We would like to thank the reviewer for their comments on our paper. Please find our answers below:

Reviewer #2:

General Comments I found the paper quite interesting and provides some substantial and important conclusions. Having said this, I think it really needs to be much more specific in the methodology, be clear on the assumptions that need to be and acknowledge a few fundamental issues with taking such an approach.

The model appears to have been calibrated for a reasonable period of time against what appears to be streamflow records. It is not clear where the streamflow records were obtained or where the locations of the gauges are. The authors should comment fifth cal/val statistics are sufficient for the analysis on climate change they propose. Also, the choice of Anuspline Homogenized (what they call CanGRD) data over perhaps other data sets for forcing is not clear.

The streamflow record was taken from Water Survey Canada and the gauges are located at the outlet of each watershed (Figure 1 in the main manuscript). Clarifications for these gauges and a reference to Figure 1 will be added to the manuscript.

As stated at the end of section 2.2, NSE and PBIAS were satisfactory (Moriasi et al., 2007). As shown in our previous study (Champagne et al., 2019), winter streamflow and snow processes were also satisfactorily simulated. A reference to Champagne et al., (2019) will be added in this section 2.2.

CanGRD meteorological dataset was used in a previous study focusing in southern Ontario (Wazneh et al., 2017). This dataset is often referred to as NRCANmet in number of other studies and is the most commonly used gridded climate dataset in Canada (Werner et al., 2019). Justification for the choice of this dataset will be added in the manuscript.

Consider the focus on snowmelt period and snowmelt simulations, the authors never discuss the appropriateness of the model physics for the snowmelt period. Does the PRMS model use an energy budget or temperature index. Is one method more appropriate for snowmlet, particularly in a climate change context, over another ? This should be at least mentioned.

The ability of PRMS to simulate snow processes will be added in the manuscript with a reference to Champagne et al. (2019). The Snowmelt algorithm uses an energy balance approach based on temperature and precipitation data. The advantage of this method is that snowmelt is better conceptualized than a temperature index approach and does not use data projections that may be difficult to obtain (e.g. radiation).

Data used to derive the physiographic information to develop the model is not described, nor are the basins, except for very cursory comments. For example, there are many small control structures in these systems. The reader needs to know that and be made aware that they have or don't have an influence on the calibration or simulations.

The data used to derive physiographic information are High Resolution Digital Elevation Model (HRDEM) and the Canadian Land Cover CIRCA 2000, both furnished by Natural

Resources Canada, and the surficial geology of southern Ontario furnished by The Ontario Ministry of Northern Development, Mines and Forestry. These data sources will be added to the new manuscript.

The control structures were not taken into consideration in the model set up. The authors are aware that these structures can play a role in the modulation of streamflow. However, our study investigates the change in average streamflow, while control dams have greater impact on specific peak flows. The dams have very limited impact on the average streamflow calculated over a 30-year period.

The value of the paper appears to be in the messaging around the ensemble members results. Also, the attribution to synoptic patters provides some very interesting insights and the methodology seems reasonable, but the author would benefit from clearer explanations in sections 4.3. and 4..4. I find this very compelling and interesting, but it seems to get lost because the methodology confounds us in trying to understand what the authors are trying to do. I believe the intent and actual contribution of this work is important and should be published, but substantial clarification and structure to the manuscript is required.

Clarification of the methodology will be addressed according to the comments below.

Section 2.2. - comments Authors should state why they used PRMS instead of other models ? What is because it is computationally efficient ? has it been used by operational agencies in the region ? Some clarification is required. This section should include 2 parts. 1. model geo-fabric setup, including details around DEM and landcover (which ones) and how HRUs and routing is derived 2. forcing variables (what is necessary and how they are derived, where they come from) is not clear

PRMS was used in this study because has been satisfactorily used in other snow dominated regions and was already applied for these same watersheds (Champagne et al., 2019). According to this study, PRMS reconstructed snow processes in Big Creek watershed. A Few sentences explaining the choice of PRMS will be added in section 2.1.

The model setup was done using Arcpy-GSFLOW as described in Gardner et al., (2011). The PRMS modules used in these watersheds have been described in Champagne et al., (2019). These two references will be added to the manuscript. We will also add more information on the datasets used for the setup (described above).

