Response to Reviewer 1.

Wang et al:

We greatly appreciate the comments which helped improve our manuscript. Our responses to the comments are listed below.

Generic comments:
1. Eq.4 does NOT assume Ss=0. It treats Ss as being composed of two parts: the part above the surface water retention capacity, which contributes to surface runoff; and (2) the part below the surface water retention capacity, which is integrated with subsurface water through soil surface infiltration and contributes to baseflow. So, when the surface runoff is equal to zero, the surface storage is not necessarily zero (but is below the surface water retention capacity, e.g., for a lake or pond without overflow), regardless in which season of the year. The text for this description is revised to make it clearer (the 2nd paragraph of Section 2).
The baseline of the TWS data, which was based on the 2004-2009 average in the original data, was readjusted to the minimum value that occurred over the study period. This information was originally given in the Supporting Information (with more details in Wang, 2019), and it is added to the main text in this revision (including re-plot of Fig 6). I apologize for missing this information in the main manuscript in the previous version.

2. The manuscript is revised by adding details for the parameter estimation. The numerical scheme for this study, coded in FORTRAN, will be made available to the public through the Canada Centre for Remote Sensing Open Data Portal. The program includes a triple-nested numerical iteration for the three parameters of k_0, T_c, and a. We used an iteration of 50 for each parameter with its given range of values. So, the total parameter combinations for a numerical experiment is 50^3. This takes about 10 CPU-minutes for a regular desktop computer. The initial values for the parameters can be estimated from either available information or expert knowledge. In this study, the values obtained in Wang (2019) were referenced. Prior numerical experiments with relatively large ranges and coarse resolutions for the parameter values could be helpful to narrow down the final ranges. The solution with maximum Nash-Sutcliffe Efficiency corresponds to a unique combination of k_0, T_c, and a, and the solutions with a range of top Nash-Sutcliffe Efficiency values include a large number of combinations of k_0, T_c, and a. This can be seen from the Figure RC1-1 attached. Numerically, it reflects the interactions of parameters in the baseflow simulations (e.g, low a compensates low k_0). Physically, it suggests the level of accuracy for parameters with the prescribed accuracy of Nash-Sutcliffe Efficiency. As such, the large number of parameter combinations is an over exaggeration, as many of them are under the prescribed accuracy level.

3. Our method has relatively low annual BFI value compared with five of the six USGS methods which have known problems of overestimating baseflow by including snowmelt runoff (Table 2). Our model may overestimate baseflow in early snowmelt season (April)

due to the impact of frozen soil which is discussed in the 3rd paragraph of Section 5. The high BFI in summer could be attributed to the large size, flat topography, and well drained soil (e.g., peatland) of the Albany watershed. Rudra et al. (2015) analysed BFI for 115 Ontario watersheds (note that some of them are very small in size). Our results showed similar magnitudes to Rudra's study for the comparable watersheds. In general, the available data and studies don't suggest baseflow overestimation by our model (except, possibly in April, as discussed). This will be further verified when more information becomes available. The above new reference was added following this comment.

4. Section 5 is revised to address the comments, including discussion on basin size (paragraph 4), impact of GRACE TWS and other uncertainties (paragraphs 5, 6 and 7), applications over different climate regions and further research recommendations (paragraph 8).

Specific comments:
1. The approach is illustrated using a cold region watershed, but it is not limited to cold regions. Paragraph 8 in Section 5 was revised following this comment.
2. Revised.
3. Revised. More discussion is given in Paragraph 9 in Section 5.
4. Baseflow vs. surface runoff, particularly for large watersheds, is somewhat conceptual and not strictly differentiable. There are studies towards the recognition of multiple baseflow components in the streamflow. We included discussions and cited relevant studies in the paper.
5, 6 and 7. Detailed examples for the baseflow/surface runoff applications in various areas and more recent review works are provided by citing available studies.
8. Revised by providing more detailed information for the manual approach and by adding references.
9. One process is that, as discussed in the following text, incorrectly identifying snowmelt runoff as groundwater discharge with the traditional methods. Addressing this problem is one of the major research objectives of this study.
10. It is due to the assumptions for these methods, for example, the estimation of duration of surface runoff (Eq. A1 in Supporting material file).
11. Revised by adding specifics, which actually are discussed in detail in the 2nd paragraph of the Introduction.
12 and 13. Revised to make the statement clear. For each month, the change of surface water above retention capacity is surface runoff, and the change of Sg contributes to baseflow. Sn has no contribution to either.
14. No. Please see clarification above.
15. A review of literatures apparently show that this parameter is called in many different ways. We revised it to rate constant to avoid any misinterpretations. Its reverse represents residence or turnover time, both are controlled by aquifer properties.
16. The Nash-Sutcliffe modelling Efficiency is used as the criteria in model calibration. Other statistical parameters are used for model evaluation.

