. Total precipitation was separated into rainfall and snowfall according to air temperature.

4 Forecast quality assessment

Both meteorological and hydrological forecasts are available at daily time steps. However, as mentioned in section 2 and 3, a lag exists between daily forecasts and observations. A monthly aggregation of the different variables is performed for verification, in order to limit the impact of the lag between forecasts (meteorological and hydrological) and observations. Moreover, ~~A~~**a**s the main goal of long-term hydro-meteorological forecasts is to provide information for seasonal to yearly dam management **and optimization of electricity production**, decision-makers are generally interested in inflow volumes to reservoirs. Consequently, monthly aggregated variables are considered. Monthly averages are computed for minimum and maximum temperature. For precipitation and streamflow, monthly cumulative values are considered using calendar months. Many other types of information derived from streamflow forecasts are useful for dam management **and reservoir operation**. Anticipating runoff volume for spring freshet is crucial, as it allows for planning the lowering of the reservoirs to avoid risks of spillage and flooding. The inflow volume for the spring freshet is calculated between 1^{st} March and May 31 for watersheds 1 and 2 and1^{st} April and June 30 for all other watersheds.

Both forecasts and reforecasts are pooled together to assess forecasts performance. Overall, 420 ensemble forecasts are available for verification purposes, as one ensemble forecast is issued on the 1^{st} of each month between 1981 and 2015 (12 month * 35 years). **Because both meteorological and streamflow observations are not available after the 31th of December 2015, 2 to 7-months lead-time have 419 to 413 forecast-observation pairs for the verification, respectively.** ~~On the one hand, the verification set should be homogeneous. However, in reality, forecasts characteristics change depending on the period of~~

~~the year and contradictory behaviors can balance each other out. On the other hand, the verification set should be as large as possible, in order to ensure statistical significance of the results.~~**The verification set should be as large as possible, in order to ensure statistical significance of the results. It should also be homogeneous. However, in reality, forecasts characteristics change depending on the period of the year and contradictory behaviors can balance each other out.** Considering these two requirements, skill scores of monthly variables are calculated over seasons. Four seasons are used, namely January-February-March (JFM), April-May-June (AMJ), July-August-September (JAS) and October-November-December (OND). For one season and one lead-time, each set of verification comprises **between 100 and** 105 monthly ensemble forecast-observation pairs.

Different numerical scores and graphical tools are used to assess the quality of the aforementioned quantities. The joint use of several ~~tools~~**diagnostic devices** is essential for different reasons. First, ensemble and probabilistic forecasts can be evaluated in terms of different attributes and no single ~~score~~**scoring rule** can simultaneously assess them all. Second, examining different attributes can help to pinpoint strengths and weaknesses of competing forecasting systems. According to Gneiting and Raftery (2007), a good probabilistic forecasting system should be reliable and sharp.- Reliability refers to the statistical consistency between the predictive distribution and the observation, while sharpness refers to the concentration of the predictive distribution. A reliable probabilistic forecasting system produces predictive distribution**s** which are unbiased and representative of the true uncertainty underlying the process.These two attributes are important in an operational context, as scenarios are used for decision making. **In the following, the term ’dispersion’ refers to the spread of the ensemble forecasts.**

Forecasts reliability is assessed using the reliability diagram. **The reliability diagram diagnostic tool compares the observed coverage frequency (effective, $1-\hat{\alpha}$) with the corresponding theoretical confidence levels (nominal, $1-\alpha$) of predictive confidence intervals calculated from ensemble forecasts. Of course, if forecasts are reliable, these values $1-\hat{\alpha}$ and $1-\alpha$ should be equal for any confidence level.** ~~Confidence intervals computed from reliable forecasts should be in agreement with their definition: the 95% confidence interval, for instance, must include on average 95 observations out of 100. For each nominal confidence level probability from 0.1 to 0.9, the effective frequency of the observation occurrence in the given nominal interval is calculated. Then, the effective frequencies are plotted against the nominal confidence level probability.~~ Moreover, the probability integral transform (PIT) histogram, which has the same interpretation as the rank histogram described in Hamill (2001) is also used to detect bias and dispersion issues in forecasts. ~~PIT histograms are preferred over rank histograms herein because of the changing number of members (see section 3).~~

Scoring rules address reliability and sharpness simultaneously. One of the most well-known probabilistic scoring rules, the Continuous Ranked Probability Score (CRPS, Matheson and Winkler, 1976) is used to assess the overall accuracy of competing forecasting systems. The mathematical expression ~~of~~**for** the CRPS is given by equation 1.

$$CRPS(p(x),y) = \int (p(x) - H(x<y))^2 dx \tag{1}$$

where $p(x)$ represents the cumulative predictive distribution of the forecast and y is the observation. H is the step function, which equals 0 when $x<y$ and 1 when $x>y$.

The CRPS skill score (CRPSS) expression is **expressed as:**~~presented in equation 2~~

$$CRPSS = 1 - \frac{CRPS_{for}}{CRPS_{ref}} \tag{2}$$

where $CRPS_{for}$ is the mean CRPS of the forecasting system and $CRPS_{ref}$ is the mean CRPS of the reference system (benchmark).

In order to evaluate the potential of ensemble meteorological forecasts, simulated streamflows were used instead of observations in the verification process. Proceeding in this way eliminates concerns about model and parametrization errors, which vary with watersheds and periods of the year. Moreover, after a spin-up period, the initial conditions are not necessarily estimated correctly by the hydrological model. Operationally, this is corrected by the forecaster, manually or by an automated data assimilation procedure, so that the simulation matches the observations closely. Since data assimilation falls outside the scope of the present study, using simulated streamflow as a benchmark eliminates this concern. Furthermore, as it is frequently the case for hydro-power reservoirs, observations are not really obtained from gauging stations but rather estimated by a water balance equation applied on each reservoir. The quality of observations also varies across watersheds. Simulated streamflow series form a complete dataset with no missing data but are subject to errors attributable in a large portion to the hydrological model itself. Consequently, the results presented in section 5 should be interpreted as the potential skill (not the operational one) of meteorological forcings, as if the hydrological model was able to reproduce the watershed's behavior perfectly.