The forcing variables are minimum and maximum temperature and precipitation at 10km spatial resolution using the NRCANmet dataset. A short explanation of the dataset and references will be added to the manuscript.

Authors should describe better how the HRUs are generated. The reviewer presumes that a single dominant land type and soil type is used for each grid cell (as per the model documentation for PRMS). Authors should define how the grid (which are the same as HRUs ?) are defined in this application of the HRU, and specify that each grid is treated as an HRU. PRMS also requires stream networks, sub-basins, lakes to be defined. a few lines around how this was done or perhaps a schematic on how PRMS was implemented here would be worthwhile. Perhaps a figure similar to Figure 4 in the PRMS user manual but for the author's Big Creek application would be useful. It is difficult to get a sense of how the model was setup for this application.

HRUs consisted of surface grid cells of 200 m² for Big Creek and Credit River watersheds and 400 m² for Grand River and Thames River. For each HRU, the percentage of each land use type (bare soil, grass, shrubs, coniferous trees and deciduous trees) and soil type (sand, loam or clay) was calculated by Arcpy-GSFLOW. Some parameters were estimated using these percentages while other PRMS calculations were based on an integer number corresponding to the most dominant land use or soil type. These precisions will be added to the manuscript.

The stream network was computed with ARCGIS using DEM and accumulation threshold was determined empirically to make the conceptual stream network match the stream positions from satellites maps. Since only one hydrometric station was used for each watershed, these watersheds were considered as one sub-basin. HRU's are considered as lakes when the entire area is covered by water as determined by land use.

The description of PRMS setup will be greatly improved in the new manuscript. Given that HRUs are grids of similar sizes, we judged it was not necessary to include a figure similar to figure 4 from the PRMS manual.

The last part of section 2.2. describing the meteorological forcing used is also quite confusing. CanGRD (according the Environment Canada) is a monthly, seasonal and annual product. Perhaps the author is referring to the homogenized data used in the development of of the product produced by McKinley, which based on the article cited which I read, does not have a formal name. Also, there are a lot of other products available, so some justification as to what this product, which is quite a bit older thanks some of the more recent published data such as WATCH or CAPA, is being used. Also, can you clarify which streamflow gauges were used ?

As previously stated, CanGRD referred to as NRCANmet in number of studies and is the most commonly used gridded climate dataset in Canada (Werner et al., 2019). This dataset was produced using station-based observations from Environment Canada and Natural Resources Canada and the gridding was accomplished using the Australian National University Spline (ANUSPLIN) with latitude, longitude and elevation as predictors (Hutchinson et al., 2009). The dataset will be renamed NRCANmet in the entire manuscript.

Lastly, you mention muskingum routing, but it is not completely clear how this was calibrated. This is likely the most sensitive parameter the the NS criteria. Can you confirm how sensitive the results were to the routing ?

The muskingum routing was calibrated by fitting the Muskingum storage coefficient (K_coef) using the Normal Root Mean Squared Error (NRMSE) between daily and monthly observed and simulated streamflow. The inter-segment variability of K_coef was estimated using the segment length and the slope. This variability was preserved during the

calibration by multiplying K_coef of each segment by the same coefficient. The results were sensitive to the routine and especially the timing of the streamflow and the amplitude of high flows.

Section 2.3 - comments. A more complete description of the data developed in CanRCM-LE would be useful. I was required to lookup what this data set contained and how the ensembles were generated. I think the authors should actually include some level of detail here.

Details and references on the development of CRCM5-LE will be added in the manuscript.

Section 2.4 - Comments This section is extremely unclear. I would recommend the authors describe what AHC is and at minimum make some reference to how the various ensembles were classified. What is the purpose of doing the ACH analysis, and is there a reference ?

The AHC will be described in the new manuscript as follows: To classify n members, the Euclidean distance between each pair of members are first calculated. The two members with the smallest Euclidean distance are merged into a single class. Then, the Euclidean distance between this class of members and the n-2 other members is calculated. The two members or classes of members with the smallest Euclidean distance are merged and a new Euclidean distance is calculated. This process is repeated n-1 times in total, until all classes of members have been merged into a single class. The dendrogram in Figure 6 shows the successive merging from the first merging using all members (bottom) to the last merging creating a single class (top) (Figure 6). Each merging is associated with a variance of Euclidean distance between the two successive merged classes (Y axis). The highest variance difference between two successive merging classes show what is the number of classes corresponding to the highest interclasses variance.