17. Revised.

18. As mentioned in our paper, the interannual variations of temperature is huge. The determination of, for example, winter or snowmelt season, is by temperature-based multi-criteria for each individual year and they varied by calendar months from year to year. Manuscript was revised by adding the information.

19. Specific number (340mm) is added.

20. Revised. From long-term statistical mean view point (Fig. 9), this is the case.

21. Revised. TWS varied from the lowest (0, the reference value) to the maximum of about 200mm.

22. Reference added. As shown in Fig 12 and 13 in Wang et al. (2013), the annual ET over the region is mainly contributed by the three months of Jun-Aug.

23. It is described in the paragraph below Eq. (3a) in Section 2.

24. Revised.

25. Fig.5 was cited. As the orange line shows, the winter season Q contribution to the annual Q is very small. The specific magnitudes were discussed in the paragraph above the figure.

26. Reference to Fig 10 is added. The discussion here is a synthesis of Fig 10, Table 2 and Fig 11.

27. Updated.

28. Revised.

Technical comments:

All comments are addressed. Note that the blue lines in Fig. 3 were removed. The transition dates varied from year to year. The blue lines may cause confusion. Given the same reason, adding average transition dates to other figures may cause confusion. We used the same color legend in Fig 9 and 10. Fig. 8 has more colour choice as the models are separated in 4 panels, and we used the same color order among the 4 panels. We used lighter color in Fig. 11 than Fig 10 as there is a lot more overlaps in Fig. 11 than Fig 10.

Reference:

Rudra, R., Ahmed, I., Khan, A. A., Singh, K. G., Goel, P. K., Khayer, M., and Dickinson, T.: Use of Baseflow Indices to Delineate Baseflow Dominated and Rapid Response Flow Dominated Watersheds. Canadian Biosystems Engineering 57, 1–11. doi.org/10.7451/CBE.2015.57.1.1, 2015.

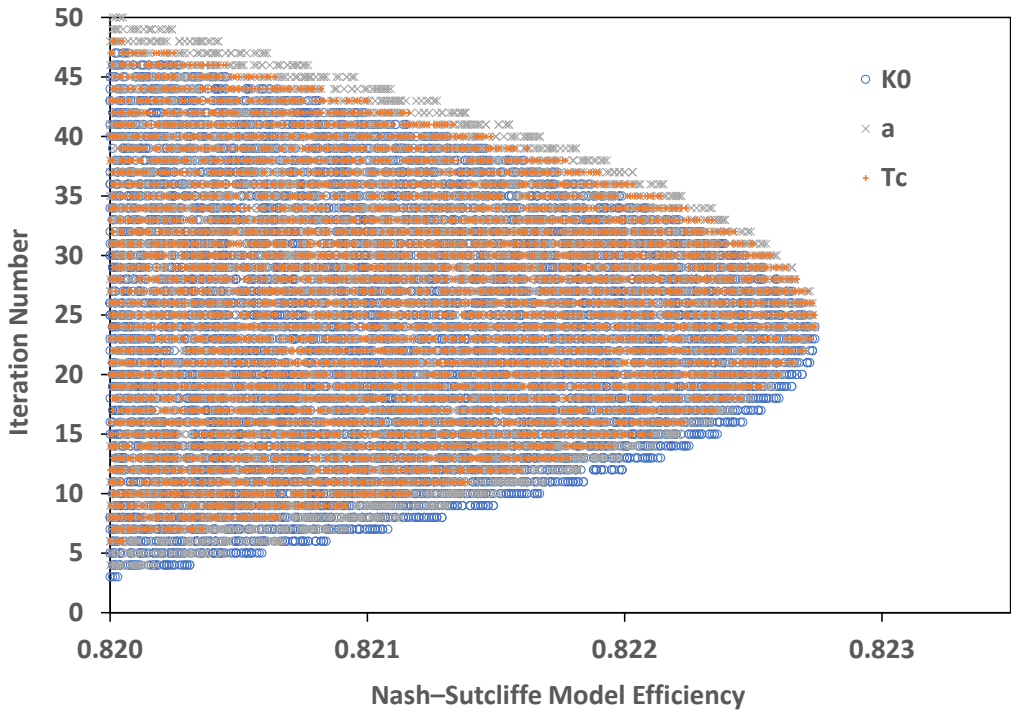Figure RC1-1: Solutions for k_0, T_c, and a using NSE as the criteria. Ranges of parameter values: $0.006 < k_0 < 0.009$, $430 < T_c < 770$, $36 < a < 55$.