In the following, **raw and bias-corrected meteorological forecasts will refer to the ensemble meteorological forecasts of the ECMWF System4.** ~~h~~**H**ydrological forecasts based on bias-corrected forecasts from System4 will be referred to as corr-DSP. Hydrological forecasts based on climatology will be referred to as "Extended Streamflow Predictions", or ESP, and hydrological forecasts based on the simulated streamflow climatology as simulated 'Historical Streamflow Prediction' or 'sim-HSP'. ~~Moreover, the lead-time refers to the time lag between the emission date of the forecast and the time at which the forecast is valid. For instance, a skill score for the January-February-March season for the 5-month lead-time correspond to the performance of the forecasts issued 5 months earlier, in August, September and October.~~**The lead-time is defined herein as the time between the date of emission of the forecast and the end of the validity period of the forecasts. For instance, the 1-month lead-time of the forecast issued on the 1st of January is the monthly volume or temperature of January.**

5 Results

5.1 Bias characterization and correction

Raw forecasts from ECMWF System4 suffer from biases. Bias is calculated as shown in equation 3 for temperature and equation 4 for precipitation.

$$Bias_{Add} = \frac{1}{N}\sum_{k=1}^{N}(\overline{x}(k) - y(k)) \tag{3}$$

$$Bias_{mult} = \frac{\frac{1}{N}\sum_{k=1}^{N}\overline{x}(k)}{\frac{1}{N}\sum_{k=1}^{N}y(k)} \tag{4}$$

where$k$ **is the day index,** $\overline{x}(k)$ is the mean of the ensemble forecast, $y(k)$ the observation and N the number of forecast-observation pairs **considered for the bias estimation.**~~in the verification set~~. **We therefore assume an additive bias**~~Bias is additive~~ for monthly mean minimal and maximal temperature: it is the mean error. For monthly precipitation, bias is **defined as** the ratio of the forecasts mean to the mean observed accumulation. **A multiplicative bias is then assumed for this meteorological variable.**

Figure 2 shows the bias for monthly forecasts by season, lead-time and watersheds.

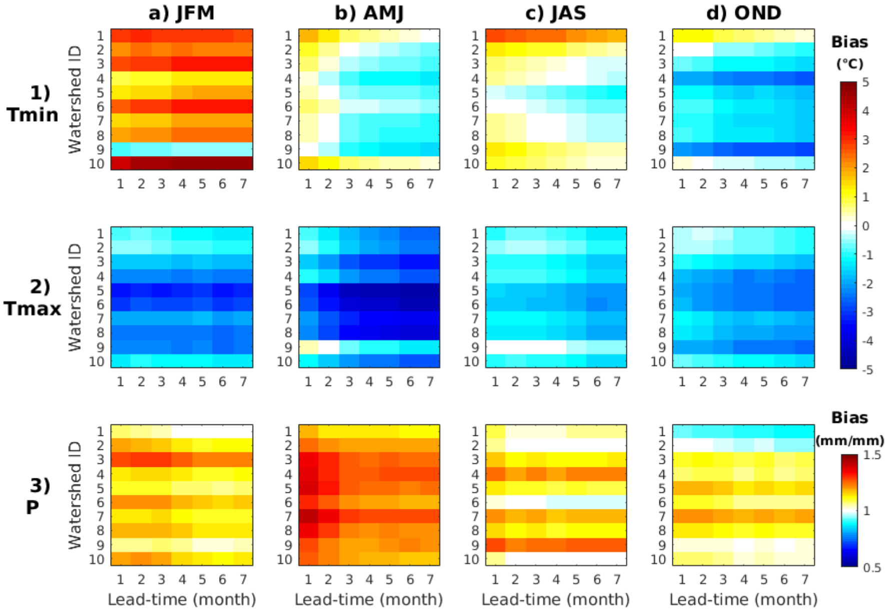

Figure 2. Bias for forecasts of 1) monthly mean minimal temperature, 2) monthly mean maximal temperature and 3) monthly accumulated precipitation for the 10 watersheds under study as a function of seasons and lead-times.

Monthly mean maximal temperature forecasts exhibit a cold bias for all watersheds, seasons and lead-times. This cold bias increases with the lead-time. Bias for the mean minimal temperatures changes depending on the season. For almost all watersheds, raw forecasts display a warm bias during winter and a cold bias during fall. During spring and summer, temperature forecasts are almost unbiased. Monthly accumulated precipitation are overestimated most of the time, especially during spring.

As shown in Figure 2, raw forecasts clearly need to be bias corrected. As mentioned above, Crochemore et al. (2016) have shown that the simple linear scaling method provides results comparable to the more complex distribution mapping **to correct the bias in precipitation ensemble forecasts**. ~~Hence,~~ ~~i~~In this study, daily precipitation forecasts **as well as minimal and**

maximal temperatures are corrected by linear scaling based on monthly bias. Bias is estimated separately for each lead-time and month of the year. A leave-one-year-out procedure is used **to calculate bias and correct the forecasts.**~~, which consists in excluding the forecast to correct from the bias evaluation process.~~ **This consists in calculating the bias based on available forecasts issued on the same month, excluding the month under correction. For instance,** ~~F~~for a given forecast, all other forecasts issued on the same day of the year are used to quantify the bias, calculating the mean of the errors between the ensemble mean and the observation. Multiplicative bias for precipitation and additive bias for temperature are calculated for each monthly forecast. The computed bias is finally used to correct the original daily forecasts.

Figure 3 presents the CRPS skill score (CRPSS) of bias corrected forecasts. ~~with~~ The raw ensemble forecasts is taken as the reference (see equation 2). A CRPSS above 0 (from yellow to red) indicates that bias correction improves the original forecasts, whereas a CRPSS below 0 (from light blue to dark blue) indicates a deterioration of the forecasts.

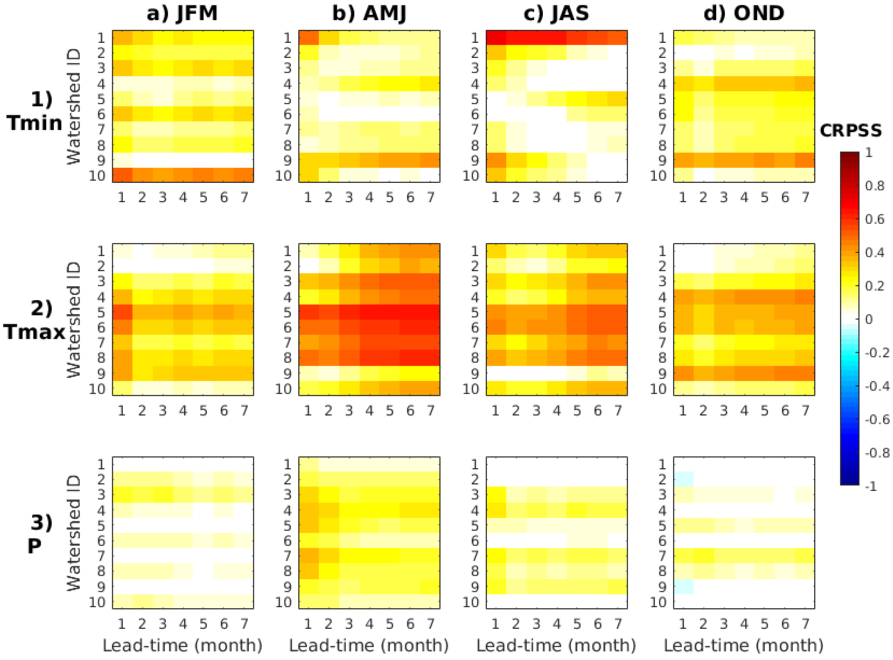

Figure 3. CRPSS of bias corrected System4 forecasts compared to raw ensemble forecasts of 1) monthly mean minimum temperature, 2) monthly mean maximum temperature, 3) monthly cumulative precipitation by watersheds, seasons and lead-times.