Section 3.1 The methodology becomes clearer after reading this section. I would encourage the authors to maybe re-write some of sections 2.2 to clarify the approach. It seems that what was done was 1. Calibrate these basins for use with PRMS using historic homogenized and gridded daily (5 years) data. 2. Using the CRCM-LE historic biased corrected forcing for the simulations and run ensembles. The authors should perhaps take a bit of time to describe why this approach was taken e.g. why not calibrate to a 10 year period. Are there any concerns about perhaps parameters values changing under a different climate regime ? Are you concerned about calibrating with Anuspline but driving the model with a different precipitation model, even if it was bias corrected. Some commentary here is necessary. The authors looked at ET, and I assume it was from the PRMS model. Why not use RCM or at least see what the RCM produces ? Since it is based on CLASS, should dit not be a bit more realistic than PRMS ?

The part 2.2 will be clarified in the new manuscript to better explain the dataset used to calibrate the model and the dataset used for the future projections. As stated in the manuscript, the calibration period was 20 years (1989-2008), not 5. The reviewer refers probably to the warm-up period (1984-1989) or the validation period (2009-2013).

The authors were concerned about calibrating with ANUSPLIN and driving the model with CRCM5-LE. This is why climate observations and historical data from CRCM5-LE

were compared and streamflow computed with ANUSPLIN and CRCM5-LE were compared as well in the historical period. Section 4.1 described the discrepancy between simulations using ANUSPLIN and CRCM5-LE in the historical period.

We appreciate the suggestion to compare ET from PRMS and RCM. However, due to the size of the dataset, the extraction and transfer of the variables are very time consuming and only temperature and precipitation were extracted. We will suggest using ET from CRCM5-LE in future studies in the new version of the manuscript.

The authors show in figure 5 increases in temp and precipitation. Can you clarify if this is the bias corrected values or original CRCM5-LE.

Temperature and precipitation shown in Figure 5 are the bias corrected values. This information will be added into the new manuscript.

Section 3.2 A paragraph describing what ACH with a reference is required either in the methodology or here. Up to this point in the text, it is unclear why the ACH approach is even necessary. It does get clarified, but should be referenced and explained in section 2. The division between hi-lo and moderate and conglomeration of weather and flow classes seems a bit subjective. The authors should be clearer on how they chose to group these. It is not clear how you have a HiT category since P and T are combined. One assumes that the change in P is simply small. Also the whole section is difficult to follow and essentially describes what is in the table and on the plots, but it doesn't really tell me what I think it is trying to tell me. It seems that this is al about attribution of the change in flows. Is it caused by increases in T, P, or both. Section 3.2 does not really assist me in understanding.

AHC will be clarified in section 2.4 as stated above. The AHC was first used to group the members into classes of similar change of streamflow. The AHC uses the Euclidean distance between members and it can be applied simultaneously using different variables (here the variables are the evolution of streamflow for each of the 4 watersheds). The AHC was applied to the standardized change of streamflow to avoid the Euclidean distance being dependent on large changes in one watershed. The AHC constructs these classes by maximizing the interclass variance. Therefore the classes are not arbitrary. The division between Hi, Low and moderate is based on the results of the AHC. Three classes are the most pertinent choice to maximize the interclass variance of streamflow change (Shown in figure 6). The variance between classes is maximal when the vertical distance between 2 successive merging is maximal. The labels Hi indicated the highest increase in streamflow while Lo indicated the lowest increase. Moderate is the class in between. High, Low and moderate are relative to other members and do not refer to an absolute high or low increase in streamflow.

The AHC was also applied to the standardized change of the two variables, temperature and precipitation (Figure 6 diagram at the bottom). The objective was to group the members so we did not have a member isolated in a single class. If we remove member #33 it is clear that the number of classes with the lowest interclass variance is 4.