Response to Reviewer 2.

Wang et al:

We greatly appreciate the comments which we helped improve our manuscript. Our responses to the comments are listed below.

Major comments:

1. The water balance closures for Canada's watersheds were studied at both short time (monthly) (Wang et al., 2014a) and long-term (30 years) (Wang, et al., 2014b) scales. The Albany watershed in this study was among the best ones with the smallest water imbalance (non-closure) in Canada. In particular with the precipitation product from Canada's meteorological station network, the water balance for the Albany watershed achieved the least water imbalance (close to 0) among all the watersheds (Fig. 11 in Wang et al., 2014a). This suggests the high data qualities over this region (including snow, or the Sn which is discussed in more detail below) and the close connections of the surface water and groundwater systems for the Albany watershed. Regardless, this study doesn't directly involve water balance assumption. It is based on the fact that the GRACE measured TWS is the sum of surface water+subsurface water+snow. The watershed doesn't have glaciers and other water components to significantly contribute to the GRACE TWS.

The uncertainties with snow (Sn), surface water (Ss), and subsurface water (Sg) were mainly from the EALCO model and the GRACE TWS errors. These uncertainties cannot be directly evaluated because no corresponding observations are available. However, they can be evaluated indirectly. We calculated the GRACE TWS measurement error, leakage error, and combined total error following Wahr et al. (2006) using the land surface model CLM4, and they were 13.2 mm, 15.8 mm, and 20.6 mm, respectively, for the watershed. The impact on the TWS error estimate due to the uncertainty in the CLM4 model was evaluated by comparing with the error estimate using a different land surface model of NOAH (Wang et al., 2014a). The magnitudes of errors from the two studies were found to be similar. Since our model is calibrated using observed baseflow measurement, systematic errors or biases in TWS would be reflected in the model calibration process and compensated in the parameter values, so their impact on the hydrograph separation results would be minimal. Random errors in TWS could directly affect the baseflow estimation. However, in cold season since the TWS change is mainly due to snow variations which doesn't contribute to river flow, the uncertainties in TWS also have minor impact on baseflow estimate. In non-frozen season, an error of 20 mm in TWS would result in an estimate error of 0.15 mm/day in baseflow, which is substantially small compared with the flow magnitude in summer of the watershed. In fact, as suggested by the water budget closure study (Wang et al., 2014a), the random error is much smaller than the measurement error of 20mm.

The uncertainties with snow can be further evaluated by comparing this study with Wang (2019). In Wang (2019), the winter time drainable water was estimated by the difference of the TWS at the winter starting time and the accumulated streamflow (from winter start to a specific month). The advantage of Wang (2019) approach is that it doesn't need snow data in the modelling, while the disadvantage is that it brings the uncertainties of winter streamflow measurement into the modelling. Also, any errors with the TWS at the winter start will propagate into the

modelling for all the months in that season. The modelling results show that this study (NSE=0.823, r=0.91) slightly improved that of Wang (2019) (NSE=0.809; r=0.903), suggesting the impact of uncertainties from EALCO snow data is less than that from the uncertainties in streamflow measurements on the baseflow modelling.

The manuscript was revised following this comment by adding paragraph 5 and 6 in Section 5 (Discussion). Some other minor changes were also made in other places to address this comment.