Bias correction is found effective, as it does improve meteorological forecasts according to the CRPSS. The effect of linear scaling for monthly ~~volume~~ forecasts is not homogeneous throughout the year and it also varies among watersheds. Bias correction is particularly efficient for periods with substantial bias, such as monthly aggregated precipitation during spring. Bias correction of monthly maximum temperature is also efficient for all months, watersheds and lead-times. However, for precipitation during winter and fall, bias correction does not improve the CRPS noticeably. This is likely because bias during those months are generally small.

5.2 Performance of ensemble forecasts

5.2.1 Bias-corrected meteorological ensemble forecasts against climatology

Figure 4 presents the CRPSS of bias-corrected ensemble forecasts with climatology taken as the reference.

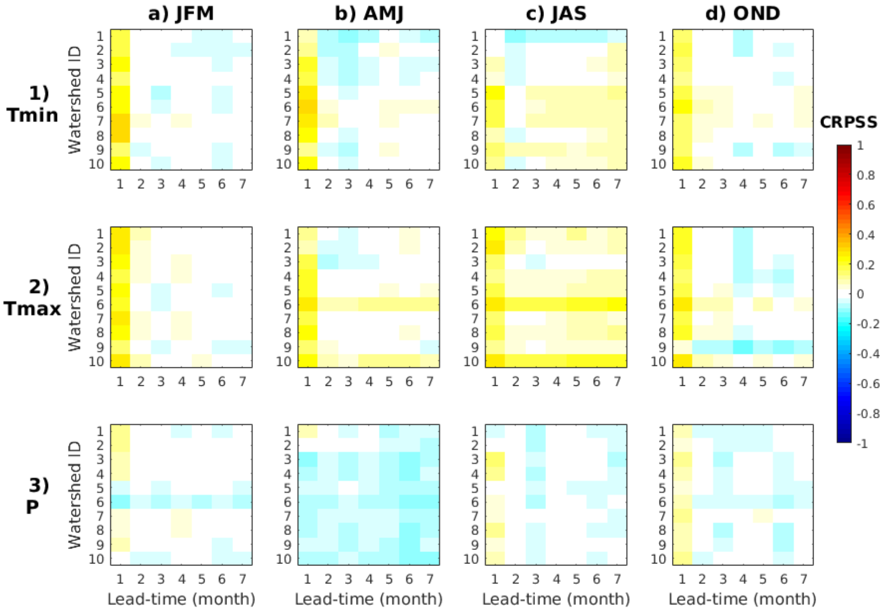

Figure 4. CRPSS of bias corrected System4 forecasts compared to climatology for 1) monthly mean minimum temperature, 2) monthly mean maximum temperature, 3) monthly cumulative precipitation by watersheds, seasons and lead-times.

For almost all watersheds and seasons, both mean minimum and maximum temperatures outperform climatology for the 1-month lead-time. However, for longer lead-times, only temperature forecasts during summer can provide a little improvement over climatology.

Precipitation is known to be less predictable than temperature and the CRPSS confirms this insight. For the 1-month lead-time, CRPSS results are mixed. Ensemble forecasts do have some skill for certain watersheds during winter, summer and fall (for instance watersheds number 3 and 4), whereas CRPSS indicate that climatology is more skillfull during spring. For longer lead-times, according to the CRPSS, climatology always outperforms forecasts.

Hydrological conditions depend mostly on precipitation. However, in a northern environment, temperatures are also important, especially during winter and spring. In fact, during these periods, temperature defines the type of precipitation (snow or rain). It also drives snow pack ~~maturation~~ **evolution** and the characteristics of the spring freshet (early/late and fast/slow). During summer, temperature controls evapotranspiration. In these conditions, it is valuable to assess the performance of hydrological forecasts produced by both systems: corr-DSP (based on bias-corrected System4 forecasts) and ESP (based on climatology).

5.2.2 Monthly inflow volume forecasts

Figure 5 shows the performance of **corr-DSP** ~~ensemble forecasts~~ for inflow volumes compared to sim-HSP (simulated streamflow climatology) and ESP. Their performance is assessed using the CRPSS.

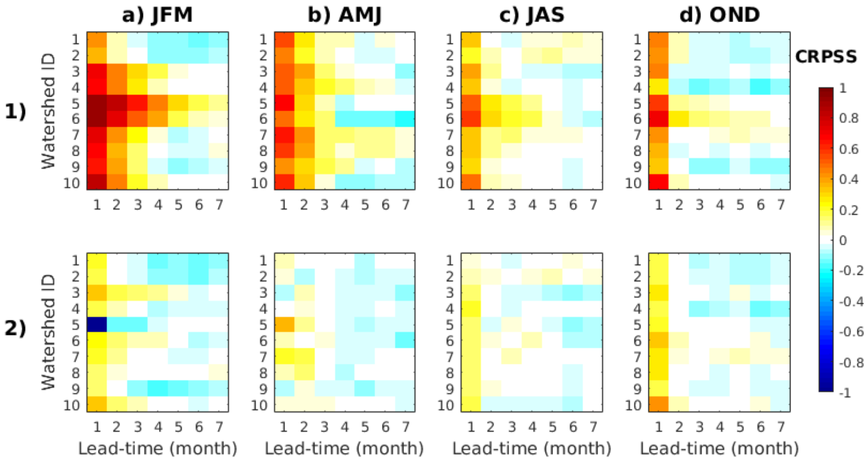

Figure 5. CRPSS of ensemble forecasts of monthly inflow volume to reservoirs, produced by corr-DSP compared to -1) sim-HSP and 2) ESP. CRPSS are shown by watersheds, seasons and lead-times.