The conglomeration of streamflow and weather classes is not subjective because it is simply done by splitting the 3 streamflow classes into weather subclasses (e.g. members with simultaneously High Q and High PT in the same class, High Q and moderate PT in another class…etc…). The reviewer is referred to the Table 4 in the main manuscript depicting the streamflow and weather classes labels.

Figure 6 (right plot) shows that the HiT grouped member with high temperature change but not high precipitation change (Orange circle). For concision we decided to call it HiT. The right panel of figure 6 shows that the construction of the weather classes was not subjective but are formed from members that are similar in term of both precipitation and temperature change.

Explanations for the causes of streamflow change was described in the discussion part. Part 3.2 will be modified to avoid repetitions from the graphs.

Section 4.1 The authors never mention issues around frozen soils, freeze that cycles or river ice formation. River ice can have a large influence on hydrometric measurements and rating curves. Often it is too dangerous to take flow measurements in the winter so many flow values are estimated that time of year. The authors need to acknowledge something on uncertainty in winter measurements.

We simulated streamflow during frozen and not frozen soil conditions in Big Creek watershed and the difference was not significant (Figure R2). We used a lag of three days between the conditions of the soil and the streamflow because rain and/or snowmelt events take 3 days to form a peak at the outlet. We also tested lags of 1 to 6 days (6 days given the best correlation between seasonal average temperature/precipitation and seasonal average streamflow) and the results were not significantly different. We can therefore conclude that frozen ground does not have a significant impact on streamflow.

River ice can have an impact in gauge measurements, and this will be acknowledged in the new version of the manuscript.

Section 4.2 and 4.3 The synoptic discussions are interesting but a bit confusing. This really need to be better explained and expanded.

These sections will be rewrite as follow:

4.3 Consistency in the weather classes

The weather classes are associated to specific changes in atmospheric conditions (Figure 9) but are composed from an average of members that have their own signature. Change in Z500 anomalies for each member are depicted in Figure 11 to investigate the variability between members. The members that comprise classes HiPT show high Z500 anomalies enhance in the east coast consistently for six members while for two members (#13 and #48) the high increase in Z500 anomalies is centered north from the Great Lakes. Eight members of the class LoPT show strong Z500 decrease in the east coast but in two members (#1 and #10) the decline is centered in the northern side of the Great Lakes. HiT shows generally

Z500 increase centered on the Great Lakes but four of the thirteen members depict a different pattern (#2, #20, #31 and #47). Finally, members from MoPT show generally a decrease in Z500 but we observe a high diversity in the change of circulation patterns. Members from MoPT depict a lower Z500 gradient compared to other classes suggesting a lower contribution of internal variability of climate to the total change in atmospheric conditions for this class. Despite the atmospheric anomalies differences between members that predict similar local weather conditions, this study gives a good probabilistic overview on how the change in regional atmospheric anomalies will impact local weather.

4.4 Lag between local climate conditions and streamflow

The results of this study show that interclass variability in the increase of streamflow is mostly driven by temperature and precipitation variability in January-February. The members with the highest increase in precipitation and temperature (HiPT) are the members associated with the highest streamflow increases and the members associated with the lowest increase in precipitation and temperature (LoPT) show the lowest streamflow increase (LoQLoPT) (Table 4). Other LoPT members show comparatively higher streamflow (MoQLoPT) but this result can be explained by more precipitation and snowfall despite a lower warming (Figure 7).

Within the other two weather classes, HiT and MoPT, a similar change in January-February weather conditions translates to a large range in streamflow projections (Table 4). These discrepancies between the evolution of weather conditions and streamflow volume in January-February can be associated to a delay between weather conditions change and streamflow change. To account for the routing delay between rain/snowmelt events and streamflow observed at the outlet, our analyses use a lag-time of 6 days between the precipitation/temperatures and the streamflow but a delay between weather conditions and streamflow may occur due to remaining snowpack from December. This hypothesis is invalidated by the low variability among MoPT and HiT members in term of change in late December snowpack (Figure 10). The delay between weather conditions and streamflow can also be due to groundwater recharge/discharge variability. The lower streamflow increase in LoQHiT is associated simultaneously with a lower increase in groundwater flow and a lower increase in November-December precipitation amount (Figure 10). To confirm the connexion between fall precipitation and winter groundwater flow we calculated the inter-members correlation between November-December change in precipitation amount and the January-February change in groundwater flow and we found a coefficient of correlation close to 0.7. This result show that winter streamflow can be modulated by atmospheric conditions occurring as early as the fall season.