2. The model was calibrated using all the available winter data during our study period. Even so, due to the limited GRACE observations and the strict selection criteria for winter months to ensure that the watershed was in frozen conditions and no rain or other events could occur to cause surface runoff, the sample number for model calibration was relatively small (~3 months/year on average). However, the calibration doesn't assume that the baseflow mechanism is stable over the years. In fact, possible variations in baseflow mechanism with watershed conditions (such as yearly weather changes as commented) were expected, and the impact, which causes deviations between modelled vs. observed flows, was reflected in the calibration results. Moreover, the model treated baseflow coefficient dynamically using the accumulated freezing temperature function to account for the seasonal (e.g., winter vs. summer) and yearly changes in watershed conditions. This represents a major innovation for this method, and the improvement and advantage over traditional methods of treating baseflow parameters as constants have been discussed in detail in Wang (2019).

Specifically for the Albany watershed, it had a large water budget surplus for soil/aquifer recharge and the yearly water condition changes were small (Wang et al., 2014a). Moreover, it is located in a flat region and the baseflow is mainly contributed by surficial aquifers including wetlands. The hydrogeological settings of the watershed are relatively simple and the baseflow mechanism is largely different with the watersheds that have deep slopes and complicated aquifer systems with large variations in water conditions, which often involve dynamic channel networks (e.g., time‑varying geometry of saturated channeled sites) and result in nonlinear storage‑baseflow relationships. Indeed, the study of Wang (2019) found that the Albany watershed follows a simple linear baseflow-storage relationship quite well when soil frost is absent (i.e., summer season), and the freezing temperature function (Eq. 3) can effectively make the model account for the watershed condition dynamic changes in winter. The Fig. RC2-1 attached shows the yearly calibrated baseflow parameter k from our model vs. that inversely calculated from observations. The monthly and interannual variations of k were mainly contributed by the variations in watershed frozen conditions (Wang 2019). After the temperature function modification, the k has a correlation coefficient of r=0.86, demonstrating the impact of interannual variations from other flow mechanisms is small.

The manuscript was revised following this comment by adding paragraph 7 in Section 5 (Discussion), and some minor changes in other places through out the manuscript.

3. Our model only yielded more conservative estimates of baseflow during the snowmelt season. The bias of overestimating baseflow in snowmelt season by the other six models due to the inclusion of snowmelt runoff is a well-known issue. Our results demonstrated improvement to

this known bias in these traditional models. Further, our data and results didn't suggest systematic bias coming from the calibration process. First, as shown in the Fig. RC2-2 attached, the range of drainable water storage used for model calibration was in 20-150 mm. It covered the variation range during the summer season fairly well. Second, our model results didn't show systematic underestimates of baseflow in summer when compared with other models (Figure 9). In fact, our model obtained higher baseflow in the mid-summer months of July-September (Figure 9) than most of the other six models. Annually, our model showed baseflow close to BFI-Standard, higher than HYSEP-Minimum, and lower than the other 4 models (Table 2). Third, our results are in good agreement with, and supported by, the results obtained in other studies, including Rudra et al. (2015) which analysed BFI for 115 Ontario watersheds, and Wang and Russell (2016) and Wang et al. (2017) which explicitly estimated the snowmelt runoff. The comment of Reviewer#1 also doesn't suggest this systematic bias.

Paragraph 1 in Section 5 (Discussion) was revised to address this comment. Some other minor changes were also made in other places following this comment.

4. I greatly appreciate the comment. However, unlike many other hydrograph studies which have large freedom in selecting study watersheds and taking the advantage of long-term (say decades) gauge measurements, the method in this study relies on GRACE data. Due to the coarse resolution (>100,000 km^2) and data availability of GRACE observations, the selection of study regions and time periods are largely constrained. In addition, our method is very new. To our knowledge, this is the first time someone has proposed a physically-based hydrograph separation model using GRACE observations. So it is important to have a robust testbed with high quality data. For example, if there are influences of glaciers or hydropower dams in flow measurements, they will affect the hypothesis tests. Large watersheds in western Canada mostly originated from the Rocky Mountains involving permanent snow/glacier influence (Wang et al., 2015), and watersheds in east Canada are mostly under the footprint of GRACE. This is the major reason for using the Albany Watershed in the Hudson Bay region for this study. We investigated other watersheds in central-east Canada. The large watersheds without much disturbances are mainly in the Hudson Bay region. Specifically, there are a total of nine large watersheds as shown in the Fig. RC2-3 attached Table RC2-1 below. Unfortunately, none of them except Albany has complete gauge measurement during the GRACE period of 2002-2016. Also, their sizes are all small and can hardly meet the requirement of GRACE resolution. So, including them would bring critical issues on (1) short data records for model fitting and (2) large uncertainties in GRACE data for under-footprint watersheds. As such, we prefer to focus this paper on presenting the innovative approach, and leave its applications (or possibly further improvement) to the worldwide community for other regions. See Reply #2 and #3 for the rest of this comment.