The first row of Figure 5 shows the improvement of corr-DSP over sim-HSP. Those results reflect the gain in performance that could be achieved by considering initial conditions as well as information about bias-corrected meteorological forecasts from System4. For all seasons and watersheds, it is valuable to use meteorological ensemble forecasts to produce monthly inflow volume forecasts instead of using simulated climatology (sim-HSP). The critical lead-time, namely the lead-time beyond which sim-HSP performs better than corr-DSP, depends on the period of the year. More specifically, inflow volume forecasts for summer and fall do not show much skill beyond the 1-month lead-time. However, forecasts for winter and spring can be predicted fairly accurately several months ahead when using System4 rather than sim-HSP.

When comparing corr-DSP with ESP (second row of Figure 5), the CRPSS reflects the advantage of integrating meteorological information from ensemble forecasts into the hydrological model. The benefit is clear for the 1-month lead-time, except during spring.The reasons why the CRPSS is not as good during spring as for other seasons could include a change in the influence of initial conditions during the different seasons and the lack of skill of precipitation forecasts. Indeed, streamflow is more variable during spring than during the rest of the year.

~~In general, corr-DSP outperforms ESP for the 1-month lead-time for watersheds number 5 and 7.~~ **In general, corr-DSP outperforms ESP for the 1-month lead-time, with some exceptions such as watershed number 5 in winter or watersheds number 3 and 9 during spring.** Predicting monthly volume during summer and fall more than one month in advance is difficult and both ESP and corr-DSP exhibit comparable skill. Finally, for some watersheds during the winter months, corr-DSP improves the predictability of monthly volumes compared to ESP. Watersheds 3, 5 and 7 reflect different CRPSS behaviors, especially for winter and spring months. Thus, for the remaining of the analysis, special attention is given to those three specific watersheds.

Figure 6 presents the reliability diagrams of corr-DSP for three specific watersheds (number 3, 5 and 7, see Figure 1) by seasons and lead-times. The same diagrams were plotted for all ten watersheds, but results are shown only for those three because, as mentioned above, they reflect specific behaviours worthy of investigation. Furthermore, in retrospect it was found that reliability diagrams for the other watersheds displayed characteristics quite similar to those that are presented. Reliability plots were also obtained for ~~S~~sim-HSP and ESP but they are, by definition, reliable. Hence, those plots are not shown.

As can be seen on Figure 6, the reliability of corr-DSP monthly inflow volume forecasts changes over seasons. The forecasts produced for the fall season are the most reliable, as the effective probabilities computed from forecasts are close to nominal probabilities for all lead-times. Forecasts do not display strong dispersion issues. However, small differences between the ~~effective probability~~**observed coverage frequency** and the ~~nominal probability~~**theoretical confidence level** remain. Indeed, even if a bias correction was applied to raw meteorological forecasts, some biases can still remain and propagate to hydrological forecasts. In addition, bias correction affects the dispersion of precipitation forecasts, and in turn the dispersion of hydrological forecasts. Figure 7 shows the PIT histograms for the three watersheds for the 1-month lead-time. **A flat PIT histogram corresponds to an accurate forecasting system whereas a higher effective frequency on one side of the histogram indicates the presence of bias (asymetric shape). Higher effective frequency in the middle of the PIT histogram is linked with too much dispersion of the ensemble (bell-shape) and, on the contrary, higher effective frequencies on both sides of the histogram is the sign of an under-dispersive behavior (U-shape, the spread of the ensemble is too small for most forecasts).** As mentioned above, some bias are visible in some cases, such as for watershed number 3 during the winter (1a) or underdispersive behavior during the spring (1b).

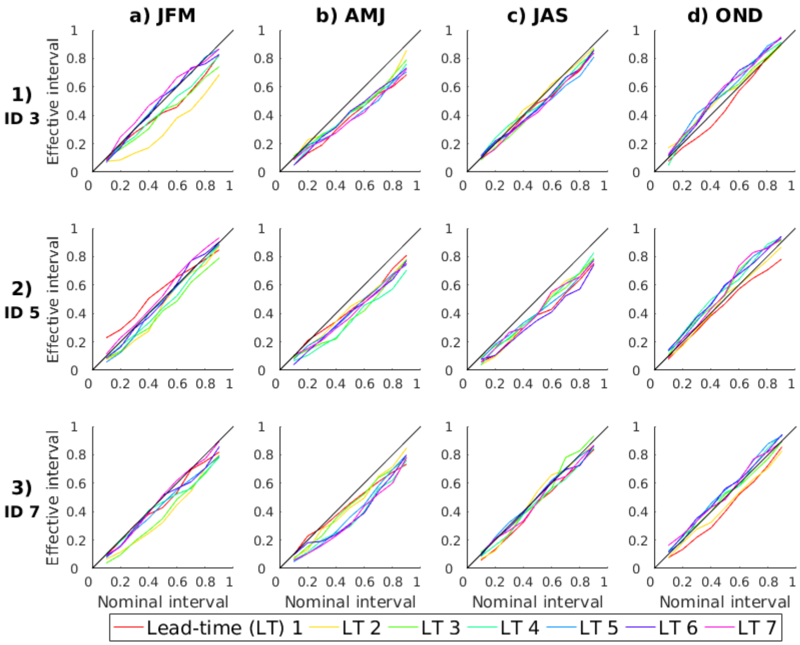

Figure 6. ~~PIT histograms~~**Reliability diagrams** for 1) watershed number 3, 2) watershed number 5 and 3) watershed number 7.

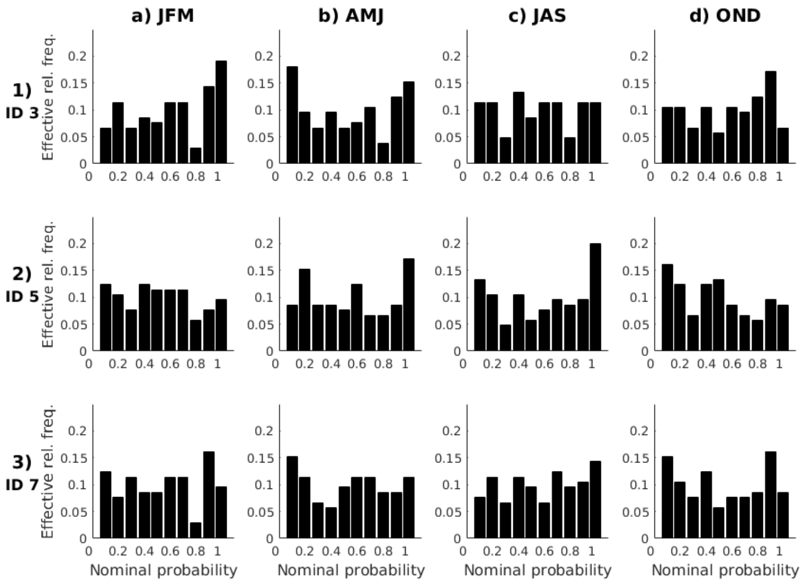

Figure 7. ~~Rank histograms~~**PIT histograms** for 1) watershed number 3, 2) watershed number 5 and 3) watershed number 7.

