Dear authors,

Thank you for your responses to the reviewer comments and for the revision of the manuscript. Both reviewers were positive about the manuscript and recommended to publish it after some considerable revisions. The results showing increases in winter discharge in key watersheds within southern Ontario under future climate are of interest, and the study helps improve understanding of the potential hydrological impacts. However, the revised manuscript suffers from a number of issues of clarity and other problems, and will require further revisions to bring it up to the quality standards for this journal. Both reviewers offered constructive and helpful comments on how this paper needs to be improved, but I find these have not yet been fully addressed. A more careful and thorough effort is required.

Major Issues

1. The model must be more fully explained and justified as to its appropriateness for use in this region and under changing climates. Why is it appropriate for use here and what are its main limitations? Simply referring to the fact that others have used it and citing your earlier paper are not sufficient. The snowmelt routine appears to be a simple temperature index approach (i.e. a snow energy balance approach explicitly accounts for turbulent and radiative energy exchanges), and here the only inputs are temperature and precipitation. So why and how can this be justified under future climates?

2. There needs to be more detail and explanation of the model setup, parameterization, and calibration/validation. Both reviewers were adamant about this. More discussion is needed on model and data uncertainties, especially given the bias in simulated winter flows as noted by reviewer #1. The comments by reviewer #2 included a number of important issues to address regarding model setup and geofabric. Why is it ok to neglect control structures and reservoirs, especially on the Grand River, where there are a number of flood control dams? (See specific comments further below.)

3. The ascending hierarchical classification needs to be better described, and perhaps better illustrated, as it remains quite unclear. The reader needs to understand this. Why focus on runoff response groupings, given the non-linear nature of runoff, as opposed to strictly synoptic climatological patterns?

Detailed Comments

I refer to page and line numbers for the “clean” (i.e. non marked-up) version of the ms. My comments here are meant to help specifically address some of the reviewer concerns and to flag other issues that need to be dealt with.

P1, L12 and throughout: what is “internal variability of climate”? This term is central to the paper, but it isn’t made entirely clear what this actual means. Does it refer to variation among the ensemble of climate model outputs?

P1, L18-22: the short summary of results is very unclear. What is meant by the terms in parentheses? The reader can’t understand this by just reading the abstract. How significant is it that 14% of the

ensemble members predict a high increase? What about the rest of the ensemble? More importantly, what is the magnitude and variation of projected flow changes?

P1, L22: what does the 16% refer to? This isn't clear.

P1, L22: what is "internal variability of hydrological projections"?

P1, L30: the "choices" are really cascading sources of uncertainty throughout the modelling process. And does this link in to the concept of internal variability of climate? There is an opportunity to explain all of this more clearly here.

P2, L15: "future climate data should not be used…" – do you mean that coarse-scale climate model outputs shouldn't be used directly?

P2, L29: does a large ensemble assess the entire range of internal variability? Does this not also depend on selection of RCP, GCM, downscaling method, bias correction, hydrological model, parameterization, etc.?

P2, L30: instead of "processes", do you mean "responses"?

Introduction section: In general, this section should be a bit more clear on the overall purpose and objectives of this study, and on how it builds on previous work to advance understanding. What is new and how and why is it important?

P3, L8-9: What about other urban areas such as Kitchener–Waterloo and others?

Section 2.1: This section should describe the major landcover types in more detail, and a bit on the climate and the hydrological regime. For example, there is a lot of deciduous forest cover. How much snow is there and when does it melt? What are the key characteristics of regional climate?

P3, L18: Is this daily forcing data? Please indicate in the ms.

P3, L25: this model is not using a snowmelt energy balance approach, and this needs to be rectified here and also justified.

P3, L30-31: The model is really using a grid, which is different than HRUs, so it should say "coarser grid". Also, what is "parameterization computation time"?

P3, L31 - P4, L1: There is a need for more detail on the model, what processes are represented, how it is run, etc., and then references can be added for further, more specific details.

P4, L4: What is the full reference for the "Natural Resources Canada" data? This is needed.

P4, L11: despite analyzing 30-year average flows, the model simulations and the calibration approach uses daily flows, so the approach to neglect flow controls needs better justification.

Section 2.2: More details are needed on how the how the model was set up and parameterized, following the advice of reviewer #2. It isn't clear how parameters were set, how they varied among basins and HRUs, and how the HRUs were defined. In fact, the approach seems to be more consistent with a grouped response unit approach, where physical landscape groupings and their proportional area are derived from a grid, and parameters are set for the GRUs. Which parameters were important in the calibration and which were the results most sensitive to? Were there ranges that certain parameters were restricted to? Table 2 indicates that 17 parameters were determined by calibration alone, which provides a high potential for model equifinality, and so a more detailed explanation is important.

P4, L5: instead of each HRU, the percentages were determined for each grid cell.

P4, L27: Table 3 provides NSE and PBIAS info, not a set of model parameters.

P4, L28: the range is less than -15 to 15%, so why say this and not the actual range?

P4, L30-31: What can be said about how well the model represents processes within the watershed, such as snow accumulation and melt, for example? This relates back to the points about physical appropriateness of the model.