The last paragraph in Section 5 (Discussion) was added to address this comment. Some other minor changes were also made throughout the manuscript following this comment.

Minor comments:

1. The sentence is revised to "The subsurface drainable water storage is one of the major drivers of baseflow for most watersheds with certain hydrogeological settings.".

2. Yes, we agree that in some cases the surface water which is under the surface water holding capacity could turn into non-dischargeable water. Since GRACE measures the changes of water, water that stays in the system and is not contributing to runoff or baseflow is not accounted for in the model. The statement is revised following this comment.

3. More details in the numerical solutions is added. The programming code will be made available to the public through the Canada Centre for Remote Sensing Open Data Portal.

4. Thank for this note. Much appreciated! The Q should be the monthly sum of daily Q, with unit of mm. I apologize for this error.

5. The July monthly land surface evapotranspiration is about 80 mm, and water surface evaporation is about 130 mm. The annual total evapotranspiration is about 350 mm, and total water surface evaporation is about 560 mm. The manuscript is revised by adding these details and reference (Wang et al., 2014a).

6. We used fairly strict criteria (Section 2) to select the winter months to ensure that the watershed was in frozen conditions and no rain and other events were present to cause surface runoff. The PART, BFI-Std and BFI-Mod models also showed BFI of 1.0 in mid-winter. Note that the water infiltrated into soil surface and later outflowed into rivers (regardless of pathways, e.g., lateral flows or groundwater flows), or the water discharged from upstream aquifers into surface water bodies and later contributed to rivers flows through surface pathways, were all accounted as baseflow contribution. The watersheds in Streletskiy et al. (2015) are over a permafrost region and have dual layers of frozen soil (permafrost table and seasonal frozen layer). Obviously, our study region is less complicated in terms of frozen dynamics and its impact on water flows. However, their findings such as the frozen soil precluding surface water infiltration and building up hydrostatic pressure within the soil column, and the months-long time-lag of late summer precipitation contributing to stream flow in winter, are in support of our model and results. The reference Streletskiy et al. (2015) is added in the revised version.

7. Thanks for this comment. We have enhanced the discussions in Section 5.

8. The description has been revised by including the discussions given in Reply #2 and 3.

9. The main constraint for applying the six USGS methods used in this study is Eq (A1). To the best of our knowledge, improvement of these methods for application over large watersheds is still limited.

10. The modelling hypothesis of this study is not limited by Equation A1, but if GRACE data is directly used for estimate TWS, the method could be limited by its large foot print. With the research advance in downscaling GRACE products, it is expected the method could be applied to small watersheds in the future. The sentence is revised following this comment.

Reference:

Rudra, R., Ahmed, I., Khan, A. A., Singh, K. G., Goel, P. K., Khayer, M., and Dickinson, T.: Use of Baseflow Indices to Delineate Baseflow Dominated and Rapid Response Flow Dominated Watersheds. Canadian Biosystems Engineering 57, 1–11. doi.org/10.7451/CBE.2015.57.1.1, 2015.

Wahr, J., Swenson, S., & Velicogna, I.: Accuracy of GRACE mass estimates. Geophysical Research Letters, 33, L06401. https://doi.org/10.1029/2005GL025305, 2006.

Wang, S., Huang, J., Li, J., Rivera, A., McKenney, D.W., and Sheffield, J.: Assessment of water budget for sixteen large drainage basins in Canada. Journal of Hydrology, 512, 1-15. doi:10.1016/j.jhydrol.2014.02.058, 2014a

Wang, S., McKenney, D.W., Shang, J., and Li, J.: A national scale assessment of long-term water budget closures for Canada's watersheds. Journal of Geophysical Research: Atmospheres, 119, 8712–8725. doi:10.1002/2014JD021951, 2014b.

Wang, S., Huang, J., Yang, D., Pavlic, G., and Li, J.: Longterm water budget imbalances and error sources for cold region drainage basins. Hydrological Processes, 29, 2125–2136. doi: 10.1002/hyp.10343, 2015.