5.2.3 The case of inflow volume forecasting during spring freshet: an example for watersheds number 3, 5 and 7

Anticipating the inflow volume to the reservoirs is crucial for winter dam management. During the winter, reservoirs are partly emptied to ensure storage space for the inflow volume that is expected during the spring freshet. Consequently, good forecasts of inflow volumes are valuable. The spring freshet volume is computed by cumulating daily forecasted volumes over three months periods. The specific time period associated to the spring freshet varies from one watershed to another, mainly because of geographical location. Watersheds have been clustered into two groups. Watersheds 1 and 2, located in the south, have earlier spring freshet. For those two watersheds, the spring freshet period is defined from March 1st to the end of May. For the other watersheds, the spring freshet occurs between the April 1st and the end of June.

Figures 8, 9 and 10 present **the boxplots of ensemble** forecasts for the spring freshet volume from (a) ESP and (b) ~~and~~ corr-DSP for three watersheds (number 3, 5 and 7). **In those figures, the 3-month lead-time forecast corresponds to the forecasts issued at the beginning of the spring freshet period with a 3-month validity period. The 4-month and the 5-month lead-time forecasts are issued respectively 1 and 2 months before the beginning of the spring freshet period, with corresponding 4- and 5-month validity period.** For ESP and corr-DSP, the dispersion of forecasts increase with the lead-time. For those three watersheds, ESP exhibit a larger dispersion at all lead-times. Extreme meteorological scenarios from past years lead to possible extreme hydrological scenarios.In some cases, such as the spring of 1993 for watershed number 3 at lead-time 1, corr-DSP provided very poor forecasts that missed the spring freshet almost entirely, whereas the ESP is much more successful (the observed volume is included in the boxplot). This issue could be explained from a bias-correction problem. Indeed, even if bias-correction was applied, some biases in corr-DSP might remain~~s~~ and further propagate to hydrological forecasts. It is also possible that the bias-correction method performs better for some years, lead-times and watersheds than others. **Moreover, single events at time scales smaller than one month might be subject to biases different than the monthly values used for bias correction.** Similar figures were obtained for the other watersheds and the general conclusions for those figures are that corr-DSP exhibit a lower dispersion than ESP, leading observations to fall outside the boundary of the predictive distribution too often. **As precipitation is most of the time overpredicted, as shown in Figure 2, bias corrected precipitation forecasts exhibit a lower dispersion than raw forecasts. This can explain the smaller dispersion of the volume forecasts.**

Figure 11 presents the boxplot**s** of the CRPS for the 35 spring freshet events between 1981 and 2015, for all watersheds and the three forecasting systems: sim-HSP, ESP and corr-DSP. In addition to the forecasts issued the first day of the spring freshet period (lead-time 0-month), three lead-times are considered: 3, 2, and 1 month(s) before the first day of the spring freshet. In most cases, the CRPS for corr-DSP displays a higher variability than the CRPS for ESP. Depending of the year, corr-DSP can have a better or worse performance than ESP and even sim-HSP. Hence, resorting to sim-HSP, the simplest "forecasting" system of all three (arguably a forecasting system at all) can be advantageous for some watersheds for which the predictibility is low. For instance, for watershed 6 at lead-times 2 and 3 months, sim-HSP outperform all other techniques. However, for lead-times 1 and 0, ESP and corr-DSP have lower CRPS in the majority of cases. Very high (poor) CRPS can result from both dispersion and bias issues.

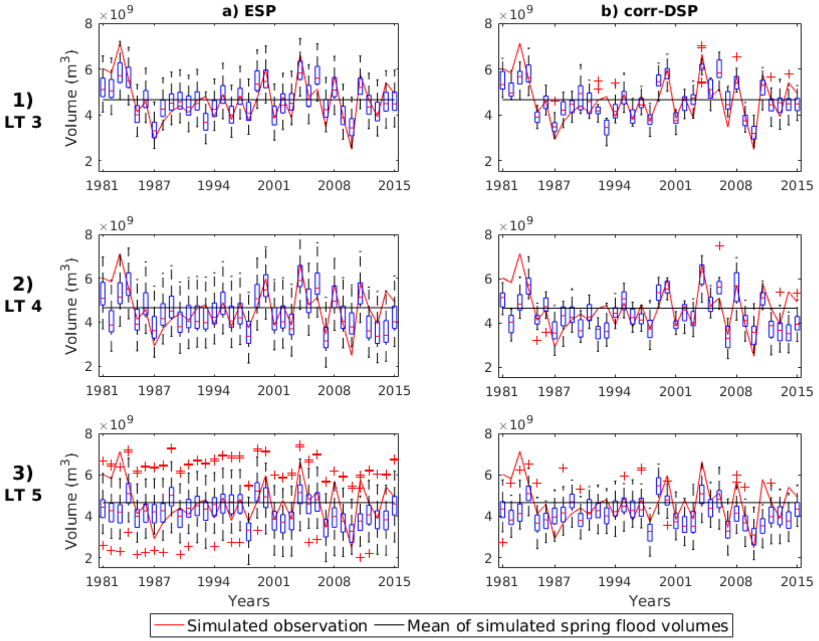

Figure 8. Spring freshet volume forecasts at 1) ~~1~~**3**-month lead-time, 2) ~~2~~**4**-month lead-time and 3) ~~3~~**5**-month lead-time for watershed number 3. The boxplots represent the ensemble forecasts for one given spring freshet and the red line the corresponding simulated observation.

6 Conclusion

The objective of this study was to compare the performance and the behavior of three hydrological forecasting systems for 10 watersheds in a northern climate (Québec, Canada). The three forecasting systems consist of HSP (streamflow climatology), ESP (forecasts based on meteorological climatology) and DSP (forecasts based on ensemble meteorological forecasts from ECMWF System4). Streamflow simulations were used to build the streamflow climatology of each watershed. Simulated streamflows were also used as pseudo observations in order to avoid considering hydrological model errors in the analys~~e~~**is**.

In the context of this study, it was found that ensemble meteorological forecasts from System4 suffer from biases (see Figure 2). However, a bias correction, performed using the linear scaling method, results in an improved performance of ensemble meteorological forecasts, as assessed by the CRPS (see Section 4). Monthly mean minimal and maximal temperature forecasts outperform climatology for the 1-month lead-time. The predictability extends to several months in specific cases (watersheds and seasons). Monthly accumulated precipitation ~~are~~**is** less predictable. In fact, ensemble meteorological forecasts do not have ~~significant~~ skill when it comes to forecasting monthly precipitation during spring. For other seasons, they slightly outperform climatology for the 1-month lead-time.