Another explanation for the discrepancy between change in weather conditions and change in streamflow may be the timing in precipitation and temperature change. Our results show that snowpack remaining at the end of January-February is decreasing at a higher rate for MoQMoPT members as compared to LoQMoPT members and for MoQHiT members compared to LoQHiT members (Figure 10). This result is likely due to more snowmelt

simulated in the MoQ members (Figure 7). In the same time, the evolution of precipitation and temperature are similar between LoQMoPT and MoQMoPT and between MoQHiT and LoQHiT (Figure 7). These results show that to explain changes in snow processes the average change in precipitation and temperature is not sufficient and their timing must be taken into consideration. An increase in precipitation simultaneously with an increase in temperature is likely to produce rain with a direct impact on streamflow. On the contrary, when increase in temperature is mostly happening in January while February is affected by wetter and cooler conditions, snowpack is likely to contribute to the streamflow later in spring.

These discussions emphasize the need to study the succession of different atmospheric patterns that occur weeks or months before a discharge event.

Specific comments

Page 2- Line 25-30 - Did you mean just limited members from CRCM5-LE or a different ensemble from Seiller and Anctil ? Same for Erler ? It would be useful if you clarified if you are using these new ensembles for the first time or you are the first to use all 50 as other authors had only used select ensemble members grin the same set. This is a bit ambiguous.

These studies used other ensembles that have only 4 or 5 members. CRCM5-LE was not used before in north-eastern North America as input in hydrological models. This sentence will be modified for clarity.

Page 2 Line 30. For readability, it would be useful to add a sentence here as to why using 50 ensemble is important.

We will add a sentence explaining that 50 members are important because it depicts a large range of internal variability of climate and is appropriate for a probabilistic approach.

Page 3: - line 22 should use “computational time” or “model computation time “ instead of model time.

We will change this part to read ‘’reduce the parametrization computation time’’.

Page 3 - Line 25-27 - The authors should expand this to either include the equation or explain this better. The reader who is not completely familiar with PRMS will not understand what the coefficients are used for or what they mean.

This part will be clarified with reference to a previous paper that used PRMS in these watersheds (Champagne et al., 2019). The reader will be referred to Markstorm for details on PRMS that are not fundamental for understanding the manuscript and are common to all watersheds using PRMS.

Page 4 - line 9 - please indicate the time step.

The timestep (daily) will be added.

References:

Champagne, O., Arain, M. A. and Coulibaly, P.: Atmospheric circulation amplifies shift of winter streamflow in Southern Ontario, Journal of Hydrology, 124051, doi:10.1016/j.jhydrol.2019.124051, 2019.

Gardner, A. S., Moholdt, G., Wouters, B., Wolken, G. J., Burgess, D. O., Sharp, M. J., Cogley, J. G., Braun, C. and Labine, C.: Sharply increased mass loss from glaciers and ice caps in the Canadian Arctic Archipelago, Nature, 473(7347), 357–360, doi:10.1038/nature10089, 2011.

Leavesley, G. H., Lichty, R. W., Troutman, B. M. and Saindon, L. G.: Precipitation-runoff modeling system; user's manual., 1983.

Moriasi, D. N., Arnold, J. G., Van Liew, M. W., Bingner, R. L., Harmel, R. D. and Veith, T. L.: Model evaluation guidelines for systematic quantification of accuracy in watershed simulations, Transactions of the ASABE, 50(3), 885–900, 2007.

Newlands, N. K., Davidson, A., Howard, A. and Hill, H.: Validation and inter-comparison of three methodologies for interpolating daily precipitation and temperature across Canada, Environmetrics, 22(2), 205–223, doi:10.1002/env.1044, 2011.

Werner, A. T., Schnorbus, M. A., Shrestha, R. R., Cannon, A. J., Zwiers, F. W., Dayon, G. and Anslow, F.: A long-term, temporally consistent, gridded daily meteorological dataset for northwestern North America, Scientific Data, 6(1), doi:10.1038/sdata.2018.299, 2019.