Wang, S., Zhou, F., and Russell, H.A.J.: Estimating snow mass and peak river flows for the Mackenzie River basin using GRACE satellite observations. *Remote Sensing*, 9, 256. doi: 10.3390/rs9030256, 2017.

Wang, S. and Russell, H.A.J.: Forecasting snowmelt-induced flooding using GRACE satellite data: A case study for the Red River watershed. *Canadian Journal of Remote Sensing*, 42, 203-213. doi: 10.1080/07038992.2016.1171134, 2016.
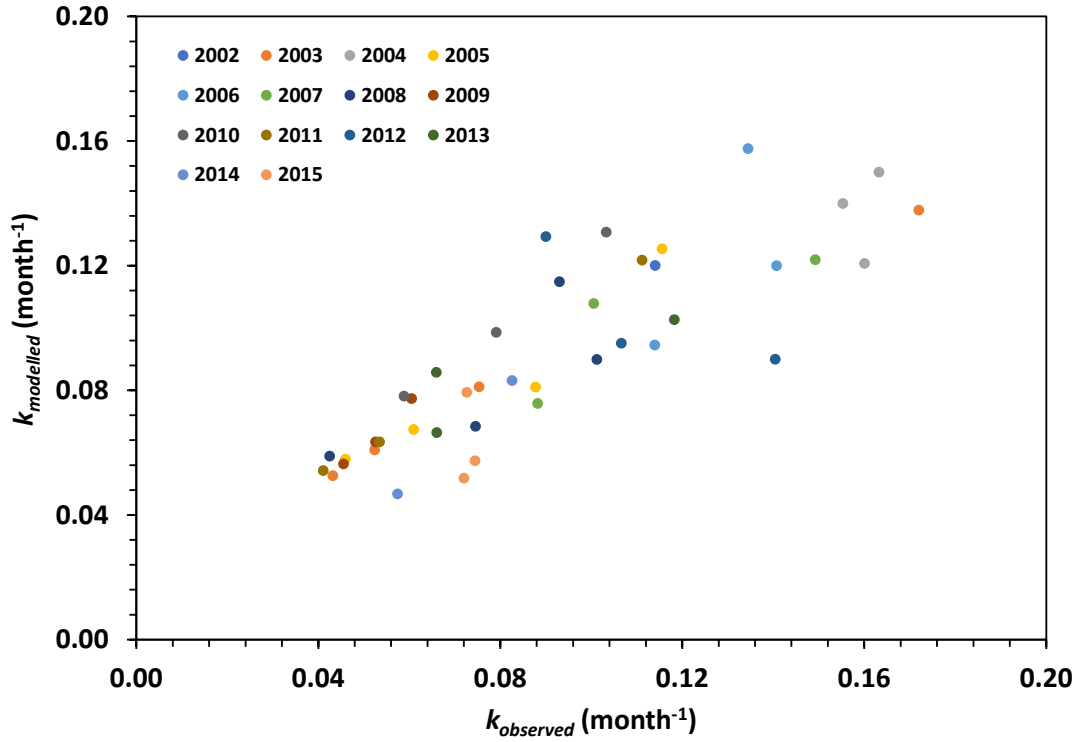
Fig. RC2-1: Yearly calibrated baseflow coefficient (k) from our model vs. that inversely calculated from observations.
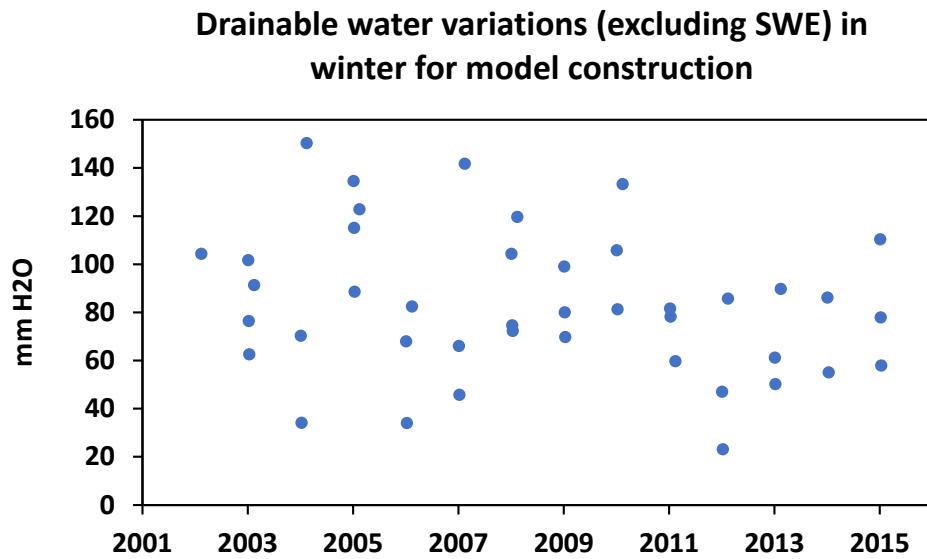