Still, according to the CRPS, bias-corrected ensemble meteorological forecasts were found to be a useful source of information to improve monthly volume forecasts (see Figure 5), especially for the 1-month lead-time. This is likely due to temperature forecasts more than precipitation forecasts, as mentioned above. Regarding the particular case of forecasts for summer and fall,

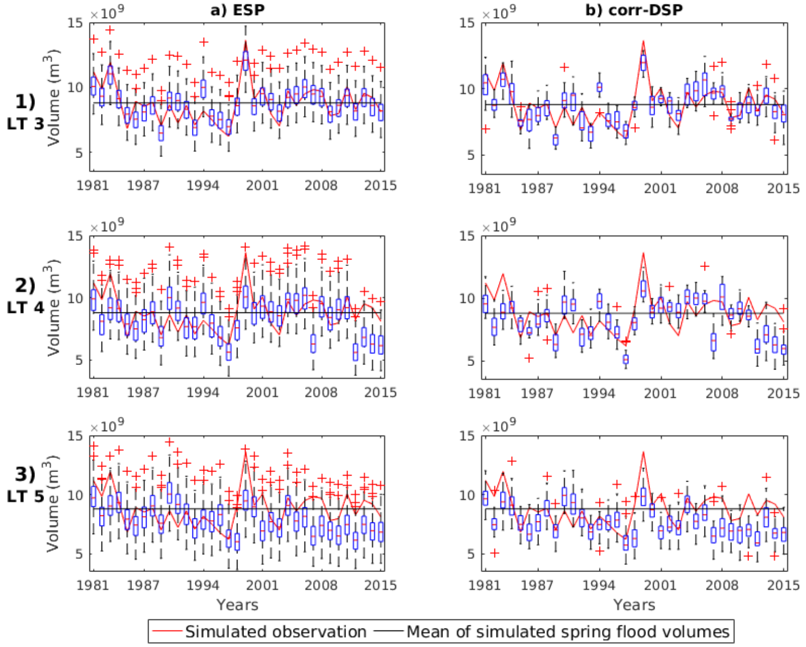

Figure 9. Spring freshet volume forecasts at 1) ~~1~~**3**-month lead-time, 2) ~~2~~**4**-month lead-time and 3) ~~3~~**5**-month lead-time for watershed number 5. The boxplots represent the ensemble forecasts for one given spring freshet and the red line the corresponding simulated observation.

the CRPS of corr-DSP outperforms the CRPS of both sim-HSP and ESP for the 1-month lead-time. Predictability of monthly **flow** volume for winter and spring months extends up to 3 months against simulated climatology. The CRPSS~~S~~ between corr-DSP and ESP is lower than the CRPSS between corr-DSP and sim-HSP. However, corr-DSP show some skill from 1-month lead-time up to 2 or 3-month lead-times for some watersheds (number 3, 7 and 10) during the winter. Monthly forecasts based on System4 are less reliable than ESP, and this possibly originate**s** from bias propagation or dispersion issues **, see Figure 7 for an example, in both meteorological and hydrological forecasts**.~~-~~ **The lack of skill of corr-DSP can originate from different sources. First, linear scaling is a rather simple bias correction method. It was performed using monthly bias and thus, there is a possibility that biases at smaller temporal scales can remain. Second, as precipitation was originally over-predicted in most cases by System4 (see Figure 2), bias correction results in a reduction of the ensemble spread for precipitation forecasts, and possibly for streamflow forecasts also.** Results for spring are mixed: the forecasting performance during spring freshet varies from one watershed to another. In general, the CRPS of corr-DSP is more variable than the CRPS of ESP.

Furthermore, skill scores are subject to sample uncertainty. In this study, each skill score computation is based on almost one hundred forecasts. The size of the verification set thus remains a limit for assessing the significance of the verification results. **A bootstrap procedure was performed to assess the significance of the difference in CRPS between corr-DSP and ESP. Results show that in most cases, results are not proven to be significant and confirm the previous insight.** For the 1-

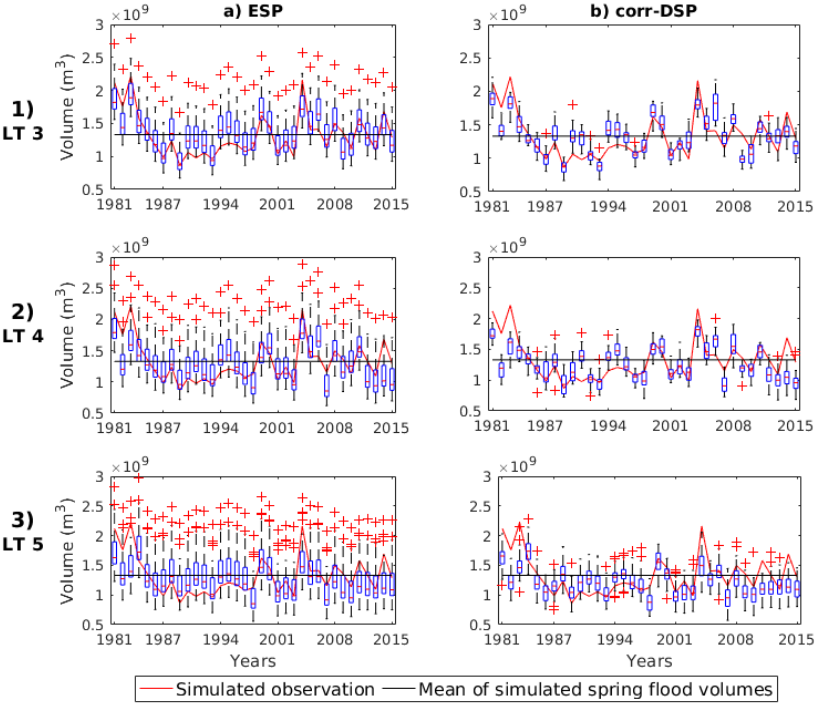

Figure 10. Spring freshet volume forecasts at 1) ~~1~~**3**-month lead-time, 2) ~~2~~**4**-month lead-time and 3) ~~3~~**5**-month lead-time for watershed number 7.The boxplots represent the ensemble forecasts for one given spring freshet and the red line the corresponding simulated observation.

month lead-time, the quality of corr-DSP is clearer and for longer lead-times, the CRPSS of corr-DSP compared to ESP tends to 0. Consequently, in this study and for long lead-times, it is not clear that bias-corrected seasonal meteorological ensemble forecasts from a dynamical model (System4) can completely replace ESP. However, they provide substantial complementary information to produce long-term hydrological forecasts, as shown by the special case of forecasting the inflow volume associated to the spring freshet (see Section 5.2.3). An analysis on the economic value of forecasts for hydro-power production, for instance using a reservoir operation model based on stochastic dynamic programming, would be the next logical step. It would allow to determine whether or not the differences observed here between the three concurrent forecasting systems are indeed ~~significant~~**notable** for water management.