Section 2.3: The bias correction procedure is not adequately described. How well did the data compare and what type of correction was necessary? There is not enough information to determine what was done.

Section 2.4: This is still very unclear and there are no further details or reference provided to give more clarity. See comments by reviewer #2. A more clear illustration or some more detail may be needed to clearly explain this to the reader.

P5, L21: What is the Euclidean distance between pairs (i.e. in what space)?

P6, L5-8: Here is where internal climate variability is defined. So is this essentially the variation in forcing among the ensemble of climate model outputs?

P6, L11: what qualifies as a high flow? Is there a specific threshold?

P6, L22-27: These are important results and should be described in more detail. For instance, what are the magnitude and variability of the changes? How do the results differ among the ensemble members?

P6, L30: higher than what? Than the range of air temperature?

P7, L8: Are you referring to Jan-Feb streamflow? The section heading indicates that, but it isn't specified.

P7, L18: Instead of “majoritarily” it would be better to simply say “mostly” or “for the most part”.

P8, L31-32: It is not clear why groundwater shows these differences, and this relates back to the need to explain how the models handles such processes.

P9, L7-8: monthly resolution of what? And what was the issue with representation of winter processes?

P9, L11-12: This is not correct. It is not clear how model structure and process representation affect the simulation of internal watershed processes, such as snowmelt and routing.

P9, L14-19: How confident can you be about the use of NRCANmet? Just because it is “widely used” isn’t justification enough. There should be some indication somewhere about how well the simulations capture other variables (i.e. the internal watershed processes – especially snow accumulation and melt). Also, if measured Q is overestimated, would that not indicate that the problems with the model are even worse? Reviewer #1 raised some important concerns around the evaluation of the model.

P9, L27-31: This is unclear and could perhaps be better written.

P10, L23: high Z500 anomalies enhance what?

Section 4.3: It is not clear what the discussion here is getting at. Is it that variability among the ensemble members still predicts similar local weather patterns? Does this relate to the internal climate variability issue? It could be more clear.

P11, L2-3: Presumably, this is because only T and P data were fed into the model as forcing variables?

P11, L24: How do you know this correlation isn’t an artifact of the model due to the representation of groundwater?

P11, L26-28: It was presumed that examining the influence of different weather patterns on streamflow regime was a purpose of the study. So what can be said about what the findings of this study suggest?

P11, L32: When it says “there will be less snow”, it isn’t clear initially that this isn’t entirely a model projection, but that under the current climate the hydrological regime is less dominated by snowmelt runoff than the other basins. This could be rephrased to be more clear, although it becomes clear further into the paragraph.

P12, L23-28: The summary of results needs to be more clear. What do the “small” and “low” in parentheses mean? What are the percentages – just the number of runs that showed a certain category of streamflow change? This should be more clear. What is the threshold for categorizing large or small changes? And more importantly, what is the magnitude and variation of the projected changes? How confident are you in the various ensemble members, and are there some that are more likely than

others (i.e. these near the median) which should be given more weight or more consideration? This is something to present more clearly in the results section, and convey briefly in the abstract.

P13, L2: what can be said about high flows from the results of this study? What insight is there into changing flood regimes?

P13, L12: There is a need to indicate where the data can be found. I think it is mostly publically available, so it should be a matter of identifying the sources.

Figure 1: It might be helpful to specify in the legend that the points are for the CRCM5-LE data.

Figure 6: is Delta Q annual? It is not clear why it varies between 0 and 2.5 when in the next figure it varies between 0 and 1. Also, to help, it could be made more clear by labelling the four weather classes in the right hand panel.

Figure 7: Instead of delta flow, this should be delta Q to be consistent with Fig. 6.

Figure 8: Units are needed throughout the figure and the legends. Is part (a) surface temperature?

Figure 9: Units are needed throughout the figure and the legends.

Figure 10: The legend is ambiguous for the different classifications. There are two red, three orange, two green and two blue categories. Although the order seems to be clear, this could be improved. Also, what are the vertical lines and hatches in the figure? Standard deviations? Please clarify.

Table 1: Where does urban landcover fit in? How much is urban?

Table 2: The ranges in parameter values do not provide enough information. What are the values for different HRUs (GRUs)?

Table 4: In the third column, what does the percentage in the top row specify? % of the ensemble? In the fourth to seventh columns, what are the units? mm/day? If so, what do the terms in brackets represent? And why are some missing?

Grammatical and Technical Issues

P2, L8: “source” should be “sources”

P8, L26: instead of “expecting” it should say “projected”

P9, L13: “incriminated”? Is there a better word choice?

P9, L16: “wrong measurements” should instead say “measurement uncertainty”

P9, L30: “which is conform”? Pease rephrase.

P10,L1: “associated to stronger”; replace “to” with “with” and subsequently where the words “associated to” are used.

P10, L12: “enhance” should say “enhancement”.

P10, L18: capital letter G for “great Lakes”.

P11, L15: replace “in the meanwhile” with “Meanwhile”

P11, L19: either say “if more snow falls” or “if there is more snowfall”

P11, L26: “connexion” should be replaced with “connection” or “link”

P12, L17: the s should be removed from “precipitations”, and where this is written subsequently.