Fig RC2-2: The variation range of drainable water storage used for model calibration
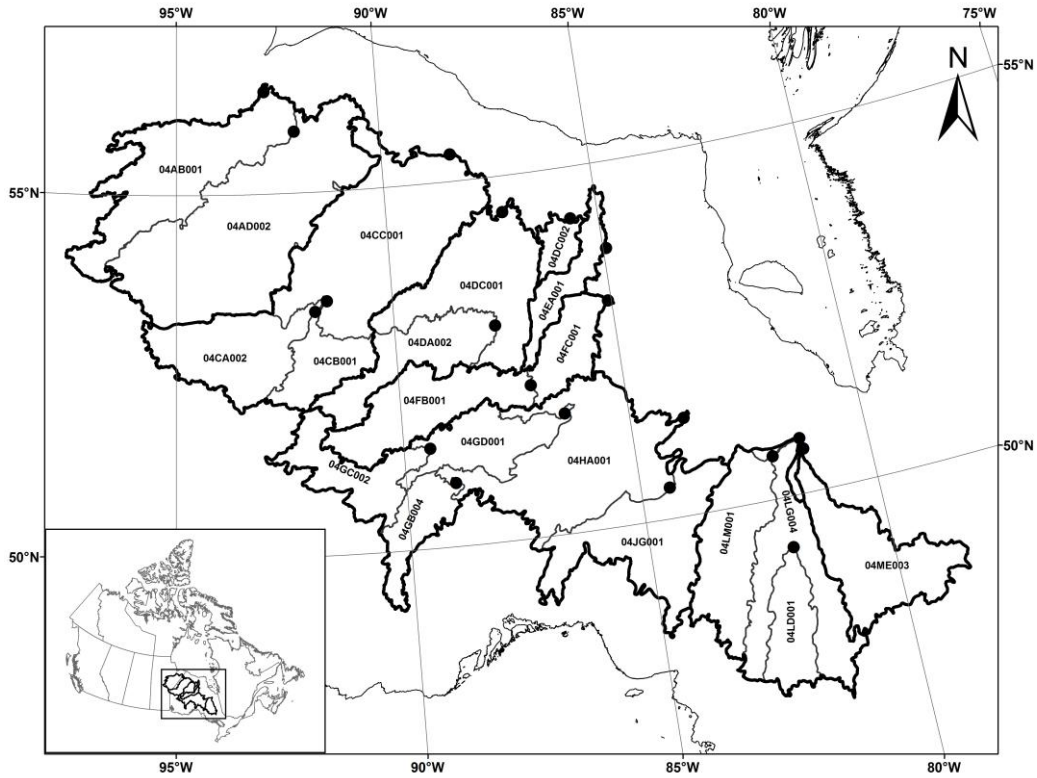
Fig RC2-3: The major watersheds in central-east Canada.

Table RC2-1. The major watersheds in central-east Canada.

Station ID	Name
04AB001	HAYES RIVER BELOW GODS RIVER
04CC001	SEVERN RIVER AT LIMESTONE RAPIDS
04DC001	WINISK RIVER BELOW ASHEWEIG RIVER TRIBUTARY
04DC002	SHAMATTAWA RIVER AT OUTLET OF SHAMATTAWA LAKE
04EA001	EKWAN RIVER BELOW NORTH WASHAGAMI RIVER
04FC001	ATTAWAPISKAT RIVER BELOW MUKETEI RIVER
04HA001	ALBANY RIVER NEAR HAT ISLAND
04LG004	MOOSE RIVER ABOVE MOOSE RIVER
04ME003	ABITIBI RIVER AT ONAKAWANA