Finally, ECMWF's System4 does not include any sea-ice model (ECMWF, 2017). This could limit to improve weather predictability in mid-latitudes. Other providers of ensemble meteorological forecasts exist and a multi-model approach could improve the skill of the seasonal forecasts. Moreover, according to our results, **it is for** the 1-month lead-time ~~is were~~**that** the most ~~significant~~**important** gain could be achieved by dynamical models over climatology (ESP). For that purpose, the newly available Sub-seasonal to Seasonal (S2S) database (Vitart et al., 2017), that gathers ensemble forecasts from different agencies at the sub-seasonal scale (from 1 up to 60 days) could be explored. In fact, the forecasts available in S2S are especially tailored for the 1 to 2-months lead time, and hence could have superior skill for hydrological applications than System4 which, as

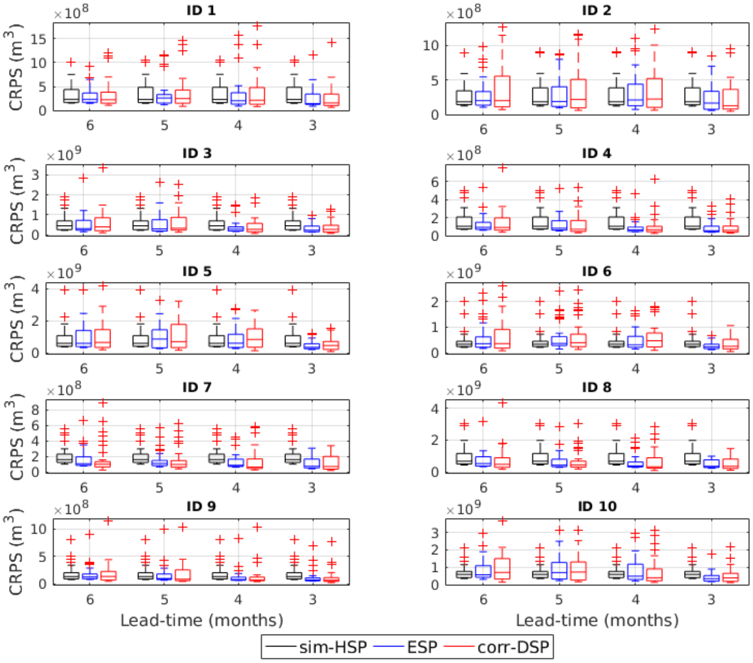

Figure 11. Boxplots of the CRPS for the 35 spring freshet events between 1981 and 2015 for all watersheds and the three forecasting systems: sim-HSP (black), ESP (blue) and corr-DSP (red).

shown here, can often lead to better 1-month ahead streamflow and volume forecasts than ESP. This new database would also enable future studies to explore multi-model forecasting approaches at long lead-times and assess the ability of such approach to extend the limits of predictability. Finally, as proposed by Yuan et al. (2014), forecasts for different lead-time would need to be efficiently joined together in a seamless way, and there is also much to explore in this regard.

5 7 Data availability

Unfortunately, the data used in this study is not publicly available. Data from the ECMWF System4 forecasts are produced by the ECMWF but are not included in the repositories of public datasets. They were provided fo the purpose of this study by Dr. Florian Pappenberger (Florian.Pappenberger@ecmwf.int). Meteorological and streamflow data as well as the watersheds delineation file and the hydrological model are the property of Hydro Québec.

10 *Author contributions.* Rachel Bazile performed all the computation and prepared all the figures. She wrote the most of the manuscript. Marie-Amélie Boucher guided the work, providing opinions about the presentation of results and analysis to be done. She also provided codes that were used in the computations. She helped writing the manuscript and revised all versions. Luc Perreault provided codes and

further guidance for analyzing the results and preparing the manuscript. Robert Leconte initiated the work, reviewed the manuscript and participated in analyzing the results.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. **The authors want to thank Kristian Förster and two anonymous reviewers for their useful comments on this article.** This work was funded by a NSERC Cooperative Research and Development grant to Robert Leconte. Rachel Bazile gratefully acknowledges a scholarship from the Fonds de Recherche du Québec Nature et Technologies. The authors wish to thank Florian Pappenberger for providing the forecasts from ECMWF System4 as well as Catherine Guay for her support with HSAMI. The team of operational forecasters at Hydro-Québec, are warmly acknowledged for their insights, especially Fabian-Tito Arandia Martinez, Éric Crobeddu and Marie-Claude Simard.

References

Bröcker, J. and Smith, L. A.: From ensemble forecasts to predictive distribution functions, Tellus A, 60, 663–678, 2008.

Cloke, H. and Pappenberger, F.: Ensemble flood forecasting: a review, Journal of Hydrology, 375, 613–626, 2009.

Crochemore, L., Ramos, M.-H., and Pappenberger, F.: Bias correcting precipitation forecasts to improve the skill of seasonal streamflow forecasts, Hydrology and Earth System Sciences, 20, 3601–3618, 2016.

Day, G. N.: Extended streamflow forecasting using NWSRFS, Journal of Water Resources Planning and Management, 111, 157–170, 1985.

DelSole, T.: Predictability and information theory. Part I: Measures of predictability, Journal of the atmospheric sciences, 61, 2425–2440, 2004.

Demargne, J., Wu, L., Regonda, S. K., Brown, J. D., Lee, H., He, M., Seo, D.-J., Hartman, R., Herr, H. D., Fresch, M., et al.: The science of NOAA's operational hydrologic ensemble forecast service, Bulletin of the American Meteorological Society, 95, 79–98, 2014.

ECMWF: https://www.ecmwf.int/en/forecasts/documentation-and-support/long-range/seasonal-forecast-documentation/user-guide/seasonal-forecasting-system, 2017.

Fortin, V.: Le modèle météo-apport HSAMI: historique, théorie et application, Institut de recherche d'Hydro-Québec, Varennes, 2000.

García-Morales, M. B. and Dubus, L.: Forecasting precipitation for hydroelectric power management: how to exploit GCM's seasonal ensemble forecasts, International Journal of Climatology, 27, 1691–1705, 2007.

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

Guay, C., Minville, M., and Braun, M.: A global portrait of hydrological changes at the 2050 horizon for the province of Québec, Canadian Water Resources Journal/Revue canadienne des ressources hydriques, 40, 285–302, 2015.

Hamill, T. M.: Interpretation of rank histograms for verifying ensemble forecasts, Monthly Weather Review, 129, 550–560, 2001.

He, M., Whitin, B., Hartman, R., Henkel, A., Fickenschers, P., Staggs, S., Morin, A., Imgarten, M., Haynes, A., and Russo, M.: Verification of Ensemble Water Supply Forecasts for Sierra Nevada Watersheds, Hydrology, 3, 35, 2016.

Hydro-Québec: Rapport annuel, http://www.hydroquebec.com/publications/fr/documents-entreprise/rapport-annuel.html, 2015.

Kim, H.-M., Webster, P. J., and Curry, J. A.: Seasonal prediction skill of ECMWF System 4 and NCEP CFSv2 retrospective forecast for the Northern Hemisphere Winter, Climate Dynamics, 39, 2957–2973, 2012.

Lepage, M.-P. and Bourgeois, G.: Le réseau québécois de stations météorologiques, p. 15, 2011.

Luo, L. and Wood, E. F.: Use of Bayesian merging techniques in a multimodel seasonal hydrologic ensemble prediction system for the eastern United States, Journal of Hydrometeorology, 9, 866–884, 2008.

Marty, R., Zin, I., Obled, C., Bontron, G., and Djerboua, A.: Toward real-time daily PQPF by an analog sorting approach: application to flash-flood catchments, Journal of Applied Meteorology and Climatology, 51, 505–520, 2012.

Matheson, J. E. and Winkler, R. L.: Scoring rules for continuous probability distributions, Management science, 22, 1087–1096, 1976.

Molteni, F., Stockdale, T., Balmaseda, M., Balsamo, G., Buizza, R., Ferranti, L., Magnusson, L., Mogensen, K., Palmer, T., and Vitart, F.: The new ECMWF seasonal forecast system (System 4), European Centre for Medium-Range Weather Forecasts, 2011.

Moradkhani, H. and Meier, M.: Long-lead water supply forecast using large-scale climate predictors and independent component analysis, Journal of Hydrologic Engineering, 15, 744–762, 2010.

Ouranos: Vers l'adaptation. Synthése des connaissances sur les changements climatiques au Québec. Partie 1 : Évolution climatique au Québec., p. 114, 2015.

Palmer, T., Doblas-Reyes, F., Hagedorn, R., Alessandri, A., Gualdi, S., Andersen, U., Feddersen, H., Cantelaube, P., Terres, J., Davey, M., et al.: Development of a European multimodel ensemble system for seasonal-to-interannual prediction (DEMETER), Bulletin of the American Meteorological Society, 85, 853–872, 2004.

Shukla, S., Sheffield, J., Wood, E. F., and Lettenmaier, D. P.: On the sources of global land surface hydrologic predictability, Hydrology and Earth System Sciences, 17, 2781–2796, 2013.

Singh, S. K.: Long-term Streamflow Forecasting Based on Ensemble Streamflow Prediction Technique: A Case Study in New Zealand, Water Resources Management, 30, 2295–2309, 2016.

Sveinsson, O. G., Lall, U., Fortin, V., Perrault, L., Gaudet, J., Zebiak, S., and Kushnir, Y.: Forecasting spring reservoir inflows in Churchill Falls basin in Quebec, Canada, Journal of Hydrologic Engineering, 13, 426–437, 2008.

Svensson, C.: Seasonal river flow forecasts for the United Kingdom using persistence and historical analogues, Hydrological Sciences Journal, 61, 19–35, 2016.

Vitart, F., Ardilouze, C., Bonet, A., Brookshaw, A., Chen, M., Codorean, C., Déqué, M., Ferranti, L., Fucile, E., Fuentes, M., Hendon, H., Hodgson, J., Kang, H.-S., Kumar, A., Lin, H., Liu, G., Liu, X., Malguzzi, P., Mallas, I., Manoussakis, M., Mastrangelo, D., MacLachlan, C., McLean, P., Minami, A., Mladek, R., Nakazawa, T., Najm, S., Nie, Y., Rixen, M., Robertson, A. W., Ruti, P., Sun, C., Takaya, Y., Tolstykh, M., Venuti, F., Waliser, D., Woolnough, S., Wu, T., Won, D.-J., Xiao, H., Zaripov, R., and Zhang, L.: The Subseasonal to Seasonal (S2S) Prediction Project Database, Bulletin of the American Meteorological Society, 98, 163–173, doi:10.1175/BAMS-D-16-0017.1, https://doi.org/10.1175/BAMS-D-16-0017.1, 2017.

Wang, Q., Robertson, D., and Chiew, F.: A Bayesian joint probability modeling approach for seasonal forecasting of streamflows at multiple sites, Water Resources Research, 45, 2009.

Weisheimer, A. and Palmer, T.: On the reliability of seasonal climate forecasts, Journal of The Royal Society Interface, 11, 20131 162, 2014.

Wood, A. W. and Lettenmaier, D. P.: An ensemble approach for attribution of hydrologic prediction uncertainty, Geophysical Research Letters, 35, 2008.

Xu, C.-y.: Modelling the effects of climate change on water resources in central Sweden, Water Resources Management, 14, 177–189, 2000.

Yang, L., Tian, F., Sun, Y., Yuan, X., and Hu, H.: Attribution of hydrologic forecast uncertainty within scalable forecast windows, Hydrology and Earth System Sciences, 18, 775, 2014.

Yossef, N. C., Winsemius, H., Weerts, A., Beek, R., and Bierkens, M. F.: Skill of a global seasonal streamflow forecasting system, relative roles of initial conditions and meteorological forcing, Water Resources Research, 49, 4687–4699, 2013.

Yuan, X., Wood, E. F., Roundy, J. K., and Pan, M.: CFSv2-based seasonal hydroclimatic forecasts over the conterminous United States, Journal of Climate, 26, 4828–4847, 2013.

Yuan, X., Wood, E. F., and Liang, M.: Integrating weather and climate prediction: Toward seamless hydrologic forecasting, Geophysical Research Letters, 41, 5891–5896, 2014.

Yuan, X., Wood, E. F., and Ma, Z.: A review on climate-model-based seasonal hydrologic forecasting: physical understanding and system development, Wiley Interdisciplinary Reviews: Water, 2, 523–536, 2015.

Yuan, X., Ma, F., Wang, L., Zheng, Z., Ma, Z., Ye, A., and Peng, S.: An experimental seasonal hydrological forecasting system over the Yellow River basin-Part I: Understanding the role of initial hydrological conditions, Hydrology and Earth System Sciences Discussions, 2016.